Eigenvalue techniques for proving bounds for convex objective, nonconvex programs[1]

Daniel Bienstock
Columbia University
New York

March 2009, version Sun.Sep.13.160130.2009

1 Introduction

We consider problems with the general form

$$
\begin{aligned}
(\boldsymbol{\mathcal{F}}) : \qquad F^z \;=\;& \min \; F(x), && (1)\\
s.t. \qquad & x \in \mathcal{P}, && (2)\\
& x \in \mathcal{K}. && (3)
\end{aligned}
$$

Here,

- $F(x)$ is a convex quadratic, i.e. $F(x) = x^T M x + v^T x$ (with $M \succeq 0$ and $v \in \mathcal{R}^n$).
- $\mathcal{P} \subseteq \mathcal{R}^n$ is a convex set over which we can efficiently optimize F,
- $\mathcal{K} \subseteq \mathcal{R}^n$ is a non-convex set with "special structure".
- Typically, n could be quite large.

A standard approach to solving this problem would start by solving a convex relaxation to $\boldsymbol{\mathcal{F}}$, thereby obtaining a lower bound on F^z. However, when $\mathcal{K}$ is complex, such a lower bound is likely to be weak. In this paper we present efficient techniques that enable us to tighten the lower bound.

Our techniques are backed by theory and also prove computationally effective – our approach yields bounds comparable to or better than those produced by sophisticated formulations, but at a very small fraction of the computational cost.

To illustrate the situation we have in mind, we focus next on the example where $\mathcal{P} = \{x \in R^n \,:\, Ax \geq b\}$ for a given matrix A and vector b, and $\mathcal{K}$ is given by a cardinality constraint, i.e.

$$\mathcal{K} \,=\, \left\{\, x \in \mathcal{R}^n_+ \,:\, \|x\|_0 \leq K \,\right\},$$

where K is a positive integer. [Here, the zero-norm $\|v\|_0$ of a vector v is used to denote the number of nonzero entries of v.] This version of problem $\boldsymbol{\mathcal{F}}$, which we denote by $\boldsymbol{\mathcal{Q}}$, has received attention in the literature. The primary contributions in this paper address problem $\boldsymbol{\mathcal{F}}$ in its general form; however some of the technical points concern the specific case of a cardinality constraint, and our computational experiments involve problem $\boldsymbol{\mathcal{Q}}$.

A specialized algorithm for $\boldsymbol{\mathcal{Q}}$ was presented in [3]; this scheme proved much more effective than general-purpose branch-and-bound, but it is clear that instances of $\boldsymbol{\mathcal{Q}}$ can prove *extremely* hard even when $Ax \geq b$, $x \in \mathcal{R}^n_+$ describes the unit simplex, a case of practical importance. The central difficulty of the problem can be summarized in a single sentence: the (typically, strict) convexity of $F(x)$ defeats traditional methods for solving discrete optimization problems. Such methods rely

[1] Partially funded by ONR Award N000140910327 and a gift from BHP Billiton

on polyhedral techniques, that is to say, cutting-plane methods used to separate points from the convex hull of $\mathcal{P} \cap \mathcal{K}$, and branch-and-bound.

In order to understand why cutting-plane and branching methods fail, write $x^* = \mathrm{argmin}\{F(x) : Ax \geq b, \, x \in \mathcal{R}^n_+\}$. Typically x^* will belong to a high-dimensional face of the feasible region, and will be "far away" from the set $\mathcal{K}$, so that a polyhedral representation of $\mathcal{K}$ will not suffice to cut-off x^*. This is clearly illustrated by the following special case of $\boldsymbol{\mathcal{Q}}$:

$$\min \left\{ \sum_{j=1}^{n} d_j x_j^2 \, : \, \sum_{j=1}^{n} x = 1, \, x \in \mathcal{R}^n_+, \; \|x\|_0 \leq K \right\}, \tag{4}$$

where $d_j > 0$ for all j. If we ignore the cardinality constraint the optimal solution is given by $x^*_j = d_j^{-1}/(\sum_i d_i^{-1})$ for all j, with objective value $1/(\sum_i d_i^{-1})$; typically a very weak lower bound to the actual value of (4). As an example, suppose that $d_j = 1$ for all j, $n = 1000$ and $K = 20$. Then the value of (4) is $1/20$, while the lower bound provided by x^* is $1/1000$. The challenge lies in how to significantly improve this lower bound. However,

$$conv \left\{ x \in \mathcal{R}^n_+ \, : \, \sum_{j=1}^{n} x = 1, \;\; \|x\|_0 \leq K \right\} \; = \; \left\{ x \in \mathcal{R}^n_+ \, : \, \sum_{j=1}^{n} x = 1 \right\}. \tag{5}$$

In other words, there is nothing to separate – a direct application of cutting-plane methods to $\boldsymbol{\mathcal{Q}}$ will prove *completely ineffective* in this case.

Branch-and-bound techniques face similar difficulties. The standard mixed-integer programming formulation for $\boldsymbol{\mathcal{F}}$ on the unit simplex is as follows:

$$\begin{aligned} F^z \; = & \;\; \min F(x) && (6) \\ \text{s.t.} \qquad & \sum_{j=1}^{n} x = 1, \;\; x \in \mathcal{R}^n_+, && (7) \\ & x_j \leq y_j, \;\; y_j = 0 \text{ or } 1 \;\; \forall j; \;\; \sum_{j=1}^{n} y_j \; \leq \; K. && (8) \end{aligned}$$

This formulation, even though correct, can prove extremely weak, and will result in a prohibitively large amount of branching in order to prove a lower bound significantly higher than the value of the continuous relaxation of (6)-(8). In the case of problem (4) with $d_j = 1$ for all j, $n = 1000$ and $K = 20$, one can prove that at least $\binom{1000}{100}$ branch-and-bound nodes need to be enumerated before the lower bound improves to $1/100$, which is still quite poor.

Stepping back from problem $\boldsymbol{\mathcal{Q}}$, we can summarize the difficulty we face: set $\mathcal{K} \cap \mathcal{P}$ is non-convex while $F(x)$ is (possibly highly) convex. Thus the solution vector x^* for a convex relaxation of problem $\boldsymbol{\mathcal{F}}$ could be (typically, will be) quite "far", in the Euclidean distance sense, from $\mathcal{K} \cap \mathcal{P}$; and $F(x^*)$ is therefore a weak lower bound for F^z.

1.1 Outline of methodology

In this paper we use the following template for proving strong lower bounds for problem $\boldsymbol{\mathcal{F}}$:

(S.1) Develop an efficient procedure that, given a point $v \in \mathcal{R}^n$, computes a strong lower bound on the distance $D(v)$ between v and the feasible region to $\boldsymbol{\mathcal{F}}$, i.e. the set

$$\{x \in \mathcal{R}^n_+ \, : \, x \in \mathcal{P}, \; x \in \mathcal{K}\}$$

for relevant sets $\mathcal{K}$, e.g. the cardinality constraint. We assume that this problem is in general "easier" than problem $\boldsymbol{\mathcal{F}}$ itself – note that we ask for a "strong lower bound" rather than an exact computation of $D(v)$. We show that this is indeed the case when $\mathcal{K}$ is defined by a cardinality constraint.

(S.2) Let x^* be the optimal solution to some convex relaxation of $\mathcal{F}$. Armed with a positive lower bound for the distance between x^* and the feasible set for $\boldsymbol{\mathcal{F}}$, use techniques from convex optimization, in particular, the S-Lemma [13] (also see [2], [5], [12]) to obtain stronger lower bounds on F^z than that provided by $F(x^*)$.

(S.1)-(S.2) provides an initial template on our approach; we will see however that this approach must be appropriately augmented in order to obtain strong bounds. Nevertheless, note that (S.1)-(S.2) can be embedded in a branch-and-bound scheme where step (S.2) is used to improve the lower bound obtained at each node.

With regards to (S.1), clearly a trivial positive lower bound on the quantity $D(v)$ is, usually, readily available (e.g. v is a fractional vector and $\mathcal{K}$ specifies that variables must take integral values). But stronger bounds can be available. As our computational experiments demonstrate, we can efficiently and very tightly approximate $D(v)$ in the cardinality constrained case. This is an empirical assessment. From a theoretical perspective we have (proof in Appendix A):

Theorem 1.1 *Let $\eta \in \mathcal{R}^n_+$, $a \in \mathcal{R}^n_+$ and $a_0 \geq 0$ be such that $a^T\eta = a_0$. Let $0 < \epsilon < 1$ and K integer. Then the minimum distance between η and the set $\{x \in \mathcal{R}^n \, : \, a^T x = a_0, \, \|x\|_0 \leq K\}$ can be approximated up to a multiplicative error of $(1+\epsilon)$ in time polynomial in n, ϵ^{-1}, and the number of bits needed to represent η and a.* ■

Concerning (S.2), a critical result that we rely on is:

S-Lemma: Let f, $g \, : \, \mathcal{R}^n \to \mathcal{R}$ be quadratic functions and suppose there exists $\bar{x} \in \mathcal{R}^N$ such that $g(\bar{x}) > 0$. Then

$$f(x) \geq 0 \quad \text{whenever} \quad g(x) \geq 0$$

if and only if there exists $\gamma \geq 0$ such that $(f \, - \, \gamma g)(x) \, \geq \, 0$ for all x.

[Remark: here, a "quadratic" may contain a linear as well as a constant term.]The S-lemma can be used as an algorithmic framework for minimizing a quadratic subject to a quadratic constraint. Let p, q be quadratic functions of a variable $x \in \mathcal{R}^n$ and let α, $\beta \geq 0$ be reals. Then

$$\min\{p(x) \, : \, q(x) \geq \beta\} \; \geq \; \alpha \tag{9}$$

if and only if there exists $\lambda \geq 0$ so that

$$p(x) \, - \, \alpha \, - \lambda q(x) + \lambda\beta \geq 0 \quad \text{for all } x. \tag{10}$$

In other words, the minimization problem in (9) can be approached as a simultaneous search for two reals α and $\lambda \geq 0$, with α largest possible such that (10) holds.

For a simple application of our template to problem $\boldsymbol{\mathcal{Q}}$, consider Figure 1.

Here we have $n = 3$ and $K = 2$; furthermore the optimizer x^* of the quadratic over the unit simplex is in the *relative interior* of the simplex, and thus $x^* = x^F$, the minimizer of the quadratic over the hyperplane $S = \{x \in R^3 \, : \, \sum_j x_j \, = \, 1\}$. If we apply the S-Lemma so as to minimize

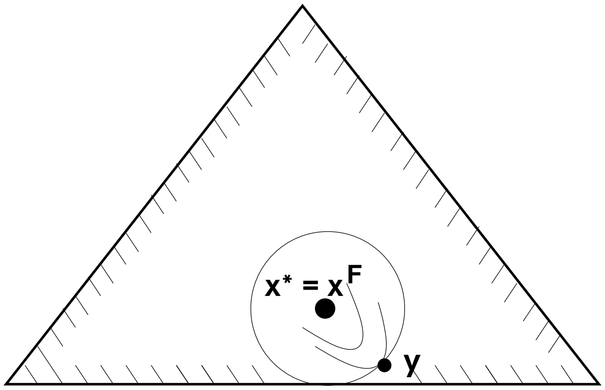

Figure 1: Simple application of S-Lemma to problem $\boldsymbol{\mathcal{Q}}$.

$F(x)$ while staying on S and on the exterior of the ball with center x^* and radius $D(x^*)$, then we obtain, as a lower bound on F^z, the value $F(y)$ [Note: we "stay on S" by appropriately changing coordinates]. In this case, of course, simple geometry would provide the same result.

Now consider the example in Figure 2, corresponding to $n = 3$, $K = 1$. Here a straightforward application of the S-Lemma will yield as a lower bound on F^z the value $F(y)$, which is weak – weaker, in fact, than $F(x^*)$. The situation in this case is that x^F, the optimizer of the quadratic over the affine hull of the feasible region, is not in the relative interior of the convex hull of the feasible region. This situation is common in practice. In summary, a direct use of the S-Lemma as outlined will prove ineffective.

To put this difficulty into a more general context, suppose we were to solve a convex relaxation to problem $\boldsymbol{\mathcal{F}}$ and let x^* denote its solution. Ideally we would want to solve a problem of the form:

$$\mathcal{V} \ = \ \min \Big\{ x^T M x \ + \ v^T x \ : \ x - x^* \in \mathcal{C}, \ (x - x^*)^T (x - x^*) \geq \delta^2 \Big\} \tag{11}$$

where $\delta > 0$ is a lower bound on the distance from x^* to $\mathcal{P} \cap \mathcal{K}$ and $\mathcal{C}$ is the cone of feasible directions (for P) at x^*. We can view this as a 'cone constrained' version of the problem addressed by the S-Lemma. Clearly, $F(x*) \leq \mathcal{V} \leq F^z$ with the first inequality in general strict. If the relaxation is polyhedral, (11) becomes

$$\min \Big\{ x^T M x \ + \ \tilde{v}^T x \ : \ Cx \geq 0, \ x^T x \geq \delta^2 \Big\} \tag{12}$$

for appropriate $\tilde{v}$ and C. However, we have:

Theorem 1.2 *For $M \succeq 0$ and $\delta > 0$, Problem (12) is NP-hard.* ■

[Proof in Appendix B].

This is unfortunate, because there is a rich literature on semidefinite programming methods used to solve problems similar to (12). The *trust region subproblem* (TRS) is a problem of the form

$$\min \ x^T Q x \ + \ c^T x, \tag{13}$$
$$s.t. \ \ x^T x \ = \ 1 \text{ (or } \leq 1). \tag{14}$$

where w is given vector. As shown in [14], this problem can be solved in polynomial time; [16] shows that polynomial solvability is maintained if we add a single (linear) constraint $a^t x \geq 0$. Also see [15], [17], [12], and references therein. As discussed in [12] the solution of problem TRS is closely related to the S-Lemma. Possibly, some of the SDP methodology could be brought to bear on problem (12), though of course the NP-hardness result will limit the scope. Potentially, the case where C has one row (or, perhaps, a fixed number of rows) might be polynomially solvable. An additional hurdle is that we are interested in very large-scale cases.

Rather than tackling problem (12) directly, we will use a computationally practicable approach that explicitly employs the S-Lemma. We will detail our approach later, but in outline we operate as follows.

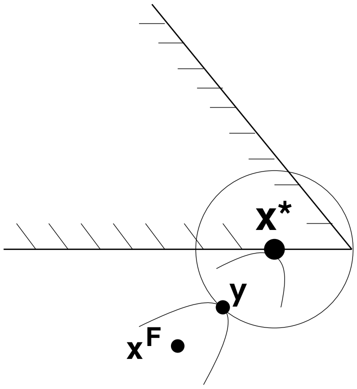

Figure 2: Direct application of S-lemma results in weak bound.

Let x^* be as above, and let $c = \nabla F(x^*)$. For $\alpha \geq 0$, let $p = x^* + \alpha c$, and let H^α be a hyperplane through p orthogonal to c. Define $L(\alpha) = \min\{F(x) : \|x - p\|_2 \geq D(p),\, x \in H^\alpha\}$. In Figure 3, $L(\alpha) = F(y)$. Clearly, $L(\alpha) \leq F^z$. Then

- Suppose $\alpha = 0$, i.e. $p = x^*$. Then x^* is a minimizer of $F(x)$ subject to $x \in H^0$. Thus $L(0) > F(x^*)$ when F is positive-definite.
- Suppose $\alpha > 0$. Since $c^T(y - x^*) > 0$, it follows that $L(\alpha) = F(y) > F(x^*)$.

Thus, $F(x^*) \leq \min_{\alpha \geq 0} L(\alpha) \leq F^z$; the first inequality being strict in the positive-definite case. Each value $L(\alpha)$ incorporates combinatorial information (through the quantity $D(x^* + \alpha c)$) and thus the computation of $\min_{\alpha \geq 0} L(\alpha)$ is not easily obtained through direct convex optimization techniques. (11):

Theorem 1.3 *If C has one row, $\mathcal{V} \leq \min_{\alpha \geq 0} L(\alpha)$.*

Proof sketch: Consider any given $0 \leq \alpha$. Let

$$x^\alpha = \text{argmin}\left\{F(x) \,:\, x \in H^\alpha, \text{ and } \|x - (x^* + \alpha c)\|_2^2 \,\geq\, D^2(x^*) - \alpha^2\|c\|_2^2\right\}.$$

Since H^α is orthogonal to c, $D^2(x^* + \alpha c) \geq D^2(x^*) - \alpha^2\|c\|_2^2$. Thus $L(\alpha) \geq F(x^\alpha) \geq \mathcal{V}$, and the result follows. ■

In order to develop a computationally practicable approach that uses these observations, let $0 = \alpha^{(0)} < \alpha^{(1)} < \ldots < \alpha^{(J)}$, such that for any $x \in P \cap \mathcal{K}$, $c^T x \leq \alpha^{(J)}\|c\|_2^2$. Then:

(1) For $0 \leq i < J$, compute tight lower bound, denoted $\tilde{L}(i)$,on

$$\min\{L(\alpha) \,:\, \alpha^{(i)} \leq \alpha \leq \alpha^{(i+1)}\}.$$

(2) Output $\min_{0 \leq i < J} \tilde{L}(i)$.

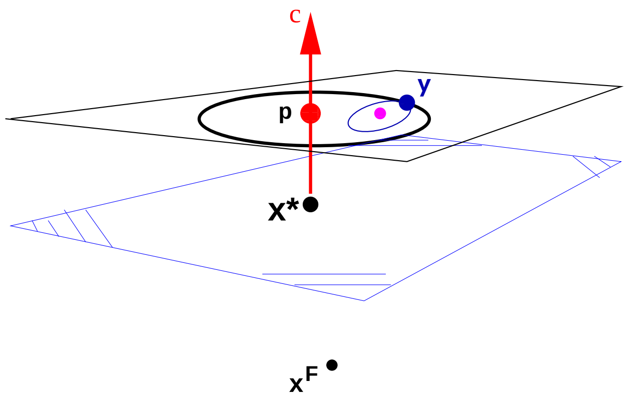

Figure 3: A better paradigm.

The idea here is that if (for all i) $\alpha^{(i+1)} - \alpha^{(i)}$ is small then $L(\alpha^{(i)}) \approx L(\alpha^{(i+1)})$ and we should be able to choose $\tilde{L}(i)$ nearby, as well. Thus the quantity output in (2) will closely approximate $\min_{\alpha \geq 0} L(\alpha)$.

To implement Step (1), we show that the computation of a single value $L(\alpha)$ can be carried out by obtaining an eigenvector decomposition of the *projection* of $F(x)$ to a hyperplane orthogonal to c (Section 2), and then appropriately changing coordinates so that the S-lemma can be efficiently applied (Section 3) on an $(n-1)$-dimensional space. Furthermore, our use of the S-lemma is such that for each $0 \leq i < J$ the computation of $\tilde{L}(i)$ amounts to an *interpolation* between $L(\alpha^{(i)})$ and $L(\alpha^{(i+1)})$ (Section 5).

As already indicated, we only perform one projection – the rest of the linear algebra, including the application of the S-Lemma, and the interpolation, is simple and fast. Thus we can choose J quite large if desired, or equivalently, make the differences $\alpha^{(i+1)} - \alpha^{(i)}$ small, thereby closely approximating $\min_{\alpha \geq 0} L(\alpha)$. The overall procedure runs fast even for large n.

In the following sections we will formalize the approach. We point out that the procedure can be adapted so that the vector c is not necessarily $\nabla F(x^*)$ (other choices can work better).

Finally, one fortuitous benefit resulting from computing the projection of the quadratic onto a hyperplane is that the eigenvalue structure changes favorably; again yielding even better bounds. In fact, if the definition of the feasible set $\mathcal{P}$ includes a set of linear equations then the projection to the hyperplane defined by these equations (or a subset thereof) *before* proceeding with step (1) above will further improve the eigenvalue structure.

2 Projecting a quadratic

Let $M = Q \Lambda Q^T$ be an $n \times n$ matrix given by its eigenvalue-eigenvector decomposition. Let c be a unit vector and denote $H = \left\{ x \in \mathcal{R}^n \, : \, c^T x = 0 \right\}$. In this section we describe an efficient algorithm for computing an eigenvalue-eigenvector decomposition of the "projected quadratic" PMP where P is the projection matrix onto H. Our approach reverse engineers, and strengthens, results from [7] (also see Section 12.6 of [8] and references therein).

Definition 2.1 *An eigenvector q of M is called* acute *if $q^T c \neq 0$. An eigenvalue λ of M is called acute if at least one eigenvector corresponding to λ is acute. An eigenvector (resp., eigenvalue) which is not acute is called* perpendicular*.*

In what follows, we will write $d = Q^T c$.

Lemma 2.2 *Let $\alpha < \beta$ be acute eigenvalues of M such that there is no acute eigenvalue in (α, β). Then the equation*

$$\sum_{i=1}^{n} \frac{d_i^2}{\lambda_i - \lambda} \;=\; 0 \tag{15}$$

has a unique solution $\tilde{\lambda}$ in (α, β). Further, $\tilde{\lambda}$ is an eigenvalue of PMP and there is a corresponding eigenvector PMP which is a linear combination of acute eigenvectors of M.

Proof. The expression on the left-hand side of (15) has a singularity at each acute eigenvalue; the choice of α and β shows that there is indeed a unique solution in (α, β).

For the second statement of the proof, note that (15), evaluated at $\tilde{\lambda}$, can be written as

$$0 \;=\; d^T(\Lambda - \tilde{\lambda} I)^{-1} d \;=\; c^T Q(\Lambda - \tilde{\lambda} I)^{-1} Q^T c \;=\; c, \tag{16}$$

thus, writing

$$w \;=\; Q(\Lambda - \tilde{\lambda} I)^{-1} Q^T c, \tag{17}$$

we have that w is a linear combination of acute eigenvectors of M and that $w \in H$, and therefore $Pw = w$. So

$$(M - \tilde{\lambda} I)w \;=\; Q(\Lambda - \tilde{\lambda} I)Q^T w \;=\; QQ^T c,$$

and therefore

$$PMPw \;=\; PMw \;=\; \tilde{\lambda} Pw \;=\; \tilde{\lambda} w,$$

as desired. ■

Notes: (1) the lemma provides an algorithmic recipe for constructing the eigenvector w, given $\tilde{\lambda}$. (2) The expression in (15) is sometimes referred to as the *secular polynomial.* (3) The computation of a zero of the secular polynomial in a given interval (α, β) can be performed by e.g. using a Newton-Raphson approach or simple golden section (or binary) search.

Lemma 2.3 *Let α be an eigenvalue of M, V^α the set of columns of Q with eigenvalue α, and $\mathcal{A} = \mathcal{A}(\alpha)$ denote the acute members of V^α. If $|\mathcal{A}| > 0$, then we can construct $|\mathcal{A}| - 1$ eigenvectors of PMP corresponding to eigenvalue α, each of which is a linear combination of elements of $\mathcal{A}$ and is orthogonal to c.*

Proof: Write $m = |\mathcal{A}|$, and let H be the $m \times m$ Householder matrix [8] corresponding to $d_\mathcal{A}$, i.e. H is a symmetric matrix with $H^2 = I_m$ such that

$$Hd_\mathcal{A} \;=\; (\|d_\mathcal{A}\|_2, 0, ..., 0)^T \;\in\; \mathcal{R}^m.$$

Let $Q_\mathcal{A}$ be the $n \times m$ submatrix of Q consisting of the columns corresponding to $\mathcal{A}$, and define

$$W \;=\; Q_\mathcal{A} H. \tag{18}$$

Then $c^T W \;=\; d_\mathcal{A}^T H \;=\; (\|d_\mathcal{A}\|_2, 0, ..., 0)$. In other words, the columns of the submatrix $\hat{W}$ consisting of the last $m-1$ columns of W are orthogonal to c. Denoting by $\hat{H}$ the submatrix of H consisting of the last $m-1$ columns of H, we therefore have

$$\hat{W} \;=\; Q_\mathcal{A} \hat{H}, \text{ and}$$

$$PMP\hat{W} \;=\; PQ\Lambda Q^T\hat{W} \;=\; PQ\Lambda Q^T Q_{\mathcal{A}}\hat{H} \;=\; \alpha PQ_{\mathcal{A}}\hat{H} \;=\; \alpha\hat{W}.$$

Finally, $\hat{W}^T\hat{W} \;=\; \hat{H}^T\hat{H} \;=\; I_m$, as desired. ■

Now suppose that

$$\alpha_1 < \alpha_2 < \ldots < \alpha_q$$

denote the distinct acute eigenvalues of M (possibly $q = 0$). Let p denote the number of columns of Q which are perpendicular eigenvectors. Writing $m_i = |\mathcal{A}(\alpha_i)| > 0$ for $1 \le i \le q$, we have that

$$n = \sum_{i=1}^{q} m_i \quad + \quad p.$$

(p.1) Using Lemma 2.2 we obtain $q-1$ eigenvectors of PMP, each of which is a linear combination of acute eigenvectors among Q. Any eigenvalue of PMP constructed in this manner is different from all acute eigenvalues of M.

(p.2) Using Lemma 2.3 we obtain, for each i, a set of $m_i - 1$ eigenvectors of PMP, orthogonal to c and with eigenvalue α_i, each of which is a linear combination of elements of $\mathcal{A}(\alpha_i)$. In total, we obtain $n - q - p$ eigenvectors of PMP.

(p.3) Let p denote the number of perpendicular vectors among Q. Any such vector v (with eigenvalue λ, say) by definition satisfies $PMPv = PMv = \lambda Pv = \lambda v$.

By construction, all eigenvectors of PMP constructed as per (p.1) and (p.2) are distinct. Those arising in (p.3) are different from those in (p.1) and (p.2) since no column of Q is a linear combination of other columns of Q. Thus, altogether, (p.1)-(p.3) account for $n-1$ distinct eigenvectors of PMP, all of them orthogonal to c, by construction. Finally, the vector c itself is an eigenvector of PMP, corresponding to eigenvalue 0.

2.1 Recursive projection

Let C be an $r \times n$ matrix of rank r. Here we show how to construct an eigenvalue-eigenvector decomposition $\tilde{Q}\tilde{\Lambda}\tilde{Q}^T$ of the projected quadratic $\tilde{P}M\tilde{P}$, where P is the projection matrix onto $H = \{x \in \mathcal{R}^n \,:\, C^T x = 0\}$, such that

(i) $n - r$ of the columns of $\tilde{Q}$ are contained in H, and

(ii) the remaining r columns of $\tilde{Q}$ form a basis for the linear space spanned by the rows of C.

The case $r = 1$ was handled in the previous section and here we complete the general case.

Proceeding inductively, assume that the construction (i)-(ii) has been carried out using the first k rows of C, obtaining a decomposition $M = \bar{Q}\bar{\Lambda}\bar{Q}^T$ which satisfies (i) and (ii) with $k = r$. Without loss of generality assume that the first k columns of $\bar{Q}$ form a basis for the linear space $\bar{S}$ spanned by the first k rows of C. Denote by c vector obtained by projecting the $(k+1)^{st}$ row of C onto the $n-k$-dimensional hyperplane orthogonal to $\bar{S}$, and then scaling so that $\|c\|_2 = 1$.

Suppose we use the construction described in section 2 using this vector c and the quadratic $\bar{Q}\bar{\Lambda}\bar{Q}^T$. By construction, the first k columns of $\bar{Q}$ are perpendicular, and any new eigenvector corresponding to case (p.1) above will be a linear combination acute columns of $\bar{Q}$ and thus orthogonal to the first k columns of $\bar{Q}$. Properties (i)-(ii) now clearly follow.

3 Application of the S-Lemma

In this section we are given an $h \times n$ matrix N of full row rank, a vector $\hat{x} \in \mathcal{R}^n$, a positive vector $(\delta_1, \ldots, \delta_n)$, a positive real β, and a vector $v \in \mathcal{R}^n$. We describe a computationally practicable approach to solving the problem

$$\min \quad x^T M x + v^T x, \quad \text{subject to} \quad \sum_{i=1}^{n} \delta_i (x_i - \hat{x}_i)^2 \, \geq \, \beta, \quad \text{and} \quad x = N\hat{x}. \tag{19}$$

By rescaling, translating, and appropriately changing notation, the problem becomes of the form

$$\min \quad x^T M x + v^T x, \quad \text{subject to} \quad \sum_{i=1}^{n} x_i^2 \, \geq \, \beta, \quad \text{and} \quad Nx = 0. \tag{20}$$

Let $H = \{x \in \mathcal{R}^n \, : \, Nx = 0\}$, and let P be the $n \times n$ matrix corresponding to projection onto H. Using the methodology in Section 2.1 we can produce a representation of PMP as $\tilde{Q}\tilde{\Lambda}\tilde{Q}^T$, where without loss of generality the last h eigenvectors in the representation are a basis for H. Thus, problem (20) becomes, for appropriately defined $\tilde{v}$,

$$\min \sum_{j=1}^{n-h} \tilde{\lambda}_j y_j^2 + 2\tilde{v}^T y, \quad \text{subject to} \quad \sum_{j=1}^{n-h} y_j^2 \, \geq \, \beta. \tag{21}$$

Using the S-lemma, the value of this problem is at least γ, if and only if there exists $\mu \geq 0$ s.t.

$$\sum_{j=1}^{n-h} \tilde{\lambda}_j y_j^2 + 2\tilde{v}^T y \, - \, \gamma \quad - \quad \mu \left(\sum_{j=1}^{n-h} y_j^2 \, - \, \beta \right) \quad \geq \quad 0 \quad \forall \, y \in \mathcal{R}^{n-h}. \tag{22}$$

A matrix form of the S-Lemma is known, but for completeness we carry out a direct derivation of a condition equivalent to (22). Defining

$$W \;=\; \begin{pmatrix} \tilde{\lambda}_1 - \mu & & & & \tilde{v}_1 \\ & \tilde{\lambda}_2 - \mu & & & \tilde{v}_2 \\ & & \cdots & & \vdots \\ & & & \tilde{\lambda}_{n-h} - \mu & \tilde{v}_{n-h} \\ \tilde{v}_1 & \tilde{v}_2 & \cdots & \tilde{v}_{n-h} & \mu\beta - \gamma \end{pmatrix},$$

(22) holds if and only if

$$z^T \, W \, z \quad \geq \quad 0 \quad \forall \, z \in \mathcal{R}^{n-h+1} \;\; \text{with } z_{n-h+1} = 1. \tag{23}$$

We claim that (23) holds if and only if $W \succeq 0$. Clearly, $W \succeq 0$ implies (23), and if there exists $z \in \mathcal{R}^{n-h+1}$ with $z^T W z < 0$ then, without loss of generality, $z_{n-h+1} \neq 0$, for if $z_{n-h+1} = 0$ then resetting $z_{n-h+1} = \epsilon$ (small) we will still have $z^T W z < 0$. Rescaling if necessary we will have $z_{n-h+1} = 1$ thus contradicting (23).
Thus, we want to choose $\mu \geq 0$ so that $W \succeq 0$, and this holds if and only if

$$\mu \, \leq \, \tilde{\lambda}_i, \quad 1 \leq i \leq n - h, \quad \text{and} \tag{24}$$

$$\mu\beta \, - \, \gamma \, - \sum_{i=1}^{n-h} \frac{\tilde{v}_i^2}{\tilde{\lambda}_i - \mu} \quad \geq \quad 0. \tag{25}$$

In other words, we want to choose $\mu \geq 0$ so as to maximize

$$\mu\beta - \sum_{i=1}^{n-h} \frac{\tilde{v}_i^2}{\tilde{\lambda}_i - \mu}, \tag{26}$$

subject to (24). This is a simple numerical task, since in the range $0 \leq \mu < \min_{1 \leq i \leq n-h} \tilde{\lambda}_i$ the expression in (26) is concave in μ.

3.1 Interpolation

Suppose we consider a family of problems of the form (19) corresponding to different vectors $\hat{x}$ of the form $\hat{x} = \hat{x}(\alpha) = p + \alpha c$, where $p, c \in \mathcal{R}^n$, $\alpha \in [\alpha^D, \alpha^U]$ and $\alpha^D \leq \alpha^U$ are given reals. We would like to obtain a lower bound on the value of problem (19) which is valid for every $\alpha \in [\alpha^D, \alpha^U]$.

Clearly, the quantities $\tilde{v}_i$ in eq. (21) are affine functions of α. We can therefore compute an upper bound on each $\tilde{v}_i^2$, valid on $[\alpha^D, \alpha^U]$. Replacing each $\tilde{v}_i^2$ by its upper bound in (26) we can proceed as above.

4 Generalized distances in the cardinality constraint case

Let $a \in \mathcal{R}^n$, $\theta \in \mathcal{R}^n$ with positive entries, and a_0 a scalar. Let $k < n$ be a positive integer, and let $\omega \in \mathcal{R}^n_+$. Here we consider the problem

$$\min \textstyle\sum_{j=1}^n \theta_j (x_j - \omega_j)^2, \quad \text{s.t.} \quad a^T x = a_0 \quad \text{and} \quad \|x\|_0 \leq k. \tag{27}$$

[Similar results hold for the version of the problem with cardinality constraint $\|x\|_0 = k$]. Problem (27) is a special case of the problem $\boldsymbol{\mathcal{F}}$ considered in this paper; consequently our general methodology applies. From an empirical perspective, these techniques produce extremely good bounds for problem (27); see Table 1 in Section 6. In our experiments, the perspective formulation used in [10] also does appear quite effective.

In this section, and Appendix A, we present a theoretical justification for why problem (27) should be "easy." We will assume, for conciseness, that $\theta > 0$, although the results extend to the general nonnegative case.

Using scaling and renaming, (27) becomes

$$\min \sum_{j=1}^n (x_j - \omega_j)^2, \quad \text{s.t.} \quad a^T x = a_0 \quad \text{and} \quad \|x\|_0 \leq k.$$

It is straightforward to show that this amounts to choosing $S \subseteq \{1, \ldots, n\}$ with $|S| \leq k$, so as to minimize

$$\frac{(a_0 - \sum_{j \in S} a_j \omega_j)^2}{\sum_{j \in S} a_j^2} \quad + \quad \sum_{j \notin S} \omega_j^2. \tag{28}$$

[We use the convention that $0/0 = 0$.] We consider three cases; we note that the first is in fact practical and we theorize that an appropriate version of the last two may prove likewise.

(a) $a_j = 1$ for all j (the unit simplex case). Here we simply let S consist of the indices of the k *largest* ω_j.

(b) Each a_j takes one of two values, say u^1 and u^2; let $U^j = \{1 \leq i \leq n \, : \, a_i = u^j\}$, for $j = 1, 2$. In this case we enumerate all possibilities for $|S \cap U^j|$, $j = 1, 2$. In each enumerated case it is easy to minimize (28) as in (a). In the general version of this idea, we would proceed in a way similar to that applied in [4] to the knapsack problem. First, we separately enumerate a case where we choose the largest a_i such that $i \in S$. For a given choice of that element let that value be a^*. Choose an accuracy parameter ϵ. We then partition the (smaller) a_i into those in the interval $(a^*(1-\epsilon), a^*]$, those in the interval $(a^*(1-\epsilon)^2, a^*(1-\epsilon)]$, and so on, until we reach those entries a_i of value smaller than ϵa^*. Altogether, for fixed ϵ, there are a fixed number of intervals. We can then proceed in a manner similar to (b), by enumerating each possible intersection pattern of S with each of the intervals. It can be shown that this gives rise an ϵ-approximation algorithm for minimizing the quantity in (28).

(c) In Appendix A we prove the following result. Assume that $a \geq 0$ and $\omega \geq 0$, and let $0 < \epsilon < 1$. Let D^{min} be the value of problem (27). Then we can compute a vector $\hat{x}$ with $\sum_j a_j \hat{x}_j = a_0$ and $\|\hat{x}\|_0 \leq K$, and such that

$$\sum_{j=1}^{n} \theta_j (\hat{x}_j - \omega_j)^2 \; \leq \; (1+\epsilon) D^{min},$$

in time at most $O(L^2(n^4K^2 + n^2K^4)\epsilon^{-4})$, where L is the total number of bits needed to represent ω and a.

4.1 Interpolation

Suppose we consider a family of problems of the form (27) corresponding to different vectors ω of the form $\omega = \omega(\alpha) = p + \alpha c$, where $p, c \in \mathcal{R}^n$, $\alpha \in [\alpha^D, \alpha^U]$ and $\alpha^D \leq \alpha^U$ are given reals. We would like to obtain a lower bound on the value of problem (27) which is valid for every $\alpha \in [\alpha^D, \alpha^U]$ and is tight when $\alpha^U - \alpha^D$ is small.

Denote by $d(S, \alpha)$ the quantity in (28), thus we want to lower bound $\min\{ \min_S d(S, \alpha) \, : \, \alpha^D \leq \alpha \leq \alpha^U \}$. For a given S, $d(S, \alpha)$ is a convex function of α; further its derivative is

$$d'(S,\alpha) \quad = \quad - \frac{2\left(a_0 - \sum_{j \in S} a_j (p_j + \alpha c_j)\right) \sum_{j \in S} a_j c_j}{\sum_{j \in S} a_j^2} \quad + \quad 2 \sum_{j \notin S} (p_j + \alpha c_j) c_j. \tag{29}$$

It is clear that for a given value of α we can compute a quantity $g(\alpha)$ such that $g(\alpha) \leq d'(S, \alpha)$ for every subset S. For example, the second term in (29) is lower bounded by the sum of the $n - k$ smallest values $2(p_j + \alpha c_j)c_j$, plus the sum of any additional $2(p_j + \alpha c_j)c_j$ that are negative. Then by convexity, for any α

$$d(S,\alpha) \; \geq \; d(S, \alpha^D) \, + \, g(\alpha^D)(\alpha - \alpha^D), \quad \text{and so}$$

$$\min_S d(S,\alpha) \; \geq \; \min_S d(S, \alpha^D) \, + \, g(\alpha^D)(\alpha - \alpha^D).$$

Note that the minima in this expression are precisely the values of problem (27) at $\omega = p + \alpha c$ and $\omega = p + \alpha^D c$, respectively. Finally,

$$\min_{\alpha^D \leq \alpha \leq \alpha^U} \min_S d(S,\alpha) \; \geq \; \min_S d(S, \alpha^D) \, + \min_{\alpha^D \leq \alpha \leq \alpha^U} \{ g(\alpha^D)(\alpha - \alpha^D) \}.$$

The quantity in the right-hand side is a lower bound with the properties we seek.

5 The overall procedure

Here we describe our implementation of the ideas described in the previous sections. As before, we consider a given convex relaxation, i.e. a convex set Ω such that $\mathcal{P} \cap \mathcal{K} \subseteq \Omega$, and we have already computed $x^* = \text{argmin}\{F(x) \,:\, x \in \Omega\}$. Unlike the description in the introduction, here we assume that the vector c is arbitrary (e.g. not necessarily $\nabla F(x^*)$). Our implementation uses the cardinality constraint only in one specific context.

(a) Choose a subset (possibly empty) of constraints $A^= x = b^=$ which hold as equalities for the affine hull of the relaxation. Let r be the number of rows of $A^=$.

(b) Compute the *projection* of the quadratic $x^T Mx$ onto the null space of $A^=$.

(c) Choose an arbitrary vector c; if $r > 0$ the vector is required to satisfy $A^= c = 0$.

(d) Choose a *finite* set of values $\alpha(0) < \alpha(2) < \ldots \alpha(J)$ (zero among them).

For any real α, denote by H^α the $(n-r-1)$-dimensional hyperplane orthogonal to c and orthogonal to the rows of $A^=$, which contains the point $p^\alpha = x^* + \alpha c$. Then

(e) For $0 \le j < J$ compute a lower bound, denoted by $\tilde{L}(j)$, on

$$\min_{\alpha(j) \le \alpha \le \alpha(j+1)} \min\{F(x) \,:\, \|x - p\|_2 \ge D(p^\alpha),\, x \in H^\alpha\}.$$

(f) Write

$$\begin{aligned} \tilde{U} &= \min\{F(x) \,:\, x \in \Omega,\, c^T x \ge \alpha(J)\|c\|_2^2\,\}, \\ \tilde{D} &= \min\{F(x) \,:\, x \in \Omega,\, c^T x \le \alpha(0)\|c\|_2^2\,\}, \end{aligned}$$

We **output**

$$\min\left\{ \min_{0 \le j < J} \tilde{L}(j) \;,\; \tilde{L} \;,\; \tilde{U} \right\}$$

Notes: step (e) is performed using the interpolation results in Sections 3.1 and 4.1. When $c = \nabla F(x^*)$ step (2) can be skipped if we choose $\alpha(0) = 0$.

6 Experiments

In our numerical experiments we considered the cardinality-constrained convex quadratic programming problem over the unit simplex. In our application of the S-lemma we always used the Euclidean ball as the bounding quadratic. In our experiments we considered, for our choice for c,

- The gradient at x^*,
- The ray from x^* to its nearest feasible neighbor (i.e. the nearest nonnegative, point on the unit simplex with at most K positive coordinates).
- An eigenvector corresponding to the minimum eigenvalue.

All three choices prove effective though the last one dominated; however more experiments are needed to decide conclusively if any is best. As we will see, however, the linear algebra is always quite fast and little is lost by trying all three variants and keeping the best bound.

We chose $\alpha(J)$ to be one-tenth as large as the maximum α such that the hyperplane H^{α} intersects the unit simplex. In general $\alpha(1)$ may be negative (this is almost always the case except when c is the gradient of F at x^*) and a similar criterion was applied in that case. Finally, we set $J = 100$, and the quantities $\alpha(j)$ equally spaced between $\alpha(1)$ and $\alpha(100)$.

In table 1 we consider examples involving separable quadratics – the coefficients of the quadratics where chosen randomly. Here, the row labeled "rMIPQP" indicates the value of the continuous relaxation of the standard MIP formulation, whereas "CpxLB50K" indicates the lower bound proved by Cplex [6] after 50K nodes of branch-and-cut; "UB" indicates the value of the best solution found by Cplex, and SLELB shows the lower bound proved by our procedure.

	n = 1000, K = 50		**n = 100, K = 50**	
rMIPQP	0.00124	0.00141	0.01229	0.01395
CpxLB50K	0.00124	0.00141	0.01410	0.01414
SLELB	0.01981	0.01969	0.02037	0.02108
CpxUB	0.02026	0.02047	0.02238	0.02416

Table 1: ***Separable quadratic examples***

Table 2 considers problems with non-separable quadratics. Here we consider a problem with $n = 2443$. The eigenvalue structure is real; however the obtained the eigenvectors by applying 5000 random rotations to an identity matrix, as a form of stress testing for our routines. In particular, the quadratic is far from separable.

K	**rQMIP**	**PRSP**	**SLE**	**UB**	**rQMIP**	**PRSP**	**SLE**
	val	val	val		sec	sec	sec
100	0.031	0.0466	0.0830	0.24	16.25	121.84	1.57
90	0.031	0.0485	0.0905	?	15.77	104.11	1.36
80	0.031	0.0509	0.1001	?	15.66	105.89	1.66
70	0.031	0.0540	0.1141	?	15.77	100.56	1.71
60	0.031	0.0581	0.1324	?	15.64	110.78	1.73
50	0.031	0.0638	0.1570	?	15.80	111.28	1.72
40	0.031	0.0725	0.1975	?	15.69	104.20	1.72

Table 2: ***Non-separable quadratic examples***

In this table "rQMIP" refers to the continuous relaxation of the (standard) mixed-integer programming formulation for the problem which we described above (equations (6)-(8)). "PRSP" refers to the *perspective* formulation; a conic formulation (see [9], [1], [10]). Finally, "SLE" is our approach. Column "UB" describes upper bounds to the problems obtained by running the binary QMIP formulation.

As we can see from the table, the QMIP formulation is quite weak – its relaxation has the same value, regardless of K. The perspective formulation is clearly stronger but the SLE is stronger still. At the same time, our approach is quite cheap from a computational standpoint.

A salient question that is raised by these experiments is whether the bound obtained by the S-lemma can be "naturally" embedded into a convex programming formulation for problem $\mathcal{F}$.

7 Continuing work

There are many ways in which our approach can be extended and improved. First of all, we are considering alternative quadratics as constraints (besides separable quadratics).

In particular, using a separable quadratic plus a sum of low-rank quadratic may be an effective way to approximate difficult eigenvalue distributions. Additionally, we are exploring the use of the more general version of the S-lemma, involving multiple quadratic constraints.

In Figure 4, point 1 is the initial solution to a relaxation. An application of the S-lemma yields the bound proved by point 2. The ball centered on point 2 represents the minimum distance from point 2 to the feasible region (of the non-convex program); we can now apply the S-lemma to lower bound the objective subject to being on the exterior of the union of both balls. This yields the bound proved by point 3. We stress that in the case of multiple quadratics the S-lemma is weaker; generally speaking such problems are viewed as much harder.

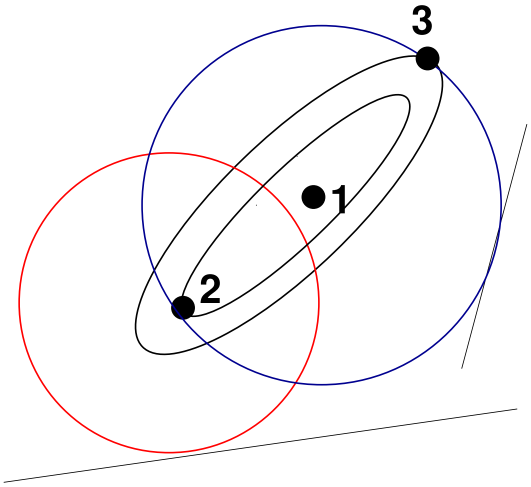

Figure 4: Using multiple quadratics

From a technical perspective, the positive-semidefiniteness argument technique culminating in equation (26) needs to be updated so as to handle multiple parameters μ (one corresponding to each quadratic constraint); this is somewhat more complex than the single parameter case but early experimentation is promising.

We also point out that our approach suggests a form of nonstandard branching. In Section 5 we described the hyperplanes $H^{\alpha}(j)$ and a technique for obtaining a bound in the "sandwich" region between $H^{\alpha(j)}$ and $H^{\alpha(j+1)}$. We could also branch: in branch j, $1 \leq j < J$ we branch by constraining the feasible points to lie in this region – and in each branch we now choose a vector c orthogonal to that used to define the $H^{\alpha(j)}$.

Finally, we are primarily interested in very large scale instances, and cases where we use a quadratic to (locally) approximate a convex (but not quadratic) objective function F; the speed of the linear algebra techniques as evidenced by Table 2 is encouraging in this context.

Acknowledgement. I thank Prof. C. van Loan for pointing out the material in Section 12.6 of [8]. I also thank Paul Tseng, Kurt Anstreicher, Sam Burer and Gabor Pátaki for useful comments.

References

[1] M. Aktürk, A. Atamtürk and S. Gürel, A strong conic quadratic reformulation for machine-job assignment with controllable processing times, *O.R. Letters* **37**, 187 – 191 (2009).

[2] A. Ben-Tal and A. Nemirovsky, em Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications (2001) MPS-SIAM Series on Optimization, SIAM, Philadelphia, PA.

[3] D. Bienstock, Computational study of a family of mixed-integer quadratic programming problems,*Math. Programming* **74** (1996), 121 – 140.

[4] D. Bienstock and B. McClosky, Tightening simple mixed-integer sets with guaranteed bounds (2008). Submitted.

[5] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory* (1994). SIAM, Philadelphia, PA.

[6] ILOG CPLEX 11.0. ILOG, Inc., Incline Village, NV.

[7] G.H. Golub, Some modified matrix eigenvalue problems, *SIAM Review* **15** (1973), 318 – 334.

[8] G.H. Golub and C. van Loan, *Matrix Computations.* Johns Hopkins University Press (1996).

[9] A. Frangioni and C. Gentile, Perspective cuts for a class of convex 0-1 mixed integer programs, *Mathematical Programming* **106**, 225 – 236 (2006).

[10] O. Günlük and J. Linderoth, Perspective Reformulations of Mixed Integer Nonlinear Programs with Indicator Variables, Optimization Tech. Report, U. of Wisconsin-Madison (2008).

[11] M. Magazine.

[12] I. Pólik and T. Terlaky, A survey of the S-lemma, *SIAM Review* **49** (2007), 371 – 418.

[13] V. A. Yakubovich, S-procedure in nonlinear control theory, Vestnik Leningrad University, 1 (1971) 62–777.

[14] F. Rendl, H. Wolkowicz, A semidefinite framework for trust region subproblems with applications to large scale minimization, *Math. Program* **77** (1997) 273 – 299.

[15] R. J. Stern and H. Wolkowicz, Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations, *SIAM J. Optim.* **5** (1995) 286 – 313.

[16] J. Sturm and S. Zhang, On cones of nonnegative quadratic functions, *Mathematics of Operations Research* **28** (2003), 246 – 267.

[17] Y. Ye and S. Zhang, New results on quadratic minimization, *SIAM J. Optim.* **14** (2003), 245 – 267.

A Appendix A

A.1 Combinatorial approximation algorithms

Here we consider the problem

$$D^{min} \;=\; \min \sum_{j=1}^{n} (x_j - x^*_j)^2, \quad \text{s.t.} \;\; a^T x = a_0 \;\; \text{and} \;\; \|x\|_0 \le K, \tag{30}$$

where $K > 0$ is an integer, and $a, x^* \in \mathcal{R}^n$. We claim that

$$D^{min} \;=\; \min \left\{ \frac{(a_0 - \sum_{j \in S} a_j x^*_j)^2}{\sum_{j \in S} a_j^2} \;\; + \;\; \sum_{j \notin S} x^{*2}_j \;:\; S \subseteq \{1, \ldots, n\} \text{ and } |S| \le K \right\}, \tag{31}$$

where we use the convention that $0/0 = 0$. To see that this is the case, let $\hat{x}$ be any feasible solution to (30). Let $S \subseteq \{1, \ldots, n\}$ with $|S| \le K$ be such that $\hat{x}_j = 0$ for all $j \notin S$, and define $y \in \mathcal{R}^n$ by $y_j = x^*_j$ for $j \in S$ and otherwise $y_j = 0$. Finally, we write $H = \left\{ x \in \mathcal{R}^n \,:\, a^T x = a_0 \right\}$. There are two cases:

(a) $a_j = 0$ for all $j \in S$. In this case we must have $a_0 = 0$. Furthermore,

$$\|\hat{x} - x^*\|_2^2 = \sum_{j \notin S} x_j^{*2} + \sum_{j \in S} (x_j^* - \hat{x}_j)^2 \geq \sum_{j \notin S} x_j^{*2} = \frac{(a_0 - \sum_{j \in S} a_j x_j^*)^2}{\sum_{j \in S} a_j^2} + \sum_{j \notin S} x_j^{*2} = \|y - x^*\|_2^2;$$

note that y is feasible for (30).

(b) Suppose $a_j \neq 0$ for at least index $j \in S$. The point $v \in H$ with $v_j = 0$ for all $j \notin S$, and closest to y, clearly has coordinates of the form

$$\begin{aligned} v_j &= y_j + t a_j, \quad j \in S \\ &= 0, \quad \text{otherwise}, \end{aligned} \tag{32}$$

for some scalar t; substituting, we obtain

$$t = \frac{a_0 - \sum_{j \in S} a_j y_j}{\sum_{j \in S} a_j^2} = \frac{a_0 - \sum_{j \in S} a_j x_j^*}{\sum_{j \in S} a_j^2}. \tag{33}$$

Note that $(x^* - y)^T (v - y) = 0$. Thus, we have

$$\|x^* - \hat{x}\|_2^2 = (x^* - y + y - \hat{x})^T (x^* - y + y - \hat{x}) = \|x^* - y\|_2^2 + \|y - \hat{x}\|_2^2 \geq \|x^* - y\|_2^2 + \|y - v\|_2^2 = \|x^* - v\|_2^2,$$

by definition of v. Wsing the value for t as in (33), we have

$$\|\hat{x} - x^*\|_2^2 \geq \sum_{j \notin S} x_j^{*2} + t^2 \sum_{j \in S} a_j^2 = \sum_{j \notin S} x_j^{*2} + \frac{(a_0 - \sum_{j \in S} a_j x_j^*)^2}{\sum_{j \in S} a_j^2}.$$

In summary, in both cases (a) and (b) we have that the right-hand side in (28) is a lower bound for D^{min}. The converse statement is similarly proved.

We next show how to approximate the minimum in (28). In order to simplify the discussion, we will assume a number of restrictions; the general case can be handled with a somewhat more complex construction. We will show that, given $0 < \epsilon < 1$, the quantity D^{min} can be approximated with a multiplicative error of $(1+\epsilon)$ in time $O(L^2(n^4K^2 + n^2K^4)\epsilon^{-4})$, where L is the total number of bits needed to represent x^* and a.

Restrictions:
(1) We assume that $\sum_j a_j x_j^* = a_0$; thus for any $S \subseteq \{1, \cdots, n\}$,

$$a_0 - \sum_{j \in S} a_j x_j^* = \sum_{j \notin S} a_j x_j^*.$$

(2) For $a_j \geq 0$ and $x_j^* \geq 0$, for all j.

Now we turn to the approximation scheme. Suppose first that $a_0 = 0$. Then a candidate for D^{min} is

$$\min \left\{ \sum_{j \notin S} x_j^{*2} \; : \; |S| \leq K, \; a_j = 0 \; \forall \; j \in S \right\},$$

which is easily computed. In the remainder we will assume that we want to approximate the right-hand side of (28) with the added stipulation that $a_j \neq 0$ for at least one index $j \in S$. Likewise, suppose that the set $Z = \{j \; : \; a_j x_j^* = 0\}$ has cardinality at least $n - K$. Then, the sum of the $n - K$ smallest x_j^{*2}, over $j \in Z$, is a candidate for D^{min}.

In what follows, therefore, we will approximately compute the minimum value $\frac{(\sum_{j \notin S} a_j x_j^*)^2}{\sum_{j \in S} a_j^2} + \sum_{j \notin S} x_j^{*2}$ over sets S with $|S| \leq K$, $a_j > 0$ for at least one index $j \in S$, and $a_j x_j^* > 0$ for at least one index $j \notin S$ – such a set S will be called *legal*.

Now, suppose that $|\{j \,:\, x^*_j = 0\}| \geq K$. Then, clearly, $D^{min} = 0$, so we will assume that fewer than K indices j are such that $x^*_j = 0$.

Denote by $0 < X^{min}$ (resp., X^{max}) the sum of the K smallest (resp., largest) x^*_j. Denote by P^{min} the smallest positive value $a_j x^*_j$, and by P^{max} the sum of the largest $n-K$ values $a_j x^*_j$. Let $0 < \epsilon < 1$. The above discussion implies that for any legal set S, without loss of generality,

$$X^{min} \;\leq\; \sum_{j \notin S} x^{*2}_j \;\leq\; X^{max}.$$

Thus, defining

$$H \;=\; \frac{\log X^{max} - \log X^{min}}{\log(1+\epsilon)} \;=\; O\left(\frac{\log X^{max} - \log X^{min}}{\epsilon}\right) \quad \text{(for ϵ small enough)}, \tag{34}$$

for any legal S there exists an integer $0 \leq h < H$ such that

$$X^{min}(1+\epsilon)^h \;\leq\; \sum_{j \notin S} x^{*2}_j \;\leq\; X^{min}(1+\epsilon)^{h+1}. \tag{35}$$

Note: the quantity H is polynomial on the size of the input data. Similarly, defining

$$Q \;=\; \frac{\log P^{max} - \log P^{min}}{\log(1+\epsilon)} \;=\; O\left(\frac{\log P^{max} - \log P^{min}}{\epsilon}\right) \tag{36}$$

for any legal S there exists an integer $0 \leq q < Q$ such that

$$P^{min}(1+\epsilon)^q \;\leq\; \sum_{j \notin S} a_j x^*_j \;\leq\; P^{min}(1+\epsilon)^{q+1}; \tag{37}$$

again the quantity Q is polynomially sized. (35) and (37) suggest the following enumerational procedure. For $0 \leq h < H$ and $0 \leq q < Q$, define:

$$\begin{aligned}
\boldsymbol{\mathcal{P}(h,q)}\textbf{ :} \quad v(h,q) \;&=\; \max_S \sum_{j \in S} a^2_j && (38)\\
\text{s.t.} \quad \sum_{j \notin S} x^{*2}_j \;&\leq\; X^{min}(1+\epsilon)^{h+1} && (39)\\
\sum_{j \notin S} a_j x^*_j \;&\leq\; P^{min}(1+\epsilon)^{q+1} && (40)\\
S \;&\; \text{legal}. && (41)
\end{aligned}$$

Then, we have:

Lemma A.1

$$\min_{h,q}\left\{ X^{min}(1+\epsilon)^{h+1} \;+\; \frac{(P^{min}(1+\epsilon)^{q+1})^2}{v(h,q)} \right\} \;\leq\; (1+O(\epsilon))D^{min}. \;\blacksquare$$

We will be interested in an approximate version of problem $\boldsymbol{\mathcal{P}(h,q)}$ – rather than approximating the obje‘ctive (38) we will instead produce superoptimal solutions that slightly violate (39) and (40). We will see that this will enable us to efficiently approximate D^{min}. To this effect, let us denote by $w(h,q)$ the quantity inside the brackets in the statement of Lemma (A.1). We have:

Lemma A.2 *Suppose that for every* $0 \le h < H$ *and* $0 \le q < Q$*, there is a set* $\bar{S}(h,q)$ *such that*

$$v(h,q) \ \le \sum_{j \in \bar{S}(h,q)} a_j^2, \tag{42}$$

$$\sum_{j \notin \bar{S}(h,q)} x_j^{*2} \ \le \ X^{min}(1+\epsilon)^{h+2}, \quad and \tag{43}$$

$$\sum_{j \notin \bar{S}(h,q)} a_j x_j^* \ \le \ P^{min}(1+\epsilon)^{q+2}. \tag{44}$$

Then, writing $w(h,q) = \sum_{j \in \bar{S}(h,q)} a_j^2$*, we have:*

$$\begin{aligned} (1+\epsilon)^{-2} \min_{h,q} \left(X^{min}(1+\epsilon)^{h+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q+1})^2}{v(h,q)} \right) \quad &\le \\ \min_{h,q} \left(X^{min}(1+\epsilon)^{h+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q+1})^2}{w(h,q)} \right) \quad &\le \\ \min_{h,q} \left(X^{min}(1+\epsilon)^{h+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q+1})^2}{v(h,q)} \right). & \end{aligned}$$

Proof. Since, for $0 \le h < H$ and $0 \le q < Q$, $w(h,q) \ge v(h,q)$ the second inequality easily follows. For the other, let $0 \le h^* < H$ and $0 \le q^* < Q$ be such that

$$X^{min}(1+\epsilon)^{h^*+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q^*+1})^2}{w(h^*,q^*)} \ = \ \min_{h,q} \left\{ X^{min}(1+\epsilon)^{h+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q+1})^2}{w(h,q)} \right\}.$$

Note that by definition of H, when $h^* = H-1$ we can tighten the right-hand side of (43) to $X^{min}(1+\epsilon)^{h^*+1}$. Similarly when $q^* = Q-1$. Hence, when $h^* = H-1$ and $q^* = Q-1$ we have that $\bar{S}(h^*,q^*)$ is optimal for (38)-(41); i.e. $w(h^*,q^*) = v(h^*,q^*)$. In what follows we assume that $h^* < H-1$ or $q^* < Q-1$.

In that case, writing $h = \min\{h^*+1, H-1\}$ and $q = \min\{q^*+1, Q-1\}$ we have that $\bar{S}(h^*,q^*)$ is certainly feasible for $\boldsymbol{\mathcal{P}(h,q)}$, and therefore $w(h^*,q^*) \le v(h,q)$, and thus

$$\begin{aligned} X^{min}(1+\epsilon)^{h^*+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q^*+1})^2}{w(h^*,q^*)} \quad &\ge \quad X^{min}(1+\epsilon)^{h^*+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q^*+1})^2}{v(h,q)} \\ &\ge \quad (1+\epsilon)^{-2} \left(X^{min}(1+\epsilon)^{h+1} \ + \ \frac{(P^{min}(1+\epsilon)^{q+1})^2}{v(h,q)} \right). \end{aligned}$$

■

Next, we turn to the (approximate) computation of the $v(h,q)$. We would like to simplify this problem so as to remove constraint (41). Note that if the value of problem (38)-(40) is positive then the optimal set satisfies one of the two requirements for legality ($a_j > 0$ for at least one index $j \in S$). And if not, then there is no legal set S satisfying (39)-(40). Consequently, we only need to guarantee that $a_j x_j^* > 0$ for at least one $j \notin S$. A simple expedient to obtain this guarantee is to *enumerate* every possible index j with this property. In summary, we have:

Corollary A.3 *The computation of* $v(h,q)$ *can be reduced to at most n problems of the form*

$$\max \sum_j a_j^2 z_j \tag{45}$$

$$s.t. \qquad \sum_j x_j^{*2}(1-z_j) \ \le \ r_1, \quad \sum_{j \notin S} a_j x_j^*(1-z_j) \ \le \ r_2 \tag{46}$$

$$\sum_j z_j \ \le \ K, \quad z_j \ = \ 0 \ or \ 1, \ \forall \, j \tag{47}$$

for appropriate values r_1, r_2*.*

This corollary handles the *exact* version of $\boldsymbol{\mathcal{P}(h, q)}$, but it is easily adaptable to the approximation version we seek, as follows. Given h and q, if for any index j we have $a_j > P^{min}(1+\epsilon)^{q+1}$, then in problem $\boldsymbol{\mathcal{P}(h, q)}$ we must have $j \in S$ (and, likewise, if $X^{min}(1+\epsilon)^{h+1} > 1$ and $K < n$ then $\boldsymbol{\mathcal{P}(h, q)}$ is infeasible). This means that the quantities r_1, r_2 in (46) are nonnegative. Similarly, we can assume that r_1 and r_2 are always at least as large as any of the left-hand side coefficients in in (46). As a consequence, we have that if we were to compute, for each problem (45)-(47), a solution which

- is feasible for (47) and superoptimal, and
- satisfies the constraints in (46) using right-hand sides $r_1(1+\epsilon)$ and $r_2(1+\epsilon)$, respectively,

then we will obtain a set $\bar{S}(h, q)$ as in Lemma A.2. In what follows, therefore, we will focus on problems of the form (45)-(47) from a superoptimality/approximate feasibility perspective. We will show that each such task can be carried out in time $O((n^3K^2+nK^4)\epsilon^{-2})$; thus, altogether, we obtain an $(1+O(\epsilon))$-approximation to D^{min} in time $O(HQ(n^4K^2+n^2K^4)\epsilon^{-2})$; using our definition of H and Q this is $O(L^2(n^4K^2+n^2K^4)\epsilon^{-4})$, where L is the total number of bits needed to represent x^* and a .

Thus, we consider the problem

$$\boldsymbol{w(K)} \ = \ \max \quad \sum_{j=1}^{n} w_j z_j, \tag{48}$$

$$s.t. \quad \sum_{j=1}^{n} a_{ij}(1-z_j) \ \le \ a_{i0}, \quad i \ = \ 1,2 \tag{49}$$

$$\sum_{j=1}^{n} z_j \ \le \ K, \quad z \in \{0,1\}^n. \tag{50}$$

Here, the vectors a_1, a_2 are integral (and here we allow them to take negative values) and $K \le n$ is an integer. Without the cardinality constraint (50) this problem is closely related to the so-called 2-knapsack problem studied by Magazine [11] (essentially, they are equivalent). In [11] it is shown that no FPTAS exists for the 2-knapsack problem unless P = NP. The proof in [11] can be adapted to show the same result for problem (48)-(50). Our approach to (48)-(50) does not yield a FPTAS – it yields superoptimal but slightly infeasible solutions, as per the discussion above.

The following constructions likely (we assume) amount to folklore, but we include them for completeness.

Lemma A.4 *Let* $0 < \epsilon < 1$. *There is an algorithm of complexity* $O((n^3K^2+nK^4)\epsilon^{-2})$ *which produces a vector* $\hat{z} \in \{0,1\}^n$ *with* $\sum_{j=1}^{n} z_j \le K$ *and such that*

(i) $\sum_j w_j \hat{z}_j \ \ge \ w(K)$,

(ii) $\sum_j a_{ij}\hat{z}_j \ \le \ a_{i0} \ + \ \epsilon \max_j |a_{ij}|$, *for* $i = 1, 2$.

Proof. To prove the result we consider a closely related problem, where constraint (50) is replaced with

$$\sum_{j=1}^{n} z_j \ = \ J, \quad z \in \{0,1\}^n. \tag{51}$$

where $0 \le J \le K$ is an integer. Denote by $\boldsymbol{\mathcal{J}_2(J)}$ the problem consisting of (48), (49) and (51), and let $\boldsymbol{v(J)}$ denote its value. Given such a problem, and given $0 < \epsilon < 1$, we modify it as follows (where $\alpha > 0$ is a parameter to be fixed later).

- Without loss of generality, assume that $\max_j |a_{1j}| = \max_j |a_{2j}| = M$, say.
- Let P be the largest integer such that $2^P \le M$. Choose integer p smallest such that $2^{P-p} \le n^{\alpha}$.
- For $i = 1, 2$, and $0 \le j \le n$, define $\hat{a}_{ij} = \lfloor a_{ij}/2^p \rfloor$.

Using these constructions, consider the problem:

$$\begin{aligned}
\boldsymbol{v'(K)} \;=\; \max \quad & \sum_{j=1}^{n} w_j z_j, && (52)\\
s.t. \quad & \sum_{j=1}^{n} \hat{a}_{ij}(1 - z_j) \;\leq\; \hat{a}_{i0}, \quad i \;=\; 1,2 && (53)\\
& \sum_{j=1}^{n} z_j \;=\; J, \qquad z \,\in\, \{0,1\}^n. && (54)
\end{aligned}$$

This problem can be solved in time $O(J\, n^{1+2\alpha})$ (using dynamic programming) since the choice of p guarantees that $\hat{a}_{ij} \leq 2n^{\alpha}$ for all i and j. Let $\hat{z}$ be an optimal solution. Since, for all i and j, $a_{ij} \leq 2^p \hat{a}_{ij} + 2^p$, it follows that for $i = 1, 2$,

$$\sum_{j=1}^{n} a_{ij}\hat{z}_j \;\leq\; a_{i0} \,+\, 2^p(n-J) \;\leq\; a_{i0} \,+\, \frac{2^{P+1}}{n^{\alpha}}(n-J) \;\leq\; a_{i0} + \frac{2(n-J)M}{n^{\alpha}}, \tag{55}$$

by definition of M and P. Now suppose we choose α such that $n^{\alpha} = 2(n-J)/\epsilon$. Then $\hat{z}$ satisfies (ii), and the complexity of the algorithm is $O(J(n-J)^2 n/\epsilon^2)$. To prove (i) note that any feasible solution for (48)-(50) is feasible for (52)-(54). ■

A Appendix B

A.1 NP-hardness

Here we outline a proof that a problem of the form

$$\min\left\{g(x) \,:\, Cx \geq 0,\; x^T x \geq \theta\right\} \tag{56}$$

where $g : \mathcal{R}^n \rightarrow \mathcal{R}$ is a positive-semidefinite quadratic, C is a matrix and $\theta > 0$, is strongly NP-hard. We will show that the Vertex Cover problem (given an undirected graph, find a minimum-cardinality set of vertices that meet all edges) can be reduced to (56).

To this effect, let G be a graph with vertex-set $\{1, 2, \ldots, n\}$ and edge-set E. Let $M = n^5$. Consider the problem with variables $x = (x_0, x_1, \ldots, x_n)$:

$$\begin{aligned}
g^* = \quad \min \quad & M(x_0 - 1)^2 \;+\; \sum_{i=1}^{n} x_i && (57)\\
s.t. \quad & x_i \,+\, x_j \;\geq\; 0 \quad\; \forall\,\{i,j\} \,\in\, E && (58)\\
& x_i - x_0 \;\leq\; 0, \;\; -x_i - x_0 \;\leq\; 0, \qquad i = 1, \ldots, n, \quad x_0 \geq 0, && (59)\\
& x_0 \geq 0, && (60)\\
& \sum_{j=0}^{n} x_j^2 \;\geq\; n+1. && (61)
\end{aligned}$$

We have:

Lemma A.1 *Let S^* be a minimum cardinality vertex cover of G. Then $\lceil g^* \rceil = 2|S^*| - n$.*

Proof. Consider the vector defined by $\hat{x}_0 = 1$ and, for $j \geq 1$, $\hat{x}_j = 1$ if $j \in S^*$ and $\hat{x}_j = -1$ otherwise. This vector is feasible for (58)-(61) and has objective value $2|S^*| - n$.

For the opposite direction, let $\tilde{x}$ be an optimal solution to (57)-(60). Write $\tilde{x}_0 = 1 + \epsilon$. Since $M\epsilon^2 \leq 2|S^*| - n \leq n$ it follows that $\epsilon \leq n^{-2}$. (59), (60) imply $(n+1)(1+\epsilon)^2 \geq n+1$, so $\epsilon \geq 0$. Let $\mu = \min_{1`\leq i \leq n} |\tilde{x}_i|$. Using (61) we have

$$n + 1 \;\leq\; n(1+\epsilon)^2 \;+\; \mu,$$

so $\mu \geq 1 - 2n\epsilon$ (for n large enough). By (58) it follows that the set $S = \{j : \tilde{x}_j \geq 1 - 2n\epsilon\}$ is a vertex cover. Moreover, the objective value of $\tilde{x}$ is at least

$$\begin{aligned} & M\epsilon^2 + |S|(1 - 2n\epsilon) - (n - |S|)(1 + \epsilon) & (62) \\ \geq \quad & 2|S| - n - 3n\epsilon \geq 2|S^*| - n - 3n\epsilon \;>\; 2|S^*| - n - 1. & (63) \end{aligned}$$

■

Note: the above proof is easily modified so that the objective is positive-definite.