Inexact Search Directions in Interior Point Methods for Large Scale Optimization

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

Outline

- *1st-* and *2nd-order* methods for optimization
- Interior Point Methods: Pros & Cons
- Accelerating IPMs
- *Exact* vs *Inexact* search directions and IPMs
  $\rightarrow$ worst-case complexity results
- Inexact Newton $\rightarrow$ Krylov subspace methods
- Preconditioner is a must
- Computational results
  - Compressed Sensing
  - Google Problem
- Conclusions

1st-order Methods for Optimization

The 1st-order methods are applied to **unconstrained** optimization

$$\begin{array}{ll} \min & f(x) + \Psi(x) \\ \text{s.t.} & x \in X, \end{array}$$

where f and Ψ are convex functions
(may be smooth, separable, strongly convex)
and X is an *easy* set ($\mathcal{R}^n$, box, hyperplane, etc)

The 1st-order methods rely on *gradients*
(or sub-gradients) of f and Ψ.

Randomization often helps.

Interior Point Methods (IPMs)

IPMs are applied to **constrained** optimization

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g(x) \leq 0, \\ & h(x) = 0, \end{array}$$

where f, g and h are convex functions.

IPMs easily deal with the *inequalities*:

LO/QO	$x \geq 0,\ x \in \mathcal{R}^n$
NLO	$g(x) \leq 0,\ g : \mathcal{R}^n \mapsto \mathcal{R}^m$
SOCO	$x \in K = K^1 \times K^2 \times \cdots \times K^k$ (cones)
SDO	$X \succeq 0,\ X \in \mathcal{SR}^{n \times n}$

IPMs rely on the 2nd-order information of f, g and h.

Observation

- First-order methods
  - complexity $\mathcal{O}(1/\varepsilon)$ or $\mathcal{O}(1/\varepsilon^2)$
  - produce a rough approx. of solution quickly
  - but ... struggle to converge to high accuracy

- IPMs are second-order methods
  (they apply Newton method to barrier subprobs)
  - complexity $\mathcal{O}(\log(1/\varepsilon))$
  - produce accurate solution in a few iterations
  - but ... one iteration may be expensive

Just think

For example, $\varepsilon = 10^{-3}$ gives
$1/\varepsilon = 10^3$ and $1/\varepsilon^2 = 10^6$, but $\log(1/\varepsilon) \approx 7$.

For example, $\varepsilon = 10^{-6}$ gives
$1/\varepsilon = 10^6$ and $1/\varepsilon^2 = 10^{12}$, but $\log(1/\varepsilon) \approx 14$.

But **ML Community** loves the 1st-order methods.

Stirring up a hornets nest:

Please give IPMs a serious consideration!

Interior Point Methods

LO & QO Problems

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

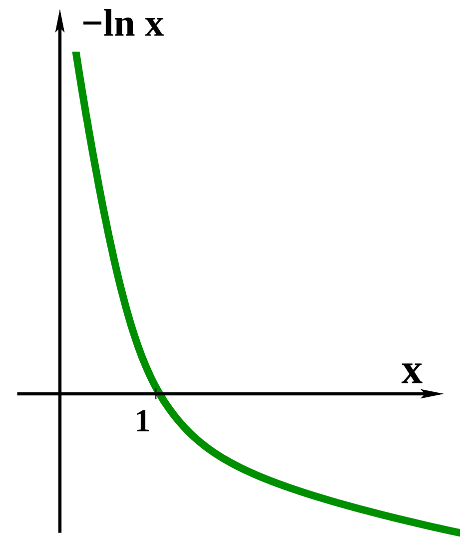
where $A \in \mathcal{R}^{m \times n}$ has full row rank
and $Q \in \mathcal{R}^{n \times n}$ is symmetric positive semidefinite.

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

Interior-Point Framework



The **log barrier** $-\log x_j$

"replaces" the inequality $x_j \geq 0$.

We derive the **first order optimality conditions** for the primal barrier problem:

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

and apply **Newton method** to solve this system of (nonlinear) equations.

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:

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

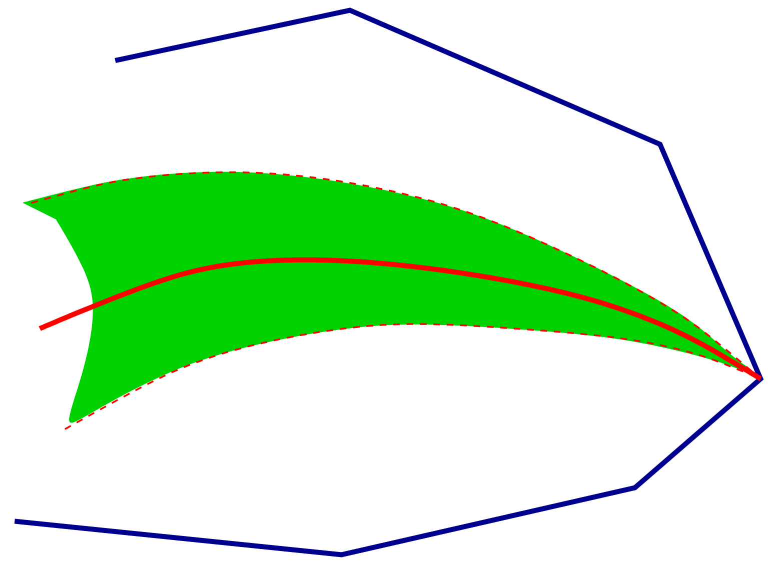
Apply Newton Method to the FOC

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

Central Path:

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

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



$N_2(\theta)$ neighbourhood of the central path

Path Following Method:

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

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

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

where

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

Standard complexity result

Theorem (Wright, Thm 5.12).

Let $\epsilon > 0$ be the required accuracy of the optimal solution. The (*short-step, feasible*) interior point method finds the ϵ-accurate solution such that

$$\mu^k \leq \epsilon$$

after at most

$$K = \mathcal{O}\big(\sqrt{n}\,\log(1/\epsilon)\big)$$

iterations.

Standard IPMs for LO/QO

We know that IPMs converge in

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

But the per-iteration cost may be high

- *practice:* between $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$

Objective: Accelerate IPMs for LO/QO

- Find an ϵ-accurate solution in

$$\mathcal{O}(\log n \; \log(1/\epsilon))$$

iterations (in practice).

- Lower the cost of a single IPM iteration

from $\mathcal{O}\big(n^3\big)$ to $\mathcal{O}\big(n\big)$.

Realistically: make only a few matrix-vector prods.

Use Inexact Newton Method

Dembo, Eisenstat & Steihaug,
SIAM J. on Num Analysis 19 (1982) 400–408.

Exact Newton Method

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

Inexact Newton Method

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

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

General Assumption

The residual r in the inexact Newton Method satisfies:

$$\|r\| \leq \delta \|\xi\|,$$

where $\delta \in (0, 1]$.

What is an acceptable δ ?

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 **inexact** Newton Method with the relative **error**

$$\|r\| \le \delta \|\xi\|.$$

Aspire to reduce duality gap:

$$\bar{\mu} = (1 - \frac{0.1}{\sqrt{n}})\mu$$

and achieve the reduction:

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

Theorem

Suppose the algorithm operates in $\mathcal{N}_2(\theta)$ neighbourhood of the central path and uses an *inexact* Newton Method with the relative precision $\delta = 0.3$.

Then it converges in at most

$$K = \mathcal{O}\left(\sqrt{n}\,\log(1/\epsilon)\right)$$

iterations.

G., Convergence Analysis of an Inexact Feasible IPM for Convex QP, *Tech Rep ERGO-2012-008*, July 2012.

Proof (key ideas)

Control the *error* in Newton Method, namely, the terms $\Delta x^T \Delta s$ and $\|\Delta X \Delta S e\|$.

Show that if the inexactness in the Newton Method is limited then the *error* satisfies

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

Use the *full* Newton step to achieve a sizeable reduction of duality gap in one step.

Conclusion

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

Allow for large residual

$$\|r\| \leq \delta \|\xi\|$$

The worst-case complexity result remains the same!

Observation

We have not made any assumption regarding the source of inexactness.

Possible sources of inexactness

- approximate Hessian Q and/or Jacobian A;
- iterative method to compute Newton direction;
- **probabilistic approach?**

From Theory to Practice

- Compressed Sensing
  with **K. Fountoulakis** and **P. Zhlobich**
- Google Problem
  with **K. Woodsend**

both exploit/rely on probabilistic arguments.

Sparse Approximations joint work with
Kimon Fountoulakis and **Pavel Zhlobich**

- Statistics: Estimate x from observations
- Wavelet-based signal/image reconstr./restoration
- Compressed Sensing (Signal Processing)

Re-cast as large dense quadratic optimization problem:

$$\min_x \; \frac{1}{2}\|Ax - b\|_2^2 + \tau\|x\|_1,$$

where $A \in \mathcal{R}^{m \times n}$.

The **ML Community** likes this problem very much.

Bayesian Statistics Viewpoint

Estimate x from observations

$$b = Ax + e,$$

where b are observations and e is the Gaussian noise.

$\rightarrow$ $\min_x \|Ax - b\|_2^2$

If the prior on x is Laplacian ($\log p(x) = -\lambda\|x\|_1 + K$) then

$$\min_x \|Ax - b\|_2^2 + \tau\|x\|_1$$

Tibshirani, *J. of Royal Stat Soc B* 58 (1996) 267-288.

Wavelet-based Signal/Image Reconstruction

A has the form $A = RW$, where

- R is the observation operator (think: tomographic projection)
  R is a *matrix* representation of this operator
- W is a wavelet basis or a redundant dictionary
  operation Wx corresponds to performing an inverse wavelet transform
- x is the vector representation coefficients of the unknown signal/image

Chen, Donoho & Saunders,
SIAM J. on Sci Comp 20 (1998) 33-61.

Compressed Sensing

> *Relatively small number of random projections of a sparse signal can contain most of its salient information.*

If a signal is sparse (or approximately sparse) in some orthonormal basis, then an accurate reconstruction can be obtained from random projections of the original signal. A has the form $A = RW$, where

- R is a low-rank randomised sensing matrix
- W is a basis over which the signal has a sparse representation

Candès, Romberg & Tao,
Comm on Pure and Appl Maths 59 (2005) 1207-1233.

LO/QO Reformulations

$$\min_x \|Ax - b\|_2^2 + \tau \|x\|_1$$

or

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \|Ax - b\|_2 \leq \varepsilon \qquad \big(\text{or} \quad Ax = b\big)$$

or

$$\min_x \ \|Ax - b\|_2^2 \quad \text{s.t.} \quad \|x\|_1 \leq t$$

that is

$$\min_x \ w^T w \quad \text{s.t.} \quad Ax - b = w \quad \text{and} \quad \|x\|_1 \leq t$$

Two-way Orthogonality of A

- *rows* of A are orthogonal to each other (A is built of a subset of rows of an othonormal matrix $U \in \mathcal{R}^{n\times n}$)

$$AA^T = I_m.$$

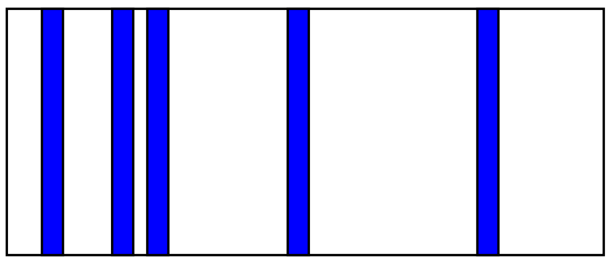
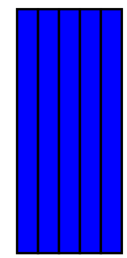
- small subsets of *columns* of A are nearly-orthogonal to each other: *Restricted Isometry Property (RIP)*

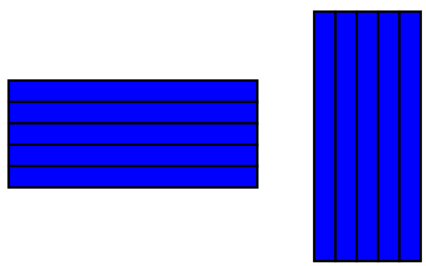
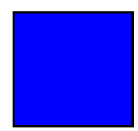
$$\|\bar{A}^T \bar{A} - \frac{m}{n} I_k\| \le \delta_k \in (0,1).$$

Candès, Romberg & Tao,
Comm on Pure and Appl Maths 59 (2005) 1207-1233.

Restricted Isometry Property

Matrix $\bar{A} \in \mathcal{R}^{m \times k}$ $(k \ll n)$ is built of a subset of columns of $A \in \mathcal{R}^{m \times n}$.

$A =$  $\longrightarrow$ $\bar{A} =$ 

$\bar{A}^T \bar{A} =$  $=$  $\approx \dfrac{m}{n} I_k.$

This yields a very well conditioned optimization problem.

Problem Reformulation

$$\min_x \ \frac{1}{2}\|Ax-b\|_2^2 + \tau\|x\|_1,$$

Replace $x = x^+ - x^-$ to be able to use $|x| = x^+ + x^-$.
Use $|x_i| = z_i + z_{i+n}$ to replace $\|x\|_1$ with $\|x\|_1 = 1_{2n}^T z$.
(Increases problem dimension from n to $2n$.)

$$\min_{z\geq 0} \ \frac{1}{2} z^T Q z + c^T z,$$

where

$$Q = \begin{bmatrix} A^T \\ -A^T \end{bmatrix} \begin{bmatrix} A & -A \end{bmatrix} = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix} \in \mathcal{R}^{2n\times 2n}$$

Preconditioner

Approximate

$$\mathcal{M} = \begin{bmatrix} A^TA & -A^TA \\ -A^TA & A^TA \end{bmatrix} + \begin{bmatrix} \Theta_1^{-1} & \\ & \Theta_2^{-1} \end{bmatrix}$$

with

$$\mathcal{P} = \frac{m}{n} \begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix} + \begin{bmatrix} \Theta_1^{-1} & \\ & \Theta_2^{-1} \end{bmatrix}.$$

We expect (*optimal partition*):

- k entries of $\Theta^{-1} \to 0, \quad k \ll 2n,$
- $2n - k$ entries of $\Theta^{-1} \to \infty.$

Spectral Properties of $\mathcal{P}^{-1}\mathcal{M}$

Theorem

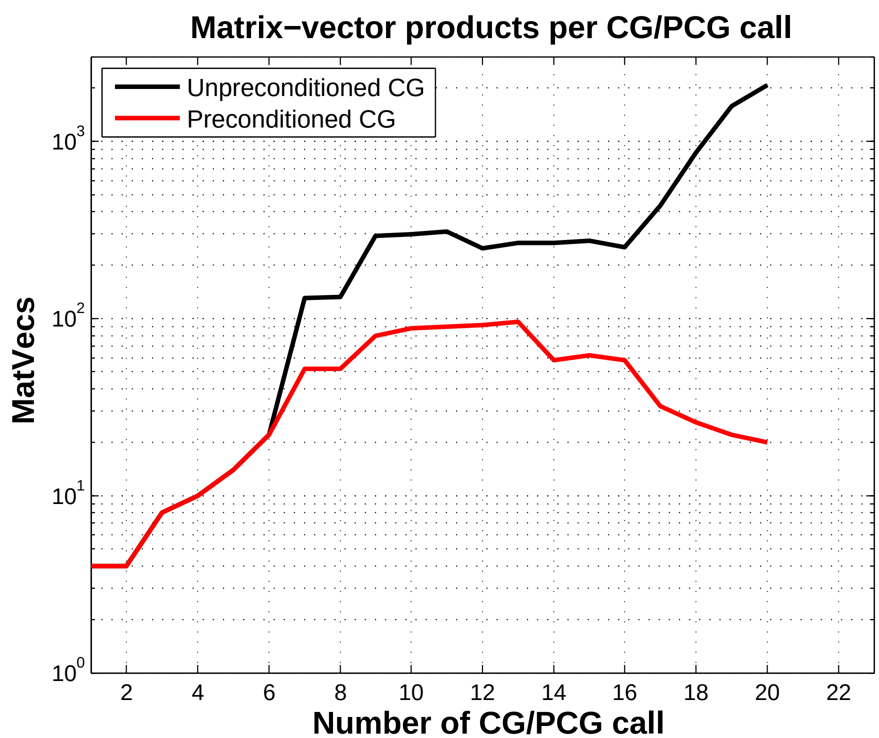
- Exactly n eigenvalues of $\mathcal{P}^{-1}\mathcal{M}$ are 1.
- The remaining n eigenvalues satisfy

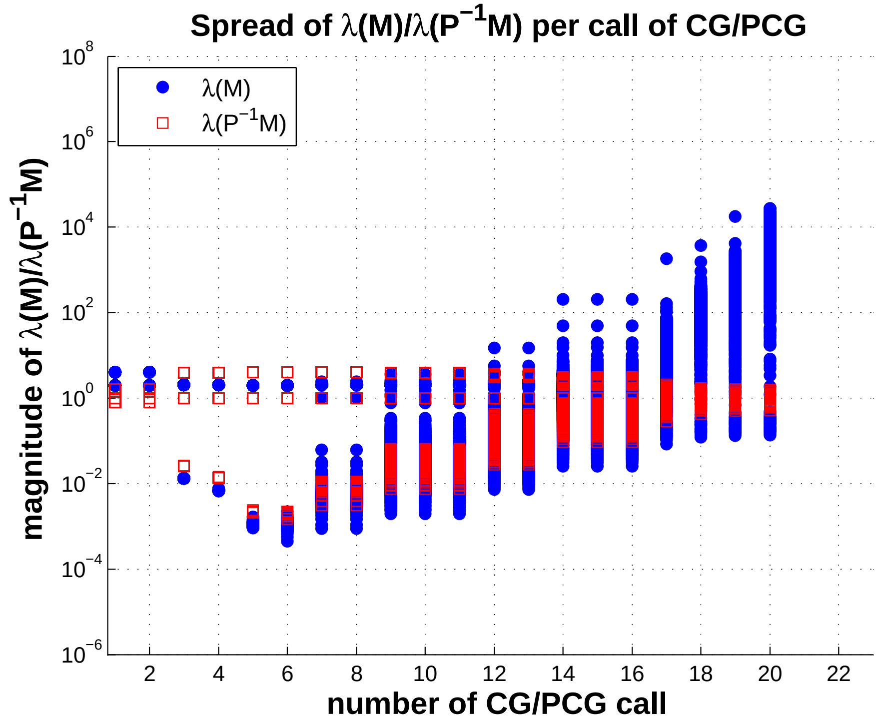
$$|\lambda(\mathcal{P}^{-1}\mathcal{M}) - 1| \leq \delta_k + \frac{n}{m\delta_k L},$$

where δ_k is the RIP-constant, and
L is a threshold of "large" $(\Theta_1 + \Theta_2)^{-1}$.

Fountoulakis, G., Zhlobich
Matrix-free IPM for Compressed Sensing Problems,
ERGO Technical Report, 2012.

Preconditioning





$\longrightarrow$ good clustering of eigenvalues

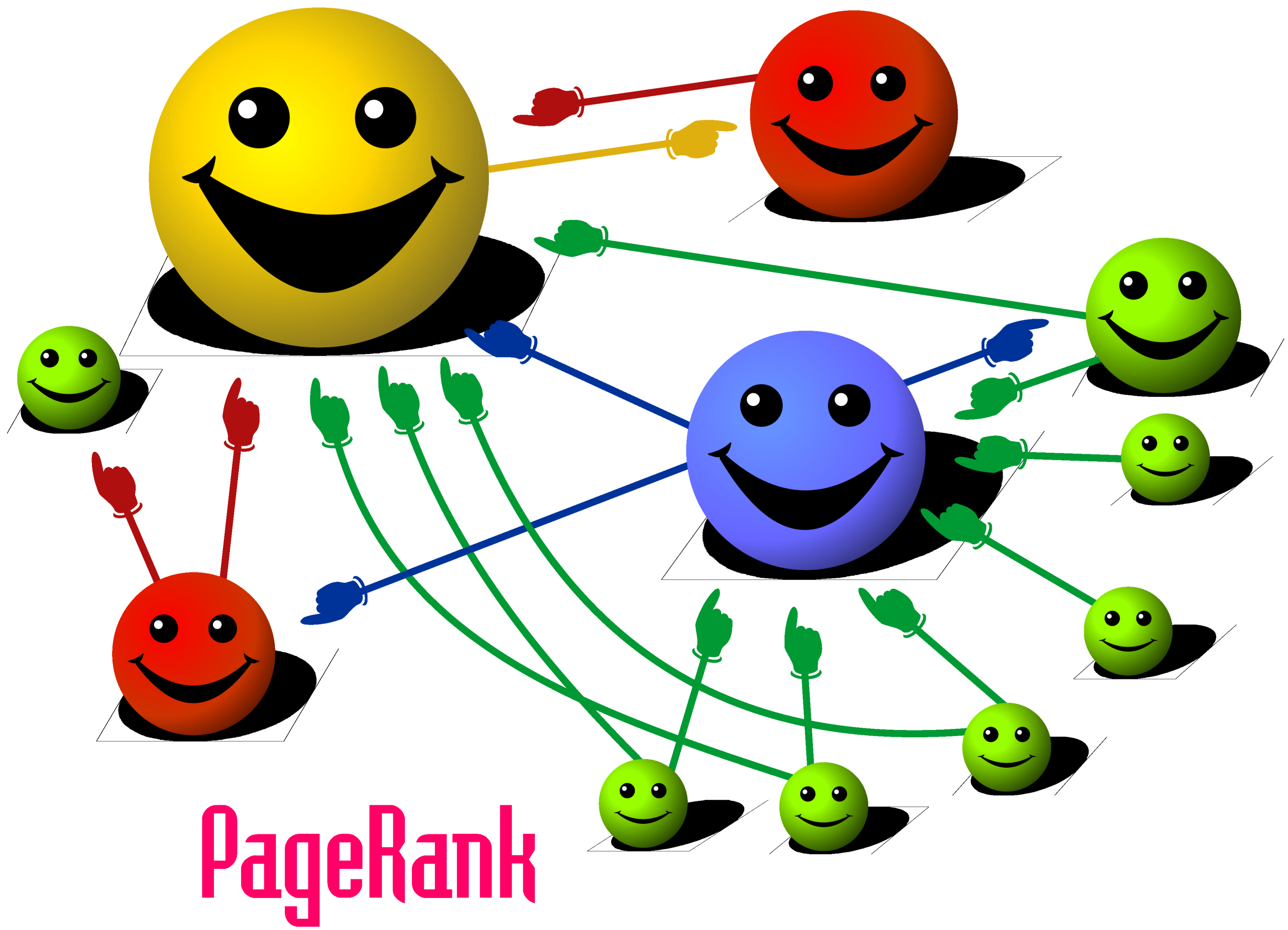
Computational Results: Comparing MatVecs

Prob size	k	**NestA**	**mf-IPM**
4k	51	424	301
16k	204	461	307
64k	816	453	407
256k	3264	589	537
1M	13056	576	613

NestA, Nesterov's smoothing gradient
Becker, Bobin and Candés,
`http://www-stat.stanford.edu/~candes/nesta/`

mf-IPM, Matrix-free IPM
Fountoulakis, G. and Zhlobich,
`http://www.maths.ed.ac.uk/ERGO/`

Ranking of nodes in networks



Google Problem joint work with
Kristian Woodsend

An adjacency matrix $G \in \mathcal{R}^{n \times n}$ of web-page links is given (web-pages are the nodes). G is *column-stochastic*.

Teleportation:

$$M = \lambda G + (1-\lambda)\frac{1}{n}ee^T,$$

with $\lambda \in (0,1)$, usually $\lambda = 0.85$.

Find the *dominant right eigenvector* x of M with eigenvalue equal to 1

$$Mx = x, \quad \text{such that} \quad e^T x = 1, \; x \geq 0.$$

and use x as a **ranking vector**.

Google Problem

$$\begin{array}{rl} \min & \frac{1}{2}\|Mx - x\|_2^2 \\ \text{s.t.} & e^T x = 1, \;\; x \geq 0 \end{array}$$

Rearrange:

$$\|Mx - x\|_2^2 = x^T(M-I)^T(M-I)x$$

to produce a standard QP formulation with

$$Q = (M-I)^T(M-I).$$

A very easy QP problem!

Preconditioner for Google Problem

Approximate

$$\mathcal{M} = \begin{bmatrix} Q + \Theta^{-1} & e \\ e^T & 0 \end{bmatrix}$$

with

$$\mathcal{P} = \begin{bmatrix} D_Q & e \\ e^T & 0 \end{bmatrix},$$

where $D_Q = diag\{Q + \Theta^{-1}\}$.

G., Woodsend
Matrix-free IPM for Google Problems,
ERGO Technical Report (in preparation) 2012.

Computational Results: mf-IPM

	Size	degree	IPM-iters	**MatVecs**
$\lambda = 0.85$	4k	20	6	13
	16k	20	5	8
	64k	20	4	5
	256k	20	3	4
	1M	20	3	11
$\lambda = 1.0$	4k	20	6	13
	16k	20	5	8
	64k	20	4	5
	256k	20	3	6
	1M	20	3	14

mf-IPM much faster than Nesterov's smoothing grad.

New IPMs:

- The *inexact* IPM enjoys the same worst-case iteration complexity as the *exact* IPM
- *Matrix-free IPM* solves many difficult problems

The **2nd order information** can (sometimes should) be used in optimization.

Inexact Newton directions in IPMs:

- little (if any) increase of iteration number
- significant reduction of per-iteration cost

Might there be a probabilistic inexact approach?

Thank You!

Matrix-Free IPM:

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

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

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 & G., *OMS* 11-12 (1999) 275-302.

General Case Normal Equation Matrix

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

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

Use diagonal pivoting to compute

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

$L = \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix}$ is trapezoidal, k columns of Cholesky;

$S \in \mathcal{R}^{(m-k)\times(m-k)}$ 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

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

and precondition $\mathcal{G}_R$ with

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

where D_S is a diagonal of S.

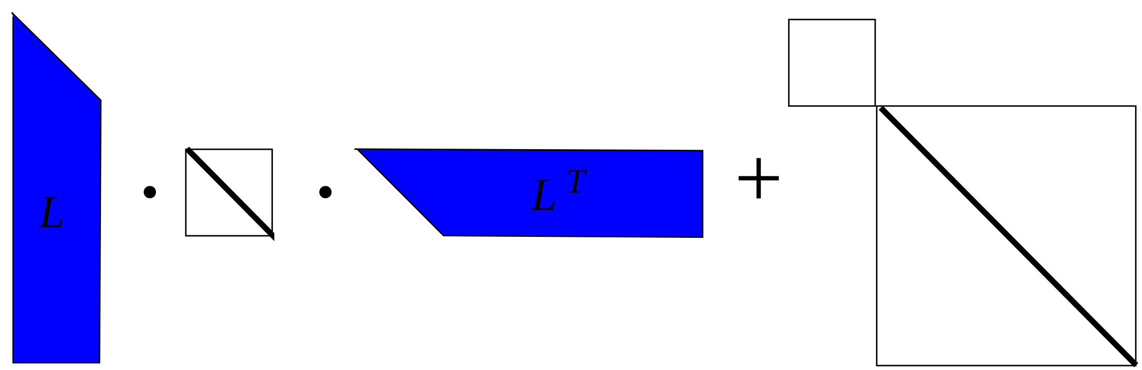
Do **not** compute S.
Update only its diagonal.

Preconditioner

Partial Cholesky of NE system

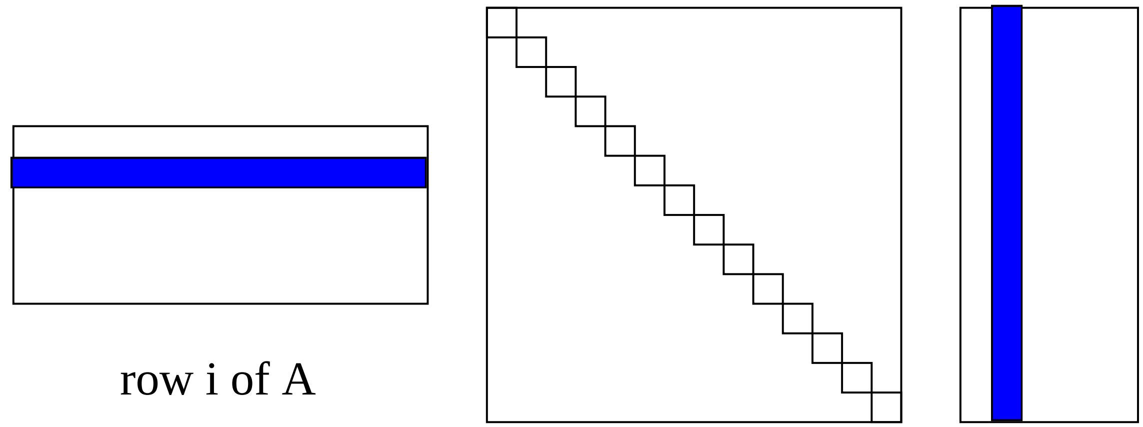
$$\mathcal{G}_R = (A(Q + \Theta^{-1} + R_p)^{-1} A^T + R_d) \approx L D_L L^T + D_S$$

$$L D_L L^T + D_S =$$



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

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).

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*

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

with **Gruca, Hall, Laskowski** and **Żukowski**

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

`Intel` Core i7 3.07GHz processor, 24 GB memory

General Case (two examples):

- Quadratic Assignment Problems (QAP)
  joint work with **Ed Smith** and **J.A.J. Hall**
- Quantum Information Theory Problems
  with **Gruca, Hall, Laskowski** and **Żukowski**

Standard approaches (Cplex Simplex and Cplex Barrier) break down on medium problems: $16K \leq m, n \leq 64K$

Matrix-free IPM solves these problems in *minutes*

MF-IPM solves large problems $m, n \geq 1M$ in *hours*