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
Outline

- Primal and dual SDP
- Examples of large SDP
- Related work
- Projection onto $S^n_+$
- Augmented Lagrangian method (ALM)
- Convergence of ALM
- Conditioning of generalized Hessian
- Numerical results
Standard primal SDP

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

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

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

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

where \( A : S^n \rightarrow R^m \) 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:

$$
\text{(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 $n = 1000$. 

For a graph $G = (V, E)$, 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:

$$
\theta(G) := \max \left\{ \langle E, X \rangle : X_{ij} = 0 \ \forall \ (i,j) \in E, \ \langle I, X \rangle = 1, \ X \succeq 0 \right\} \quad (2)
$$

$$
\theta_+(G) := \frac{n(n+1)}{2} \text{ additional constraints } X \succeq 0 \quad (3)
$$

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

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

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

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

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

$$\min_{X \in S^n} \Phi(X) := \max\{L(y; X) : y \in R^m\}$$  \hspace{1cm} (4)

\textbf{ALM:} Input $X^0 \in S^n_+$, $\sigma_0 > 0$, iterate:

$$y^k \approx \arg\max\{L(y; X^k) : y \in R^m\}$$  \hspace{1cm} (5)

$$X^{k+1} = X^k - \sigma_k \nabla_X L(y^k; X^k) = \Pi_+(U(y^k; X^k))$$

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

If $\|R^k_d := C - AX^k - Z^{k+1}\| \leq \epsilon$; stop; else; update $\sigma_k$; end

- $R^k_d = \nabla_X L(y^k; X^{k+1})$.

- For the inner subproblem (5), optimality condition is $\nabla_y L(y; X^k) = 0$. $R^k_p = b - AX^{k+1} = \nabla_y L(y^k; X^k) \approx 0$ if (5) is solved accurately.
For the inner problem (5), 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

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

Note: larger \(\sigma_{\text{max}} := \max_k \{\sigma_k\}\) leads to faster convergence. But inner subproblem (5) is harder to solve.
A semismooth Newton-CG method for solving inner subproblem

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

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

$$\mathcal{H}_y := \sigma AP_+(U^k(y))A^T, \quad \mathcal{H}_y \Delta y = -\nabla_y L(y; X^k).$$

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

$$P_+(U^k(y))[H] = P(\Omega \circ (P^T HP))P^T,$$

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 (7)!
Conditioning of generalized Hessian

Assume that (P) is strictly feasible and \( A \) is surjective, then inner problem (5) has a solution \( \hat{y} \) and the dual of (5) 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

\[
\Pi'_+(\hat{U})[H] = P(\Omega \circ (P^T HP))P^T.
\]

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

\[
\mathcal{H}_{\hat{y}} = \tilde{A}_1\tilde{A}_1^T + \tilde{A}_2D_2\tilde{A}_2^T + \tilde{A}_3D_3\tilde{A}_3^T,
\]

where \( \tilde{A}_1 = \tilde{A}P_\gamma \otimes P_\gamma, \tilde{A}_2 = \tilde{A}P_{\gamma} \otimes P_{\bar{\gamma}}, D_2 = \text{vec}(\Omega_{\bar{\gamma}\gamma}) \), etc.
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 (5).

  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 (5):

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

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

<table>
<thead>
<tr>
<th></th>
<th>boundary point method</th>
<th>NCG-ALM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rn8m100p3</strong></td>
<td>135</td>
<td>11</td>
</tr>
<tr>
<td>$m = 100K, \ n = 800$</td>
<td>17$m$</td>
<td>27$m$</td>
</tr>
<tr>
<td><strong>QAP: lipa40a</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 1.28\times10^6, \ n = 1600$</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td><strong>θ+ : 1zc.2048</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 2.14\times10^6, \ n = 2048$</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td><strong>θ : 2dc.512</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 54896, \ n = 512$</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2400s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.2e-5</td>
</tr>
</tbody>
</table>
Summary

- We have tested NCG-ALM on about 400 SDPs from $\theta$, $\theta_+$, 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!