A semismooth Newton-CG augmented Lagrangian method for solving large scale SDPs

Kim-Chuan Toh

National University of Singapore
Department of Mathematics, and Singapore-MIT Alliance

Based on joint work with Defeng Sun and Xinyuan Zhao

SJOM 2008, Taiwan
Outline

- Primal and dual SDP
- Examples of large SDP
- Related work
- Augmented Lagrangian method (ALM)
- Convergence of ALM
- Conditioning of generalized Hessian
- Numerical results
Standard primal SDP

\[ S^n = \{ n \times n \text{ symmetric matrices} \}, \quad \langle P, Q \rangle = \sum_{i,j} P_{ij}Q_{ij} = \text{Trace}(PQ). \]

\[ Q \succeq 0 \ (Q \succ 0) \implies Q \text{ is symmetric positive semidefinite (definite)}. \]

Given data: \( C, A_1, \cdots, A_m \in S^n, b \in R^m \)

\[
\begin{align*}
(P) \quad & \min \quad \langle C, X \rangle \\
& \text{s.t.} \quad A(X) = b, \quad X \succeq 0, \quad X \in S^n \\
& \quad \text{(convex)}
\end{align*}
\]

where \( A : S^n \to R^m \) is the linear map s.t.

\[
A(X) = \left[ \langle A_1, X \rangle, \ldots, \langle A_m, X \rangle \right]^T.
\]

Assume \( (P) \) is feasible.

Problem dimension: \( n = \text{dimension of } X, \quad m = \text{number of linear constraints} \)

We consider SDPs with large \( m \geq 10,000 \), but moderate \( n \leq 2000 \).
(D) \[ \max \ b^T y \]
\[ \text{s.t. } A^T(y) + Z = C, \quad Z \succeq 0, \quad y \in R^m, \ Z \in S^n \]

where $A^T : R^m \rightarrow S^n$ is the adjoint of $A$ and $A^T(y) = \sum_{k=1}^{m} y_k A_k$.
Assume (D) is feasible.
Examples of SDP: nearest correlation matrix

Nearest correlation matrix problem: Given an estimated correlation matrix $C$, we want to find a valid correlation matrix $X$ that is nearest to the data:

$$(NCM) \quad \min \left\{ \sum_{ij} |X_{ij} - C_{ij}| : \text{diag}(X) = 1, \ X \succeq 0 \right\}$$

$$\Downarrow$$

$$\sum_{ij} v_{ij}^+ + v_{ij}^- : X_{ij} - C_{ij} = v_{ij}^+ - v_{ij}^-, \ v_{ij}^+, v_{ij}^- \geq 0$$

$n(n+1)/2$ equality constraints

$m = n + n(n+1)/2$, which is about 500K when $n = 1000$. 
Examples of SDP: sparse maximum eigenvalue

Sparse maximum eigenvalue [d’Aspremont, El Ghaoui, Jordan, Lanckriet]
Given $B \in S^n$, find a "maximal eigenvector" with at most $k$ non-zeros:

$$\max\{\langle B, xx^T \rangle : \|x\|_2 = 1, \text{card}(x) \leq k\}.$$  

SDP relaxation based on $X = xx^T \Rightarrow X \succeq 0$ gives

$$\text{(spmaxeig) } \max\{\langle B, X \rangle : \langle I, X \rangle = 1, \langle E, |X| \rangle \leq k, X \succeq 0\}. \quad (2)$$

$$m = 2 + n(n + 1)/2.$$

Example: SDP relaxation of the maximum stable set problem of a graph

For a graph $G = (V, \mathcal{E})$, a stable set $S$ is subset of $V$ such that no vertices in $S$ are adjacent. The problem is to find a stable set with maximum cardinality.

The standard SDP relaxation of the maximum stable set problem is:

$$\theta(G) := \max \left\{ \langle E, X \rangle : X_{ij} = 0 \ \forall \ (i,j) \in \mathcal{E}, \ \langle I, X \rangle = 1, \ X \succeq 0 \right\}$$ (3)

$$\theta_+(G) := n(n+1)/2 \text{ additional constraints } X \succeq 0$$ (4)

$\theta(G)$: number of constraints $m = |\mathcal{E}| + 1$.

$\theta_+(G)$: number of constraints $m = |\mathcal{E}| + 1 + n(n+1)/2$. 
Related work

Number of constraints $m$ is large: $m \geq 10,000 \Rightarrow m \times m$ dense Schur complement matrix cannot be stored explicitly. For $m = 10^5$, needs 100GB RAM memory.

- Parallel computation [Benson, Borchers, Kojima et al., de Klerk]
- First-order gradient methods (low accuracy):
  - NLP reformulation [Burer-Monteiro]
  - Saddle-point mirror-prox [Lu-Nemirovski-Monteiro]
- Inexact IPM $\leftarrow$ compute direction via iterative solvers [Kojima, Toh]
- Generalized Lagrangian method on barrier-penalized (D) [Kocvara-Stingl]
- Augmented Lagrangian method for primal SDPs from relaxation of lift-and-project scheme [Burer-Vandenbussche]
- Boundary-point method: based on augmented Lagrangian method for (D) [Rendl et al. ]
Projection onto positive semidefinite cone $S^n_+$

Given $Y \in S^n$, find

$$\min\{\|Y - X\|^2 : X \succeq 0\},$$

where $\| \cdot \|$ is the Frobenius norm.

Eigenvalue decomposition: $Y = QDQ^T$ with $Q$ orthogonal, $D = \text{diag}(d)$.

Write $d = d_+ - d_-$, where $d_\pm = \max(0, \pm d)$.

Let $\Pi_+(Y) = Q\text{diag}(d_+)Q^T$, $\Pi_-(Y) = Q\text{diag}(d_-)Q^T$. Then

$$Y = \Pi_+(Y) - \Pi_-(Y), \quad \Pi_+(Y)\Pi_-(Y) = 0, \quad \Pi_+(Y), \Pi_-(Y) \succeq 0,$$

$$\min\{\|Y \mp X\|^2 : X \succeq 0\} = \|\Pi_+(Y)\|^2.$$
Augmented Lagrangian function for \((D)\)

\[
(D) \quad \max \left\{ b^T y : C - A^T(y) - Z = 0, \ y \in \mathbb{R}^m, \ Z \succeq 0 \right\}
\]

Let \(X \in S^n\) be the multiplier associated with the equality constraint. Construct

\[
\tilde{L}(y, Z; X) = b^T y + \langle X, C - A^T y - Z \rangle - \frac{\sigma}{2} \|C - A^T y - Z\|^2
\]

\[
= b^T y + \frac{1}{2\sigma} \|X\|^2 - \frac{1}{2\sigma} \|U(y; X) + \sigma Z\|^2.
\]

where \(U(y; X) = X - \sigma (C - A^T y)\).

\[
L(y; X) := \max \left\{ \tilde{L}(y, Z; X) : Z \succeq 0 \right\}
\]

\[
= b^T y + \frac{1}{2\sigma} \|X\|^2 - \frac{1}{2\sigma} \|\Pi_+(U(y; X))\|^2
\]
Augmented Lagrangian method

Under Slater’s condition, solving (D) is equivalent to

\[
\min_{X \in S^n} \Phi(X) := \max \{ L(y; X) : y \in \mathbb{R}^m \} \tag{5}
\]

**ALM:** Input \( X^0 \in S^n_+ \), \( \sigma_0 > 0 \), iterate:

\[
y^k \approx \arg\max \{ L(y; X^k) : y \in \mathbb{R}^m \} \tag{6}
\]

\[
X^{k+1} = X^k - \sigma_k \nabla X L(y^k; X^k) = \Pi_+ (U(y^k; X^k))
\]

\[
Z^{k+1} = \frac{1}{\sigma_k} \Pi_- (U(y^k; X^k))
\]

If \( \| R_d^k := C - A^T y^k - Z^{k+1} \| \leq \epsilon \); stop; else; update \( \sigma_k \); end

- \( R_d^k = \nabla X L(y^k; X^{k+1}) \).

- For the inner subproblem (6), optimality condition is \( \nabla_y L(y; X^k) = 0 \).

\[ R_p^k = b - AX^{k+1} = \nabla_y L(y^k; X^k) \approx 0 \quad \text{if (6) is solved accurately.} \]
Convergence of ALM

For the inner problem (6), if we use the stopping condition below:
\[ \| \nabla_y L(y^k; X^k) \| \leq (\delta_k / \sigma_k) \| X^{k+1} - X^k \|, \quad \delta_k \to 0, \]
then we get the following theorem based on [Rockafellar, MOR, 76].

**Theorem:** Assuming that (P) and (D) are strictly feasible, and constraint non-degeneracies hold at the optimal solution $X^*$ for (P) and $y^*$ for (D), then the iterates $\{X^k\}, \{y^k\}$ generated by ALM converges to $X^*$ and $y^*$, respectively. Moreover, there exist constants $\theta, \theta'$ such that for $k$ large, we have

\[
\| X^{k+1} - X^* \| \leq \frac{\theta}{\sqrt{\theta^2 + \sigma_{max}^2}} \| X^k - X^* \| \]
\[
\| y^{k+1} - y^* \| \leq \frac{\theta'}{\sigma_{max}} \| X^k - X^* \|. \]

Note: larger $\sigma_{max} := \max_k \{\sigma_k\}$ leads to faster convergence. But inner subproblem (6) is harder to solve.
A semismooth Newton-CG method for solving inner subproblem

Aim: solve $\nabla_y L(y; X^k) = b - \mathcal{A}\Pi_+ (U^k(y)) = 0$, $U^k(y) = X^k - \sigma(C - \mathcal{A}^T y)$.

$\nabla_y L(y; X^k)$ is not differentiable, but is strongly semismooth. At a current iterate $y$, we have a generalized Newton equation:

$$\mathcal{H}_y := \sigma \mathcal{A}\Pi'_+ (U^k(y)) \mathcal{A}^T, \quad \mathcal{H}_y \Delta y = -\nabla_y L(y; X^k). \tag{7}$$

From eigenvalue decomp. $U^k(y) = PDP^T$ with $d_1 \geq \cdots \geq d_r > 0 \geq d_{r+1} \geq \cdots \geq d_n$, we can choose

$$\Pi'_+ (U^k(y))[M] = P(\Omega \circ (P^T MP))P^T, \tag{8}$$

where $\Omega_{ij} = (d_i^+ - d_j^+) / (d_i - d_j)$.

For $\gamma = \{1, \ldots, r\}$ and $\bar{\gamma} = \{r+1, \ldots, n\}$, we have

$$\Omega = \begin{bmatrix}
E_{\gamma\gamma} & \Omega_{\gamma\bar{\gamma}} \\
\Omega_{\bar{\gamma}\gamma} & 0
\end{bmatrix}.$$ 

The (1,1) and (2,2) blocks in $\Omega$ allows for efficient computation of rhs of (8)!
Assume that \((P)\) is strictly feasible and \(A\) is surjective, then inner problem (6) has a solution \(\hat{y}\) and the dual of (6) has a unique solution \(\hat{Z} \succeq 0\).

Let \(\hat{U} = U(\hat{y}; X^k)\). Consider the eigenvalue decomp. \(\hat{U} = PDP^T\) as before, and

\[
\Pi_+'(\hat{U})[M] = P(\Omega \circ (P^T MP))P^T.
\]

Let \(P_\gamma, P_\bar{\gamma}\) be the eigenvectors associated with positive and negative eigenvalues, respectively. Then

\[
\mathcal{H}_\hat{y} = \tilde{A}_1 \tilde{A}_1^T + \tilde{A}_2 D_2 \tilde{A}_2^T + \tilde{A}_3 D_3 \tilde{A}_3^T,
\]

where \(\tilde{A}_1 = AP_\gamma \otimes P_\gamma, \tilde{A}_2 = AP_\gamma \otimes P_\bar{\gamma}, D_2 = \text{vec}(\Omega_{\bar{\gamma}\gamma})\), etc.
Conditioning of generalized Hessian

**Theorem:** If constraint nondegeneracy holds at \( \hat{Z} \), then \( \mathcal{H}_{\hat{y}} \succ 0 \), and

\[
\text{cond}(\mathcal{H}_{\hat{y}}) = \sigma \Theta(1) \text{cond}([\tilde{A}_1, \tilde{A}_2, \tilde{A}_3])^2.
\]

In contrast, for IPM, the Schur complement matrix \( \mathcal{M} \) at a point on the central path with parameter \( \nu \downarrow 0 \) has

\[
\text{cond}(\mathcal{M}) \geq \frac{1}{\nu} \Theta(1) \text{cond}([\tilde{A}_1, \tilde{A}_2, \tilde{A}_3])^2.
\]

Moreover,

\[
\text{cost}(\mathcal{H}_{\hat{y}} \Delta y) = 8 \min\{r, n - r\} n^2 + \text{cost}(\mathcal{A}(\cdot)) + \text{cost}(\mathcal{A}^T(\cdot))
\]

\[
\text{cost}(\mathcal{M} \Delta y) = 4n^3 + \text{cost}(\mathcal{A}(\cdot)) + \text{cost}(\mathcal{A}^T(\cdot))
\]
Practical Newton-CG ALM

- Solve $H_y \Delta y = \text{rhs}$ by CG with diagonal preconditioner.
  Stop when relative-residual $\leq 0.01$.
- Stop the inner iteration when $\|\nabla_y L(y^k; X^k)\| \leq 0.2\|X^{k+1} - X^k\|$.
- Typically ALM needs 30-50 outer iterations, and each requires 5 – 30 Newton steps to solve the inner subproblem (6).

  In contrast, IPM requires about 30-50 iterations each uses only 1 Newton step.
For the boundary-point method of Rendl et al., one step of modified gradient method is used to solve the inner subproblem (6):

\[ y^k = y^{k-1} + (\sigma_k AA^T)^{-1} \nabla_y L(y^{k-1}; X^k). \]
Numerical results

want: \( \text{rel-err} = \max \left\{ \frac{\|R_p\|}{1+\|b\|}, \frac{\|R_d\|}{1+\|C\|}, \frac{\langle X, Z \rangle}{1+|\langle C, X \rangle|+|b^T y|} \right\} \leq 10^{-6} \).

PC: Intel Xeon 3.2GHz with 4G RAM, MATLAB 7.3

<table>
<thead>
<tr>
<th>\theta</th>
<th>parallel IPM 64 nodes 2.4GHz PC</th>
<th>boundary point method</th>
<th>NCG-ALM</th>
</tr>
</thead>
</table>
| \( \theta: \text{theta62} \)  
\( m = 13390, \ n = 300 \) | 459s | 223 | 20 |
|  |  | 95s | 32s |
| \( \theta: \text{theta82} \)  
\( m = 23872, \ n = 400 \) | 2403s | 236 | 21 |
|  |  | 228s | 73s |
| \( \theta: \text{G43} \)  
\( m = 9991, \ n = 1000 \) |  | 2000  
7.5h  
1.2e-5 | 16  
15m |
| \( \text{NCM: 400H1} \)  
\( m = 80.6K, \ n = 400 \) |  | 2000  
1944s  
3.1e-6 | 22  
539s |
### Numerical results

<table>
<thead>
<tr>
<th>Problem</th>
<th>Boundary Point Method</th>
<th>NCG-ALM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rn8m100p3</td>
<td>135 (17m)</td>
<td>11 (27m)</td>
</tr>
<tr>
<td>(m = 100K, n = 800)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QAP: lipa40a</td>
<td></td>
<td>22 (19h)</td>
</tr>
<tr>
<td>(m = 1.28 \times 10^6, n = 1600)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\theta_+: 1zc.2048)</td>
<td></td>
<td>11 (3.6h)</td>
</tr>
<tr>
<td>(m = 2.14 \times 10^6, n = 2048)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\theta: 2dc.512)</td>
<td></td>
<td>27 (2400s)</td>
</tr>
<tr>
<td>(m = 54896, n = 512)</td>
<td></td>
<td>2.2e-5</td>
</tr>
</tbody>
</table>
We have tested NCG-ALM on about 400 SDPs from $\theta, \theta_+, \text{NCM, QAP, binary QP}$.

When the SDPs are primal-dual nondegenerate, NCG-ALM can efficiently solve large SDPs to rather high accuracy.

For SDPs with degeneracies, relative primal infeasibilities can range from $10^{-6}$ to $10^{-3}$, while relative dual infeasibilities are $< 10^{-6}$. 
Thank you!