ESI Summer Institute, Nonlinear Methods in Combinatorial Optimization

Florian Jarre,

Univ. Düsseldorf

Klagenfurt, Aug. 20 – Sept. 3, 2010

Accelerated Projection Methods for Semidefinite Programs

Outline

- **The problem and assumptions**
- A semismooth approach
  for Solving Semidefinite Programs
- Further theoretical results
- Numerical experiments

Semidefinite Program

$$\text{minimize } C \bullet X \mid \mathcal{A}(X) = \bar{b}, \quad X \succeq 0.$$

Here,

$$C \bullet X := \langle C, X \rangle := \sum_{i,j} C_{i,j} X_{i,j} = \text{trace } (C^\top X)$$

and

$$\mathcal{A}(X) = (A^{(1)} \bullet X; \ldots; A^{(m)} \bullet X) \in \mathbb{R}^m.$$

Notation

Let $\mathcal{L} := \{X \mid \mathcal{A}(X) = 0\}$ and $\mathcal{A}^*(y) := \sum_{i=1}^m y_i A^{(i)}$ then,

$$\mathcal{L}^\perp = \{S \mid S = \mathcal{A}^*(y) \text{ for some } y \in I\!R^m\}$$

and the dual problem can be written as

$$\text{maximize } \bar{b}^T y \mid \mathcal{A}^*(y) + S = C, \quad S \succeq 0$$

or

$$\text{minimize } B \bullet S \mid S \in \mathcal{L}^\perp + C, \quad S \succeq 0$$

where B is some matrix with $\mathcal{A}(B) = \bar{b}$.

More general format

Let K be a pointed closed convex cone with nonempty interior in some Euclidean space E
and let $\mathcal{L}$ be a subpace of E.

(For semidefinite programs $K := \{X = X^T \ | \ X \succeq 0\}$.)

We formulate a convex conic program in general form:

minimize $\langle c, x\rangle \mid x \in K \cap (\mathcal{L} + b)$.

Normalization of the data

One can easily normalize the data and assume (without loss of generality) that

- $b \in \mathcal{L}^{\perp}$ and $\|b\|_2 = 1$.
- $c \in \mathcal{L}$ and $\|c\|_2 = 1$.

Moreover, we assume (with slight loss of generality) that the interior point condition holds:

$$\exists x \in \text{int}(K) \cap \mathcal{L} + b, \qquad \exists s \in \text{int}(K^D) \cap \mathcal{L}^{\perp} + c.$$

Optimality conditions (Nesterov, Nemirovski 1994)

Then,

(P) minimize $\langle c, x \rangle \mid x \in K \cap (\mathcal{L} + b)$

and its dual

(D) minimize $\langle b, s \rangle \mid s \in K^D \cap (\mathcal{L}^\perp + c)$

satisfy strong duality, i.e. x is optimal for (P) if, and only if, there exists a point s feasible for (D) with

$$\langle c, x \rangle + \langle b, s \rangle = 0.$$

We denote such x and s by x^{opt} and s^{opt}.

Outline

- The problem and assumptions
- **A semismooth approach for Solving Semidefinite Programs**
- Further theoretical results
- Numerical experiments

Augmented Primal Dual Approach (APD)

Let the affine subspace $\mathbf{A} \subset E \times E$ be defined as

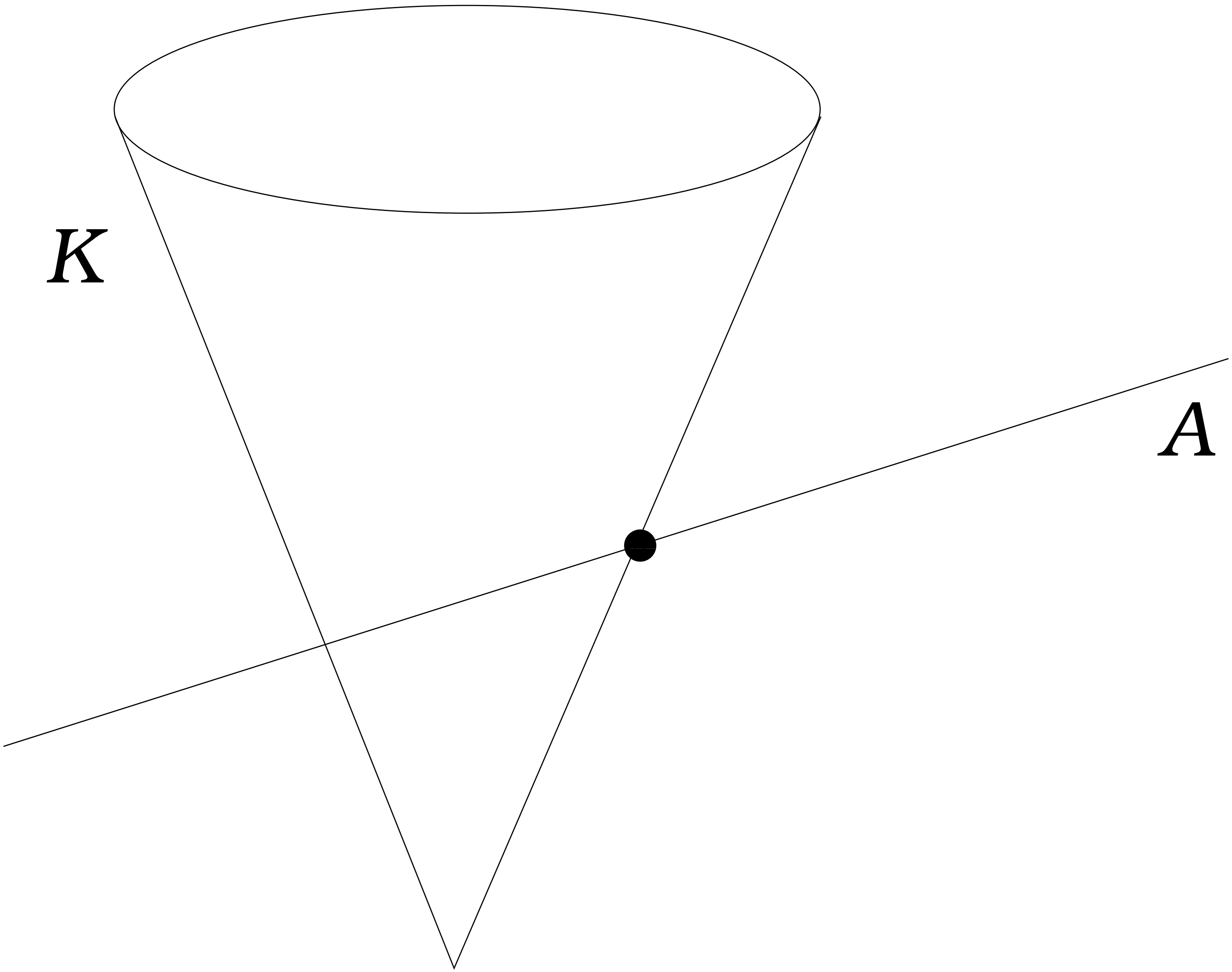
$$\mathbf{A} := (\mathcal{L} + b) \times (\mathcal{L}^{\perp} + c) \cap \{(x; s) \mid \ \langle c, x \rangle + \langle b, s \rangle = 0\}$$

and the full dimensional closed convex cone $\mathbf{K} \subset E \times E$ as

$$\mathbf{K} := K \times K^{D}.$$

Solving (P) is equivalent to finding $z := (x; s) \in \mathbf{A} \cap \mathbf{K}$.

Intersection, cone and affine subspace



Using projections?

Given $z \in E \times E$, it is often very cheap to compute the orthogonal projection of z onto $\mathbf{A}$ or onto $\mathbf{K}$.

Projection onto $\mathbf{K}$:

LP: order n. ($x \to x^+$.)

SOCP: order n. (Straightforward, 3 cases...)

SDP: order n^3. (Set negative eigenvalues to zero.)

Projection onto $\mathbf{A}$

Let

$$\mathcal{L} + b = \{x \mid Ax = Ab\} \subset \mathbb{R}^n.$$

Then,

$$\Pi_{\mathcal{L}+b}(x) = x - A^T(AA^T)^{-1}A(x-b)$$

and

$$\Pi_{\mathcal{L}^\perp+c}(s) = s - (I - A^T(AA^T)^{-1}A)(s-c).$$

Cholesky factor of AA^T computed once during the overall algorithm. (Often by orders of magnitude cheaper than one interior-point iteration.)

(Once AA^T is factored, it is cheap to replace b with $\Pi_{\mathcal{L}^\perp}(b)$ and c with $\Pi_{\mathcal{L}}(c)$.)

Computation of the projection onto $\mathbf{A}$

Let

$$\mathbf{A}_1 := (\mathcal{L} + b) \times (\mathcal{L}^\perp + c)$$

and

$$\mathbf{A}_2 := \{(x; s) \mid \ \langle c, x\rangle + \langle b, s\rangle = 0\}$$

Then $\mathbf{A} = \mathbf{A}_1 \cap \mathbf{A}_2$.

Since

$$b \in \mathcal{L}^\perp \qquad \text{and} \qquad c \in \mathcal{L}$$

we have

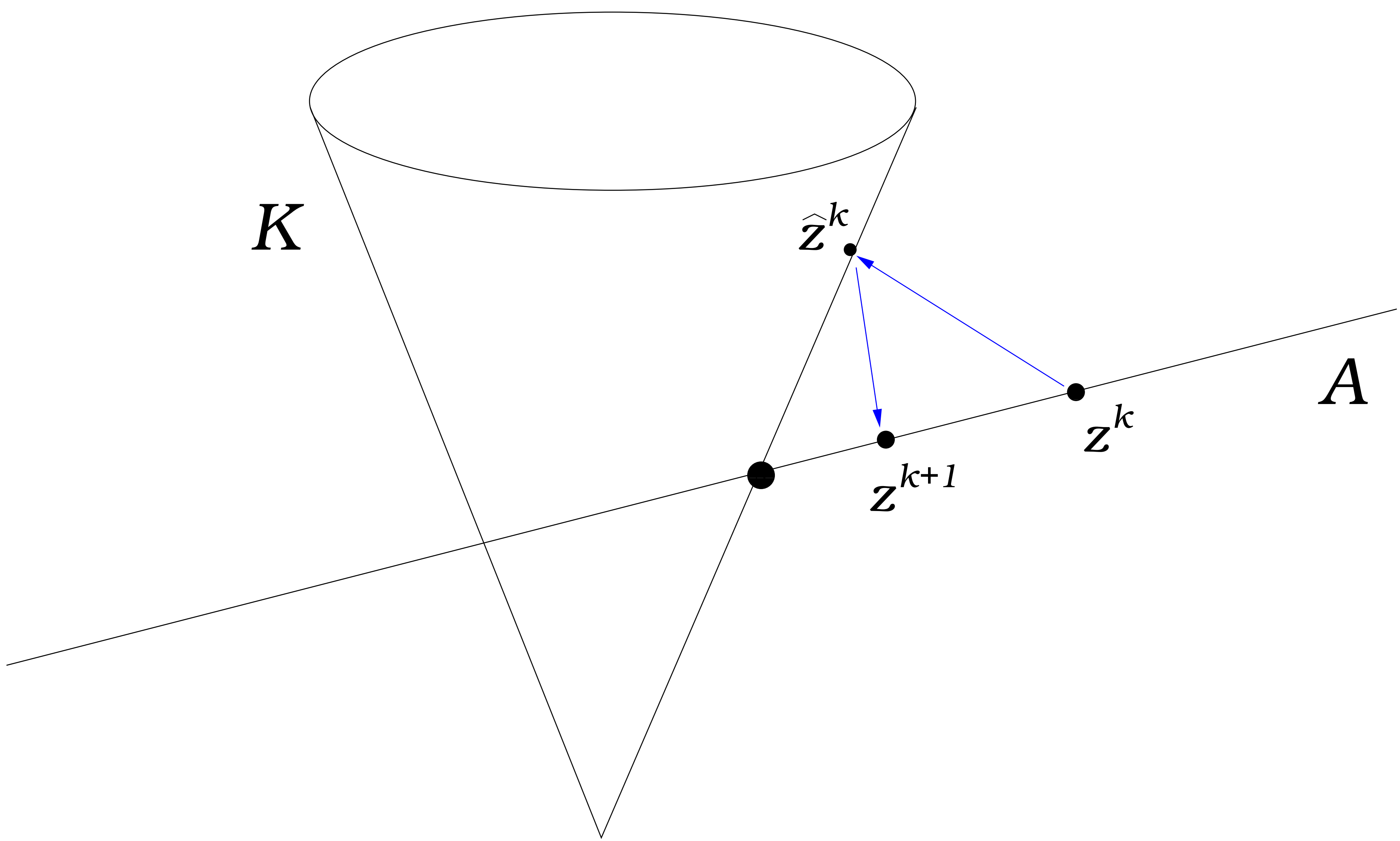
$$\Pi_{\mathbf{A}} = \Pi_{\mathbf{A}_1}\Pi_{\mathbf{A}_2} = \Pi_{\mathbf{A}_2}\Pi_{\mathbf{A}_1}.$$

Simple projection method

Let $z^0 \in \mathbf{A}$ be given. Set $k = 0$.

1. Set $\hat{z}^k := \Pi_{\mathbf{K}}(z^k)$.
2. Set $z^{k+1} := \Pi_{\mathbf{A}}(\hat{z}^k)$.
3. Set $k = k + 1$. Go to Step 1.

Simple projection method



Minimizing a differentiable convex function

For a closed set $\mathcal{C}$ and a vector $\bar{z}$ we denote the distance of $\bar{z}$ to $\mathcal{C}$ by

$$d(\bar{z}, \mathcal{C}) := \min\{\|z - \bar{z}\|_2 \mid z \in \mathcal{C}\}.$$

All we need is a point in $\mathbf{A}$, i.e. a point z such that

$$\phi(z) := \frac{1}{2} d(z, \mathbf{K})^2 = 0,$$

i.e. such that the differentiable convex function ϕ is minimized.

Differentiating ϕ

Let $\mathcal{C}$ be a closed convex set and let $\Pi_{\mathcal{C}}$ be the orthogonal projection (with respect to the Euclidean norm) onto $\mathcal{C}$. Then,

$$d(z, \mathcal{C}) = \|z - \Pi_{\mathcal{C}}(z)\|_2,$$

and the gradient of the differentiable function $f_{\mathcal{C}}(z) := \frac{1}{2} d(z, \mathcal{C})^2$ is given by

$$\nabla f_{\mathcal{C}}(z) = z - \Pi_{\mathcal{C}}(z).$$

Restriction to $\mathbf{A}$

Let

$$\tilde{\phi}(\tilde{z}) := \phi(\tilde{z}) = \frac{1}{2} d(\tilde{z}, \mathbf{K})^2 \qquad \text{for } \tilde{z} \in \mathbf{A}.$$

Then,

$$\nabla \tilde{\phi}(\tilde{z}) = \tilde{z} - \Pi_{\mathbf{A}}(\Pi_{\mathbf{K}}(\tilde{z})).$$

A steepest descent step with step length 1 for minimizing $\tilde{\phi}$ starting at a point $\tilde{z} = z^k \in \mathbf{A}$ is the same as the computation of z^{k+1} with the projection algorithm.

L-BFGS-algorithm

Let $\tilde{z}^0 \in \mathbf{A}$ be given. Let $\Delta\tilde{z}^0 := -\nabla\tilde{\phi}(\tilde{z}^0)$. Set $k = 0$.

1. Let $\lambda_k :=$ argmin$\{\tilde{\phi}(\tilde{z}^k + \lambda\Delta\tilde{z}^k) \mid \lambda > 0\}$.
2. Set $\tilde{z}^{k+1} := \tilde{z}^k + \lambda_k\Delta\tilde{z}^k$.
3. Compute $\Delta\tilde{z}^{k+1}$ from $\Delta\tilde{z}^k$ and $\nabla\tilde{\phi}(\tilde{z}^{k+1})$ with L-BFGS update formula.
4. Set $k := k + 1$. Go to Step 1.

Handicap for SDP-case

Hessian of $\tilde{\phi}$ at the optimal solution is typically singular, even when the primal-dual optimal solution is unique and strictly complementary. (2$\times$2-example)
More precisely, the Hessian does not exist, but the generalized Hessian contains singular matrices.

Result observed in preliminary experiments

The L-BFGS-method for minimizing $\tilde{\phi}$ converges rapidly in the inital stage of the algorithm, and then slows down.

A local acceleration

Let

$$\tilde{f}(Z) = \tilde{f}(X, S) := \|XS - SX\|_F^2.$$

The non convex function $\tilde{f}$ is minimized at Z^{opt}. It is differentiable and the derivative can be computed with three matrix-matrix multiplications.

Second order growth condition (J', Rendl, 2007)

The gradient of $\tilde{f} + \tilde{\phi}$ is strongly semismooth and – when Z^{opt} is a unique strictly complementary solution of the semidefinite program – there is an $\epsilon > 0$ such that

$$\tilde{f}(Z^{opt} + \Delta Z) + \tilde{\phi}(Z^{opt} + \Delta Z) \geq \epsilon \|\Delta Z\|^2$$

for all sufficiently small $\|\Delta Z\|$ with $Z^{opt} + \Delta Z \in \mathbf{A}$.

Note, if γ is some function in C^2, then the second order growth condition at some point x^* implies that $\nabla^2 \gamma(x^*) \succ 0$.

Outline

- The problem and assumptions
- A semismooth approach
  for Solving Semidefinite Programs
- **Further theoretical results**
- Numerical experiments

Second order growth condition for semi-smooth functions

If γ is in C^1, $\nabla\gamma$ is locally Lipschitz-continuous and strongly semismooth at x^*, then even if γ satisifies the second order growth condition at x^*, the generalized Hessian of γ at x^* may contain singular elements or elements with negative eigenvalues: Let $\gamma : \mathbb{R}^2 \to \mathbb{R}$

$$\gamma(x,y) := \begin{cases} x^2 & \text{if } x \geq 0, x \geq |y|, \\ x^2 + (y-x)^2 & \text{if } y > 0, y > |x|, \\ x^2 + (y+x)^2 & \text{if } y < 0, -y > |x|, \\ 3x^2 + 2y^2 & \text{if } x < 0, -x \geq |y|. \end{cases}$$

Second order growth condition...

Here, $\nabla\gamma$ is Lipschitz-continuous ($L = 4$), $\nabla\gamma$ is strongly semismooth, and $\gamma(x,y) \geq \frac{1}{4}(x^2 + y^2)$.

Nevertheless, $\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \in \partial^2\gamma(x^*)$, $x^* = (0,0)$.

Moreover, $\gamma(x,y) - \frac{1}{8}(x^2 + y^2)$ still satisfies the second order growth condition at x^*, and we have

$$\frac{1}{4}\begin{pmatrix} 7 & 0 \\ 0 & -1 \end{pmatrix} \in \partial^2\gamma(x^*).$$

How about $\tilde{\phi} + \tilde{f}$*?*

Let

$$C := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \bar{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

and

$$\mathcal{A}(X) = \begin{pmatrix} A^{(1)} \bullet X \\ A^{(2)} \bullet X \\ A^{(3)} \bullet X \end{pmatrix}$$

with

$$A^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \ A^{(2)} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \ A^{(3)} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

The unique and strictly complementary optimal solution is given by

$$X^{opt} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad S^{opt} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Consider the pair

$$X_\varepsilon = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix}, \quad S_\varepsilon = \begin{pmatrix} -2\varepsilon & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

for $\varepsilon > 0$. Here, $(X_\varepsilon, S_\varepsilon) \in \mathbf{A}$.

For

$$H = \left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right)$$

we have

$$(\tilde{\phi} + \tilde{f})((X_\varepsilon, S_\varepsilon) + \delta H) \equiv 2\varepsilon^2 \quad \forall \delta \in [-\varepsilon, \varepsilon].$$

Therefore, $\nabla^2(\tilde{\phi} + \tilde{f})(X_\varepsilon, S_\varepsilon)[H, H] = 0$ for all $\varepsilon > 0$.
Moreover, $\Theta = \lim_{\varepsilon \searrow 0} \nabla^2(\tilde{\phi} + \tilde{f})(X_\varepsilon, S_\varepsilon)$ exists.
Here, $\Theta[H, H] = 0$ and, $\Theta \in \partial^2(\tilde{\phi} + \tilde{f})(X^{opt}, S^{opt})$.

Θ is singular.

Stronger second order growth condition (2010)

For the function

$$\tilde{\hat{f}}(X,S) := \|XS\|_F^2 = \frac{1}{4}\left(\tilde{f}(X,S) + \|XS + SX\|_F^2\right)$$

the stronger result

$$\partial^2(\tilde{\phi} + \tilde{\hat{f}})(X^{opt}, S^{opt}) \succ 0$$

holds true (under the same assumptions of uniqueness and strict complementarity).

(In numerical experiments, the convergence results with this function were best.)

Consequence

We solve (P) and (D) in two stages, the first one minimizing $\tilde{\phi}$ for $\tilde{Z} \in \mathbf{A}$, and when convergence of this stage is slow, starting a second stage minimizing $\tilde{\phi} + \tilde{\hat{f}}$ for $\tilde{Z} \in \mathbf{A}$.
For both stages we may use a L-BFGS-method.

Note

The function $\tilde{\phi} + \tilde{\hat{f}}$ may (sometimes does!) have local minimizers.

$\Longrightarrow$ Minimize $\tilde{\phi} + \alpha\tilde{\hat{f}}$ for $\alpha > 0$ and control α.

Outline

- The problem and assumptions
- A semismooth approach
  for Solving Semidefinite Programs
- Further theoretical results
- **Numerical experiments**

Preliminary numerical results –

– for general SDP’s
http://www.math.uni-klu.ac.at/or/Software

L-BFGS
(Line search with only one extra function evaluation per iteration.)

LBFGS, General random SDPs

Examples with $n \geq 400$ and $m \geq 30000$ (50 iterations)

dim	m	sec	lg(phi)	lg(fhat)	err_P	err_D
400	30k	133.6	-5.896	-6.447	-6	-7
500	30k	172.4	-5.366	-6.133	-9	-21
600	40k	278.5	-5.334	-6.209	-7	-22
700	50k	418.9	-5.204	-6.132	-8	-25
800	70k	610.9	-5.294	-6.296	-7	-20
900	100k	857.1	-5.431	-6.490	-6	-15
1000	100k	1139.5	-5.168	-6.285	-8	-22

$err_P = \frac{\lambda_{min}(X^{it})}{1+\|X^{it}\|_F} \cdot 10^5, \qquad err_D = \frac{\lambda_{min}(S^{it})}{1+\|S^{it}\|_F} \cdot 10^5$

BFGS vs. Nesterov's method

The regularization term is chosen $\alpha = 15$ for both methods. (With a safeguard to prevent convergence to local minimizer.)

Without regularization the Lipschitz constant can be chosen $L = 1$ for Nesterov's method. ($L = 0.5$ still works in our experiments, but $L = 0.495$ leads to divergence.)

With regularization the Lipschitz constant $L := 1 + \max\{\lambda_{max}(X), \lambda_{max}(S)\}$ seems overly pessimistic.

The line search in LBFGS eliminates the need for estimating the Lipschitz constant – but it costs one extra function evaluation per step.

LBFGS vs. Nesterov's method (continued)

10 Examples with $n \geq 400$ and $m \geq 30000$ (average values)

Method	it	sec	err_P	err_D
LBFGS	300	970	-0.11	-0.14
Nest (L=1)	300	583	-0.32	-0.42
Nest (L=2)	300	587	-1.15	-1.49
Nest (L=1)	480	1004	-0.20	-0.26

$$err_P = \frac{\lambda_{min}(X^{it})}{1+\|X^{it}\|_F} \cdot 10^5, \qquad err_D = \frac{\lambda_{min}(S^{it})}{1+\|S^{it}\|_F} \cdot 10^5$$

(Other experiments are quite similar.)

Other Modifications

- Use Newton-cg for $\tilde{\phi} + \alpha \tilde{\hat{\hat{f}}}$ in the final stage after LBFGS with regularization turns slow as well.
- Numerical results give some improvement – but not conclusive.
- High number of cg-iterations needed and even when cg is run up to machine precision, the observed rate of convergence of Newton's method is not the expected quadratic rate. (Rounding errors?)

Discussion

- The function $\tilde{\phi}$ contains the normal equations. Solving the normal equations by an iterative method generally is a bad idea.
- Here, the normal equations are “preconditioned” in some form as we assume that the projection onto $\mathbf{A}$ is carried out exactly, but still, the Hessian of $\tilde{\phi}$ being based on the sum of two projections may (and usually does) have a poor condition number.
- Use QMR on the AHO-System (plain primal-dual system without centering).

AHO-QMR

- Use complementary starting point:
  Set $W := X - S$ and decompose $W = UDU^T$,
  then $U^T XU$ and $U^T SU$ are nearly diagonal.
  Project onto nearest complementary diagonal matrix pair. In the transformed space, the complementarity operators are diagonal.
- Use further transformations to make AHO symmetric.
  (Number of iterations and work per iteration!)
- Use Cholesky factor of $\mathcal{A}\mathcal{A}^*$ as preconditioner.
- AHO-QMR typically fails if started without Phase 1.
  (Some interior-point approach would be needed.)

LBFGS, AHO-QMR

Example with $n = 400$ and $m = 30000$

Method	it	sec	lg(phi)	lg(fhat)
LBFGS	100	195.6	-7.290	-7.729
LBFGS	500	935.3	-10.802	-10.904
LBFGS$\backslash$QMR	100$\backslash$6	867.7	-17.158	-16.628

Summary

Simple concept minimizing squared distance to $\mathbf{K}$ within $\mathbf{A}$.

Regularization and accelerations, such as L-BFGS or truncated Newton-cg.

Phase 1 suitable for AHO-QMR.

Many applications that require low accuracy e.g. in combinatorial optimization and completely positive programming.

Implementation still (always!?) has room for improvement.