An inexact primal-dual path-following algorithm for convex quadratic SDP

Kim-Chuan TOH

Department of Mathematics, NUS

http://www.math.nus.edu.sg/~mattohkc

FOCM 2008, Hong Kong

Convex quadratic semidefinite programming (QSDP)

$\mathcal{S}^n = \{n \times n$ symmetric matrices$\}$ with inner product $X \bullet S = \mathrm{trace}(XS)$.

$$\text{(P)} \quad \min_X \ \tfrac{1}{2} X \bullet \mathcal{Q}(X) + C \bullet X$$
$$\mathcal{A}(X) \ = \ b, \quad X \succeq 0,$$

$\mathcal{Q} : \mathcal{S}^n \to \mathcal{S}^n$ is a given self-adjoint psd linear operator

$\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ is a linear map s.t. $\mathcal{A}(X) = [A_1 \bullet X, \ldots, A_m \bullet X]^T$.

Problem dimension: $n =$ dimension of X; $m =$ number of constraints.
Assume $n \leq 2000$, $m \leq 5000$.

$$\text{(D)} \quad \max_{X,y,Z} \ -\tfrac{1}{2} X \bullet \mathcal{Q}(X) + b^T y$$
$$\mathcal{A}^T(y) - \mathcal{Q}(X) + Z \ = \ C, \quad Z \ \succeq \ 0,$$

$\mathcal{A}^T : \mathbb{R}^m \to \mathcal{S}^n$ is defined by $\mathcal{A}^T y = \sum_{k=1}^m y_k A_k$.

Outline

- Applications
- Newton system at each interior-point iterations
- Computation of Newton directions via KKT systems
- Conditioning of KKT systems
- Preconditioner I, II
- Computational results

Applications

Semidefinite least squares problem: given $K \in \mathcal{S}^n$, solve

$$\text{(SDLS)} \quad \min_X \left\{ \|\mathcal{L}(X) - \mathcal{L}(K)\|_F \,:\, \mathcal{A}(X) = b, \;\; X \succeq 0 \right\},$$

where $\mathcal{L} : \mathcal{S}^n \to \mathcal{S}^n$ is a given linear operator.

Special case: Nearest correlation matrix problem

$$\text{(NCM)} \quad \min_X \left\{ \|\mathcal{L}(X) - \mathcal{L}(K)\|_F \,:\, \mathrm{diag}(X) = e, \;\; X \succeq 0 \right\}.$$

Often $\mathcal{Q}(X) = \mathcal{L}^T\mathcal{L}(X) = \Omega \circ X \;\equiv\; \mathrm{diag}(\mathrm{vec}(\Omega))\,\mathrm{vec}(X)$. The matrix representation of $\mathcal{Q}$ w.r.t the standard basis of $\mathcal{S}^n$ is a diagonal matrix.

- Nearest Euclidean distance matrix problem [Alfakih et al]
- Compliance estimation in deformable object modelling [Krislock]
- Correlation stress testing for value-at-risk [Qi,Sun]
- Inverse quadratic eigenvalue problems with partial eigenstructure [Bai,Chu,Sun]

Related Work

(QSDP) is a generalization of a linearly constrained convex QP:

$$\text{(LCCQP)} \quad \min \left\{ \frac{1}{2} x^T Q x + c^T x \;:\; Ax = b,\; x \geq 0 \right\}.$$

[Bergamaschi et al, 04] proposed preconditioners for the KKT systems arising from IPM for LCCQP. Given a primal-dual iterate (x, y, z), the KKT system is given by:

$$(*) \qquad \begin{bmatrix} -\mathrm{diag}(x)^{-1}\mathrm{diag}(z) - Q & A^T \\ A & 0 \end{bmatrix}.$$

Key difference: KKT systems for LCCQP is sparse if Q and A are sparse.

KKT systems for QSDP is dense even if $\mathcal{Q}$ and $\mathcal{A}$ are sparse.

Focus of this talk: how to solve large dense ill-conditioned KKT systems arising from IPM for QSDP.

Perturbed optimality conditions

Assumption: $\mathcal{A}$ is surjective; (P) and (D) are strictly feasible.

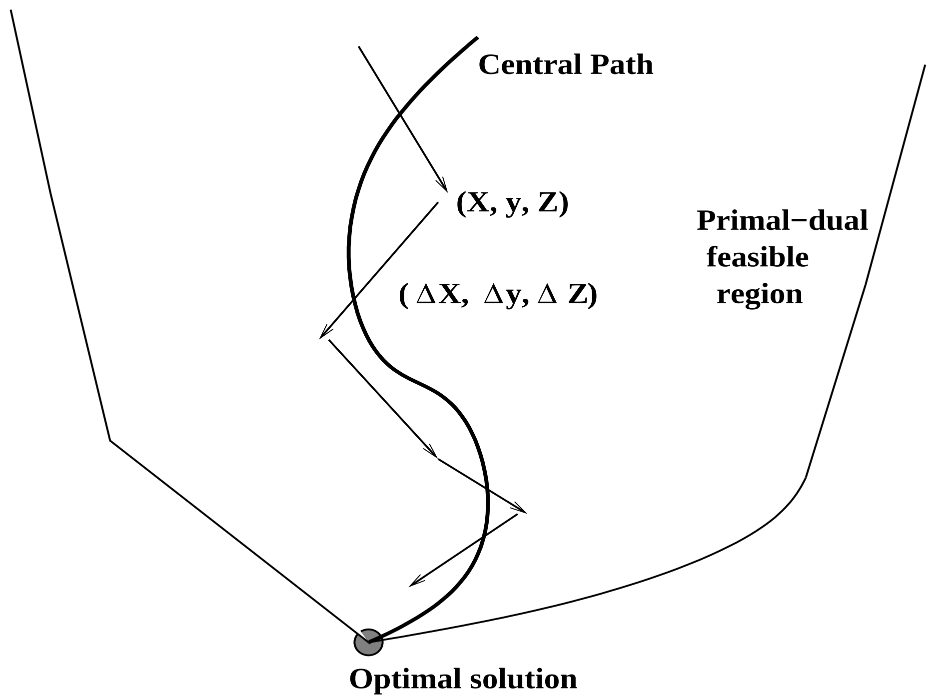
$$\begin{array}{rcl} -\mathcal{Q}(X) \; + \; \mathcal{A}^T(y) \; + \; Z & = & C, \quad Z \succeq 0 \\ \mathcal{A}(X) & = & b, \quad X \succeq 0 \\ XZ & = & \nu I, \end{array}$$

parameter $\nu > 0$ is to be driven to 0 explicitly.

The set of solutions $\{(X^\nu, y^\nu, Z^\nu) : \nu > 0\}$ is called the central path.

As $\nu \downarrow 0$, it converges to an optimal solution (X^*, y^*, Z^*) satisfying $X^* Z^* = 0$.

A primitive primal-dual path-following IPM



Polynomial iteration complexity of such IPM for QSDP has been established by Kojima et al. by treating it as a monotone SDLCP.

Newton system

At each IPM iteration, given current iterate (X, y, Z) with $X, Z \succ 0$, let W (dense!) be the scaling matrix satisfying $WXW = Z$, compute the search direction from the Newton system:

$$
\begin{array}{rcccccl}
-\mathcal{Q}(\Delta X) & + & \mathcal{A}^T(\Delta y) & + & \Delta Z & = & R_d := C - Z - \mathcal{A}^T y + \mathcal{Q}(X) \\
\mathcal{A}(\Delta X) & & & & & = & R_p := b - \mathcal{A}(X) \\
W \otimes W(\Delta X) & & & + & \Delta Z & = & R_c := \nu X^{-1} - Z,
\end{array}
$$

where $\nu > 0$ is the barrier parameter.

Note: $\dim(W \otimes W) = n^2 \times n^2$.

Computation of search directions $(\Delta X, \Delta y, \Delta Z)$

(a) Eliminate ΔZ, solve the following KKT system:

$$\text{(KKT)} \qquad \begin{matrix} n^2 \\ m \end{matrix} \underbrace{\begin{bmatrix} -W \otimes W - \mathcal{Q} & \mathcal{A}^T \\ \mathcal{A} & 0 \end{bmatrix}}_{\mathcal{B} \text{ dense!}} \begin{bmatrix} \Delta X \\ \Delta y \end{bmatrix} = \text{rhs},$$

For $n = 1000$, $\dim(\mathcal{B}) \geq 10^6$!

- Impossible to solve (KKT) by direct methods for $n \geq 100$, iterative solver is the only choice.

(b) Further eliminate ΔX, solve the $m \times m$ Schur complement equation:

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

- For a general $\mathcal{Q}$ (even diagonal ones), computing $(W \otimes W + \mathcal{Q})^{-1}$ need $\Theta(n^6)$ flops, so (SCE) is not viable!

Partitioning the KKT matrix $\mathcal{B}$

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

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

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

$$\mathcal{D} = \mathrm{diag}(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3), \quad \text{where}$$

$$\mathrm{diag}(\mathcal{D}_1) = \Theta(\nu),\ \ \mathrm{diag}(\mathcal{D}_2) = \Theta(1),\ \ \mathrm{diag}(\mathcal{D}_3) = \Theta(1/\nu).$$

Let $\widetilde{\mathcal{Q}} := \mathcal{P}^T \mathcal{Q} \mathcal{P}$, $\widetilde{\mathcal{A}} = \mathcal{A}\mathcal{P}$. Partition $\mathcal{P}$, $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{Q}}$ according to that of $\mathcal{D} = \mathrm{diag}(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$:

$$\mathcal{P} = [\mathcal{P}_1,\ \mathcal{P}_2,\ \mathcal{P}_3], \quad \widetilde{\mathcal{A}} = [\widetilde{\mathcal{A}}_1,\ \widetilde{\mathcal{A}}_2,\ \widetilde{\mathcal{A}}_3], \quad \widetilde{\mathcal{Q}}_{ij} = \mathcal{P}_i^T \mathcal{Q} \mathcal{P}_j, \quad i, j = 1, 2, 3.$$

Note that

$$\mathcal{B} \;=\; \begin{bmatrix} \mathcal{P} & \\ & I \end{bmatrix} \begin{bmatrix} -(\mathcal{D} + \widetilde{\mathcal{Q}}) & \widetilde{\mathcal{A}}^T \\ \widetilde{\mathcal{A}} & 0 \end{bmatrix} \begin{bmatrix} \mathcal{P}^T & \\ & I \end{bmatrix}.$$

Later: construction of preconditioners and analysis are based on these partitions in the middle matrix.

Conditioning of $\mathcal{B}^\nu$ along the central path

Let

$$X^*, Z^* = \lim_{\nu\to 0} X^\nu, Z^\nu.$$
$$\mathcal{P}^*, \widetilde{\mathcal{A}}^*, \widetilde{\mathcal{Q}}^* = \lim_{\nu\to 0} \mathcal{P}^\nu, \widetilde{\mathcal{A}}^\nu, \widetilde{\mathcal{Q}}^\nu.$$

Partition $\mathcal{P}^*, \widetilde{\mathcal{A}}^*, \widetilde{\mathcal{Q}}^*$ as before.

Assume that the optimal solution (X^*, Z^*) is strictly complementary ($X^* + Z^* \succ 0$), X^* is said to be primal non-degenerate if $[\widetilde{\mathcal{A}}_1^*, \widetilde{\mathcal{A}}_2^*]$ has full row rank.

Theorem: $\|\mathcal{B}^\nu\| \geq \Theta(1/\nu)$ as $\nu \downarrow 0$.
Suppose $\widetilde{\mathcal{Q}}_{11}^* \succ 0$ and X^* is primal non-degenerate. Then $\|(\mathcal{B}^\nu)^{-1}\| = \Theta(1)$.

$\mathcal{B}$ is ill-conditioned with condition number $\geq \Theta(1/\nu)$ as $\nu \downarrow 0$.
$\mathcal{B}$ is dense, and cannot be stored $\Rightarrow$ difficult to construct preconditioners based on many existing techniques.

Preconditioner I: "Block diagonal" preconditioner

Recall (1,1) block of $\mathcal{B} = -\mathcal{P}(\mathrm{diag}(\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3) + \widetilde{\mathcal{Q}})\mathcal{P}^T$. Consider

$$\Phi_{\pm} = \begin{bmatrix} \pm\mathcal{P}\mathrm{diag}(\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3)\mathcal{P}^T & 0 \\ 0 & \widehat{\mathcal{M}} \end{bmatrix},$$

where $\mathcal{H}_j$ is a pd approximation of $\mathcal{D}_j + \widetilde{\mathcal{Q}}_{jj}$, and $\widehat{\mathcal{M}}$ is a pd approximation of $\mathcal{M}$.

Theorem: Suppose there exist constants $\underline{\sigma}$, $\overline{\sigma} > 0$ s.t.

$$\underline{\sigma} I \preceq \mathcal{H}_1^\nu,\ \mathcal{H}_2^\nu,\ \widehat{\mathcal{M}}^\nu \preceq \overline{\sigma} I, \qquad \mathcal{D}_3^\nu \preceq \mathcal{H}_3^\nu.$$

Suppose $\widetilde{\mathcal{Q}}_{11}^* \succ 0$ and X^* is primal non-degenerate. Then there exist constants $c_1, c_2, c_3, c_4 > 0$ s.t.

$$\mathrm{eig}((\Phi_+^\nu)^{-1}\mathcal{B}^\nu) \subset [-c_1,\ -c_2] \cup [c_3,\ c_4] \quad \forall\ \nu \ll 1.$$

Weaker asymptotic result for $\mathrm{eig}((\Phi_-^\nu)^{-1}\mathcal{B}^\nu)$, but Φ_- performs better in practice!

Example 1

QSDP arising from finding the nearest correlation matrix of K with $\mathcal{Q}(X) = \Omega \circ X$.

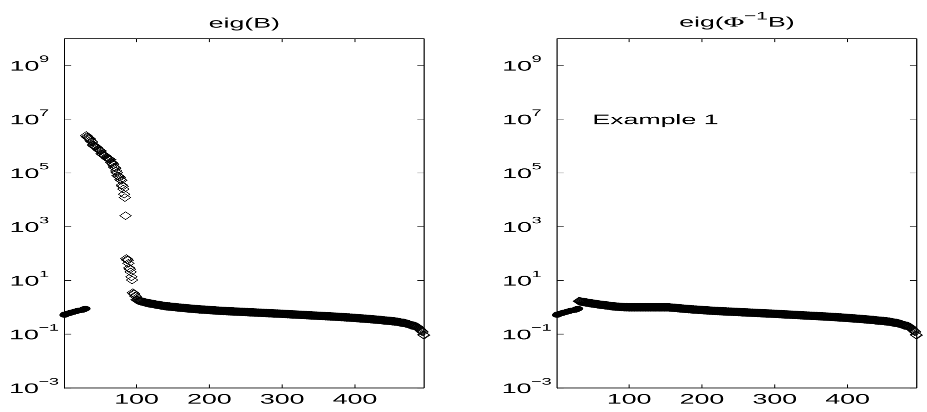
Data generation: Let Σ be an $n \times n$ random correlation matrix with $\lambda_{\max}(\Sigma) = 1$.

Set $K = \Sigma +$ unit random.

Set $\Omega = n \times n$ symmetric nonnegative random matrix.

Take $\mathcal{H}_1 = \mathcal{D}_1 + \|\mathcal{Q}\| I, \quad \mathcal{H}_2 = \mathcal{D}_2 + \|\mathcal{Q}\| I, \quad \mathcal{H}_3 = \mathcal{D}_3, \quad \widehat{\mathcal{M}} = I_m.$

Spectra of $\mathcal{B}^\nu$ and $(\Phi_+^\nu)^{-1}\mathcal{B}^\nu$ for $\nu \approx 10^{-9}$.



Preconditioner II

$$\Psi = \begin{bmatrix} -\mathcal{P}\mathrm{diag}(\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3)\mathcal{P}^T & \mathcal{A}^T \\ \mathcal{A} & 0 \end{bmatrix},$$

$\mathcal{H}_j$ must be easy to invert so that (1,1) block is easy to invert.

Theorem: Suppose there exist constants $\underline{\sigma}$, $\overline{\sigma} > 0$ s.t.

$$\underline{\sigma} I \preceq \mathcal{H}_1^\nu, \mathcal{H}_2^\nu \preceq \overline{\sigma} I, \quad \mathcal{D}_3^\nu \preceq \mathcal{H}_3^\nu.$$

If $\widetilde{\mathcal{Q}}_{11}^* \succ 0$, then there exist constants $0 < c_1 < 1 < c_2$ s.t.

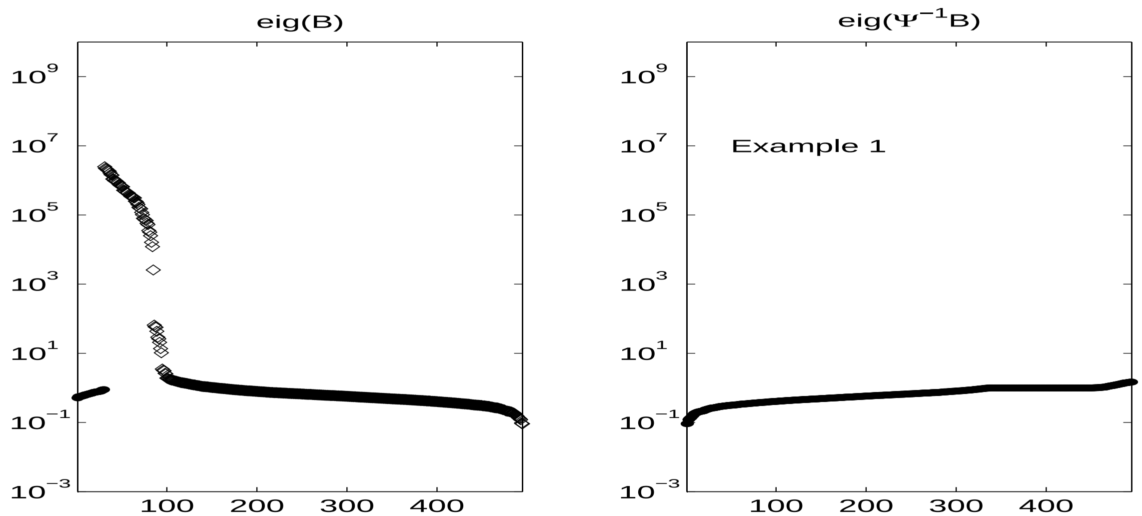
$$\text{eig}((\Psi^\nu)^{-1}\mathcal{B}^\nu) \subset [c_1, c_2] \quad \forall \ \nu \ll 1.$$

Note: non-degeneracy conditions are not needed. But 2 times more expensive to evaluate compared to $\Phi_\pm$.

Take

$$\mathcal{H}_1 \;=\; \mathcal{D}_1 + \|\mathcal{Q}\| I, \quad \mathcal{H}_2 \;=\; \mathcal{D}_2 + \|\mathcal{Q}\| I, \quad \mathcal{H}_3 \;=\; \mathcal{D}_3.$$

Spectra of $\mathcal{B}^\nu$ and $(\Psi^\nu)^{-1}\mathcal{B}^\nu$ for $\nu \approx 10^{-9}$ in Example 1.



Efficient computation of $\Phi^{-1}[X; y]$, $\Psi^{-1}[X; y]$

Depends on:

(1) The block diagonal matrix $\mathcal{H} := \mathrm{diag}(\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3)$ must be easy to invert so that (1,1) block ($= \mathcal{P}\mathcal{H}\mathcal{P}^T$) is easy to invert.

(2) Computational cost of Schur complement $S = \mathcal{A}\mathcal{P}\mathcal{H}^{-1}\mathcal{P}^T\mathcal{A}^T$ must be moderate.

(1a) Consider the choice $\mathcal{H}_1 = H_1 \otimes H_1$ such that $\|\mathcal{D}_1 + \widetilde{\mathcal{Q}}_{11} - H_1 \otimes H_1\|_F = \min$; similarly for $\mathcal{H}_2, \mathcal{H}_3$.

(1b) Consider the diagonal approximation: $\mathcal{H} = \mathcal{D} + \mathrm{diag}(\widetilde{\mathcal{Q}})$. Computing $\mathrm{diag}(\widetilde{\mathcal{Q}})$ need at least $\Theta(n^4)$ flops even if $\mathcal{Q}$ is diagonal! Fortunately, when $\mathcal{Q}$ is diagonal, we can compute a good approximation of $\mathrm{diag}(\widetilde{\mathcal{Q}})$ in $4n^3$ flops.

(2a) Computing $S = S_1 + S_2 + S_3$ needs $O(mn^3) + O(m^2n^2)$ flops,
where $S_1 = \mathcal{A}(PH_1P^T) \otimes (PH_1P^T)\mathcal{A}^T$, similarly for S_2, S_3.

(2b) Computing S needs $\Theta(mn^3) + O(m^2n^2)$ flops in general, because it is difficult to exploit sparsity in $\mathcal{A}$ if the middle matrix is not in Kronecker product form. Remedy: approximate $S \approx \mathcal{A}U \otimes U\mathcal{A}^T$, where $U \otimes U \approx \mathcal{D} + \mathrm{diag}(\widetilde{\mathcal{Q}})$.

Preconditioners for initial phase of IPM

When $\nu \geq 1$, use

$$\mathcal{V} = \begin{bmatrix} -V \otimes V & \mathcal{A}^T \\ \mathcal{A} & 0 \end{bmatrix},$$

where $V \otimes V$ is a pd approximation of $W \otimes W + \mathcal{Q}$, such as

$$\|W \otimes W + \mathcal{Q} - V \otimes V\|_F = \min.$$

Another obvious choice is:

$$\begin{bmatrix} -\mathrm{diag}(W) \otimes \mathrm{diag}(W) - \mathrm{diag}(\mathcal{Q}) & \mathcal{A}^T \\ \mathcal{A} & 0 \end{bmatrix}.$$

It is sparse if $\mathcal{A}$ is sparse, but is not as effective as $\mathcal{V}$.

Numerical experiments

We test the following variants of IPMs on 8 classes of problems.

A0. by PSQMR with no preconditioning.

A1. by PSQMR preconditioned by $\mathcal{V}$, Φ_{-}.

A2. by PSQMR preconditioned by $\mathcal{V}$, Ψ.

We generated problems for $n = 100, 200, 400, 800, 1600$.

Linear map $\mathcal{A}$	Data as in Eg1	Data as in Eg2	cond($\mathcal{Q}$) in Eg1 increased by 10	cond($\mathcal{Q}$) in Eg2 increased by 10
$\mathrm{diag}(X) = e$	E1	E2	E3	E4
$\mathrm{diag}(X) = e$ and $X_{ij} = 0$ for some ij	E5	E6	E7	E8

Stop IPMs when $\mathbf{err} \leq 10^{-7}$.

Stopping condition for PSQMR: residual $\leq 10^{-3} \max\{\|R_p\|, \|R_d\|_F, \|R_c\|_F\}$

		A0 (I)				A1 $(\mathcal{V}, \Phi_-)$			A2 $(\mathcal{V}, \Psi)$		
	$m \approx n$	it	$\mathbf{err}$	time (h:m)	qmr	it	time (h:m)	qmr	it	time (h:m)	qmr
E3	n=800	10	1.0 -6	0:40	179.4	11	0:22	32.0	11	0:22	21.0
	1600					11	2:10	26.6	11	2:21	19.9
E5	n=800	12	1.0 -5	0:54	200.7	13	0:25	30.5	13	0:20	15.3
	1600					13	2:53	30.6	13	2:01	13.0

Each linear system (with dimension $\geq 10^6$) is solved (approximately) in about 10 mins

Ongoing work

- Extend IPM to linearly constrained convex SDP:

$$\min\{f(X) \,:\, \mathcal{A}(X) = b,\ X \succeq 0\}.$$

von Neumann entropy: $f(X) = \mathrm{Tr}(X(\log X - \log Y) - X + Y)$.

Each iteration needs to solve a linear system with matrix:

$$\begin{bmatrix} -W \otimes W - \nabla^2 f(X) & \mathcal{A}^T \\ \mathcal{A} & 0 \end{bmatrix}.$$

- Augmented Lagrangian method for QSDP.