Inexact search directions in interior point methods for very large scale optimization

Jacek Gondzio

Email: `J.Gondzio@ed.ac.uk`
URL: `http://www.maths.ed.ac.uk/~gondzio/`

Outline

- Motivation: Make IPMs *faster*
- Inexact Newton directions
  - Krylov subspace methods
  - Preconditioner is a must
- Theory
- Difficult problems
  - Quadratic Assignment
  - Quantum Information
- Computational results
- Conclusions

Brazil 2014 view of LP/QP:

How to solve LP/QP problems?

If we asked **Neymar Jr**, the likely answer would be: *"go through the interior of the polytope"*.

Objective: Accelerate IPMs for LO/QO

We know that IPMs converge in

- *theory:* $\mathcal{O}(\sqrt{\boldsymbol{n}} \log(1/\varepsilon))$ iterations
- *practice:* $\mathcal{O}(\mathbf{log}\, \boldsymbol{n} \log(1/\varepsilon))$ iterations

but the per-iteration cost may be high.

Redesign IPMs:

- make a single iteration as fast as possible
  replace *exact* Newton Method
  with *inexact* Newton Method
- work in *matrix-free* and *limited-memory* regime

Splitting ("Simplex-type") Preconditioner

Oliveira, *PhD Thesis*, Rice University, 1997

Oliveira & Sorensen, *Linear Algebra and its Applications* 394 (2005) 1-24.

O, OS show that *all preconditioners for the NE have an equivalent for the AS but the opposite is not true*.

→ it is better to precondition AS
→ guess the basis matrix.

Al-Jeiroudi, G. & Hall, *Optimization Methods and Software* 23 (2008) 345-363.

Al-Jeiroudi & G., *J. of Optimization Theory and Applications* 141 (2009) 231-247.

Many people use iterative methods in IPMs...

Adler et al. (1989a,b)
Karmarkar and Ramakrishnan (1991)
Gill et al. (1992) (indefinite systems)
Resende and Veiga (1993) (network flows)
Oliveira (1997)
Freund and Jarre (1997)
Lukšan and Vlček (1998)
Bellavia (1998)
Mizuno and Jarre (1999)
Baryamureeba, Steihaug and Zhang (1999)
Castro (2000) (network flows)
Wang and O'Leary (2000)
Bergamaschi and Zilli (2000)

Iterative Methods

Normal Equations or **Augmented System**:

- NE is positive definite: use conjugate gradients;
- AS is indefinite: use BiCGSTAB, GMRES, QMR;

G. & Toraldo (eds.),
Comp. Optimization & Appls, 36 (2007), No 2/3.
Special issue on "Linear Algebra in Interior P. Methods",
8 out of 10 papers about iterative methods.

D'Apuzzo, De Simone & di Serafino,
Comp. Optimization & Appls, 45 (2010) No 2.
Survey on lin. algebra in interior p. methods.

LO & QO Problems

$$\begin{aligned} \min \quad & c^T x + \tfrac{1}{2} x^T Q\, x \\ \text{s.t.} \quad & Ax = b, \\ & x \geq 0, \end{aligned}$$

where $A \in \mathcal{R}^{m \times n}$ has full row rank
and $Q \in \mathcal{R}^{n \times n}$ is symmetric positive semidefinite.

m **and** n **may be large.**

Assumption: A and Q are *"operators"* $A \cdot u$, $A^T \cdot v$, $Q \cdot u$

Expectation: Low complexity of these operations

Applications: LPs, QPs constructed implicitly

- problems generated by an algebraic mod. language
- problems too large to be stored
  (but generated by some "simple" process)
- LP relaxations of combinatorial (integer) problems
- sparse approximations (compressed sensing)

Assumption: A and Q as "operators" $A \cdot u$, $A^T \cdot v$, $Q \cdot u$

Expectation: Low complexity of these operations

The First Order Optimality Conditions

$$\begin{aligned} Ax &= b, \\ -Qx + A^T y + s &= c, \\ XSe &= \mu e, \\ (x, s) &> 0. \end{aligned}$$

Assume primal-dual feasibility:

$$Ax = b \qquad \text{and} \qquad -Qx + A^T y + s = c$$

Apply Newton Method to the FOC

$$\begin{bmatrix} A & 0 & 0 \\ -Q & A^T & I \\ S & 0 & X \end{bmatrix} \cdot \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta s \end{bmatrix} = \begin{bmatrix} b - Ax \\ c - A^T y - s + Qx \\ \sigma \mu e - XSe \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \xi \end{bmatrix}.$$

Central Path:

A set of all solutions to the optimality conds for $\mu > 0$.

Path Following Method:

Stay in the **neighbourhood** (of the central path)

$$\begin{aligned}
\mathcal{N}_{\mathbf{2}}(\theta) &:= \{(x,y,s) \in \mathcal{F}^0 : \|XSe - \mu e\|_2 \le \theta\mu\} \\
\mathcal{N}_{-\infty}(\gamma) &:= \{(x,y,s) \in \mathcal{F}^0 : x_i s_i \ge \gamma\mu\} \\
\mathcal{N}_{\boldsymbol{S}}(\gamma) &:= \{(x,y,s) \in \mathcal{F}^0 : \gamma\mu \le x_i s_i \le (1/\gamma)\mu\}
\end{aligned}$$

where

$$\mathcal{F}^0 := \{(x,y,s) : c - A^T y - s + Qx = 0, Ax = b, x, s > 0\}.$$

Standard complexity result

Let $\epsilon > 0$ be the required accuracy of the optimal sol., that is, we stop when $\mu^k \leq \epsilon$.

The (*short-step, feasible*) IPM operates in $\mathcal{N}_2(\theta)$ and finds the ϵ-accurate solution after at most

$$K = \mathcal{O}(\sqrt{\boldsymbol{n}} \ln(1/\epsilon))$$

iterations.

The (*long-step, feasible*) IPM operates in $\mathcal{N}_S(\gamma)$ and finds the ϵ-accurate solution after at most

$$K = \mathcal{O}(\boldsymbol{n} \ln(1/\epsilon))$$

iterations.

Exact Newton Method

$$\begin{bmatrix} A & 0 & 0 \\ -Q & A^T & I \\ S & 0 & X \end{bmatrix} \cdot \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \boldsymbol{\xi} \end{bmatrix}.$$

Inexact Newton Method

$$\begin{bmatrix} A & 0 & 0 \\ -Q & A^T & I \\ S & 0 & X \end{bmatrix} \cdot \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \boldsymbol{\xi} + \boldsymbol{r} \end{bmatrix}$$

allows for an error in the (linearized) complementarity condition only.

General Assumption

The residual $\boldsymbol{r}$ in the inexact Newton Method satisfies:

$$\|\boldsymbol{r}\| \leq \boldsymbol{\delta}\|\boldsymbol{\xi}\|,$$

where $\boldsymbol{\delta} \in (0, 1]$.

What happens to the complexity result?

Short-step (Feasible) Algorithm

Stay in the **small** neighbourhood of the central path

$$\mathcal{N}_2(\theta) := \{(x,y,s) \in \mathcal{F}^0 : \|XSe - \mu e\|_2 \le \theta\mu\}$$

Use $\boldsymbol{\sigma} = (1 - \frac{0.1}{\sqrt{n}})$.
Set $\boldsymbol{\delta} = 0.3$ to achieve the reduction:

$$\bar{\mu} = (1 - \frac{0.02}{\sqrt{n}})\mu.$$

$\Rightarrow$ Convergence in $\mathcal{O}(\sqrt{\boldsymbol{n}} \ln(1/\epsilon))$ iterations.

Long-step (Feasible) Algorithm

Stay in the **large** neighbourhood of the central path

$$\mathcal{N}_S(\gamma) := \{(x, y, s) \in \mathcal{F}^0 : \gamma\mu \le x_i s_i \le (1/\gamma)\mu\}$$

Use $\boldsymbol{\sigma} = 0.5$.
Set $\boldsymbol{\delta} = 0.05$ to achieve the reduction:

$$\bar{\mu} = (1 - \frac{0.002}{n})\mu.$$

$\Rightarrow$ Convergence in $\mathcal{O}(\boldsymbol{n} \ln(1/\epsilon))$ iterations.

Theorem

Suppose the algorithm uses the **inexact** Newton Method.

- If $(x, y, s) \in \mathcal{N}_2(\theta)$ and $\boldsymbol{\sigma} = (1 - \frac{0.1}{\sqrt{n}})$, $\boldsymbol{\delta} = 0.3$ then the algorithm converges in at most
$$K = \mathcal{O}(\sqrt{\boldsymbol{n}} \ln(1/\epsilon))$$
iterations.

- If $(x, y, s) \in \mathcal{N}_S(\gamma)$ and $\boldsymbol{\sigma} = 0.5$, $\boldsymbol{\delta} = 0.05$ then the algorithm converges in at most
$$K = \mathcal{O}(\boldsymbol{n} \ln(1/\epsilon))$$
iterations.

Proof (key ideas)

For the Short-step Algorithm, show that the *error*

$$\|\Delta X \Delta S e\| = \mathcal{O}(\mu).$$

Use the *full* Newton step.
The proof requires 3 pages of maths.

For the Long-step Algorithm, show that the *error*

$$\|\Delta X \Delta S e\| = \mathcal{O}(n\mu).$$

Use the *damped* Newton step with $\alpha = \mathcal{O}(1/n)$.
The proof requires 5 pages of maths.

JG, Convergence Analysis of an Inexact Feasible IPM for Convex QP, *SIAM J. on Optimization* 23 (2013) No 3, pp. 1510-1527.

Conclusion

Replace the **Exact** Newton Method
with the **Inexact** Newton Method

Allow for large residual

$$\|r\| \leq \boldsymbol{\delta} \|\xi\|$$

The worst-case complexity result remains the same!

From Theory to Practice

Solve augmented system

$$\begin{bmatrix} -Q-\Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}.$$

Use an iterative method with a suitable **preconditioner** which must work in a **matrix-free** regime.

A good **preconditioner** depends on the problem. Finding it may be a challenge.

Preconditioning

Do it in two steps:

1. Improve the conditioning of the linear system
   $\longrightarrow$ use primal-dual regularization
   (bounded condition number of KKT system)
2. Precondition the (easier) system

Augmented System Matrix

Original: $\mathcal{H} = \begin{bmatrix} -Q - \Theta^{-1} & A^T \\ A & 0 \end{bmatrix}$

and *regularized*: $\mathcal{H}_R = \begin{bmatrix} -(Q + \Theta^{-1} + R_p) & A^T \\ A & R_d \end{bmatrix}.$

Normal Equation Matrix

Original: $\mathcal{G} = (A(Q + \Theta^{-1})^{-1} A^T)$

and *regularized*: $\mathcal{G}_R = (A(Q + \Theta^{-1} + R_p)^{-1} A^T + R_d).$

Altman & JG, Regularized symmetric indefinite systems in IPMs for linear and quadratic optimization, *Optim. Methods and Software* 11-12 (1999) 275-302.

Original NE system

$$\underbrace{\left(A(Q+\Theta^{-1})^{-1}A^T\right)}_{\mathcal{G}}\Delta y = g$$

Consider a (difficult) LP case: $Q = 0 \quad \longrightarrow \quad \mathcal{G} = A\Theta A^T$

Theorem. The condition number of $\mathcal{G}$ satisfies:

$$\kappa(\mathcal{G}) \leq [\kappa(A)]^2 \cdot \mathcal{O}(\mu^{-2}).$$

Proof:

The largest eigenvalue of $\mathcal{G}$ $\qquad \lambda_{max} \leq \sigma_m^2 \cdot \mathcal{O}(\mu^{-1})$

The smallest eigenvalue of $\mathcal{G}$ $\qquad \lambda_{min} \geq \sigma_1^2 \cdot \mathcal{O}(\mu)$

Hence: $$\kappa(\mathcal{G}) \leq [\kappa(A)]^2 \cdot \mathcal{O}(\mu^{-2})$$

$\square$

Regularized NE system

$$\underbrace{\big(A(Q+\Theta^{-1}+R_p)^{-1}A^T+R_d\big)}_{\mathcal{G}_R}\Delta y = g$$

Theorem. Assume: $R_p=\gamma^2 I_n$ and $R_d=\delta^2 I_m$.

The condition number of $\mathcal{G}_R$

$$\kappa(\mathcal{G}_R) \le \frac{\sigma_m^2 \cdot \gamma^{-2}+\delta^2}{\delta^2} = 1+\frac{\sigma_m^2}{\gamma^2\delta^2} \approx \frac{\sigma_m^2}{\gamma^2\delta^2}.$$

is **bounded** and **independent of** μ.

Proof:

Recall the regularized NE system:

$$\underbrace{\left(A(Q+\Theta^{-1}+R_p)^{-1}A^T+R_d\right)}_{\mathcal{G}_R}\Delta y = g$$

and the assumptions: $R_p = \gamma^2 I_n$ and $R_d = \delta^2 I_m$.

The largest eigenvalue of $\mathcal{G}_R$ $\quad \lambda_{max} \leq \sigma_m^2 \cdot \gamma^{-2} + \delta^2$

The smallest eigenvalue of $\mathcal{G}_R$ $\quad \lambda_{min} \geq \delta^2$

Hence

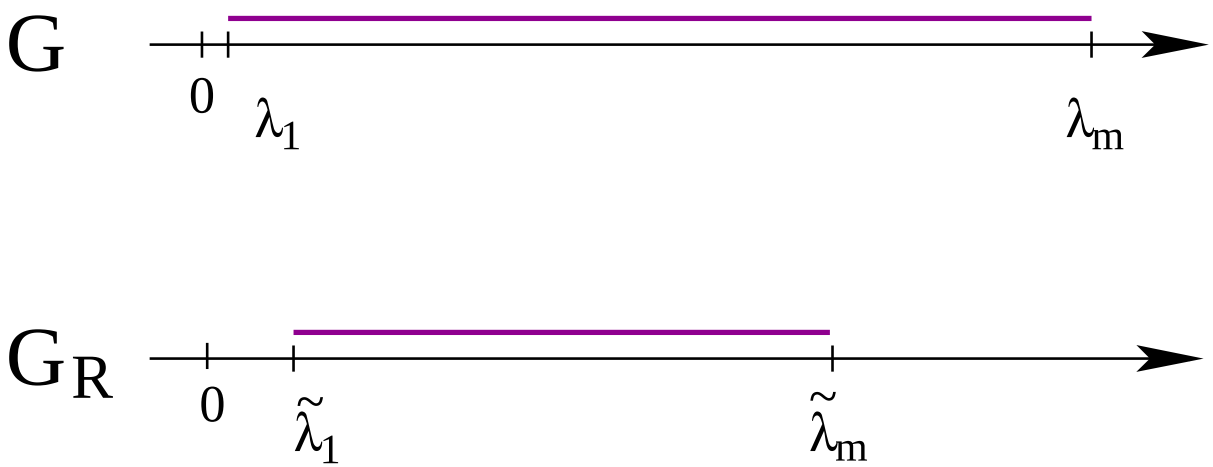
$$\kappa(\mathcal{G}_R) \leq \frac{\sigma_m^2 \cdot \gamma^{-2} + \delta^2}{\delta^2} = 1 + \frac{\sigma_m^2}{\gamma^2\delta^2} \approx \frac{\sigma_m^2}{\gamma^2\delta^2}$$

$\square$

Summary: Eigenvalues of NE

Regularizations:

$$\gamma^2 \leq R_{pj} \leq \Gamma^2, \ \forall j = 1..n, \quad \text{and} \quad \delta^2 \leq R_{di} \leq \Delta^2, \ \forall i = 1..m.$$



CG's rate of convergence

$$e^{k+1} \;\leq\; \frac{\kappa^{1/2}-1}{\kappa^{1/2}+1}\, e^k$$

For regularized NE system, we have:

$$\frac{\kappa^{1/2}-1}{\kappa^{1/2}+1} \approx \frac{\frac{\sigma_m}{\gamma\delta}-1}{\frac{\sigma_m}{\gamma\delta}+1} = \frac{1-\frac{\gamma\delta}{\sigma_m}}{1+\frac{\gamma\delta}{\sigma_m}} \approx 1-2\frac{\gamma\delta}{\sigma_m}.$$

now: **precondition the regularized system**

Decompose the regularized NE system

Use diagonal pivoting to compute

$$\mathcal{G}_R = \begin{bmatrix} L_{11} & \\ L_{21} & I \end{bmatrix} \begin{bmatrix} D_L & \\ & S \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{21}^T \\ & I \end{bmatrix},$$

where $L = \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix}$ is a trapezoidal matrix:
(the first k columns of Cholesky factor of $\mathcal{G}_R$);
$S \in \mathcal{R}^{(m-k)\times(m-k)}$ is the corresp. **Schur complement**.

Order diagonal elements of D_L and $D_S = diag(S)$:

$$\underbrace{d_1 \geq d_2 \geq \cdots \geq d_k}_{D_L} \;\geq\; \underbrace{d_{k+1} \geq d_{k+2} \geq \cdots \geq d_m}_{D_S}.$$

Preconditioner

Use the decomposition

$$\mathcal{G}_R = \begin{bmatrix} L_{11} & \\ L_{21} & I \end{bmatrix} \begin{bmatrix} D_L & \\ & \boldsymbol{S} \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{21}^T \\ & I \end{bmatrix}$$

and precondition $\mathcal{G}_R$ with

$$P = \begin{bmatrix} L_{11} & \\ L_{21} & I \end{bmatrix} \begin{bmatrix} D_L & \\ & \boldsymbol{D_S} \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{21}^T \\ & I \end{bmatrix},$$

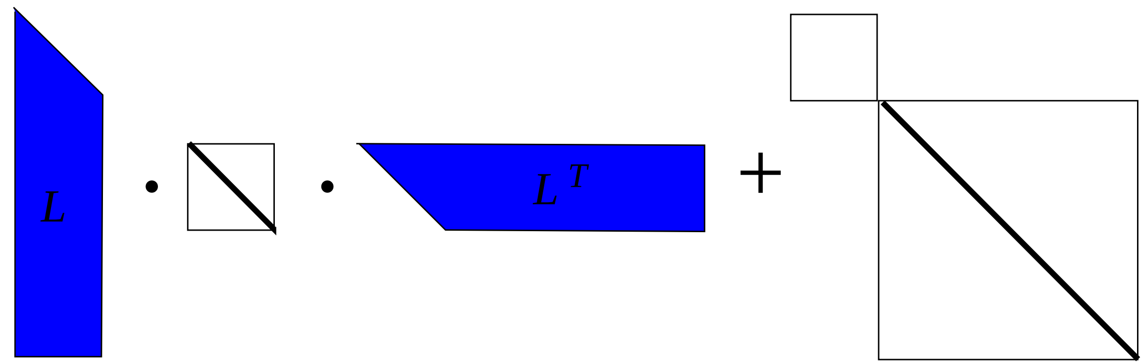
where $\boldsymbol{D_S}$ is a diagonal of $\boldsymbol{S}$.

Do **not** compute $\boldsymbol{S}$.
Update only its diagonal.

Preconditioner: Partial Cholesky of NE system

$$\mathcal{G}_R = (A(Q+\Theta^{-1}+R_p)^{-1}A^T+R_d) \approx LD_LL^T+D_S$$

$LD_LL^T + D_S =$

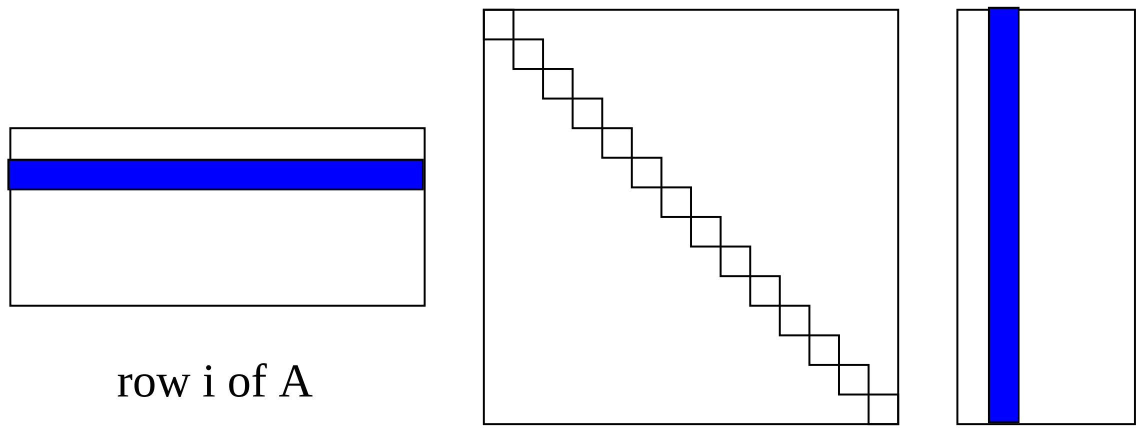


- low rank matrix L: $k \ll m$
- D_L contains k largest pivots of $\mathcal{G}_R$

JG, Matrix-Free Interior Point Method, *Computational Optimization and Applications*, vol. 51 (2012) 457–480.

Matrix-Free Implementation

$A\Theta A^T =$



To build the preconditioner we need only:

- a complete diagonal of $A\Theta A^T$ $\rightarrow$ $d_{ii} = r_i^T \Theta\, r_i$
- a column i of $A\Theta A^T$ $\rightarrow$ $(A\Theta) \cdot r_i$

both operations are **easy** if we access r_i^T (row i of A).

Two examples of difficult LPs

- Quadratic Assignment Problem, Nugent et al.
  with **Ed Smith and J.A.J. Hall**
- Quantum Information Problems
  with **Gruca, Hall, Laskowski and Żukowski**

use Matrix-Free IPM

Quadratic Assignment Problem, Nugent et al.

LP relaxations of size $m \approx 2 \times N^3$ and $n \approx 8 \times N^3$

joint work with **Ed Smith** and **J.A.J. Hall**

Prob	**Cplex 11.0.1**				**mf-IPM**			
	Simplex		Barrier		rank=200		rank=500	
	its	time	its	time	its	time	its	time
nug12	96148	187	13	10	7	**2**	7	15
nug15	387873	2451	16	71	7	**10**	7	34
nug20	$2.9 \cdot 10^6$	79451	18	1034	6	**35**	5	122
nug30	?	>28 *days*	-	*OoM*	5	**1272**	5	4465

mf-IPM solves large problems $N = 40, 50, \ldots, 100$ in *hours*

Quantum Information Problems

- model Quantum Entanglement (quBit, quNit)
- need solving a sequence of LPs

Features

- very sparse A
  $\longrightarrow$ inexpensive **MatVec** operation
- completely dense AA^T
  $\longrightarrow$ factorization of AS or NE is prohibitive

Einstein-Podolsky-Rosen Paradox, 1935

Following Wikipedia:
"[EPR paradox] refutes the dichotomy that *either* the measurement of a physical quantity in one system must affect the measurement of a physical quantity in another, spatially separate, system *or* the description of reality given by a wave function must be incomplete."

Quantum Entanglement:

The measurements performed on spatially separated parts of quantum systems may instantaneously influence each other.

Bell, *Physics*, 1 (1964) proposed inequalities which allow to capture situations when this happens.

Quantum Information Problems

Prob	**Cplex 12.0**				**mf-IPM**	
	Simplex		Barrier		rank=200	
	its	time	its	time	its	time
4kx4k	5418	0.8	20	15	6	4
16kx16k	62772	57	10	399	5	15
64kx64k	$2.6 \cdot 10^6$	6h51m	-	*OoM*	8	3m22s
256kx256k		>48*h*	-	*OoM*	9	28m38s
1Mx1M		-	-	*OoM*	9	1h34m19s
4Mx4M		-	-	*OoM*	10	9h14m49s

JG, Gruca, Hall, Laskowski and Żukowski,
Solving LSO Problems Related to Bell's Theorem,
J. of Comput and Appl Maths, 263C (2014) 392–404.

Conclusions

Theory:

The **inexact** IPM enjoys the same worst-case iteration complexity as the **exact** IPM.

Computational practice:

Matrix-free IPM solves otherwise intractable problems.
It needs:

- $\boldsymbol{\mathcal{O}(\log n)}$ iterations
- with $\boldsymbol{\mathcal{O}(nz(A))}$ cost per iteration
- it involves only **MatVec** operations.

Thank You!

Inexact Newton, Matrix-Free IPM:

JG, Convergence Analysis of an Inexact Feasible IPM for Convex QP,
SIAM J. on Optimization 23 (2013) pp. 1510-1527.

JG, Matrix-Free Interior Point Method,
Computational Optimization and Applications,
51 (2012) 457–480.

JG, Interior Point Methods 25 Years Later,
European Journal of Operational Research,
218 (2012) 587–601.

13th EUROPT Workshop

on Advances in Continuous Optimization

EUROPT, **Edinburgh**, 8-10 July 2015

`http://www.maths.ed.ac.uk/hall/EUROPT15/index.html`

EURO, **Glasgow**, 12-16 July 2015

Augmented system

$$\begin{bmatrix} -Q-\Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} -X^{-1}\xi \\ 0 \end{bmatrix}.$$

Inexact solution $(\Delta\tilde{x}, \Delta\tilde{y})$ satisfies

$$\begin{bmatrix} -Q-\Theta^{-1} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta\tilde{x} \\ \Delta\tilde{y} \end{bmatrix} = \begin{bmatrix} -X^{-1}\xi + r_x \\ r_y \end{bmatrix}.$$

Using

$$-X^{-1}\xi + r_x = -X^{-1}(\xi + \mathbf{r})$$

the practical stopping criteria for an iterative method is:

$$r_y = 0 \quad \text{and} \quad \|r\| = \|X r_x\| \leq \boldsymbol{\delta}\|\xi\|.$$