Inexact search directions and matrix-free second-order methods for optimization

Jacek Gondzio

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

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 $\log(1/\varepsilon) \approx 14$.

Stirring up a hornets nest:

Give 2nd-order/IPMs a serious consideration!

Motivation

Large problems are there:

- too large to store
- direct methods (factorizations) impossible
- matrices are available in some "simple" form: very sparse, or fast `MatVec` operators

If such problems are *easy* (many of them are), then the 1st-order methods may be used

But what if the problems are *not so easy*?

Outline

- Motivation: Make *2nd-order* methods faster
- Inexact Newton directions
- Homotopy: IPMs and Continuation
- $\boldsymbol{\ell_1}$-regularization
  - use *smoothing* (pseudo-Huber function)
    $\rightarrow$ we need the 2nd-order information
  - use *continuation*
    $\rightarrow$ to improve the pseudo-Huber approx
  - work *matrix-free*
- Computational results
- Conclusions

ℓ_1-regularization

Convex optimization problem:

$$\min_x \quad \tau \|x\|_1 + \phi(x),$$

where $\|.\|_1$ is the ℓ_1 norm, and
$\phi : \mathcal{R}^n \mapsto \mathcal{R}$ is a convex function (often strongly convex).

Usual example:

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

where $A \in \mathcal{R}^{m \times n}$ (often $m \geq n$ or $m \gg n$).

ℓ_1-regularization

$$\min_{x} \quad \tau \|x\|_1 + \phi(x).$$

Unconstrained optimization $\Rightarrow$ easy(?)

Serious Issue: nondifferentiability of $\|.\|_1$

Two possible tricks:

- Splitting $x = u - v$ with $u, v \geq 0$
- Huber or pseudo-Huber regression

Splitting: $\boldsymbol{x} = \boldsymbol{u} - \boldsymbol{v},\ \boldsymbol{u} \geq \boldsymbol{0},\ \boldsymbol{v} \geq \boldsymbol{0}$

Replace $x_i = u_i - v_i$,
where $u_i = \max\{x_i, 0\}$ and $v_i = \max\{-x_i, 0\}$.

Then $x_i = u_i - v_i$ and $|x_i| = u_i + v_i$.

Hence $\|x\|_1 = \sum_{i=1}^{n} (u_i + v_i)$.

Removes nondifferentiability, but:

- doubles the dimension,
- introduces inequality constraints (fine for IPMs).

Huber: **Replace $\|\mathbf{x}\|_1$ with $\boldsymbol{\psi_\mu(x)}$**

Huber approximation: replaces $\|x\|_1$ with $\sum_{i=1}^n \Big[\psi_\mu(x)\Big]_i$

$$\Big[\psi_\mu(x)\Big]_i = \begin{cases} \frac{1}{2}\mu^{-1}x_i^2, & \text{if } |x_i| \le \mu \\ |x_i| - \frac{1}{2}\mu, & \text{if } |x_i| \ge \mu \end{cases} \qquad i = 1, 2, \dots, n$$

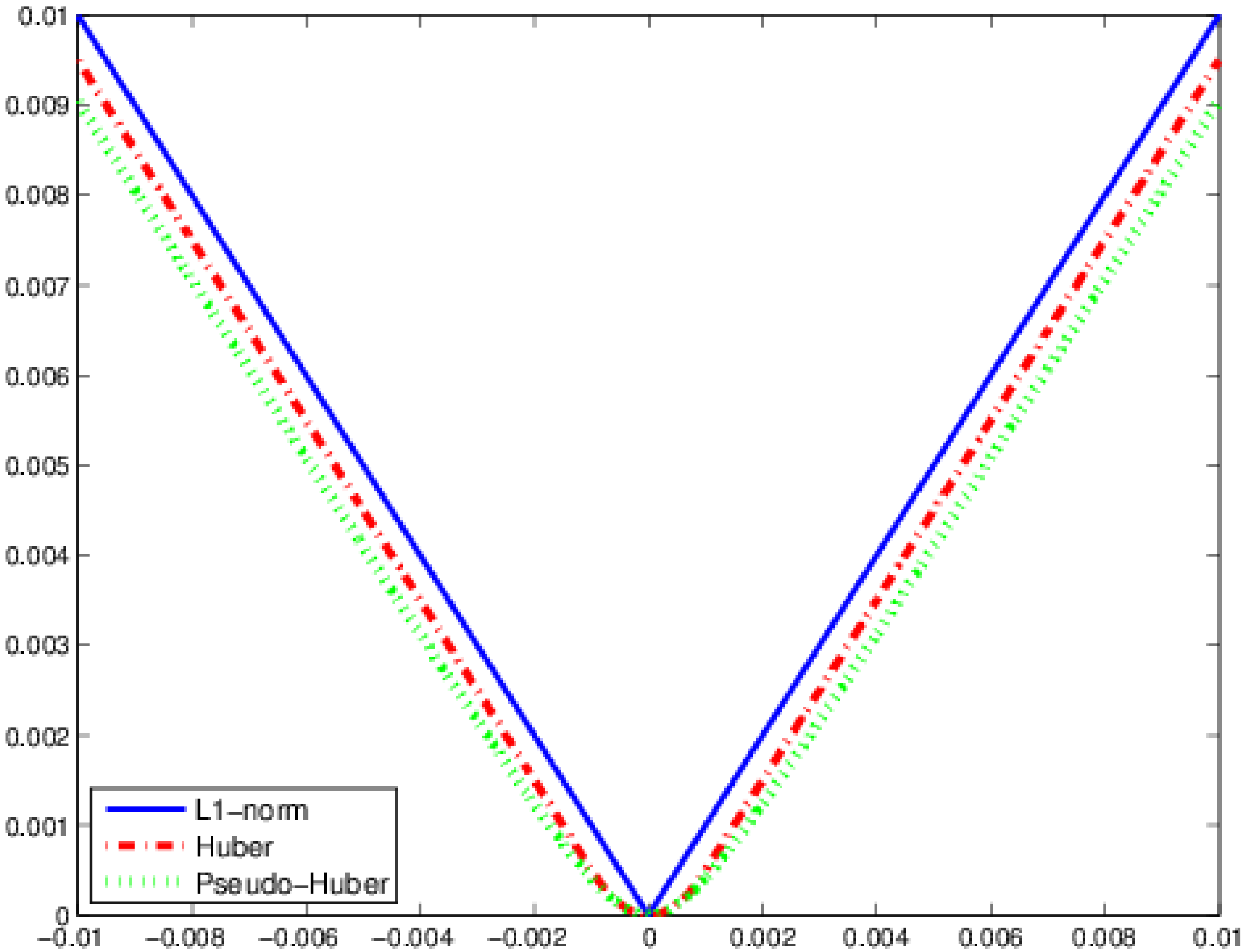
where $\mu > 0$. Only first-order differentiable.

Pseudo Huber approximation: replaces $\|x\|_1$ with

$$\psi_\mu(x) = \mu \sum_{i=1}^n \Big(\sqrt{1 + \frac{x_i^2}{\mu^2}} - 1\Big)$$

Smooth function, has derivatives of any degree.

Huber:



0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
−0.01
−0.008
−0.006
−0.004
−0.002
0
0.002
0.004
0.006
0.008
0.01
L1−norm
Huber
Pseudo−Huber

Continuation

Embed inexact Newton Meth into a *homotopy* approach:

- Inequalities $u \geq 0, v \geq 0 \quad \longrightarrow$ use **IPM**

  replace $z \geq 0$ with $-\mu \, log z$ and drive μ to zero.

- pseudo-Huber regression $\quad \longrightarrow$ use **continuation**

  replace $|x_i|$ with $\mu(\sqrt{1+\frac{x_i^2}{\mu^2}}-1)$ and drive μ to zero.

Theory ???

Inexact Newton Direction in IPMs

Replace an *exact* Newton direction

$$\nabla^2 f(x)\Delta x = -\nabla f(x)$$

with an *inexact* one:

$$\nabla^2 f(x)\Delta x = -\nabla f(x) + \boldsymbol{r},$$

where the *error* $\boldsymbol{r}$ is small: $\|\boldsymbol{r}\| \leq \boldsymbol{\eta}\|\nabla f(x)\|$, $\boldsymbol{\eta} \in (0,1)$.

The NLP community usually writes it as:

$$\|\nabla^2 f(x)\Delta x + \nabla f(x)\|_2 \leq \boldsymbol{\eta}\|\nabla f(x)\|_2, \quad \boldsymbol{\eta} \in (0,1).$$

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

Theorem: Suppose the feasible IPM for QP is used.
If the method operates in the *small* neighbourhood

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

and uses the *inexact* Newton direction with $\boldsymbol{\eta} = 0.3$, then it converges in at most

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

If the method operates in the *symmetric* neighbourhood

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

and uses the *inexact* Newton direction with $\boldsymbol{\eta} = 0.05$, then it converges in at most

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

Theory for IPM:

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

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

Computational practice:

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

- $\mathcal{O}(\log n)$ iterations
- with $\mathcal{O}(nz(A))$ cost per iteration.

Quantum Information Problems

Prob	**Cplex 12.0**				**mf-IPM**	
	Simplex		Barrier		rank=200	
	its	time	its	time	its	time
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

G., Gruca, Hall, Laskowski and Żukowski, Solving Large-Scale Optimization Problems Related to Bell's Theorem, *J. of Computational and Applied Maths*, 263C (2014) 392–404.

ℓ_1-Regularization and Continuation

Use Pseudo-Huber approximation to replace $\|\mathbf{u}(\mathbf{x})\|_1$ with

$$\psi_\mu(u(x)) = \mu \sum_{i=1}^{n} \left(\sqrt{1 + \frac{(u(x))_i^2}{\mu^2}} - 1\right)$$

Hence replace

$$\min_x \quad \tau \|u(x)\|_1 + \phi(x)$$

with

$$\min_x \quad \tau \psi_\mu(u(x)) + \phi(x)$$

Solve approximately a family of problems for a (short) decreasing sequence of μ's: $\boldsymbol{\mu_0 > \mu_1 > \mu_2 \cdots}$

Three examples of ℓ_1-regularization

- Compressed Sensing
  with **K. Fountoulakis and P. Zhlobich**

- Compressed Sensing (Coherent and Redundant Dict.)
  with **I. Dassios and K. Fountoulakis**

- Machine Learning Problems
  with **K. Fountoulakis**

Example 1: Compressed Sensing

with **K. Fountoulakis and P. Zhlobich**

Large dense quadratic optimization problem:

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

where $A \in \mathcal{R}^{m \times n}$ is a **very special matrix**.

Fountoulakis, G., Zhlobich
Matrix-free IPM for Compressed Sensing Problems,
Math. Prog. Computation 6 (2014), pp. 1–31.

Software available at `http://www.maths.ed.ac.uk/ERGO/`

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 $U \in \mathcal{R}^{n \times n}$)

$$AA^T = I_m.$$

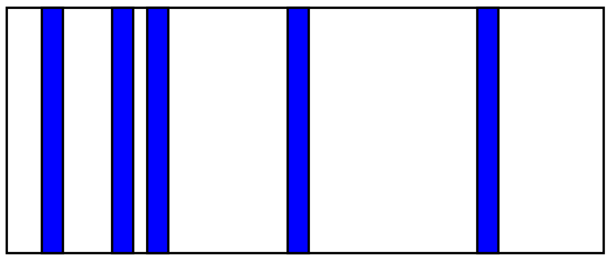
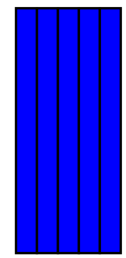
- small subsets of *columns* of A are nearly-orthogonal to each other: *Restricted Isometry Property (RIP)*

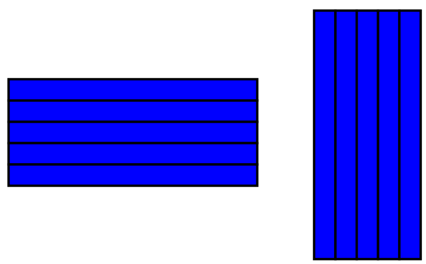
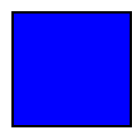
$$\|\bar{A}^T \bar{A} - \frac{m}{n} I_k\| \leq \delta_k \in (0, 1).$$

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

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

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

This yields a very well conditioned optimization problem.

Problem Reformulation

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

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 $2n$.)

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

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

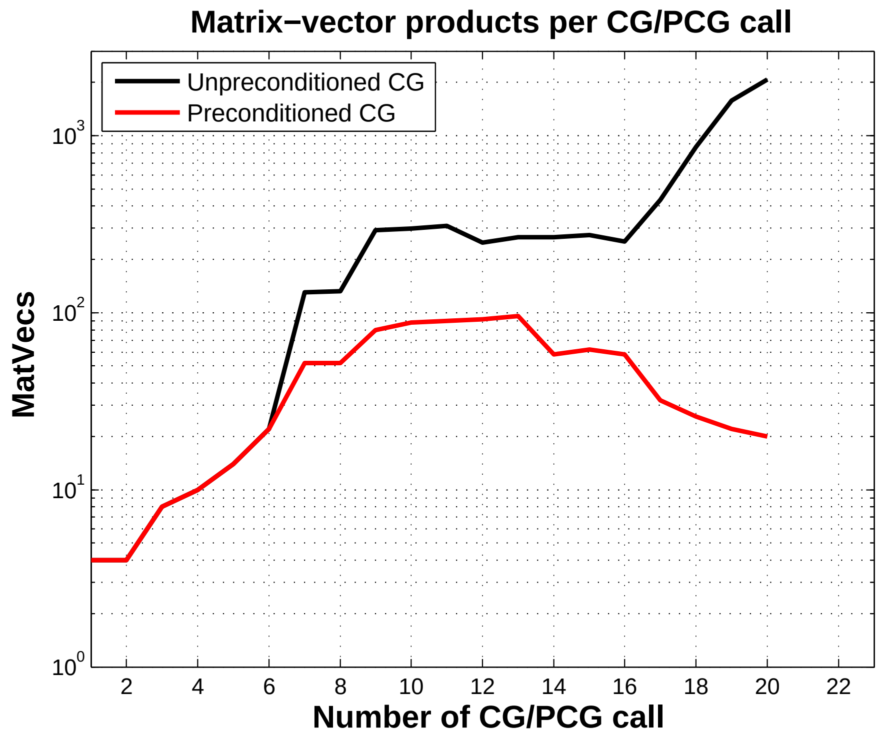
- Exactly n eigenvalues of $\mathcal{P}^{-1}\mathcal{M}$ are 1.
- The remaining n eigenvalues satisfy

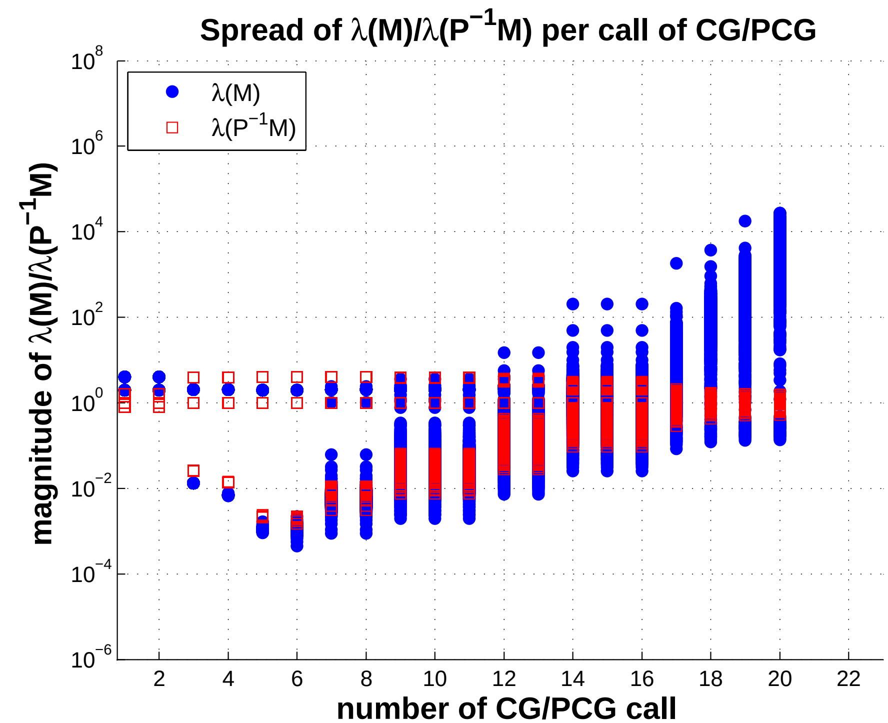
$$|\lambda(\mathcal{P}^{-1}\mathcal{M}) - 1| \le \delta_k + \frac{n}{m\delta_k L},$$

  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,
Math. Prog. Computation 6 (2014), pp. 1–31.

Preconditioning





$\longrightarrow$ good clustering of eigenvalues

mfIPM compares favourably with NestA on easy probs (NestA: Becker, Bobin and Candés).

Computational Results: Comparing MatVecs

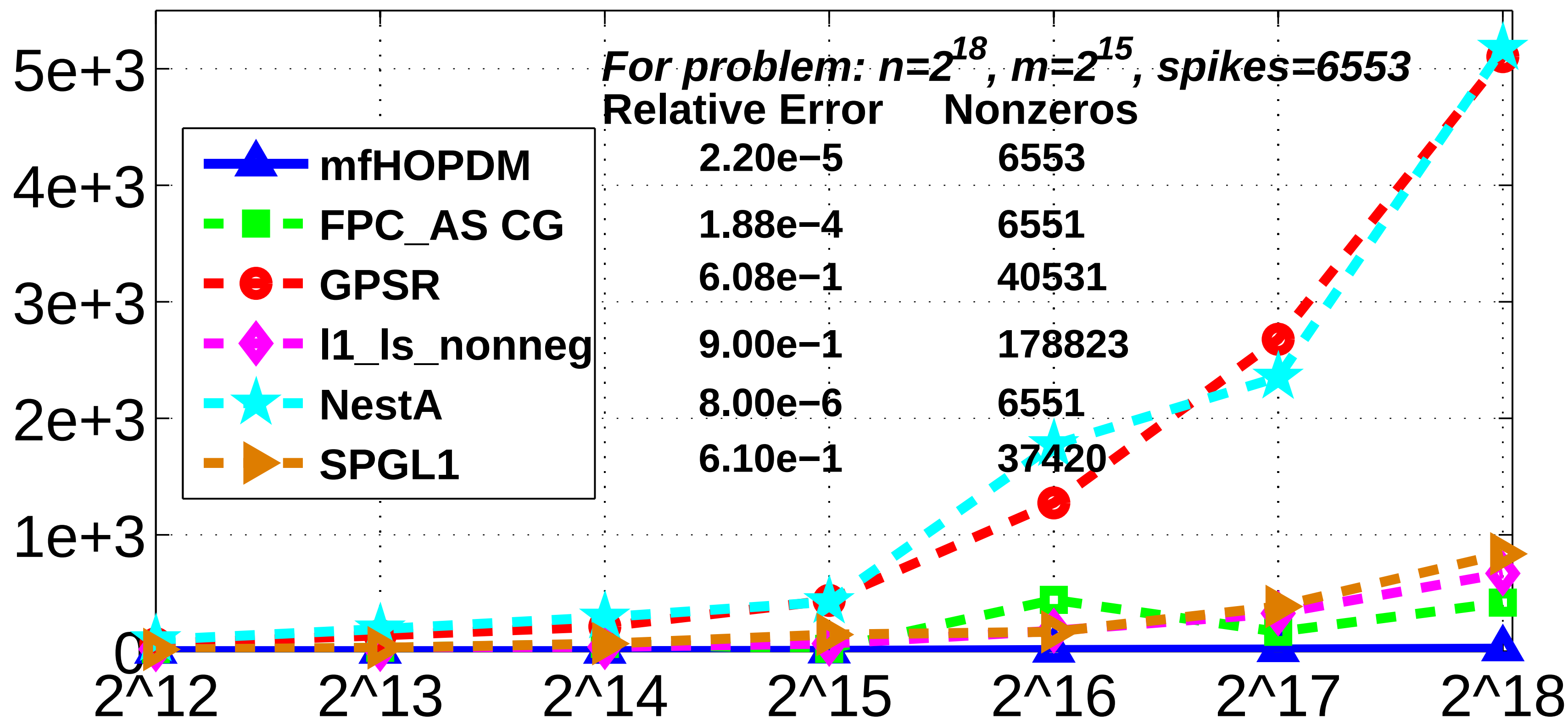
Prob size	k	**NestA**	**mfIPM**
4k	51	424	301
16k	204	461	307
64k	816	453	407
256k	3264	589	537
1M	13056	576	613

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

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

Nontrivial Reconstruction Problems

Sparse vector: entries zero or 10^5. Gaussian noise $\sigma = 0.1$



SPARCO problems

Comparison on 18 out of 26 classes of problems
(all but 6 complex and 2 installation-dependent ones).

Solvers compared:

PDCO, *Saunders and Kim*, Stanford,
$\boldsymbol{\ell}_1\text{–}\boldsymbol{\ell}_s$, *Kim, Koh, Lustig, Boyd, Gorinevsky*, Stanford,
FPC-AS-CG, *Wen, Yin, Goldfarb, Zhang*, Rice,
SPGL1, *Van Den Berg, Friedlander*, Vancouver, and
mf-IPM, *Fountoulakis, G., Zhlobich*, Edinburgh.

On 36 runs (noisy and noiseless problems), **mf-IPM**:

- is the fastest on 11,
- is the second best on 14, and
- overall is very robust.

Example 2: CS, Coherent & Redundant Dict.

with **I. Dassios and K. Fountoulakis**.

Large dense quadratic optimization problem:

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

where $A \in \mathcal{R}^{m\times n}$ and $W \in \mathcal{C}^{n\times l}$ is a *dictionary*.

Dassios, Fountoulakis and G.
A Second-order Method for Compressed Sensing Problems with Coherent and Redundant Dictionaries, *Tech Rep ERGO-2014-007*, May 2014.

Software available at `http://www.maths.ed.ac.uk/ERGO/`

Compressed Sensing and Continuation

Replace

$$\min_x \quad f(x) = \tau \|W^* x\|_1 + \frac{1}{2}\|Ax - b\|_2^2, \qquad \longrightarrow \boldsymbol{x_\tau}$$

with

$$\min_x \quad f_\mu(x) = \tau \psi_\mu(W^* x) + \frac{1}{2}\|Ax - b\|_2^2, \qquad \longrightarrow \boldsymbol{x_{\tau,\mu}}$$

Solve approximately a family of problems for a (short) decreasing sequence of μ's: $\boldsymbol{\mu_0 > \mu_1 > \mu_2 \cdots}$

Theorem (Brief description)

There exists a $\tilde{\mu}$ such that $\forall \mu \leq \tilde{\mu}$ the difference of the two solutions satisfies

$$\|x_{\tau,\mu} - x_\tau\|_2 = \mathcal{O}(\mu^{1/2}) \quad \forall\, \tau, \mu$$

Convergence of the primal-dual Newton CG

Use inexact Newton directions:

$$\|\nabla^2 f_\mu(x)\Delta x + \nabla f_\mu(x)\|_2 \leq \boldsymbol{\eta}\|\nabla f_\mu(x)\|_2, \quad \boldsymbol{\eta} \in (0,1)$$

computed by the Newton CG method.

Theorem (Primal convergence)

Let $\{x^k\}_{k=0}^\infty$ be a sequence generated by pdNCG. Then the sequence $\{x^k\}_{k=0}^\infty$ converges to the primal (perturbed) solution $x_{\tau,\mu}$.

Theorem (Rate of convergence)

If the forcing factor η^k satisfies $\lim_{k\to\infty}\eta^k = 0$, then pdNCG converges superlinearly.

W-Restricted Isometry Property (W-RIP)

- *rows* of A are nearly-orthogonal to each other, i.e., there exists a small constant δ such that

$$\|AA^T - I_m\| \leq \delta.$$

- *W-Restricted Isometry Property (W-RIP)*: there exists a constant δ_q such that

$$(1-\delta_q)\|Wz\|_2^2 \leq \|AWz\|_2^2 \leq (1+\delta_q)\|Wz\|_2^2$$

for all at most q-sparse $z \in \mathcal{C}^n$.

Candès, Eldar & Nedell,
Appl and Comp Harmonic Anal 31 (2011) 59-73.

Preconditioner

Approximate

$$\mathcal{H} = \tau \nabla^2 \psi_\mu (W^* x) + A^T A$$

with

$$\mathcal{P} = \tau \nabla^2 \psi_\mu (W^* x) + \rho I_n .$$

We expect (*optimal partition*):

- k entries of $W^* x \gg 0, \quad k \ll l,$
- $l - k$ entries of $W^* x \approx 0.$

The preconditioner approximates well the 2nd derivative of the pseudo-Huber regularization.

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

Theorem

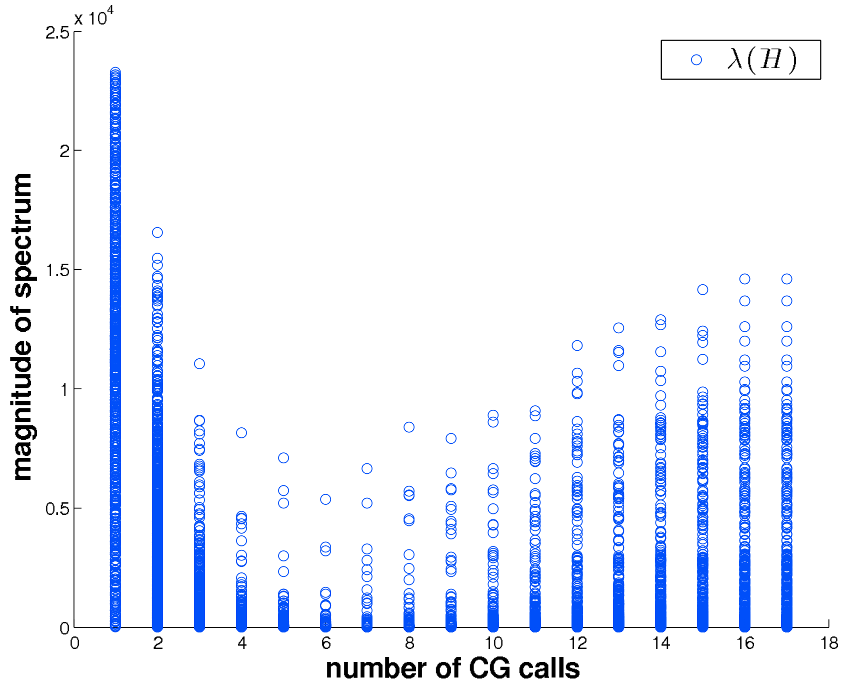
- The eigenvalues of $\mathcal{P}^{-1}\mathcal{H}$ satisfy

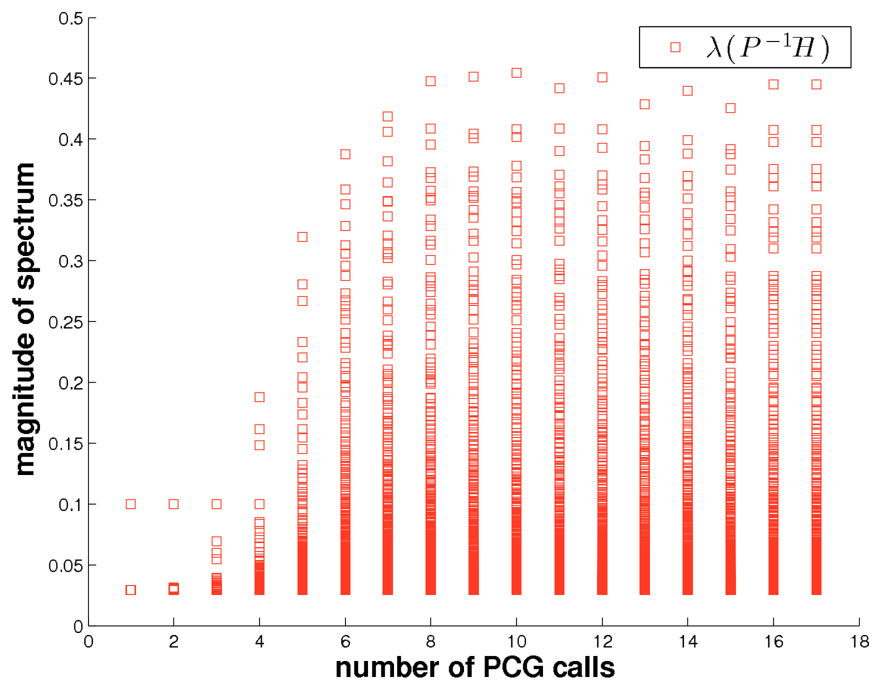
$$|\lambda(\mathcal{P}^{-1}\mathcal{H}) - 1| \leq \frac{\eta(\delta, \delta_q, \rho)}{\rho},$$

  where δ_q is the W-RIP constant,
  δ is another small constant, and
  $\eta(\delta, \delta_q, \rho)$ is some simple function.

Dassios, Fountoulakis and G.
A Second-order Method for Compressed Sensing Problems with Coherent and Redundant Dictionaries, *Tech Rep ERGO-2014-007*, May 2014.

CS: Coherent and Redundant Dictionaries



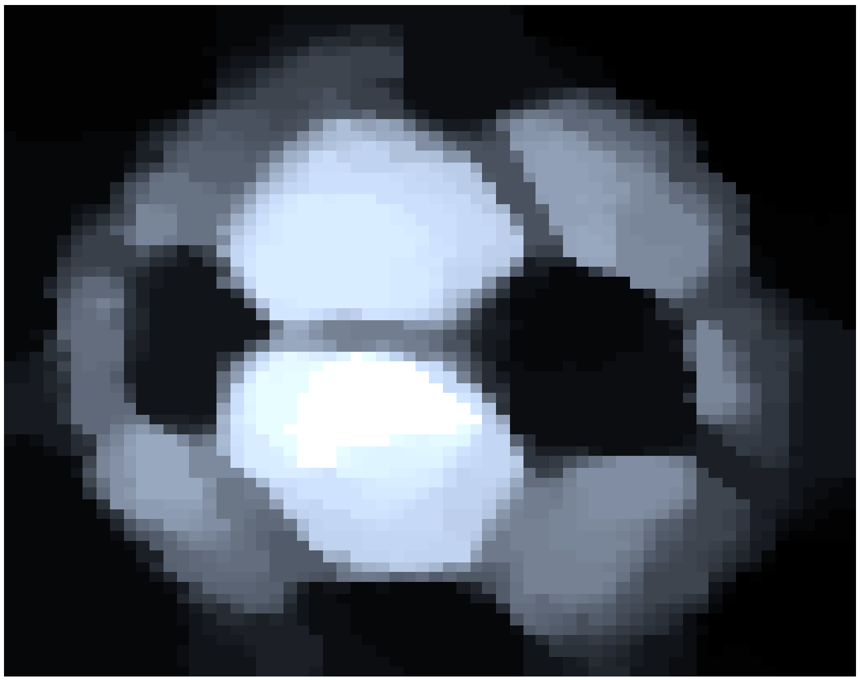


$\longrightarrow$ good clustering of eigenvalues

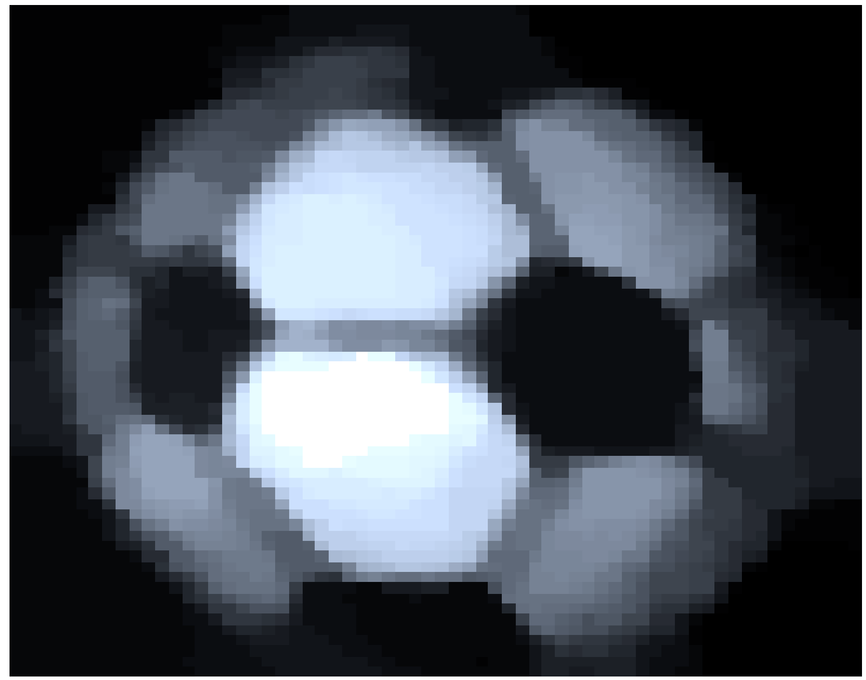
pdNCG outperforms TFOCS on several examples
(TFOCS: Becker, Candés and Grant).

Brazil 2014 A 64×64 resolution example:

Single pixel camera problem set:
`http://dsp.rice.edu/cscamera`



TFOCS, 24 sec.



pdNCG, 15 sec.

Example 3: Machine Learning Problems

with **K. Fountoulakis**

Large dense quadratic optimization problem:

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

where $A \in \mathcal{R}^{m \times n}$ is: **very sparse** and **unstructured**

Fountoulakis and G.
A Second-order Method for Strongly Convex ℓ_1 Regularization, *Tech Rep ERGO-2014-005*, April 2014.

Software available at `http://www.maths.ed.ac.uk/ERGO/`

Amazing efficiency of the 1st order methods

Nesterov, *Math Prog*, 103 (2005) 127-152.

Nesterov, Gradient methods for minimizing composite objective function. *CORE Discussion Papers 2007076*, September 2007.

Richtárik and Takáč, Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math Prog*, 2012.

Richtárik and Takáč, Parallel coordinate descent methods for big data optimization. *Tech Rep ERGO-2012-013*, November 2012.

Problem with $\boldsymbol{n = 2 \times 10^9}$ solved in **37 MatVecs**!

What is going on here?

If we ignore the nondifferentiable $\|x\|_1$ term, then the minimization of $\|Ax - b\|_2^2$ is equivalent to solving

$$\left(A^T A\right) x = A^T b.$$

The *conjugate gradient method* applied to solve this system has the following rate of convergence:

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

where e^k is the error at iteration k and κ is the condition number of $A^T A$.

Inverse engineering exercise:
For $\epsilon = 10^{-2}$, $\kappa \approx 300$, for $\epsilon = 10^{-4}$, $\kappa \approx 64$, for $\epsilon = 10^{-6}$, $\kappa \approx 29$.

Toy Problem (used by 1st-order community)

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

where $A \in \mathcal{R}^{m \times n}$ ($m = 2n$: overdetermined system).
Dimensions: $\boldsymbol{m = 4 \times 10^9}$, $\boldsymbol{n = 2 \times 10^9}$.
Very sparse: 20 nonzero entries per column.

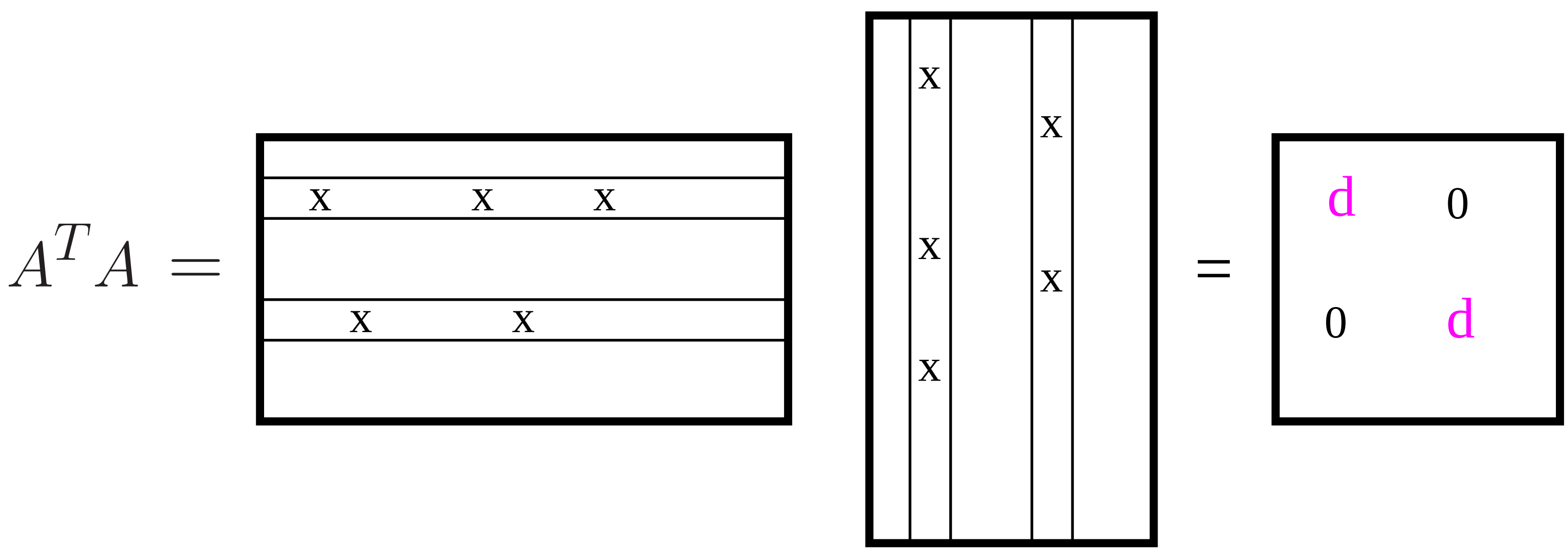
- **Parallel RCD (Richtárik and Takáč)**
  solves it doing 34-37 scans through the matrix
  35 iterations, CPU time: 10779s;
- **Inexact Newton (Fountoulakis and G.)**
  Replace $A^T A$ with $diag\{A^T A\}$
  solves it using 12-13 matrix-vector multiplications
  13 iterations, CPU time: 5079s.

Trivial problem

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

where $A \in \mathcal{R}^{m \times n}$. Highy overdetermined system: $m = 2n$.

Massive diagonal in matrix A^TA.



What is going on?

The 1st-order method (coordinate descent) uses:

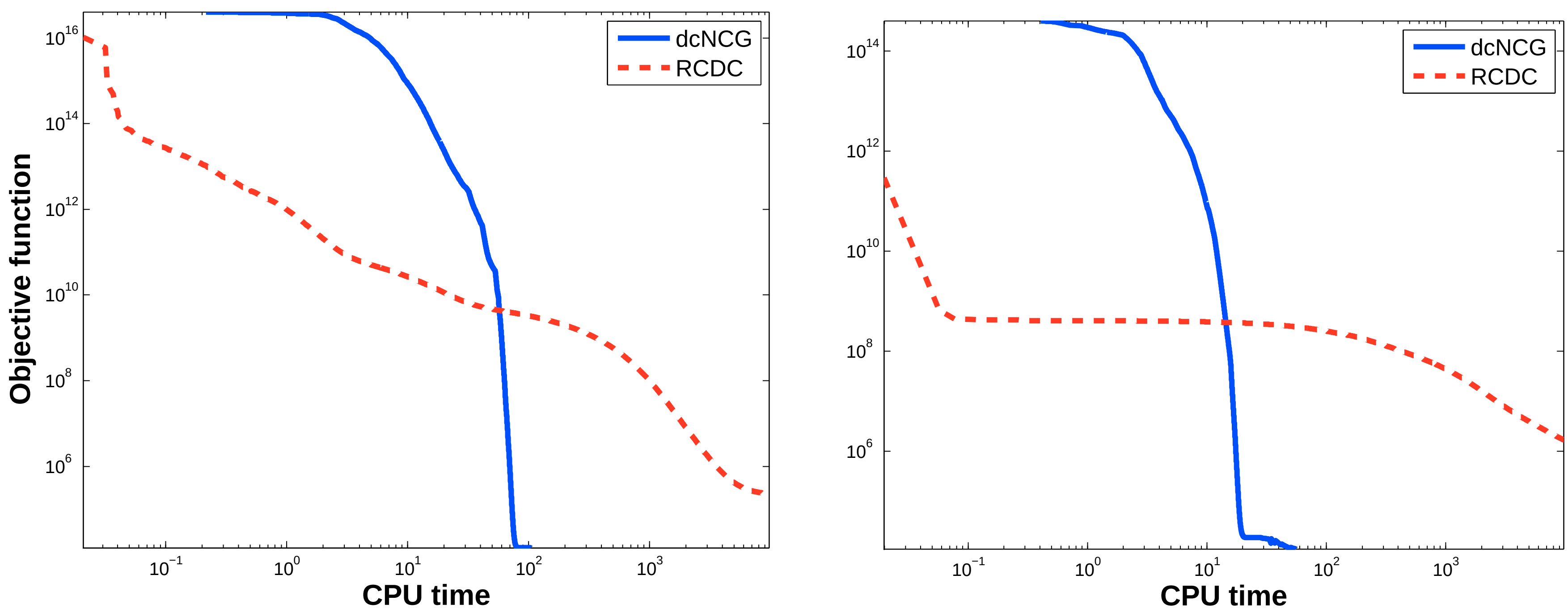
$$d_i = \text{argmin}_{p_i} \tau |x_i + p_i| + [\nabla \phi(x)]_i \, p_i + \frac{\beta}{2} p_i^T [diag(\nabla^2 \phi(x))]_{ii} \, p_i$$

If $\nabla^2 \phi(x)$ is a **diagonal** matrix (or well approximated by a diagonal matrix), then

$$\boldsymbol{d_{CD} \approx d_N}$$

hence the **1st-order** method is in fact the **2nd-order** method.

More realistic test example: RCDC vs dcNCG



Dimensions: $\boldsymbol{m = 4 \times 10^3}$, $\boldsymbol{n = 2 \times 10^3}$.

$\boldsymbol{x}^*$ has 50 non-zero elements randomly positioned.

RCDC interrupted after **10^9 iterations**, **31 hours**.

Newton-CG: Summary

Theory:

The primal-dual Newton Conjugate Gradient (pdNCG) enjoys good convergence properties.

Computational practice:

The primal-dual Newton-CG

- provides reliability
- outperforms the 1st-order methods

Software available at http://www.maths.ed.ac.uk/ERGO/
Edinburgh Research Group on Optimization

Conclusions

The **2nd-order information** can (sometimes should) be used also in very large scale optimization.

Use **inexact Newton directions** in:

- IPMs,
- primal-dual NCG.

Then the **2nd-order methods** offer an attractive alternative to the **1st-order methods**.