Algorithms for Large Scale Structured Optimization Problems

Complexity of a first-order augmented Lagrangian (A.L.) method and A.L. based algorithms for semidefinite programming

(Third Lecture)

Renato D.C. Monteiro
(Georgia Tech)

Fields 09 – Toronto
October 20-24, 2009

Outline of third lecture

- Complexity of a first-order AL Method
  - Problem of interest and its duals
  - Exact augmented Lagrangian Method
  - Background and objectives
  - Termination Criteria
  - Solving of A.L. subproblem
  - Inexact augmented Lagrangian method
  - Iteration-complexity
  - Improving the complexity
- SDP algorithms based on the A.L. method
  - Low-rank method (Burer and M.)
  - Boundary point SDP method (Pohv-Rendl-Wiegele)
  - Newton-CG A.L. method (Zhao, Sun and Toh)
- Other efficient methods

Problem of Interest and its duals

Consider the convex program

$$\mathbf{(CP)} \qquad \mathbf{f}^* := \inf\{\mathbf{f}(\mathbf{x}) : \mathbf{h}(\mathbf{x}) = \mathbf{0},\ \mathbf{x} \in \mathbf{X}\},$$

where $\mathbf{X} \subseteq \Re^{\mathbf{n}}$ is a compact convex set and $\mathbf{f} : \mathbf{X} \to \Re$ is convex and has $\mathbf{L_f}$-Lipschitz-continuous gradient, and $\mathbf{h}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ for some $\mathbf{A} \in \Re^{\mathbf{m}\times\mathbf{n}}$ and $\mathbf{b} \in \Re^{\mathbf{m}}$.

Define the dual function $\mathbf{d} : \Re^{\mathbf{m}} \to \Re$ as

$$\mathbf{d}(\lambda) := \min_{\mathbf{x}\in\mathbf{X}} \left\{ \mathcal{L}(\mathbf{x}, \lambda) := \mathbf{f}(\mathbf{x}) + \lambda^{\mathbf{T}}\mathbf{h}(\mathbf{x}) \right\} \qquad (*)$$

The Lagrangian dual is

$$\max\{\mathbf{d}(\lambda) : \lambda \in \Re^{\mathbf{m}}\}$$

Note: The set of optimal sol's $\mathbf{X}^*$ of CP is nonempty.

Assumption: The set of Lagrange multipliers

$$\mathbf{\Lambda}^* := \{\lambda^* \in \Re^{\mathbf{m}} : \mathbf{d}(\lambda^*) = \mathbf{f}^*\} \neq \emptyset,$$

ρ-augmented dual function: For $\rho > \mathbf{0}$, let

$$\mathbf{d}_\rho(\lambda) := \min_{\mathbf{x}\in\mathbf{X}} \left\{ \mathcal{L}_\rho(\mathbf{x}, \lambda) := \mathbf{f}(\mathbf{x}) + \lambda^{\mathbf{T}}\mathbf{h}(\mathbf{x}) + \frac{\rho}{\mathbf{2}}\|\mathbf{h}(\mathbf{x})\|^{\mathbf{2}} \right\}$$

ρ-augmented dual: $\max\{\mathbf{d}_\rho(\lambda) : \lambda \in \Re^{\mathbf{m}}\}$ has the same optimal value and solution set as $(*)$.

Exact augmented Lagrangian method

Recall that

$$\mathbf{d}_\rho(\lambda) := \min_{\mathbf{x}\in\mathbf{X}} \left\{ \mathcal{L}_\rho(\mathbf{x},\lambda) := \mathbf{f}(\mathbf{x}) + \lambda^{\mathbf{T}}\mathbf{h}(\mathbf{x}) + \frac{\rho}{\mathbf{2}}\|\mathbf{h}(\mathbf{x})\|^{\mathbf{2}} \right\} \qquad (*)$$

Proposition: $\mathbf{d}_\rho(\cdot)$ is concave and has $\mathbf{1}/\rho$-Lipschitz-continuous gradient

$$\nabla\mathbf{d}_\rho(\lambda) = \mathbf{h}(\mathbf{x}^*_\lambda)$$

where $\mathbf{x}^*_\lambda$ denotes an arbitrary optimal sol. of $(*)$.

Augmented Lagrangian Method: Steepest ascent method applied to $\max_\lambda \mathbf{d}_\rho(\lambda)$. For every $\mathbf{k} \geq \mathbf{0}$:

$$\begin{aligned} &\mathbf{x}^*_{\lambda_\mathbf{k}} \in \mathrm{Argmin}_{\mathbf{x}\in\mathbf{X}} \mathcal{L}_\rho(\mathbf{x},\lambda_\mathbf{k}) \\ &\lambda_{\mathbf{k+1}} = \lambda_\mathbf{k} + \rho\nabla\mathbf{d}_\rho(\lambda_\mathbf{k}) = \lambda_\mathbf{k} + \rho\mathbf{h}(\mathbf{x}^*_{\lambda_\mathbf{k}}) \end{aligned}$$

Remark: Can also be view as proximal point method applied to the regular Lagrangian dual, from which convergence follows.

Background and objectives

Background: **1) The augmented Lagrangian method is a classical alg. for nonlinear programming [Bertsekas (04), Ruszczynski(06)].**

2) Recently, it regained a lot of interest due to its efficiency in solving large-scale SDPs and its reformulations [Burer and Monteiro (03, 05), Burer and Vandenbussche (2004), Jarre and Rendl (07), Pohv, Rendl and Wiegele (2006), Zhao, Sun and Toh (08)].

Goal: **Study the complexity of a first-order inexact A.L. method for (CP)**

Issues:

- **How accurately should the Lagrangian subproblem $\min_{\mathbf{x}\in\mathbf{X}} \mathcal{L}_\rho(\mathbf{x}, \lambda_\mathbf{k})$ be solved? Which algorithm to use for that?**
- **How to choose the penalty parameter ρ?**
- **What is the complexity for obtaining a near-optimal solution of CP in terms of total # of inner iterations?**

Termination Criteria

It is well-known that $\mathbf{x}^* \in \mathbf{X}^*$ and $\lambda^* \in \mathbf{\Lambda}^*$ if, and only if, $(\tilde{\mathbf{x}}, \tilde{\lambda}) = (\mathbf{x}^*, \lambda^*)$ satisfies

$$\mathbf{h}(\tilde{\mathbf{x}}) := \mathbf{A}\tilde{\mathbf{x}} - \mathbf{b} = \mathbf{0},$$

$$\nabla \mathbf{f}(\tilde{\mathbf{x}}) + \mathbf{A}^{\mathbf{T}}\tilde{\lambda} \in -\mathcal{N}_{\mathbf{X}}(\tilde{\mathbf{x}}),$$

where $\mathcal{N}_{\mathbf{X}}(\tilde{\mathbf{x}}) := \{\mathbf{s} \in \Re^{\mathbf{n}} : \langle \mathbf{s}, \mathbf{x} - \tilde{\mathbf{x}} \rangle \leq \mathbf{0}, \forall \mathbf{x} \in \mathbf{X}\}$ denotes the normal cone of $\mathbf{X}$ at $\tilde{\mathbf{x}}$

Definition: For $(\epsilon_{\mathbf{p}}, \epsilon_{\mathbf{d}}) \in \Re^{\mathbf{2}}_{++}$, the pair $(\tilde{\mathbf{x}}, \tilde{\lambda}) \in \mathbf{X} \times \Re^{\mathbf{m}}$ is called an $(\epsilon_{\mathbf{p}}, \epsilon_{\mathbf{d}})$-*primal-dual* solution of (CP) if

$$\|\mathbf{h}(\tilde{\mathbf{x}})\| \leq \epsilon_{\mathbf{p}},$$
$$\nabla \mathbf{f}(\tilde{\mathbf{x}}) + \mathbf{A}^{\mathbf{T}}\tilde{\lambda} \in -\mathcal{N}_{\mathbf{X}}(\tilde{\mathbf{x}}) + \mathcal{B}(\epsilon_{\mathbf{d}}),$$

where $\mathcal{B}(\eta) := \{\mathbf{x} \in \Re^{\mathbf{n}} : \|\mathbf{x}\| \leq \eta\}$ for every $\eta \geq \mathbf{0}$.

Solving the Lagrangian subproblem

Note that $\mathcal{L}_\rho(\cdot, \lambda_k)$ has M_ρ-Lipschitz-continuous gradient with $M_\rho := L_f + \rho\|A\|^2$. Hence, the A.L. subproblem $d_\rho(\lambda_k) = \min_{x \in X} \mathcal{L}_\rho(\cdot, \lambda_k)$ can be solved by a first-order algorithm such as Nesterov's optimal method

An inexact A.L. method then consists of two types of iterations:

- **the inner iterations for solving the subproblems**
- **the outer iterations to update λ_k**

The outer iteration is $\lambda_{k+1} = \lambda_k + \rho h(x_k)$, where x_k is an approximate solution of the k-th A.L. subpr.

Proposition: Assume that $x_k \in X$ is such that $\mathcal{L}_\rho(x_k, \lambda_k) - d_\rho(\lambda_k) \le \eta$. Then,

$$\|h(x_k) - \nabla d_\rho(\lambda_k)\| \le \sqrt{\frac{2\eta}{\rho}}$$

Moreover, x_k can be found by Nesterov's optimal method in

$$\mathcal{O}\left(D_X \sqrt{\frac{2M_\rho}{\eta}}\right)$$

iterations, where $D_X := \max\{\|x - \tilde{x}\| : x, \tilde{x} \in X\}$.

INEXACT A.L. METHOD

I-AL Method: Given $\lambda_0 \in \Re^m$, $(\epsilon_p, \epsilon_d) \in \Re^2_{++}$ and $\{\eta_k\} \subseteq \Re_{++}$. Set $k = 0$.

1) find $x_k \in X$ such that $\mathcal{L}_\rho(x_k, \lambda_k) - d_\rho(\lambda_k) \le \eta_k$

2) if $\|h(x_k)\| > 3\epsilon_p/4$, set $\lambda_{k+1} = \lambda_k + \rho h(x_k)$, increment k by 1 and go to step 1

3) find $\tilde{x} \in X$ such that $\mathcal{L}_\rho(\tilde{x}, \lambda_k) - d_\rho(\lambda_k) \le \zeta$, where

$$\zeta := \min\left\{\frac{\rho\epsilon_p^2}{128}, \frac{\epsilon_d^2}{8M_\rho}\right\}$$

4) Stop and output the pair $(\tilde{x}^+, \tilde{\lambda}^+)$ given by

$$\begin{aligned} \tilde{x}^+ &:= \Pi_X(\tilde{x} - \nabla_x \mathcal{L}_\rho(\tilde{x}, \tilde{\lambda})/M_\rho) \\ \tilde{\lambda}^+ &:= \tilde{\lambda} + \rho h(\tilde{x}^+) \end{aligned}$$

Proposition: If the method terminates, then it outputs an (ϵ_p, ϵ_d)-primal-dual solution of CP.

Iteration Complexity

Proposition (Lan and M. (2008): Let

$$\mathbf{N} := \lceil 16\mathbf{D}_{\mathbf{\Lambda}}^{2}/(\rho^{2}\epsilon_{\mathbf{p}}^{2})\rceil$$

where $\mathbf{D}_{\mathbf{\Lambda}} := \min_{\lambda^* \in \mathbf{\Lambda}^*} \|\lambda_{\mathbf{0}} - \lambda^*\|$. If

$$\sum_{\mathbf{k}=\mathbf{0}}^{\mathbf{N}-\mathbf{1}} \eta_{\mathbf{k}} \leq \frac{\rho\epsilon_{\mathbf{p}}^{2}}{\mathbf{128}},$$

then an $(\epsilon_{\mathbf{p}}, \epsilon_{\mathbf{d}})$ -primal-dual solution of CP is found within at most $\mathbf{N}$ outer iterations.

Theorem (Lan and M.): Assume $\mathbf{D}_{\mathbf{\Lambda}}$ is known. If

$$\rho = \frac{\mathbf{4}\,\mathbf{D}_{\mathbf{\Lambda}}^{\frac{3}{4}}\epsilon_{\mathbf{d}}^{\frac{1}{4}}}{\|\mathbf{A}\|^{\frac{1}{4}}\epsilon_{\mathbf{p}}} + \frac{\mathbf{L}_{\mathbf{f}}}{\|\mathbf{A}\|^{\mathbf{2}}}, \qquad \eta_{\mathbf{k}} := \frac{\rho\epsilon_{\mathbf{p}}^{2}}{\mathbf{128N}}, \quad \mathbf{k} = \mathbf{0}, \ldots, \mathbf{N}-\mathbf{1},$$

then the I-AL method computes an $(\epsilon_{\mathbf{p}}, \epsilon_{\mathbf{d}})$-primal-dual solution in at most $\mathcal{O}(\mathcal{I}_{\mathbf{pd}})$ inner iterations, where

$$\mathcal{I}_{\mathbf{pd}} := \left\lceil \mathbf{D}_{\mathbf{X}} \left(\frac{\|\mathbf{A}\|^{\frac{7}{4}}\mathbf{D}_{\mathbf{\Lambda}}^{\frac{3}{4}}}{\epsilon_{\mathbf{p}}\epsilon_{\mathbf{d}}^{\frac{3}{4}}} + \frac{\|\mathbf{A}\|}{\epsilon_{\mathbf{p}}} + \frac{\mathbf{L}_{\mathbf{f}}}{\epsilon_{\mathbf{d}}} \right) + \left(\frac{\mathbf{D}_{\mathbf{\Lambda}}\|\mathbf{A}\|}{\epsilon_{\mathbf{d}}} \right)^{\frac{1}{2}} \right\rceil,$$

and $\mathbf{D}_{\mathbf{X}} = \max_{\mathbf{x_1},\mathbf{x_2}\in\mathbf{X}} \|\mathbf{x_1} - \mathbf{x_2}\|$

Remark: It is possible to develop a scheme which consists of guessing an upper bound $\mathbf{t}$ on $\mathbf{D}_{\mathbf{\Lambda}}$ and then applying the I-AL algorithm with $\mathbf{D}_{\mathbf{\Lambda}}$ replaced by $\mathbf{t}$. Its overall complexity is the same as in the above theorem.

Improving the complexity: Consider the perturbation problem

$$\mathbf{f}^*_\gamma := \min\{\mathbf{f}_\gamma(\mathbf{x}) := \mathbf{f}(\mathbf{x}) + \frac{\gamma}{\mathbf{2}}\|\mathbf{x}-\mathbf{x_0}\|^{\mathbf{2}} : \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in \mathbf{X}\}.$$

where $\mathbf{x_0}$ is a point in $\mathbf{X}$ and $\gamma := \epsilon_{\mathbf{d}}/(\mathbf{2D_X})$. Let

$$\mathcal{L}_{\rho,\gamma}(\mathbf{x},\lambda) := \mathbf{f}(\mathbf{x}) + \frac{\gamma}{\mathbf{2}}\|\mathbf{x}-\mathbf{x_0}\|^{\mathbf{2}} + \lambda^{\mathbf{T}}\mathbf{h}(\mathbf{x}) + \frac{\rho}{\mathbf{2}}\|\mathbf{h}(\mathbf{x})\|^{\mathbf{2}}$$

and

$$\mathbf{d}_{\rho,\gamma}(\lambda) := \min_{\mathbf{x}\in\mathbf{X}} \mathcal{L}_{\rho,\gamma}(\mathbf{x},\lambda) \qquad (*)$$

Denote the set of optimal dual multipliers associated with $(*)$ by $\mathbf{\Lambda}^*_\gamma$.

Exploiting strong-convexity:

Note that the function $\mathcal{L}_{\rho,\gamma}(\cdot,\lambda)$ has $\mathbf{M}_{\rho,\gamma}$-Lipschitz continuous gradient with

$$\mathbf{M}_{\rho,\gamma} := \mathbf{L_f} + \rho\|\mathbf{A}\|^{\mathbf{2}} + \gamma$$

and it is γ-strongly-convex.

Modifications:

- Use a variant of Nesterov's optimal method that takes advantage of the strong convexity;
- Apply the "warm-start" strategy, by which the approximate solution $\mathbf{x_k}$ is used as starting point for solving the Lagrangian subproblem at the next iteration.

Better iteration-complexity

Theorem: Assume that $\mathbf{D}_{\boldsymbol{\Lambda}}^{\gamma} := \inf_{\lambda_\gamma \in \boldsymbol{\Lambda}_\gamma^*} \|\lambda_{\mathbf{0}} - \lambda^*\|$ is known. Then, the I-AL method applied to the perturbed problem with

$$\begin{aligned} \rho &:= \frac{4\mathbf{D}_{\boldsymbol{\Lambda}}^{\gamma}}{\epsilon_{\mathbf{p}}(\log \mathcal{T})^{\frac{1}{2}}} + \frac{\mathbf{L_f} + \gamma}{\|\mathbf{A}\|^{\mathbf{2}}}, \\ \eta_{\mathbf{k}} &:= \frac{\rho\epsilon_{\mathbf{p}}^{\mathbf{2}}}{\mathbf{128N}}, \qquad \mathbf{k} = \mathbf{0}, \ldots, \mathbf{N} - \mathbf{1}, \end{aligned}$$

where $\mathbf{N} := \lceil \mathbf{16}(\mathbf{D}_{\boldsymbol{\Lambda}}^{\gamma})^{\mathbf{2}}/(\rho^{\mathbf{2}}\epsilon_{\mathbf{p}}^{\mathbf{2}}) \rceil$ and

$$\mathcal{T} := \sqrt{\frac{\mathbf{D_X}\mathbf{D}_{\boldsymbol{\Lambda}}^{\gamma}\|\mathbf{A}\|^{\mathbf{2}}}{\epsilon_{\mathbf{p}}\epsilon_{\mathbf{d}}}} + \sqrt{\frac{\mathbf{D_X}\mathbf{L_f}}{\epsilon_{\mathbf{d}}}} + \sqrt{\frac{\mathbf{D_X}\|\mathbf{A}\|}{\epsilon_{\mathbf{p}}}} + \mathbf{4}$$

finds an $(\epsilon_{\mathbf{p}}, \epsilon_{\mathbf{d}})$-primal-dual solution of CP in at most

$$\mathcal{O}(\mathcal{T} \cdot \log \mathcal{T} \cdot \log\log \mathcal{T})$$

inner iterations.

Remark: Same complexity holds even if $\mathbf{D}_{\boldsymbol{\Lambda}}^{\gamma}$ is not known.

Low-rank method for SDP

Consider the SDP

$$(\mathbf{P}) \quad \min\{\langle \mathbf{C}, , \mathbf{X}\rangle : \mathcal{A}\mathbf{X} = \mathbf{b},\ \mathbf{X} \succeq \mathbf{0}\}$$

For a fixed integer $\mathbf{1 \leq r \leq n}$, we have:

$$\mathbf{X} \succeq \mathbf{0},\ \mathbf{rank}(\mathbf{X}) \leq \mathbf{r} \iff \mathbf{X} = \mathbf{V}\mathbf{V}^{\mathbf{T}},\ \mathbf{V} \in \Re^{\mathbf{n}\times\mathbf{r}}$$

Hence,

$$(\mathbf{P_r}) \qquad \min\{\langle \mathbf{C}, \mathbf{X}\rangle : \mathcal{A}\mathbf{X} = \mathbf{b},\ \mathbf{X} \succeq \mathbf{0},\ \mathbf{rank}(\mathbf{X}) \leq \mathbf{r}\},$$

$$\Updownarrow$$

$$(\tilde{\mathbf{P}}_{\mathbf{r}}) \qquad \min\{\langle \mathbf{C}, \mathbf{V}\mathbf{V}^{\mathbf{T}}\rangle : \mathcal{A}(\mathbf{V}\mathbf{V}^{\mathbf{T}}) = \mathbf{b},\ \mathbf{V} \in \Re^{\mathbf{n}\times\mathbf{r}}\},$$

Proposition: $\mathbf{V}$ is a local (resp., global) minimum of $(\tilde{\mathbf{P}}_{\mathbf{r}})$ iff $\mathbf{V}\mathbf{V}^{\mathbf{T}}$ is a local (resp., global) minimum of $(\mathbf{P_r})$.

- $(\tilde{\mathbf{P}}_{\mathbf{n}})$ is equivalent to $(\mathbf{P_n}) = (\mathbf{P})$;
- drawback of $(\tilde{\mathbf{P}}_{\mathbf{n}})$ is its large number of variables, namely $\mathbf{n^2}$ variables.

Key idea: Choose $r \ll n$ so that (P_r) is still equivalent to (P).

Theorem: (Barvinok 1995 and Pataki 1998)
If (P) has an optimal solution then it has one whose rank r^* satisfies

$$r^*(r^* + 1) \leq 2m.$$

As a consequence, if $r \geq \lfloor \sqrt{2m} \rfloor$ then (P_r) is equivalent to (P).

Implementation: (Burer and M. 2003)

- Augmented Lagrangian applied to $(\tilde{P}_r)$. The number of variables is relatively low, namely $nr \ll n^2$;
- r is chosen dynamically; generally, there is no need to have $r \geq \lfloor \sqrt{2m} \rfloor$;
- sparsity is nicely exploited.

Conclusions:

- Number of iterations is large since the method is a first-order method, but the work per iteration is relatively very low.
- No convergence proof is available to support the method, but it never seems to fail in practice.
- It performs very well in practice particularly on problems where the dimension n of X is very large (e.g., SDP relaxations of maxcut problems).

Flops per Iteration:

$$\mathcal{O}\left(\mathbf{r}\,\mathbf{nz}(\mathbf{S}) + \mathbf{nz}(\mathbf{C}) + \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{m}} \mathbf{nz}(\mathbf{A}_{\mathbf{i}})\right)$$

Boundary Point Meth. (Pohv et al.)

Consider the pair of dual SDPs:

$$(\mathbf{P}) \quad \min \quad \mathbf{b^T y} \quad \text{s.t.} \quad \mathcal{A}^*\mathbf{y} - \mathbf{S} = \mathbf{C}, \quad \mathbf{S} \succeq \mathbf{0} \qquad\qquad (\mathbf{D}) \quad \max \quad \langle\, \mathbf{C}, \mathbf{X}\,\rangle \quad \text{s.t.} \quad \mathcal{A}\mathbf{X} = \mathbf{b}, \quad \mathbf{X} \succeq \mathbf{0}$$

Given $\mathbf{X} \in \mathcal{S}^{\mathbf{n}}$ and $\rho > \mathbf{0}$, the augmented Lagrangian function $\mathcal{L}_\rho(\cdot,\cdot;\mathbf{X})$ for $(\mathbf{P})$ is

$$\mathcal{L}_\rho(\mathbf{y},\mathbf{S};\mathbf{X}) := \mathbf{b^T y} + \langle\, \mathbf{X}, \mathbf{C} - \mathcal{A}^*\mathbf{y} + \mathbf{S}\,\rangle + \frac{\rho}{\mathbf{2}}\|\mathbf{C} - \mathcal{A}^*\mathbf{y} + \mathbf{S}\|^{\mathbf{2}}$$

$\forall(\mathbf{y},\mathbf{S}) \in \Re^{\mathbf{m}} \times \mathcal{S}^{\mathbf{n}}_+$, and the assoc. dual function is

$$\mathbf{d}_\rho(\mathbf{X}) = \min\{\mathcal{L}_\rho(\mathbf{y},\mathbf{S};\mathbf{X}) : (\mathbf{y},\mathbf{S}) \in \Re^{\mathbf{m}} \times \mathcal{S}^{\mathbf{n}}_+\} \quad (*)$$

Proposition: **If $\exists\, \mathbf{y} \in \Re^{\mathbf{m}}$ such that $\mathcal{A}^*\mathbf{y} \succ \mathbf{C}$, then**

$$\mathrm{val}(\mathbf{P}) = \mathrm{val}(\mathbf{D}) = \max\{\mathbf{d}_\rho(\mathbf{X}) : \mathbf{X} \in \mathcal{S}^{\mathbf{n}}\}$$

Proposition: **If $(\tilde{\mathbf{y}},\tilde{\mathbf{S}}) = (\tilde{\mathbf{y}}(\mathbf{X}),\tilde{\mathbf{S}}(\mathbf{X}))$ is the optimal solution of $(*)$, then $\nabla\mathbf{d}_\rho(\mathbf{X}) = \mathbf{C} - \mathcal{A}^*\tilde{\mathbf{y}} - \tilde{\mathbf{S}}$.**

Augmented Lagrangian iteration: **Given $\mathbf{X} \in \mathcal{S}^{\mathbf{n}}$, obtain $(\tilde{\mathbf{y}},\tilde{\mathbf{S}}) = (\tilde{\mathbf{y}}(\mathbf{X}),\tilde{\mathbf{S}}(\mathbf{X}))$ by solving $(*)$ and set $\mathbf{X} \leftarrow \mathbf{X} + \rho(\mathbf{C} - \mathcal{A}^*\tilde{\mathbf{y}} - \tilde{\mathbf{S}})$.**

Dual viewpoint: The dual of

$$\mathbf{d}_\rho(\mathbf{X}) = \min\{\mathcal{L}_\rho(\mathbf{y},\mathbf{S};\mathbf{X}) : (\mathbf{y},\mathbf{S}) \in \Re^{\mathbf{m}} \times \mathcal{S}^{\mathbf{n}}_+\} \quad (*)$$

is the problem

$$\max_{\tilde{\mathbf{X}}} \left\{ \langle \mathbf{C}, \tilde{\mathbf{X}} \rangle - \frac{\mathbf{1}}{\mathbf{2}\rho} \|\tilde{\mathbf{X}} - \mathbf{X}\|^{\mathbf{2}} : \mathcal{A}\tilde{\mathbf{X}} = \mathbf{b},\ \tilde{\mathbf{X}} \succeq \mathbf{0} \right\} \quad (**)$$

Notation: For $\mathbf{U} \in \mathcal{S}^{\mathbf{n}}$, let $\mathbf{U}_+$ denote the orthogonal projection onto $\mathcal{S}^{\mathbf{n}}_+$ and $\mathbf{U}_- = (-\mathbf{U})_+$. Clearly, $\mathbf{U} = \mathbf{U}_+ - \mathbf{U}_-$.

Proposition: For $\mathbf{X} \in \mathcal{S}^{\mathbf{n}}$, the following are equivalent:

a) $(\tilde{\mathbf{y}}, \tilde{\mathbf{S}})$ and $\tilde{\mathbf{X}}$ are optimal sol's of $(*)$ and $(**)$

b) if $\mathbf{W} := \mathbf{X}/\rho + \mathbf{C} - \mathcal{A}^*\tilde{\mathbf{y}}$, then $\tilde{\mathbf{X}} = \rho\mathbf{W}_+$, $\tilde{\mathbf{S}} = \mathbf{W}_-$ and

$$\tilde{\mathbf{y}} = (\mathcal{A}\mathcal{A}^*)^{-\mathbf{1}} \left[\mathcal{A}\left(\frac{\mathbf{X}}{\rho} + \mathbf{C} + \tilde{\mathbf{S}} \right) - \frac{\mathbf{b}}{\rho} \right]$$

Remark: Note that $\tilde{\mathbf{X}}, \tilde{\mathbf{S}} \succeq \mathbf{0}$ and $\langle \tilde{\mathbf{X}}, \tilde{\mathbf{S}} \rangle = \mathbf{0}$. The method preserves this property while trying to obtain $\mathcal{A}\tilde{\mathbf{X}} = \mathbf{b}$ and $\mathcal{A}^*\tilde{\mathbf{y}} + \tilde{\mathbf{S}} = \mathbf{C}$.

Boundary Point Method: Choose $\rho > 0$, sequence $\{\eta_k\} \downarrow 0$ and $\epsilon > 0$.

```
0) Set $k = 0$ and $X_0 = S_0 = 0$;
1) Set $\tilde{S} = S_k$;
    2) Compute
        $\tilde{y} = (\mathcal{A}\mathcal{A}^*)^{-1}\left[\mathcal{A}\left(\frac{X_k}{\rho} + C + \tilde{S}\right) - \frac{b}{\rho}\right]$   (1)
        $W = X_k/\rho + C - \mathcal{A}^*\tilde{y}$
     and set $\tilde{S} = W_-$ and $\tilde{X} = \rho W_+$.
    3) if $\|\mathcal{A}\tilde{X} - b\| > \rho\eta_k$, then go to step 2.
4) set $(X_{k+1}, S_{k+1}) = (\tilde{X}, \tilde{S})$, $y_{k+1} = \tilde{y}$ and $k \leftarrow k+1$
5) if $\|C + S_k - \mathbf{A}^* y_k\| > \epsilon$, then go to 1); else *stop*.
```

Alternating direction viewpoint:

$(1) \iff \tilde{y} = \mathrm{argmin}\{\mathcal{L}_\rho(y, \tilde{S}; X_k) : y \in \Re^m\}$

$\tilde{S} = W_- \iff \tilde{S} = \mathrm{argmin}\{\mathcal{L}_\rho(\tilde{y}, S; X_k) : S \succeq 0\}$

$\tilde{X} = \rho W_+ \iff \tilde{X} = X_k + \rho(C + \tilde{S} - \mathcal{A}^*\tilde{y})$

Simplified Boundary Point Method: Choose $\rho > 0$ and $\epsilon > 0$.

0) Set $\mathbf{k = 0}$ and $(\mathbf{y_0}, \mathbf{S_0}, \mathbf{X_0}) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$;

1) If $\max\{\|\mathcal{A}\mathbf{X_k} - \mathbf{b}\|, \|\mathbf{C} + \mathbf{S_k} - \mathcal{A}^*\mathbf{y_k}\|\} \le \epsilon$, then *stop*.

2) Compute

$$\begin{aligned} \mathbf{y_{k+1}} &= (\mathcal{A}\mathcal{A}^*)^{-1}\left[\mathcal{A}\left(\frac{\mathbf{X_k}}{\rho} + \mathbf{C} + \mathbf{S_k}\right) - \frac{\mathbf{b}}{\rho}\right] \quad (1) \\ \mathbf{W} &= \mathbf{X_k}/\rho + \mathbf{C} - \mathcal{A}^*\mathbf{y_{k+1}} \end{aligned}$$

and set $\mathbf{S_{k+1}} = \mathbf{W_-}$ and $\mathbf{X_{k+1}} = \rho\mathbf{W_+}$.

3) Set $\mathbf{k} \leftarrow \mathbf{k+1}$ and go to step 1).

Proposition: Assume that set of optimal solutions of $(\mathbf{P})$ is non-empty and that $(\mathbf{P})$ satisfies the Slater condition. Assume also that $\mathcal{A}^*$ is one-to-one. Then, $\{(\mathbf{y_k}, \mathbf{S_k}, \mathbf{X_k})\}$ converges to a primal-dual optimal solution.

Work per iteration: Discarding the processing of the factoriztion of $\mathcal{A}\mathcal{A}^*$, storage is $\mathcal{O}(\mathbf{n^2})$ and number of flops is

$$\mathcal{O}\left(\mathbf{n^3} + \mathbf{nz(C)} + \sum_{\mathbf{i=1}}^{\mathbf{m}} \mathbf{nz(A_i)}\right)$$

Conclusions:

- Being a first-order method, its number of iterations is usually large compared to second-order methods but can be considerably lower than that of the low-rank method on some classes of SDP problems.

- Its work per iteration and amount of storage is at least $\mathcal{O}(\mathbf{n^3})$ and $\mathcal{O}(\mathbf{n^2})$, respectively, making it prohibitively expensive for problems with large variable matrix $\mathbf{X}$ (i.e., with $\mathbf{n \geq 5,000}$).

- It performs very well in practice particularly on problems where the dimension $\mathbf{n}$ of $\mathbf{X}$ is not so large (e.g., $\mathbf{\leq 3,000}$). It is generally efficient for problems with large $\mathbf{m}$ (i.e., the size of $\mathbf{y}$).

- It has been reported that the method enconters more difficulty on extremely sparse or extremely dense problems or on instances where either the optimal $\mathbf{X}$ or $\mathbf{S}$ has small rank.

- Convergence proof is available to support the method but no iteration-complexity is known for it yet.

Newton-CG A.L. method (Zhao et al.)

Zhao et al.'s method is an implementation of the augmented Lagrangian method. Consider the same SDP as in the previous method and recall that

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{y},\mathbf{S};\mathbf{X_k}) &:= \mathbf{b^T y} + \langle\, \mathbf{X_k}, \mathbf{C} - \mathcal{A}^*\mathbf{y} + \mathbf{S}\,\rangle + \tfrac{\rho}{2}\|\mathbf{C} - \mathcal{A}^*\mathbf{y} + \mathbf{S}\|^2 \\ &= \mathbf{b^T y} + \tfrac{1}{2\rho}\left(\|\rho\mathbf{S} + \mathbf{W_k(y)}\|^2 - \|\mathbf{X_k}\|^2\right)\end{aligned}$$

where $\mathbf{W_k(y)} := \mathbf{X_k} - \rho(\mathbf{A^*y} - \mathbf{C})$. Hence,

$$\begin{aligned}\mathbf{d}_\rho(\mathbf{X_k}) &= \min\{\mathcal{L}_\rho(\mathbf{y},\mathbf{S};\mathbf{X_k}) : (\mathbf{y},\mathbf{S}) \in \Re^{\mathbf{m}} \times \mathcal{S}^{\mathbf{n}}_+\} \\ &= \quad \min_{\mathbf{y}} \mathbf{L_k(y)} := \mathbf{b^T y} + \frac{\mathbf{1}}{\mathbf{2}\rho}\left(\|[\mathbf{W_k(y)}]_+\|^2 - \|\mathbf{X_k}\|^2\right) \quad (*)\end{aligned}$$

Clearly, if $\tilde{\mathbf{y}}_\mathbf{k}$ is an optimal solution of (*) and $\tilde{\mathbf{S}}_\mathbf{k} = \rho^{-1}[\mathbf{W_k}(\tilde{\mathbf{y}}_\mathbf{k})]_-$, then

$$\mathbf{X_{k+1}} = \mathbf{X_k} + \rho(\mathbf{C} - \mathcal{A}\tilde{\mathbf{y}}_\mathbf{k} - \tilde{\mathbf{S}}_\mathbf{k}) = [\mathbf{W_k}(\tilde{\mathbf{y}}_\mathbf{k})]_+$$

A semi-smooth Newton-CG algorithm is used to solve subproblem $(*)$. The gradient of the o.f. of $(*)$ is

$$\nabla_\mathbf{y}\mathbf{L_k(y)} = \mathbf{b} - \mathcal{A}[\mathbf{W_k(y)}]_+, \quad \forall \mathbf{y} \in \Re^{\mathbf{m}},$$

which is almost everywhere Frechet-differentiable.

An approximate sol $\mathbf{y_k}$ of $(*)$ satisfying $\mathbf{L_k(y_k)} - \mathbf{d}_\rho(\mathbf{X_k}) \leq \eta_\mathbf{k}$ is computed and then the update $\mathbf{X_{k+1}} = [\mathbf{W_k(y_k)}]_+$ is performed, where $\{\eta_\mathbf{k}\}$ satisfies $\sum_{\mathbf{k=0}}^{\infty} \eta_\mathbf{k} < \infty$.

Work per iteration:

- Each CG step is on the other of

$$\mathcal{O}\left(\beta_{\mathbf{y}}\mathbf{n}^{\mathbf{2}} + \sum_{\mathbf{i=1}}^{\mathbf{m}} \mathbf{nz}(\mathbf{A_i})\right)$$

  where $\beta_{\mathbf{y}} = \min\{\gamma_{\mathbf{y}}, \mathbf{n} - \gamma_{\mathbf{y}}\}$, $\gamma_{\mathbf{y}} = \mathbf{rank}[\mathbf{W_k}(\mathbf{y})]_-$ and $\mathbf{y}$ is the current Newton iterate.

- Each Newton step requires a new eigenvalue factorization.

Conclusions:

- Its work per iteration and amount of storage is at least $\mathcal{O}(\beta_{\mathbf{y}}\mathbf{n}^{\mathbf{2}})$ and $\mathcal{O}(\mathbf{n}^{\mathbf{2}})$, respectively, making it prohibitively expensive for problems with large variable matrix $\mathbf{X}$ (i.e., with $\mathbf{n} \geq \mathbf{5,000}$).

- It performs very well in practice particularly on problems where the dimension $\mathbf{n}$ of $\mathbf{X}$ is not so large (e.g., $\leq \mathbf{3,000}$). It is generally efficient for problems with large $\mathbf{m}$ (i.e., the size of $\mathbf{y}$).

- Convergence proof is available to support the method but no iteration-complexity is known for it yet.

Other efficient methods

Methods based on barrier functions:

1) Toh and Kojima (2002), Toh (2004) and the SPDA code developed by Kojima and his collaborators.

These are IP methods based on iterative solvers, applied to either an augmented system or its Schur complement (normal) system.

2) Kocvara and Stingl (2003, 2005) - Modified barrier method (PENNON)

For $(\mathbf{X_k}, \sigma_\mathbf{k}) \in \mathcal{S}^\mathbf{n} \times \Re_{++}$, it is based on the modified barrier subproblem

$$\min_{\mathbf{y}} \ \mathbf{b}^\mathbf{T}\mathbf{y} + \langle\, \mathbf{X_k}, [\sigma_\mathbf{k}^\mathbf{2}(\mathcal{A}^*\mathbf{y} - \mathbf{C} + \sigma_\mathbf{k}\mathbf{I})^{-\mathbf{1}} - \sigma_\mathbf{k}\mathbf{I}]\,\rangle$$

which is solved by a Newton-CG approach. The amount of work per iteration and storage is quite similar to those for IP methods based on iterative solvers.

THANK YOU!

AND

THE END