

# Interior-point and augmented Lagrangian methods for semidefinite programming

Kim-Chuan Toh

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

CDO Program, MIT, 2008

## Focus of this talk

IPM: interior-point method

ALM: augmented Lagrangian method

- I. IPMs for solving medium scale SDP
- II. Inexact IPMs for solving large scale SDP
- III. ALMs for solving large scale SDP

How to efficiently solve **large dense ill-conditioned linear systems** arising from IPMs/ALMs for SDP.

**Solve  $Mx = b$  by Cholesky factorization** ▷p.3:

$m = \dim(M) = 10,000$  and dense

$\Theta(m^3)$  flops  $\Rightarrow$  CPU time = 2 mins in MATLAB, memory = 2 GB.

## Part I: **IPM for medium scale SDP**

Based on work done with M.J. Todd and R.H. Tutuncu.

- Primal and dual SDP
- Examples of SDP
- Optimality conditions, central path, and path-following IPM
- Newton step for central path
- Computation of search direction
- Practical IPMs and software

## I. Standard primal SDP

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

$Q \succeq 0$  ( $Q \succ 0$ )  $\Rightarrow$   $Q$  is symmetric positive semidefinite (definite).

Given data:  $C, A_1, \dots, A_m \in \mathcal{S}^n, b \in \mathbb{R}^m$

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

where  $\mathcal{A} : \mathcal{S}^n \rightarrow \mathbb{R}^m$  is the linear map s.t.

$$\mathcal{A}(X) = \left[ \langle A_1, X \rangle, \dots, \langle A_m, X \rangle \right]^T.$$

Assume (P) is feasible.

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

For medium scale SDP:  $n \leq 3000, m \leq 10000$ .

## I. Special case: LP

Suppose data is diagonal:  $C = \text{diag}(c)$ ,  $A_k = \text{diag}(a_k) \forall k$ .

Restrict to  $X = \text{diag}(x)$ , (P) reduces to standard primal LP:

$$\min\{\langle c, x \rangle : Ax = b, \quad x \geq 0, x \in R^n\}$$

with  $A = [a_1, \dots, a_m]^T$ .

## I. Standard dual SDP

$$\begin{aligned} \text{(D)} \quad & \max \quad b^T y \\ & \text{s.t.} \quad \mathcal{A}^T(y) + Z = C, \quad Z \succeq 0, \quad y \in R^m, \quad Z \in S^n \end{aligned}$$

where  $\mathcal{A}^T : R^m \rightarrow S^n$  is the adjoint of  $\mathcal{A}$  and  $\mathcal{A}^T(y) = \sum_{k=1}^m y_k A_k$ .

Assume (D) is feasible.

---

For “diagonal data”, restrict to  $Z = \text{diag}(z)$ , (D) reduces to standard dual LP:

$$\max\{b^T y : \mathcal{A}^T y + z = c, \quad z \geq 0, \quad y \in R^m\}$$

## I. More general forms of SDP

- May include: multiple SDP blocks
- second-order cone constraints for vectors:  $\|x\|_2 \leq \alpha$
- non-negative constraints for vectors:  $x \geq 0$
- unrestricted vectors
- log-determinant terms

$$\min \sum_j \langle C_j^s, X_j^s \rangle - w_j \log \det(X_j^s) + \sum_k \langle c_k^q, x_k^q \rangle + \langle c^l, x^l \rangle + \langle c^u, x^u \rangle$$

$$\text{s.t. } \sum_j A_j^s(X_j^s) + \sum_k A_k^q x_k^q + A^l x^l + A^u x^u = b$$

$$X_j^s \succeq 0, \quad x_k^q \succeq 0, \quad x^l \geq 0, \quad x^u \text{ free.} \quad (\text{SQLP})$$

## I. 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:

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

$\Downarrow$

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

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

## I. Examples of SDP: maximum stable set 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 maximum stable set problem is to find the stable set with maximum cardinality. Let

$$x_i = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad |S| = \sum_{i=1}^n x_i.$$

A common formulation of the max-stable-set problem:

$$\alpha(G) := \max \left\{ |S| = \frac{1}{|S|} \sum_{ij} x_i x_j : x_i x_j = 0 \forall (i, j) \in \mathcal{E}, x \in \{0, 1\}^n \right\}$$

$$\Downarrow X := xx^T / |S|$$

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

SDP relaxation:  $X = xx^T / |S| \Rightarrow X \succeq 0$ , get

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

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

## I. Optimality conditions and central path

Path-following IPMs are based solving the following central-path equations, where  $\nu > 0$  is a parameter:

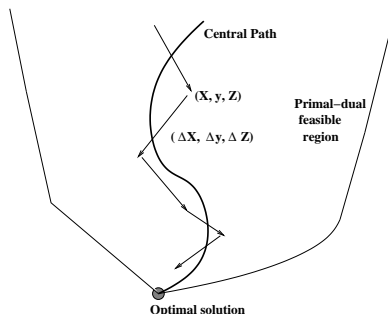
$$(CPE)_\nu \quad 0 = \begin{pmatrix} R_p & := & b - AX \\ R_d & := & C - Z - A^T y \\ R_c & := & \nu I - XZ \end{pmatrix}, \quad X, Z \succ 0$$

Note:  $\nu = 0$  corresponds to optimality conditions for (P), (D).

**Theorem:** Suppose (P) and (D) are strictly feasible, and  $\{A_1, \dots, A_m\}$  linearly independent. Then the solutions  $\{(X^\nu, y^\nu, Z^\nu) : \nu > 0\}$  to (CPE) exist (the set is called central-path). As  $\nu \downarrow 0$ , this path converges to an optimal solution  $(X^*, y^*, Z^*)$  of (P) and (D) satisfying  $X^* Z^* = 0$ .

# I. Path-following interior-point method

Based on tracing the central path by a Newton-type method.



Theoretical infeasible path-following IPMs need at most  $O(n^2 \log(1/\epsilon))$  iterations to get an  $\epsilon$ -optimal solution satisfying  $\text{rel-err} \leq \epsilon$ , where

$$\text{rel-err} = \frac{\max\{\|R_p\|, \|R_d\|, \langle X, Z \rangle\}}{\text{initial value}} \quad (4)$$

## I. Newton step for (CPE)

At the current iterate  $(X, y, Z)$  with  $X, Z \succ 0$ ,  
let  $W \succ 0$  (dense!) be the scaling matrix satisfying  $WXW = Z$ .

$W = X^{-1/2}Z^{1/2}$  if  $X, Z$  commute. Generally  
 $W = X^{-1/2}(X^{1/2}ZX^{1/2})^{1/2}X^{-1/2}$ , computable via eigenvalue and Cholesky  
decomp.

Compute search direction from **symmetrized Newton equation**:

$$\left. \begin{array}{lcl} \mathcal{A}(\Delta X) & = & R_p \\ \mathcal{A}^T(\Delta y) + \Delta Z & = & R_d \\ W \otimes W(\Delta X) + \Delta Z & = & \hat{R}_c \end{array} \right\} 2n^2 + m \text{ equations}$$

---

For LP:  $W \otimes W \Rightarrow W^2 = X^{-1}Z$  with  $X = \text{diag}(x), Z = \text{diag}(z)$ .

## I. Computation of search direction $(\Delta X, \Delta y, \Delta Z)$

Eliminate  $\Delta Z$ , solve:

$$(AE) \quad \begin{matrix} n^2 \\ m \end{matrix} \begin{bmatrix} -W \otimes W & \mathcal{A}^T \\ \mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta y \end{bmatrix} = \text{rhs}, \quad (5)$$

or further eliminate  $\Delta X$ , solve:

$$(SCE) \quad \underbrace{\mathcal{A}(W^{-1} \otimes W^{-1})\mathcal{A}^T}_{\mathcal{M} \quad (m \times m \text{ dense, pd})} \Delta y = \text{rhs}$$

---

For LP: the matrices are

$$\begin{matrix} n \\ m \end{matrix} \begin{bmatrix} -W^2 & A^T \\ A & 0 \end{bmatrix}, \quad AW^{-2}A^T$$

where  $W^2 = X^{-1}Z$ . They are sparse if  $A$  is sparse.

## I. Computation via (SCE) by a direct solver

- Compute and store  $\mathcal{M}$ :  $\mathcal{M}_{ij} = \langle A_i, T := W^{-1}A_jW^{-1} \rangle$ ,  $i, j = 1, \dots, m$   
 $\Rightarrow$  cost =  $m(\Theta(n^3) + \Theta(mn^2))$  flops if ignoring sparsity in  $\{A_k\}$ .

Suppose  $A_k = e_k e_k^T$ . For  $m, n = 500$ ,  $\begin{cases} 0.5 \text{ secs} & \text{if sparsity is exploited} \\ 180 \text{ secs} & \text{if sparsity is ignored} \end{cases}$

- Factor  $\mathcal{M}$  and compute  $\Delta y$ :  $\Rightarrow$  cost =  $\Theta(m^3)$
- Compute  $\Delta Z$  and  $\Delta X$ :  $\Delta Z = R_d - \mathcal{A}^T \Delta y$ ,  $\Delta X = W^{-1}(\widehat{R}_c - \Delta Z)W^{-1}$ .

Cost per iteration =  $O(mn^3) + O(m^2n^2) + \Theta(m^3) + \Theta(n^3)$ .

Memory  $\geq 8m^2 + \Theta(n^2)$  bytes

## I. Major computations at each iteration

Major computation time per IPM iteration:

Parts	(m, n=2000)	(m, n=1596, 165)	(m, n=4375, 300)
Total time	385 secs	225 secs	187 secs
Compute M	1 %	76 %	15 %
Factor M	4	11	76
W, rhs, dX	80	13	5

## I. Practical interior-point methods

Many application problems are modelled as SDPs because they can be solved efficiently.

Main factors contributing to very efficient SDP solvers:

- Practical IPMs need much fewer iterations than the worst-case complexities of  $O(n^2)$ . For 360 test cases with  $n$  up to few thousands, SDPT3 need less than 100 iterations ( $95\% \leq 50$ ) to get  $10^{-6}$  accuracy (see 4).
- Very powerful numerical linear algebra routines are available in LAPACK for matrix-matrix multiplication, eigenvalue decomposition, Cholesky/LU factorization, etc.
- Ability to exploit structures (sparsity, low-rank, etc) in SDP data to cut cost per iteration.

## Softwares for medium scale SDPs

medium scale:  $n \leq 3000$ ,  $m \leq 6000$

NEOS SDP Solvers: <http://www-neos.mcs.anl.gov/>

software	lang.	method	authors
CSDP	c	p-d ipm	Borchers
SDPA	c++	p-d ipm	Kojima et al.
SDPA-c	c++	p-d ipm, pd matrix compl.	Kojima et al.
SDPT3	matlab+c	p-d ipm +self-dual embed.	Toh-Todd-Tutuncu
SeDuMi	matlab+c	p-d ipm, self-dual embed.	Sturm-->McMaster
DSDP	c	dual ipm	Benson-Ye
PENSDP	matlab	generalized Lag. on (D)	Kocvara-Stingl
SDPLR	c	aug. Lag. on (P), $X = RR'$	Burer-Monteiro

SDPT3 and SeDuMi: can solve general SQLPs.

Benchmarks by Mittelmann: [http://plato.asu.edu/ftp/sparse\\_sdp.html](http://plato.asu.edu/ftp/sparse_sdp.html)

## Part II: **Inexact IPM for large SDP**

- Existing work
- Inexact IPM
- Condition number of Schur complement matrix
- Computation of direction via reduced AE
- Numerical results

## II. Existing work

- $n \times n$  matrix variable large:  $n \geq 5,000$ .  $X$  is dense.  
dual scaling IPM [Benson-Ye] ← avoid  $X$   
 $pd$  matrix compl. [Kojima et al., Burer] ← since  $X^{-1}$  may be sparse  
low-rank factorization [Burer-Monteiro] ← set  $X = RR^T$   
Spectral bundle [Helmberg-Rendl, Nayakkankuppam]
- Number of constraints  $m$  is large:  $m \geq 10,000$ .  
 $\mathcal{M}$  cannot be stored:  $m = 10^5 \Rightarrow$  need 100G Bytes.  
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]Generalized Lagrangian method [Kocvara-Stingl]  
Inexact IPM ← compute direction via iterative solvers [Kojima, Toh]  
Augmented Lagrangian method [Rendl et al., Sun-Toh-Zhao]

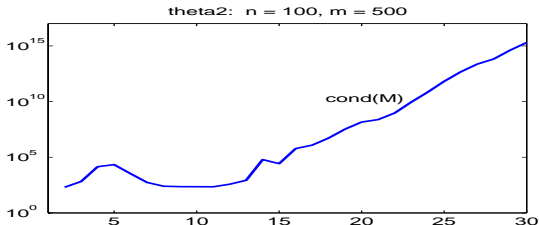
Assume:  $n \leq 2000$  so that  $X$  can be stored.

## II. Inexact IPM

Inexact IPM: solve  $\mathcal{M}\Delta y = h$  approximately by conjugate gradient (CG) method.

- $\mathcal{M}$  need not be computed since CG only need matrix-vector multiply.
- Matrix-vector multiply  $\mathcal{M}y = \mathcal{A}(W^{-1}(\mathcal{A}^T y)W^{-1})$  can be computed at moderate cost of  $\Theta(n^3)$  flops.
- If  $\text{cond}(\mathcal{M}) := \lambda_{\max}/\lambda_{\min}$  is moderate, CG gives good solution in number of steps  $\ll m$ . Roughly, number of CG steps needed to achieve a desired accuracy is proportional to  $\sqrt{\text{cond}(\mathcal{M})}$  when  $\text{cond}(\mathcal{M}) \gg 1$ .

- $\mathcal{M}$  is ill-conditioned with condition number  $\geq \Theta(1/\nu)$  as  $\nu \downarrow 0$ .



- Preconditioning:  $\widehat{\mathcal{M}}^{-1}\mathcal{M}x = \widehat{\mathcal{M}}^{-1}h$ .  
Desirable preconditioner  $\widehat{\mathcal{M}}$ : cheap to invert, improve the eigenvalue distribution of  $\mathcal{M}$ .
- $\mathcal{M}$  is dense, and not stored  $\Rightarrow$  difficult to construct preconditioners.  
 $\widehat{\mathcal{M}} = \text{diag}(\mathcal{M})$  is the only obvious and reasonably good choice.

## II. Conditioning of $\mathcal{M}^\nu$ along the central path

Consider the eigenvalue decomposition:  $W = PDP^T$ . Then

$$W \otimes W = \mathcal{P} \mathcal{D} \mathcal{P}^T, \quad \text{with } \mathcal{P} = P \otimes P, \quad \mathcal{D} = D \otimes D.$$

For  $(X, y, Z)$  on the central path with  $XZ = \nu I$  and  $\nu \ll 1$ , the diagonal entries of  $\mathcal{D}$  must separate into 3 groups with orders  $\nu, 1, 1/\nu$ :

$$\mathcal{D} = \text{diag}(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3).$$

Let  $\tilde{\mathcal{A}} := \mathcal{A}\mathcal{P}$ , and partition  $\mathcal{P}, \tilde{\mathcal{A}}$  accordingly:

Then 
$$\mathcal{P} = [\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3], \quad \tilde{\mathcal{A}} = \mathcal{A}\mathcal{P} = [\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2, \tilde{\mathcal{A}}_3].$$

$$\mathcal{M}^\nu = \mathcal{A}(W \otimes W)^{-1} \mathcal{A}^T = \tilde{\mathcal{A}} \mathcal{D}^{-1} \tilde{\mathcal{A}}^T = \sum_{j=1}^3 \tilde{\mathcal{A}}_j \mathcal{D}_j^{-1} \tilde{\mathcal{A}}_j^T$$

$$\Rightarrow \|\mathcal{M}^\nu\| = \Theta(1/\nu).$$

---

For LP:  $\tilde{\mathcal{A}} = [\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_3] = [A_B, A_N]$ ,  $\mathcal{D}_1 = \nu X_B^{-2}$ ,  $\mathcal{D}_3 = Z_N^2/\nu$

## II. Condition number of Schur complement matrix

**Theorem:**  $\|\mathcal{M}^\nu\| = \Theta(1/\nu)$ .

$$\|(\mathcal{M}^\nu)^{-1}\| = \begin{cases} \Theta(1) & \text{if } X^* \text{ is non-degenerate} = [\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2] \text{ has full row rank} \\ \Theta(1/\nu) & \text{otherwise} \end{cases}$$

---

For LP:  $x^*$  is non-degenerate if the basic matrix  $A_B$  has full row-rank.

## II. Computation of direction via reduced AE

Using  $W \otimes W = \mathcal{P} \mathcal{D} \mathcal{P}^T$ , the AE (5) is equivalent to

$$\begin{matrix} n^2 \\ m \end{matrix} \begin{bmatrix} -\mathcal{D} & \tilde{\mathcal{A}}^T \\ \tilde{\mathcal{A}} & 0 \end{bmatrix} \begin{bmatrix} \Delta \tilde{X} \\ \Delta y \end{bmatrix} = \text{rhs}$$

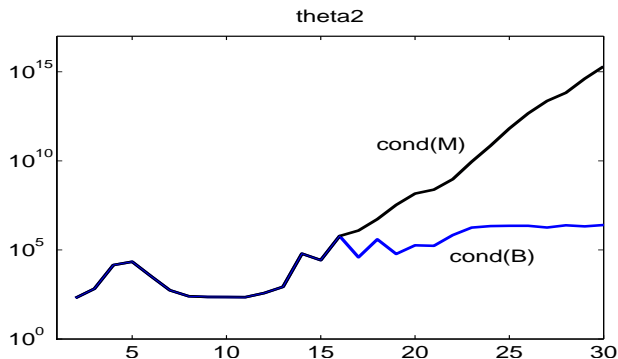
Partition  $\Delta \tilde{X} := \mathcal{P}^T(\Delta X) = [\Delta \tilde{X}_1, \Delta \tilde{X}_2, \Delta \tilde{X}_3]$  according to those of  $\mathcal{D} = [\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3]$ . After some simple algebra and eliminating  $\Delta \tilde{X}_2, \Delta \tilde{X}_3 \Rightarrow$

$$\text{(RAE)} \quad \begin{matrix} p \\ m \end{matrix} \underbrace{\begin{bmatrix} -\mathcal{D}_1 & \tilde{\mathcal{A}}_1^T \\ \tilde{\mathcal{A}}_1 & \tilde{\mathcal{A}} \text{diag}(I, \mathcal{D}_2^{-1}, \mathcal{D}_3^{-1}) \tilde{\mathcal{A}}^T \end{bmatrix}}_{\mathcal{B}} \begin{bmatrix} \Delta \tilde{X}_1 \\ \Delta y \end{bmatrix} = \text{rhs}$$

Note:  $p \leq m$ .

## II. Conditioning of RAE

**Theorem:** If optimal  $(X^*, Z^*)$  is primal and dual non-degenerate, and strictly complementary, then  $\text{cond}(B)$  is bounded independent of  $\nu$ .



## Numerical results

Find  $X, y, Z$  such that

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

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

	direct	inexact-SCE	inexact-RAE
$\theta$ : brock200-4 $m = 6812, n = 200$	12 342s	16 210s	15 59s
$\theta$ : theta62 $m = 13390, n = 300$	15 2280s	17 727s	16 204s
NCM: 200H1 $m = 20300, n = 200$	$\approx 20$ $\geq 5.3h$	28 954s (9e-6)	22 190s
NCM: 400H1 $m = 80600, n = 400$	$\approx 20$ $\geq 350h$	31 5h (4e-6)	26 2176s

Red numbers are estimated total time to perform Cholesky factorization of  $\mathcal{M}$ .

## Part III: **Augmented Lagrangian methods for large SDP**

Based on work done with D.F. Sun and X.Y Zhao.

- Projection on  $\mathcal{S}_+^n$
- Augmented Lagrangian method
- Convergence of ALM
- A semi-smooth Newton-CG method for inner subproblem
- Practical ALM
- Numerical results

### III. Projection onto $S_+^n$

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

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

$$\min\{\|Y \mp X\|^2 : X \succeq 0\} = \|\Pi_{\mp}(Y)\|^2.$$

### III. Augmented Lagrangian function for (D)

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

Let  $X \in \mathcal{S}^n$  be the multiplier associated with the equality constraint. Construct

$$\begin{aligned} \tilde{L}(y, Z; X) &= b^T y + \langle X, C - \mathcal{A}^T y - Z \rangle - \frac{\sigma}{2} \|C - \mathcal{A}^T y - Z\|^2 \\ &= b^T y + \frac{1}{2\sigma} \|X\|^2 - \frac{1}{2\sigma} \|U(y; X) + \sigma Z\|^2. \end{aligned}$$

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

$$\begin{aligned} L(y; X) &:= \max \left\{ \tilde{L}(y, Z; X) : Z \succeq 0 \right\} \\ &= b^T y + \frac{1}{2\sigma} \|X\|^2 - \frac{1}{2\sigma} \|\Pi_+(U(y; X))\|^2 \end{aligned}$$

### III. Augmented Lagrangian method

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

$$\min_{X \in \mathcal{S}^n} \mathcal{L}(X) := \max\{L(y; X) : y \in R^m\} \quad (7)$$

**ALM:** Input  $X^0 \in \mathcal{S}_+^n$ ,  $\sigma_0 > 0$ , iterate:

$$y^k \approx \operatorname{argmax}\{L(y; X^k) : y \in R^m\} \quad (8)$$

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

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

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

- $R_d^k = \nabla_X L(y^k; X^{k+1})$ .
- For the inner subproblem (8), optimality condition is  $\nabla_y L(y; X^k) = 0$ .  
 $R_p^k = b - \mathcal{A}X^{k+1} = \nabla_y L(y^k; X^k) \approx 0$  if (8) is solved accurately.

### III. Convergence of ALM

For the inner problem (8), if we use the stopping condition below:

$$\|\nabla_y L(y^k; X^k)\| \leq (\delta_k/\sigma_k)\|X^{k+1} - X^k\|, \quad \delta_k \rightarrow 0,$$

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  $\sigma_{\max}$  leads to faster convergence. But inner subproblem (8) is harder to solve.

### III. A semismooth Newton-CG method for solving inner subproblem

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

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

$$\mathcal{H}_y := \sigma \mathcal{A} \Pi'_+(U(y)) \mathcal{A}^T, \quad \mathcal{H}_y \Delta y = -\nabla_y L(y; X^k). \quad (9)$$

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

$$\Pi'_+(U(y))[H] = Q(\Omega \circ (Q^T H Q))Q^T, \quad (10)$$

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

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

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

Such a structure in  $\Omega$  allows for efficient computation of the rhs of (10).

### III. Conditioning of generalized Hessian

Assume (P) is strictly feasible and  $\mathcal{A}$  is surjective, then inner problem (8) has a solution  $\hat{y}$  and the dual of (8) has a unique solution  $\hat{Z} \succeq 0$ .

**Theorem:** If constraint nondegeneracy holds at  $\hat{Z}$ , then  $\mathcal{H}_{\hat{y}}$  is positive definite, and

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

▷p.6

In contrast, for IPM where  $\mathcal{M} = \mathcal{A}W^{-1} \otimes W^{-1}\mathcal{A}^T$ , we have

$$\text{cond}(\mathcal{M}) \leq \frac{\|X\|^2}{\nu} \Theta(1) \text{cond}([\tilde{\mathcal{A}}_1, \tilde{\mathcal{A}}_2])^2.$$

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

$$\text{cost}(\mathcal{M} \Delta y) = 4n^3 + \text{cost}(\mathcal{A}(\cdot)) + \text{cost}(\mathcal{A}^T(\cdot))$$

### III. Practical ALMs

- 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 (8).

In contrast, IPM requires about 30-50 iterations each uses only 1 Newton step.

### III. Numerical results

	parallel IPM 64 nodes 2.4GHz PC	inexact RAE	NCG-ALM
$\theta$ : theta62 $m = 13390, n = 300$	459s	16 204s	20 32s
$\theta$ : theta82 $m = 23872, n = 400$	2403s	16 547s	21 73s
NCM: 400H1 $m = 80600, n = 400$		26 2176s	22 539s
QAP: lipa40a $m = 1.28 \times 10^6, n = 1600$			22 19h
$\theta_+$ : 1zc.2048 $m = 2.14 \times 10^6, n = 2048$			11 3.6h
$\theta$ : 2dc.512 $m = 54896, n = 512$			27 2400s (2.2e-5)

Thank you!

## Cholesky factorization

Given  $M \in \mathcal{S}^m$  and positive definite,  $M$  can be factorized as

$$M = R^T R, \quad R \text{ is upper triangular.}$$

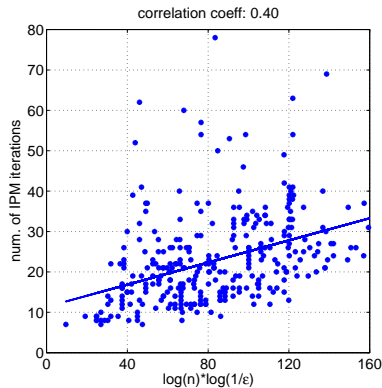
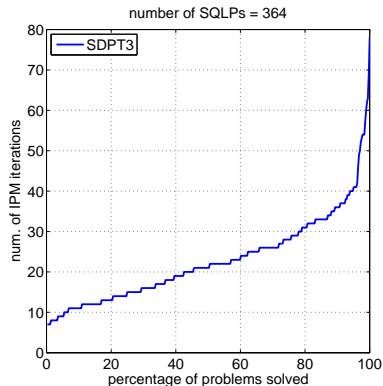
To solve  $Mx = b$ , solve 2 triangular linear systems:

$$R^T y = b, \quad Rx = y.$$

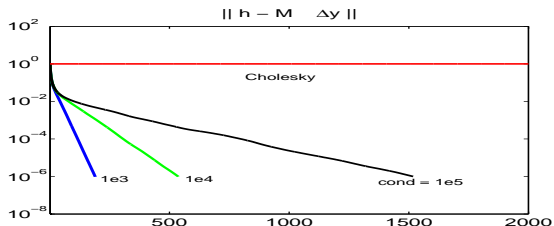
<p.2

## Performance of SDPT3

Based on 360 SDPs collected from various sources: Mittelmann, Borchers, Fukuda, Didier, ...



## Conjugate gradient method

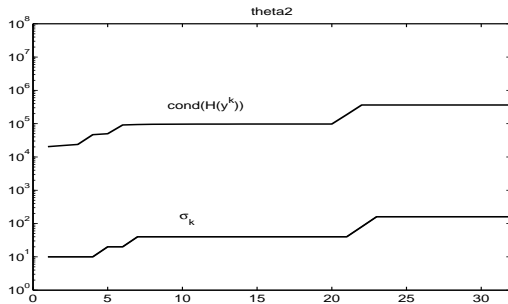


Convergence rate of CG depends on the eigenvalue distribution of  $\mathcal{M}$ :

$$\text{rate} \approx 1 - \frac{2}{\sqrt{\text{cond}(\mathcal{M})}} \quad \text{when } \text{cond}(\mathcal{M}) \gg 1.$$

<p.1

## Condition number of generalized Hessian in ALM



◀p.2

For the boundary-point method of Rendl et al., one step of modified gradient method is used to solve (8):

$$y^k = y^{k-1} + (\sigma_k \mathcal{A} \mathcal{A}^T)^{-1} \nabla_y L(y^{k-1}; \mathcal{X}^k).$$

◁p.1