Eigenvalue techniques for convex objective, nonconvex optimization problems

Daniel Bienstock, Columbia University, New York

November, 2009

Abstract

Consider a minimization problem given by a nonlinear, convex objective function over a nonconvex feasible region. Traditional optimization approaches will frequently encounter a fundamental difficulty when dealing with such problems: even if we can efficiently optimize over the convex hull of the feasible region, the optimum will likely lie in the interior of a high dimensional face, "far away" from any feasible point. As a result (and in particular, because of the nonconvex objective) the lower bound provided by a convex relaxation will typically be extremely poor. Furthermore, we will tend to see very large branch-and-bound (or -cut) trees with little or no improvement over the lower bound.

In this work we present theory and implementation for an approach that relies on three ingredients: (a) the S-lemma, a major tool in convex analysis (b) efficient projection of quadratics to lower dimensional hyperplanes, and (c) efficient computation of combinatorial bounds for the minimum distance from a given point to the feasible set, in the case of several signficant optimization problems.

Altogether, our approach strongly improves lower bounds at a small computational cost, even in very large examples.

1 Introduction

We consider problems with the general form

$$(\mathcal{F}): \qquad \bar{F} := \min\ F(x), \tag{1}$$
$$s.t. \qquad x \in \mathcal{P}, \tag{2}$$
$$x \in \mathcal{K}. \tag{3}$$

Here,

- $F(x)$ is a convex function; in this abstract, 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".
- In the applications we consider, n can be quite large.

We assume that a given convex relaxation of the set described by (2), (3) is under consideration. For example, when dealing with a particularly complex set $\mathcal{K}$, this may be the *only* sensible relaxation that we are able to produce. Or, it may be the only *computationally practicable* relaxation we have, especially in the case of large n. Or, as it is often in the case of practice, the relaxation is one from which it is easy to quickly produce (perhaps, good) *heuristic* solutions to problem $\mathcal{F}$.

Regardless of the case, one common fundamental difficulty is likely to be encountered: because of the convexity of F, the optimum solution to a convex relaxation will frequently be attained in the interior of a high-dimentional face of the relaxation, and far from the set $\mathcal{K}$. Thus, the lower bound proved by the relaxation will be weak (often, very weak) compared to $\bar{F}$. What is more, if one were to rely on branch-and-cut (the major tool of modern mixed-integer programming) the proved lower bound may improve little if at all when n is large, even after massive amounts of branching and extremely long computational time.

This *stalling* of the lower bounding procedure is commonly encountered in practice and constitutes a significant challenge, the primary subject of our study. We present a body of techniques that are designed to alleviate this difficulty. After obtaining the solution x^* to the given relaxation for problem $\mathcal{F}$, our methods will use techniques of convex analysis, of eigenvalue optimization, and combinatorial estimations, in order to quickly obtain a valid lower on $\bar{F}$ which is strictly larger (often, significantly so) than $F(x^*)$.

We will describe an important class of problems where our method, applied to a "cheap" but weak formulation, produces bounds comparable to or better than those produced by much more sophisticated formulations, and at a small fraction of the computational cost.

To motivate our discussion, we introduce two significant examples.

Cardinality constrained optimization problems. Here, for some integer $0 < K \leq n$, $\mathcal{K} = \{\, x \in \mathcal{R}^n : \|x\|_0 \leq K \,\}$, where the zero-norm $\|v\|_0$ of a vector v is used to denote the number of nonzero entries of v. A classical example of such a constraint arises in portfolio optimization (see e.g. [2]) but modern applications involving this constraint arise in statistics, machine learning [12],

and, especially, in engineering and biology [18]. Problems related to *compressive sensing* have an explicit cardinality constraint (see www.dsp.ece.rice.edu/cs for material). Also see [7].

The simplest canonical example of problem $\mathcal{F}$ is as follows:

$$\bar{F} = \min\ F(x), \tag{4}$$

$$s.t. \quad \sum_j x_j = 1, \quad x \geq 0, \tag{5}$$

$$\|x\|_0 \leq K. \tag{6}$$

This example is significant because (a) one can show that this problem is strongly NP-hard, and (b) it does arise in practice, exactly as stated. Note that modulo a simple change in variables, the linear constraint in (5) is simply a stand-in for a general non-homogeneous linear equation with nonzero coefficients.

In spite of its difficulty, this example already incorporates the fundamental difficulty alluded to above: clearly, $conv\left\{x \in \mathcal{R}^n_+ : \sum_j x_j = 1,\ \|x\|_0 \leq K\right\} = \left\{x \in \mathcal{R}^n_+ : \sum_j x_j = 1\right\}$. In other words, from a convexity standpoint the cardinality constraint disappears. Moreover, if the quadratic in F is positive definite and dominates the linear term, then the minimizer of F over the unit simplex will be an interior point, i.e., a point with all coordinates positive (and in real-life examples, of similar orders of magnitude); whereas $K \ll n$ in practice.

Multi-term disjunctive inequalites. Consider vectors vectors $c^i \in \mathcal{R}^n$ and reals β^i $(1 \leq i \leq m)$. We are interested in a non-convex set of the form

$$\bigcap_{i=1}^{m} \mathcal{K}^i, \quad \text{where } \mathcal{K}^i = \{x \in \mathcal{R}^n : c^{iT}x \leq \beta^i \ \textbf{or} \ c^{iT}x \geq \beta^i + 1\}, \text{ for } 1 \leq i \leq m. \tag{7}$$

More generally, we are interested in (and our methodology applies to) cases where each $\mathcal{K}^i$ is a disjunction among multiple polyhedral sets. The classical *split cuts* [6] provide an example of disjunctions based on inequalities as in (7), with integral c^i and β^i. However, disjunctive inequalities naturally arise in numerous settings as valid strong statements on a combinatorial set (see e.g. [3], [4]). Note that in general, given a polyhedral set $\mathcal{P}$, the convex hull of the intersection of $\mathcal{P}$ and a set (7) will have exponentially many facets (even the case $m = 1$ is not trivial). Thus, most likely, one would work with a relaxation of $\mathcal{P} \cap \bigcap_{i=1}^m \mathcal{K}^i$. Moreover, in the typical application of disjunctions (or split cuts) the disjunctions are found sequentially as a result of a cutting procedure.

For simplicity, consider the case $m = 1$ – the very first disjunction is added because it is violated by the solution to an initial relaxation. But thinking about the resulting geometry in the case of a (strictly) convex $F(x)$, it is clear that, quite possibly, $\text{argmin}\{F(x) : x \in conv(\mathcal{P} \cap \mathcal{K}^1)\}$ will *still* violate the first disjunction. For $m > 1$ the likelihood of such a "stall" in the cutting procedure increases.

To the extent that disjunctive sets are a general-purpose technique for formulating combinatorial constraints, the methods in this paper apply to a wide variety of optimization problems, and should prove effective when the objective is strictly convex.

1.1 Techniques

Our methods embody two primary techniques:

(a) The S-lemma (see [19], also [1], [5], [14]). Let $f, g : \mathcal{R}^n \rightarrow \mathcal{R}$ 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 $\mu \geq 0$ such that $(f - \mu g)(x) \geq 0$ 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 and let α, β be reals. Then

$$\min\{p(x) : q(x) \geq \beta\} \geq \alpha, \quad \text{iff } \exists\, \mu \geq 0 \text{ s.t. } p(x) - \alpha - \mu q(x) + \mu\beta \geq 0 \ \ \forall\, x. \tag{8}$$

In other words, the minimization problem in (8) can be approached as a simultaneous search for two reals α and $\mu \geq 0$, with α largest possible such that the last inequality in (8) holds. The S-lemma is significant in that it provides a good characterization (i.e. polynomial-time) for a a usually non-convex optimization problem. See [13], [15], [16], [17], [20] and the references therein, in particular regarding the connection to the trust-region subproblem.

(b) Consider a given nonconvex set $\mathcal{K}$. We will assume, as a primitive, that (possibly after an appropriate change of coordinates), given a point $\hat{x} \in \mathcal{R}^n$, we can efficiently compute a strong (combinatorial) lower bound for the Euclidean distance between $\hat{x}$ and the nearest point in $\mathcal{P} \cap \mathcal{K}$. We will make this assumption more precise in Section 1.2, but we will show that this is indeed the case for the cardinality constrained case (see Section 1.4; in the full paper we will show how to do so for the multi-term disjunctive case). Roughly speaking, it is the "structure" of a set $\mathcal{K}$ of interest that makes the assumption possible. In the rest of this section we will denote by $D(\hat{x})$ our lower bound on the minimum distance from $\hat{x}$ to $\mathcal{P} \cap \mathcal{K}$.

We can put together (a) and (b) into a simple template for proving lower bounds for $\bar{F}$:

S.1 Compute an optimal solution x^* to the given relaxation to problem $\mathcal{F}$.
S.2 Next we obtain the quantity $D(x^*)$.
S.3 Finally, we apply the S-lemma as in (8), using $F(x)$ for $p(x)$, and (the exterior of) the ball centered at x^* with radius $D(x^*)$ for $q(x) - \beta$.

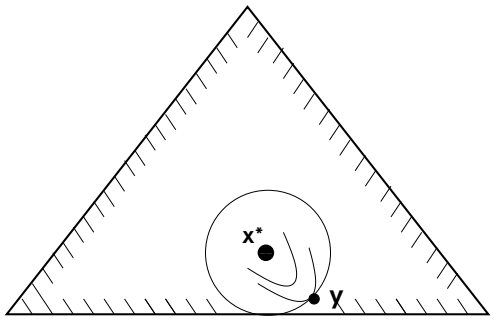

Figure 1: A simple case.

For a simple application of this template, consider Figure 1. This shows an instance of problem (4)-(6), with $n = 3$ and $K = 2$ where all coordinates of x^* are positive. The figure also assumes that $D(x^*)$ is exact – it equals the minimum distance from x^* to the feasible region. If we minimize $F(x)$, subject to being on the *exterior* of this ball (in other words, if we apply the S-lemma using $F(x)$ as the objective and the ball as the constraint) the optimum will be attained at y.

Thus, $F(y)$ is a valid lower bound on $\bar{F}$; we have $F(y) = F(x^*) + \tilde{\lambda}_1 R^2$, where R is the radius of the ball and $\tilde{\lambda}_1$ is the minimum eigenvalue of the *restriction* of $F(x)$ to the unit simplex. Note: $\tilde{\lambda}_1$ is lower bounded by the smallest eigenvalue of $F(x)$.

Now consider the example in Figure 2, corresponding to the case of a single disjunctive inequality. Here, x^F is the optimizer of $F(x)$ over the affine hull of the set $\mathcal{P}$. A straightforward application of the S-Lemma will yield as a lower bound (on $\bar{F}$) the value $F(y)$, which is weak – weaker, in fact, than $F(x^*)$. The problem is caused by the fact that x^F is not in the relative interior of the convex hull of the feasible region. In summary, a direct use of our template will not work.

1.2 Adapting the template

In order to correct the general form of the difficulty depicted by Figure 2 we would need to solve a problem of the form:

$$\mathcal{V} \ := \ \min\Big\{F(x) \ : \ x - x^* \in \mathcal{C}, \ (x - x^*)^T(x - x^*) \geq \delta^2\Big\} \tag{9}$$

where $\delta > 0$, and $\mathcal{C}$ is the cone of feasible directions (for $\mathcal{P}$) at x^*. We can view this as a 'cone constrained' version of the problem addressed by the S-Lemma. Clearly, $F(x^*) \leq \mathcal{V} \leq \bar{F}$ with the first inequality in general strict. If we are dealing with polyhedral sets, (9) becomes (after some renaming):

$$\min\Big\{F(\omega) \ : \ C\omega \geq 0, \ \omega^T\omega \geq \delta^2\Big\} \tag{10}$$

where C is an appropriate matrix. However, we have (proof in full paper):

Theorem 1.1 *Problem (10) is strongly NP-hard.* ■

We stress that the NP-hardness result is *not* simply a consequence of the nonconvex constraint in (10) – without the *linear* constraints, the problem becomes polynomially solvable (i.e., it is handled by the S-lemma, see the references).

To bypass this negative result, we will adopt a different approach. We assume that there is a positive-definite quadratic function $q(x)$ such that for any $y \in \mathcal{R}^n$, in polynomial time we can produce a (strong, combinatorial) lower bound $D^2_{min}(y, q)$ on the quantity

$$\min\{q(y - x) \ : \ x \in \mathcal{P} \cap \mathcal{K}\}.$$

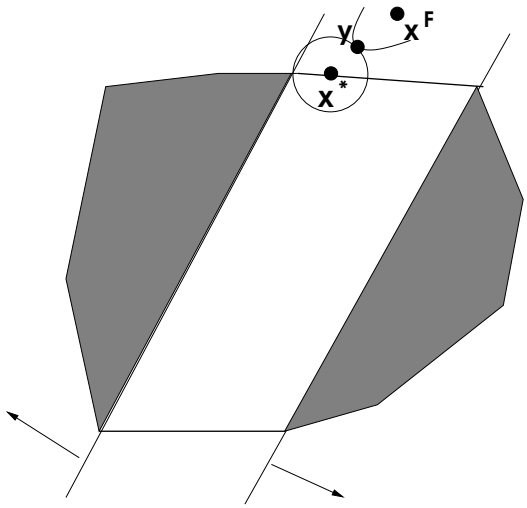

Figure 2: The simple template fails.

In Section 1.4 we will address how to produce the quadratic $q(x)$ and the value $D^2(y, q)$ when $\mathcal{K}$ is defined by a cardinality constraint (and a similar construct exists in the disjunctive inequalities case).

Let $c = \nabla F(x^*)$ (other choices for c discussed in full paper). Note that for any $x \in \mathcal{P} \cap \mathcal{K}$, $c^T(x - x^*) \geq 0$. For $\alpha \geq 0$, let $p^\alpha = x^* + \alpha c$, and let H^α be the hyperplane through p^α orthogonal to c. Finally, define

$$V(\alpha) \ := \ \min\{F(x) \ : \ q(x - p^\alpha) \geq D^2(p^\alpha, q), \ x \in H^\alpha\}, \tag{11}$$

and let y^α attain the minimum. Note: computing $V(\alpha)$ entails an application of the S-lemma, "restricted" to H^α. See Figure 3. Clearly, $V(\alpha) \leq \bar{F}$. Then

- Suppose $\alpha = 0$, i.e. $p^\alpha = x^*$. Then x^* is a minimizer of $F(x)$ subject to $x \in H^0$. Thus $V(0) > F(x^*)$ when F is positive-definite.

- Suppose $\alpha > 0$. Since $c^T(y^\alpha - x^*) > 0$, by convexity $V(\alpha) = F(y) > F(x^*)$.

Thus, $F(x^*) \leq \inf_{\alpha \geq 0} V(\alpha) \leq \bar{F}$; the first inequality being strict in the positive-definite case. [It can be shown that the "inf" is a "min"]. Each value $V(\alpha)$ incorporates combinatorial information (through the quantity $D^2(p^\alpha, q)$) and thus the computation of $\min_{\alpha \geq 0} V(\alpha)$ cannot be obtained through direct convex optimization techniques. As a counterpoint to Theorem 1.1, we have (proof in full paper):

Theorem 1.2 *In (10), if C has one row and $q(x) = \sum_j x_j^2$ then $\mathcal{V} \leq \inf_{\alpha \geq 0} V(\alpha)$.* ■

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:

Updated Template

1. For $0 \leq i < J$, compute a value $\tilde{V}(i) \leq \min\{ V(\alpha) : \alpha^{(i)} \leq \alpha \leq \alpha^{(i+1)} \}$.
2. Output $\min_{0 \leq i < J} \tilde{V}(i)$.

The idea here is that if (for all i) $\alpha^{(i+1)} - \alpha^{(i)}$ is small then $V(\alpha^{(i)}) \approx V(\alpha^{(i+1)})$. Thus the quantity output in (2) will closely approximate $\min_{\alpha \geq 0} V(\alpha)$.

In our implementation, we compute $\tilde{V}(i)$ by appropriately interpolating between $V(\alpha^{(i)})$ and $V(\alpha^{(i+1)})$ (details, full paper). Thus our approach reduces to computing quantities of the form $V(\alpha)$. We need a fast procedure for this task (since J may be large). Considering eq. (11) we see that this involves an application of the S-lemma, "restricted" to the hyperplane H^α. An efficient realization of this idea, which allows for additional leveraging of combinatorial information, is obtained by computing the *projection* of the quadratic $F(x)$ to H^α. This is the subject of the next section.

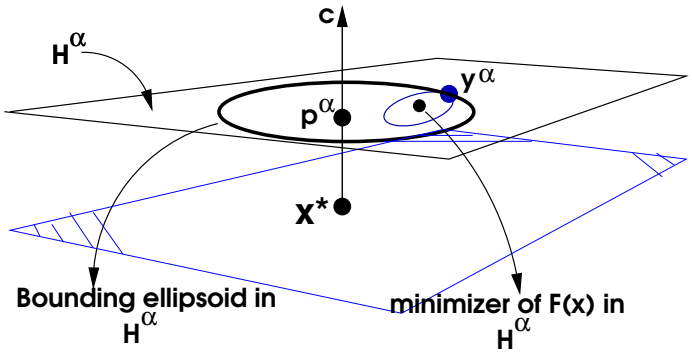

Figure 3: A better paradigm.

1.3 Projecting a quadratic

Let $M = Q\Lambda Q^T$ be a $n \times n$ matrix. Here the columns of Q are the eigenvectors of M and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ where the λ_i are the eigenvalues of M. We assume $\lambda_1 \leq \ldots \leq \lambda_n$. Let $c \neq 0$, denote $H = \left\{ x \in \mathcal{R}^n : c^T x = 0 \right\}$, and let P be the projection matrix onto H. In this section we describe an efficient algorithm for computing an eigenvalue-eigenvector decomposition of the "projected quadratic" PMP. Note that if $x \in H$, $x^T PMP x = x^T M x$. The vector c could be dense (is dense in important cases) and Q could also be dense. Our approach reverse engineers, and extends, results from [8] (also see Section 12.6 of [9] and references therein).

Clearly, c is an eigenvector of PMP (corresponding to eigenvalue 0). The remaining eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{n-1}$ are known to satisfy $\lambda_1 \leq \tilde{\lambda}_1 \leq \lambda_2 \leq \tilde{\lambda}_2 \leq \ldots \leq \lambda_{n-1} \leq \tilde{\lambda}_{n-1} \leq \lambda_n$.

Definition 1.3 *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.*

In (e.2) below we will use the convention $0/0 = 0$.

Lemma 1.4 *Let $\alpha_1 < \alpha_2 < \ldots < \alpha_q$ be the acute eigenvalues of M. Write $d = Q^T c$. Then, for $1 \le i \le q-1$,*

(e.1) The equation $\sum_{j=1}^{n} \frac{d_j^2}{\lambda_j - \lambda} = 0$ has a unique solution $\hat{\lambda}_i$ in (α_i, α_{i+1}).

(e.2) Let $w^i = Q(\lambda - \hat{\lambda}_i I)^{-1} d$. Then $c^T w^i = 0$ and $PMPw^i = \hat{\lambda}_i w^i$. ■

Altogether, Lemma 1.4 produces $q-1$ eigenvalue/eigenvector pairs of PMP. The vector in (e.2) should not be explicitly computed; rather the factorized form in (e.2) will suffice. The root to the equation in (e.1) can be quickly obtained using numerical methods (such as golden section search) since the expression in (e.1) is monotonely increasing in (α_i, α_{i+1}).

A different construction, which handles the non-acute eigenvectors and the eigenvalues of M with multiplicity greater than one, produces $n - q + 1$ additional distinct eigenvalues/eigenvector pairs for PMP orthogonal to c which are also distinct from those obtained through Lemma 1.4 (details in full paper).

To conclude this section, we note that it is straightforward to iterate the procedure in this section, so as to project a quadratic to hyperplanes of dimension less than $n-1$. More details will be provided in the talk and in the full paper.

1.4 Combinatorial bounds on distance functions

Here we take up the problem of computing strong lower bounds on the Euclidean distance from a point to the set $\mathcal{P} \cap \mathcal{K}$. In this abstract we will focus on the cardinality constrained problem, but results of a similar flavor hold for the disjunctive inequalities case.
Let $a \in \mathcal{R}^n$, $b \in \mathcal{R}$, $K < n$ be a positive integer, and $\omega \in \mathcal{R}^n$. Consider the problem

$$D^2_{min}(\omega, a) \;:=\; \min \left\{ \sum_{j=1}^{n} (x_j - \omega_j)^2, \;:\; a^T x = b \;\text{ and }\; \|x\|_0 \le K \right\}. \tag{12}$$

Clearly, the sum of smallest $n-K$ values ω_j^2 constitutes a ("naive") lower bound for problem (12). But it is straightforward to show that an exact solution to (12) is obtained by choosing $S \subseteq \{1, \ldots, n\}$ with $|S| \le K$, so as to minimize

$$\frac{(b - \sum_{j \in S} a_j \omega_j)^2}{\sum_{j \in S} a_j^2} \;+\; \sum_{j \notin S} \omega_j^2. \tag{13}$$

[We use the convention that $0/0 = 0$.] Empirically, the naive bound mentioned above is very weak since the first term in (13) is typically at least an order of magnitude larger than the second; and it is the bound, rather than the set S itself, that matters.

Suppose $a_j = 1$ for all j. It can be shown, using (13), that the optimal set S has the following structure: $S = P \cup N$, where $|P| + |N| \le K$, and P consists of the indices of the $|P|$ smallest nonnegative ω_j (resp., N consists of the indices of the $|N|$ smallest $|\omega_j|$ with $\omega_j < 0$). The optimal S can be computed in $O(K)$ time, after sorting the ω_j. When $\omega \ge 0$ or $\omega \le 0$ we recover the naive procedure mentioned above (though again we stress that the first term in (13) dominates). In general, however, we have:

Theorem 1.5 *(a) It is NP-hard to compute $D^2_{\min}(\omega, a)$. (b) Let $0 < \epsilon < 1$. We can compute a vector $\hat{x}$ with $\sum_j a_j \hat{x}_j = b$ and $\|\hat{x}\|_0 \le K$, and such that*

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

in time polynomial in n, ϵ^{-1}, and the number of bits needed to represent ω and a. ■

In our current implementation we have not used the algorithm in part (b) of the Lemma, though we certainly plan to evaluate this option. Instead, we proceed as follows. Assume $a_j \neq 0$ for all j. Rather than solving problem (12), instead we consider

$$\min \left\{ \sum_{j=1}^{n} a_j^2 (x_j - \omega_j)^2 \quad : \quad a^T x = b \ \text{ and } \ \|x\|_0 \leq K \right\}.$$

Writing $\bar{\omega}_j = a_j \omega_j$ (for all j), this becomes $\min \left\{ \sum_{j=1}^{n} (x_j - \bar{\omega}_j)^2 \ : \ \sum_j x_j = b \ \text{ and } \ \|x\|_0 \leq K \right\}$, which as noted above can be efficiently solved.

1.5 Application of the S-Lemma

Let $M = Q \Lambda Q^T \succeq 0$ be a matrix given by its eigenvector factorization. Let H be a hyperplane through the origin, $\hat{x} \in H$, $v \in \mathcal{R}^n$, $\delta_j > 0$ for $1 \leq j \leq n$, $\beta > 0$, and $v \in \mathcal{R}^n$. Here we solve the problem

$$\min \quad x^T M x + v^T x, \quad \text{subject to} \quad \sum_{i=1}^{n} \delta_i (x_i - \hat{x}_i)^2 \geq \beta, \ \text{ and } \ x \in H. \tag{14}$$

By rescaling, translating, and appropriately changing notation, the problem becomes:

$$\min \quad x^T M x + v^T x, \quad \text{subject to} \quad \sum_{i=1}^{n} x_i^2 \geq \beta, \ \text{ and } \ x \in H. \tag{15}$$

Let P be the $n \times n$ matrix corresponding to projection onto H. Using Section 1.3 we can produce a representation of PMP as $\tilde{Q} \tilde{\Lambda} \tilde{Q}^T$, where the the n^{th} eigenvector $\tilde{q}_n$ is orthogonal to H, and $\tilde{\lambda}_1 = \min_{i<n} \{\tilde{\lambda}_i\}$. Thus, problem (15) becomes, for appropriately defined $\tilde{v}$,

$$\Gamma \ := \ \min \sum_{j=1}^{n-1} \tilde{\lambda}_j y_j^2 + 2 \tilde{v}^T y, \quad \text{subject to} \quad \sum_{j=1}^{n-1} y_j^2 \geq \beta. \tag{16}$$

Using the S-lemma, we have that $\Gamma \geq \gamma$, iff there exists $\mu \geq 0$ s.t.

$$\sum_{j=1}^{n-1} \tilde{\lambda}_j y_j^2 + 2 \tilde{v}^T y - \gamma \ - \ \mu \left(\sum_{j=1}^{n-1} y_j^2 - \beta \right) \ \geq \ 0 \quad \forall \, y \in \mathcal{R}^{n-1}. \tag{17}$$

Using some linear algebra, this is equivalent to

$$\Gamma \ = \ \max \left\{ \mu \beta - \sum_{i=1}^{n-1} \frac{\tilde{v}_i^2}{\tilde{\lambda}_i - \mu} \ : \ 0 \leq \mu < \tilde{\lambda}_1 \right\}. \tag{18}$$

This is a simple numerical task, since in $[0, \tilde{\lambda}_1)$ the objective in (18) is concave in μ.

Remarks:
(1) Our updated template in Section 1.2 requires the solution of multiple problems of the form 18 (for different β and $\tilde{v}$) but just *one* computation of $\tilde{Q}$ and $\tilde{\Lambda}$.
(2) Consider any integer $1 \leq p < n-1$. When $\mu < \tilde{\lambda}_1$, the expression maximized in (18) is lower bounded by $\mu\beta - \sum_{i=1}^{p} \frac{\tilde{v}_i^2}{\tilde{\lambda}_i - \mu} - \frac{\sum_{i=p+1}^{n-1} \tilde{v}_i^2}{\lambda_{p+1} - \mu}$. This, and related facts, yield an approximate version of our approach which only asks for the first p elements of the eigenspace of PMP (and M).

1.5.1 Capturing the second eigenvalue

We see that $\Gamma < \tilde{\lambda}_1 \beta$ (and frequently this bound is close). In experiments, the solution y^* to (15) often "cheats" in that y_1^* is close to zero. We can then improve on the bound if the second projected eigenvalue, $\tilde{\lambda}_2$, is significantly larger than $\tilde{\lambda}_1$. Assuming that is the case, pick a value θ with $y_1^{*2}/\beta < \theta < 1$.

(a) If we assert that $y_1^2 \geq \theta\beta$ then we can strengthen the constraint in (14) to $\sum_{i=1}^n \delta_i (x_i - \hat{x}_i)^2 \geq \gamma$, where $\gamma = \gamma(\theta) > \beta$. This is certainly the case for the cardinality constraint and for the disjunctive inequalities case (details, full paper). So the assertion amounts to applying the S-lemma, but using γ in place of β.

(b) Otherwise, we have that $\sum_{i=2}^{n-1} y_i^2 \geq (1-\theta)\beta$. In this case, instead of the right-hand side of (18), we will have

$$\max \left\{ \mu(1-\theta)\beta \, - \, \sum_{i=2}^{n-1} \frac{\tilde{v}_i^2}{\tilde{\lambda}_i - \mu} \; : \; 0 \leq \mu \leq \tilde{\lambda}_2 \right\}. \tag{19}$$

The minimum of the quantities obtained in **(a)** and **(b)** yields a valid lower bound on Γ; we can evaluate several candidates for θ and choose the strongest bound. When $\tilde{\lambda}_2$ is significantly larger than $\tilde{\lambda}_1$ we often obtain an improvement over the basic approach as in Section 1.5.

Note: the approach in this section constitutes a form of *branching* and in our testing has proved very useful when $\lambda_2 > \lambda_1$. It is, intrinsically, a combinatorial approach, and quite distinct from the S-lemma and related problems (e.g. the trust region subproblem). It is thus not easily reproducible using convexity arguments alone. Also see remark (2) of the previous section.

2 Computational experiments

For the sake of brevity, we describe a partial set of experiments (more in the talk and in the full paper). The purpose of our experiments is to (a) investigate the speed of our numerical algebra routines, in particular the projection of quadratics, for large n and large number of nonzeroes in the quadratic, and (b) to study the strength of the bound our method produces.

We study problems of the form

$$\min\{ \, x^T M x + v^T x \; : \; \sum_j x_j = 1, \, x \geq 0, \, \|x\|_0 \leq K \, \}.$$

The matrix M is given in its eigenvector/eigenvalue factorization $Q \Lambda Q^T$. To stress-test our linear algebar routines, we construct Q as the product of random rotations: as the number of rotations increases, so does the number of nonzeroes in Q, and the overall "complexity" of M.

In our experiments, we ran our procedure after computing the solution to the (diagonalized) "weak" formulation

$$\min\{ \, y^T \Lambda y + v^T x \; : \; Q^T x = y, \; \sum_j x_j = 1, \, x \geq 0 \}.$$

We compare our bounds to those obtained by running the (again, diagonalized) *perspective formulation* [10], [11], a strong conic formulation (here, λ_{min} is the minumum λ_i):

$$
\begin{aligned}
\min \quad & \lambda_{min} \sum_j w_j \; + \; \sum_j (\lambda_j - \lambda_{min}) y_j^2 \\
s.t. \quad & Q^T x \; = \; y, \;\; \sum_j x_j = 1 \\
& x_j^2 \; - \; w_j z_j \; \leq \; 0, \quad 0 \; \leq \; z_j \; \leq \; 1 \quad \forall \; j, \\
& \sum_j z_j \; \leq \; k, \quad x_j \; \leq \; z_j \;\; \forall \; j, \\
& x, w \; \in \; \mathcal{R}^n_+.
\end{aligned}
\tag{20}
$$

For our experiments, we used Cplex 12.1 on a single-core 2.66GHz Xeon machine with 16 Gb of physical memory, which was never exceeded, even in the largest examples.

For the set of tests in Table 1, we used $n = 2443$ and a vector or real-world eigenvalues from a finance application. Q is the product of 5000 random rotations, resulting in 142712 nonzeros in Q (and thus, not particularly large).

K	**rQMIP** LB	**PRSP** LB	**SLE** LB	**rQMIP** sec	**PRSP** sec	**SLE** sec
200	0.031	0.0379	0.0382	14.02	59.30	5.3
100	0.031	0.0466	0.0482	13.98	114.86	5.8
90	0.031	0.0485	0.0507	14.08	103.38	5.9
80	0.031	0.0509	0.0537	14.02	105.02	6.2
70	0.031	0.0540	0.0574	13.95	100.06	6.2
60	0.031	0.0581	0.0624	15.64	111.63	6.4
50	0.031	0.0638	0.0696	13.98	110.78	6.4
40	0.031	0.0725	0.0801	14.03	104.48	6.5
30	0.031	0.0869	0.0958	14.17	104.48	6.8
20	0.031	0.1157	0.1299	15.69	38.13	6.9
10	0.031	0.2020	0.2380	14.05	43.77	7.2

Table 1: ***Examples with few nonzeroes***

Here, **rQMIP** refers to the weak formulation, **PRSP** to the perspective formulation, and **SLE** to the approach in this paper. "LB" is the **lower bound** produced by a given approach, and "sec" is the CPU time in seconds. We see that that **rQMIP** is quite weak especially for smaller K. Both **PRSP** and **SLE** substantially improve on **rQMIP**, though **PRSP** is significantly more expensive.

In Table 2 we consider examples with $n = 10000$ and random Λ. In the table, **Nonz** indicates the number of nonzeroes in Q; as this number increases the quadratic becomes less diagonal dominant.

Nonz in Q	**rQMIP** LB	**PRSP** LB	**SLE** LB	**rQMIP** sec	**PRSP** sec	**SLE** sec
5.3e+05	2.483e-03	1.209e-02	1.060e-02	332	961.95	57.69
3.7e+06	2.588e-03	1.235e-02	1.113e-02	705	2299.75	57.55
1.8e+07	2.671e-03	1.248e-02	1.117e-02	2.4e+03	1.3e+04	57.69
5.3e+07	2.781e-03	1.263e-02	1.120e-02	1.1e+04	8.5e+04	58.44
8.3e+07	2.758e-03	1.262e-02	1.211e-02	2.3e+04	1.4e+05	57.38

Table 2: ***Larger examples***

As in Table 1, **SLE** and **PRSP** provide similar improvements over **rQMIP** (which is clearly extremely weak). Moreover, **SLE** proves uniformly fast – essentially, *free* compared to **rQMIP**.

In the examples in Table 2, the smallest ten (or so) eigenvalues are approximately equal, with larger values after that. The techniques in Section 1.5.1 should extend to this situation in order to obtain an even stronger bound – we hope to present results on this in the talk.

Also note that the perspective formulation quickly proves impractical. A cutting-plane procedure that replaces the conic constraints in (20) with (outer approximating) linear inequalities is outlined in [10], [11] and tested on random problems with $n \leq 400$ (which we will also test in the full paper). Such a procedure begins by solving **rQMIP** and then iteratively adds the inequalities; or it could simply solve a formulation consisting of **rQMIP**, augmented with a set of pre-computed inequalities. In either case the running time will be slower than that for **rQMIP**. In our initial experiments with this linearized approximation, we found that (a) it can provide a very good lower bound to the conic perspective formulation, (b) it can run significantly faster than the full conic formulation, but, (c) it proves significantly *slower* than **rQMIP**, and, in particular, still significantly slower than the combination of **rQMIP** and **SLE**.

The discussion regarding the perspective formulation is of interest because *it* is precisely an example of the paradigm that we consider in this paper: a convex formulation for a nonconvex problem with a convex objective. In our tests involving large cases, and using branch-and-cut, the lower bound proved by the perspective formulation exhibited the expected stalling behavior. Figure 3 concerns the $K = 70$ case of Table 1 where we ran the mixed-integer programming version of several formulations using Cplex 12.1 on a faster machine (on which rQMIP requires 4.35 seconds and our method, 3.54 seconds, to prove the lower bound of 0.0574). The runs in the table used four parallel threads of execution.

formulation	**nodes**	**time**	**LB**	**UB**
QPMIP	124300	85320 (2.75 sec/node)	0.0312	0.337
PRSP-MIP	6100	85467 (56.04 sec/node)	0	0.712
LPRSP-MIP	39000	109333 (11.21 sec/node)	0.0554 root 0.0540	0.305

Table 3: ***Detailed analysis of*** $K = 70$ ***case of Table 1***

In Table 3, **QPMIP** is the weak formulation, **PRSP-MIP** is the perspective formulation [Comment: on conic MIPs, Cplex appears to keep the lower bound fixed at 0 for a very long time]. **LPRSP-MIP** it the linearized perspective version. The figures in parentheses indicate CPU seconds per branch-and-cut node. "time" indicates the observed time (i.e. total CPU time will be, roughly, four times larger). Note the stalling of the lower bound in **LPRSP-MIP** at a value strictly smaller than the 0.0574 bound SLE proved with no branching. **QPMIP** with all Cplex cuts suppressed (not shown in the table) yielded an upper bound of 0.281 in 200 seconds – this underlines the point concerning fast heuristics that we made in the introduction.

Our thechniques apply to the solution computed by the perspective formulation (after projecting out the auxiliary variables). This approach can only be desirable if there is a fast and tight approximation to the perspective formulation. However, Table 2 suggests that for large problems **rQMIP** is already too expensive. It might simply better to approximate its solution, quickly, perhaps using a first-order method (and apply our techniques to the resulting solution vector). We plan to investigate these issues in the full paper.

A more significant focus of our upcoming work concerns problems in complex applications where auxiliary binary variables cannot be naturally employed. A major issue that we plan to investigate is how to incorporate our techniques within branch-and-cut, so that the bound proved at each node strictly improves on the bound at its parent node – no stalling. In this context, the single most important idea that we will investigate and report on concerns branching based on eigenvector structure, along the lines of Section 1.5.1.

References

[1] 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.

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

[3] D. Bienstock and M. Zuckerberg, Subset algebra lift algorithms for 0-1 integer programming, *SIAM J. Optimization* **105** (2006), 9 – 27.

[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] W. Cook, R. Kannan and A. Schrijver, Chv'atal closures for mixed ineger programs, *Math. Programming* **47** (1990), 155 – 174.

[7] I. De Farias, E. Johnson and G. Nemhauser, A polyhedral study of the cardinality constrained knapsack problem, *Math. Programming* **95** (2003), 71 – 90.

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

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

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

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

[12] B. Moghaddam, Y. Weiss, S. Avidan, Generalized spectral bounds for sparse LDA, *Proc. 23rd Int. Conf. on Machine Learning* (2006), 641 – 648.

[13] J.J. Moré and D.C. Sorensen, Computing a trust region step, *SIAM J. Sci. Statisr Comput* **4** (1983), 553 – 572.

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

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

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

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

[18] W. Miller, S. Wright, Y. Zhang, S. Schuster and V. Hayes, Optimization methods for selecting founder individuals for captive breeding or reintroduction of endangered species, manuscript (2009).

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

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