Preconditioners for inexact Newton method in big data optimization

Jacek Gondzio

thanks to my collaborator:

K. Fountoulakis

Outline

- Sparse Approximations $\longrightarrow$ ℓ_1-regularization
  - Machine Learning
  - Signal/Image processing problems
- Easy (?) unconstrained optimization problems
  - Near orthogonality of matrices
  - Sparsity of solution
- **2nd-order** methods for optimization
  - *Inexact* Newton method
  - *Preconditioners* needed
  - *Matrix-free* methods
- Numerical results
- Big Data problem ($\mathbf{2^{40} \approx 10^{12}}$ variables)
- Conclusions

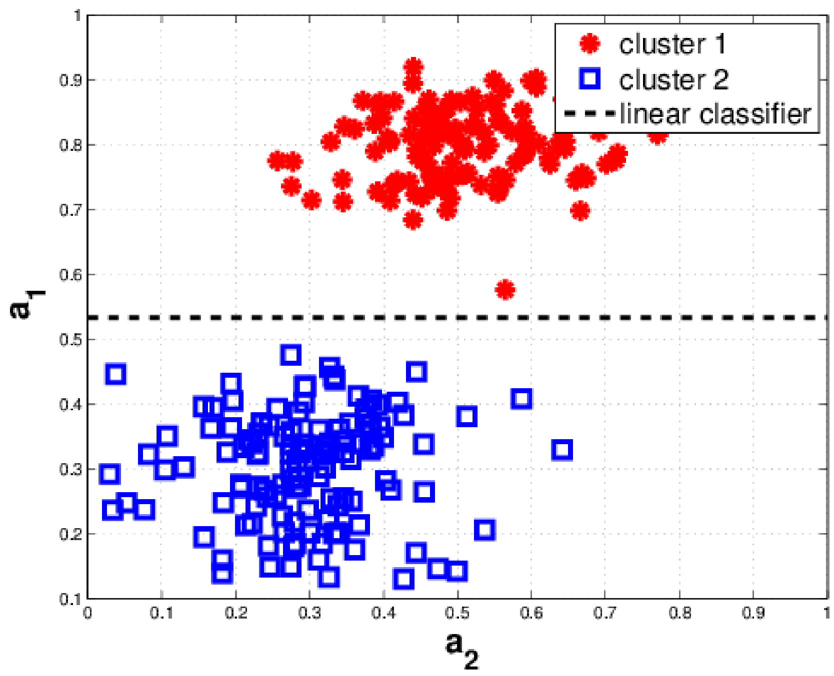
Sparse Approximation

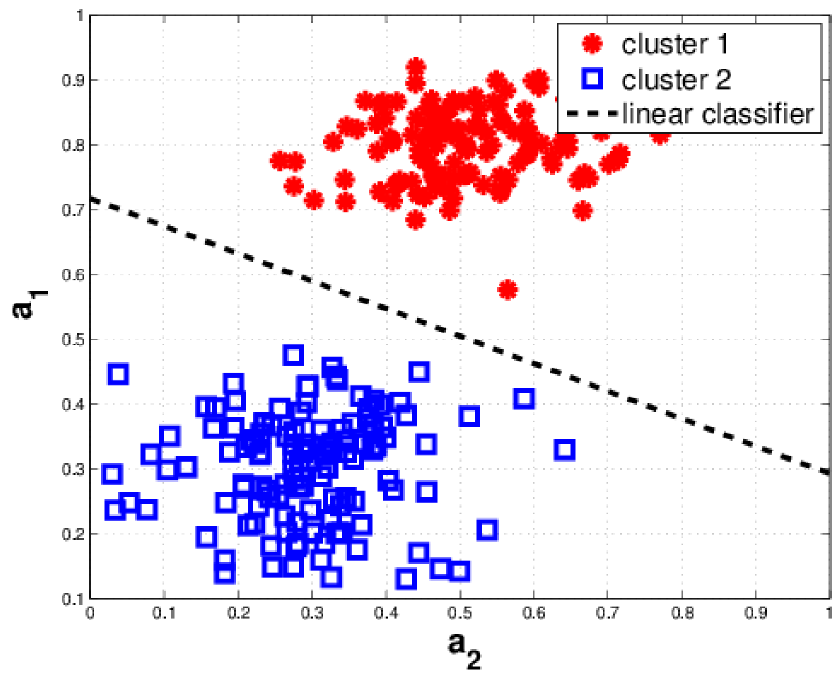
- Machine Learning: Classification with SVMs
- Statistics: Estimate x from observations
- Wavelet-based signal/image reconst. & restoration
- Compressed Sensing (Signal Processing)

All such problems lead to the same dense, potentially very large QP.

Binary Classification

$$\min \tau \|x\|_1 + \sum_{i=1}^{m} \log(1 + e^{-b_i x^T a_i}) \qquad \min \tau \|x\|_2^2 + \sum_{i=1}^{m} \log(1 + e^{-b_i x^T a_i})$$





Bayesian Statistics Viewpoint

Estimate x from observations

$$y = Ax + e,$$

where y are observations and e is the Gaussian noise.

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

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

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

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

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 (columns of W form this basis, for example wavelet basis)

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

Interesting feature

$A = RW$, where $A, R \in \mathcal{R}^{m \times n}$ and $W \in \mathcal{R}^{n \times n}$

- m may be $10^6 - 10^8$
- n may be $10^8 - 10^9$

→ no way to store A.

However, the operations

$$\begin{aligned} Ax &= R(Wx) \\ A^T y &= W^T(R^T y) \end{aligned}$$

can be executed very efficiently in many applications.

We need to solve a problem with an **implicit** A.

→ **Iterative Method** is the only hope!

ℓ_1-regularization

$$\min_x \; f(x) = \tau \|x\|_1 + \|Ax - b\|_2^2$$

Thousands of **1st-order** methods exist ...
gradient-descent, coordinate-descent
"block-, mini-batch, randomized, parallel, accelerated, (a)synchronous, proximal, robust, etc", you name it.

However, the **1st-order** methods:

- struggle with accuracy, and
- work only for trivial, well conditioned problems.

This talk will demonstrate why the **2nd-order** methods are a better option.

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 $u, v \geq 0$
- Smoothing with pseudo-Huber approximation replaces $\|x\|_1$ with $\psi_\mu(x) = \sum_{i=1}^n (\sqrt{\mu^2 + x_i^2} - \mu^2)$

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.

Questions:

- How?
- Theory?

Main Tool: Inexact Newton Method

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.

Inexact Newton Method:

- relies on **iterative** solvers, and
- needs **preconditioners**

Continuation

(three examples of ℓ_1-regularization)

Three examples of ℓ_1-regularization

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

$$\min_x \ \tau\|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2, \quad A \in \mathcal{R}^{m\times n}$$

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

$$\min_x \ \tau\|W^*x\|_1 + \frac{1}{2}\|Ax - b\|_2^2, \quad W \in \mathcal{C}^{n\times l},\ A \in \mathcal{R}^{m\times n}$$

think of Total Variation

- Big Data optimization (Machine Learning)
  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/`

Restricted Isometry Property (RIP)

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

$$AA^T = I_m.$$

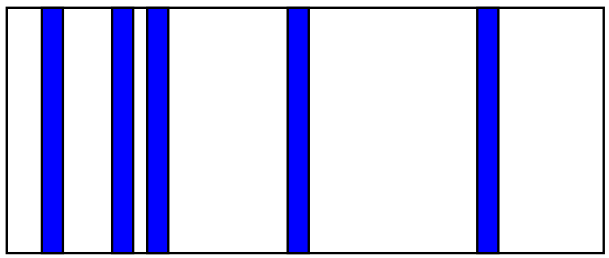
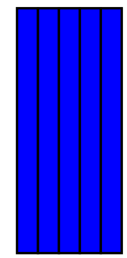
- small subsets of *columns* of A are nearly-orthogonal to each other: *Restricted Isometry Property (RIP)*

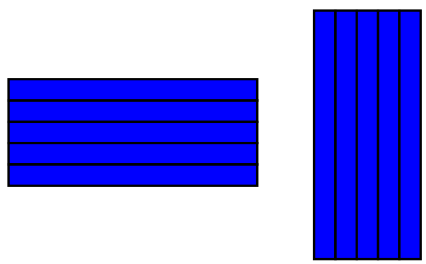
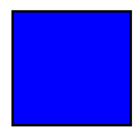
$$\|\bar{A}^T \bar{A} - \frac{m}{n} I_k\| \leq \delta_k \in (0, 1).$$

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.

Restricted Isometry Property?

Football ball

Rugby ball

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

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

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

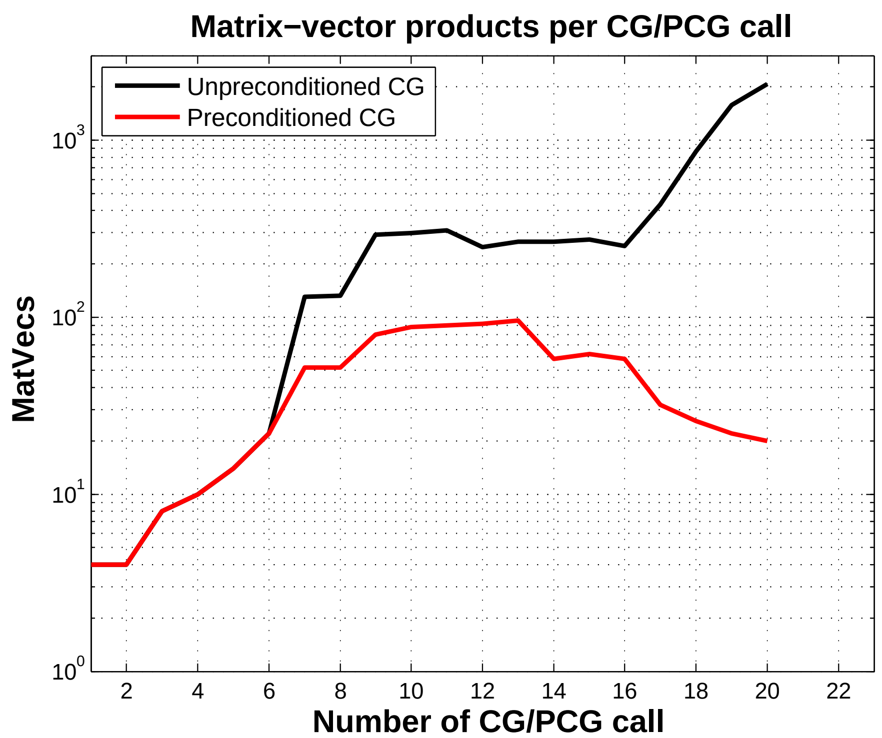
- Exactly n eigenvalues of $\mathcal{P}^{-1}\mathcal{M}$ are 1.
- The remaining n eigenvalues satisfy

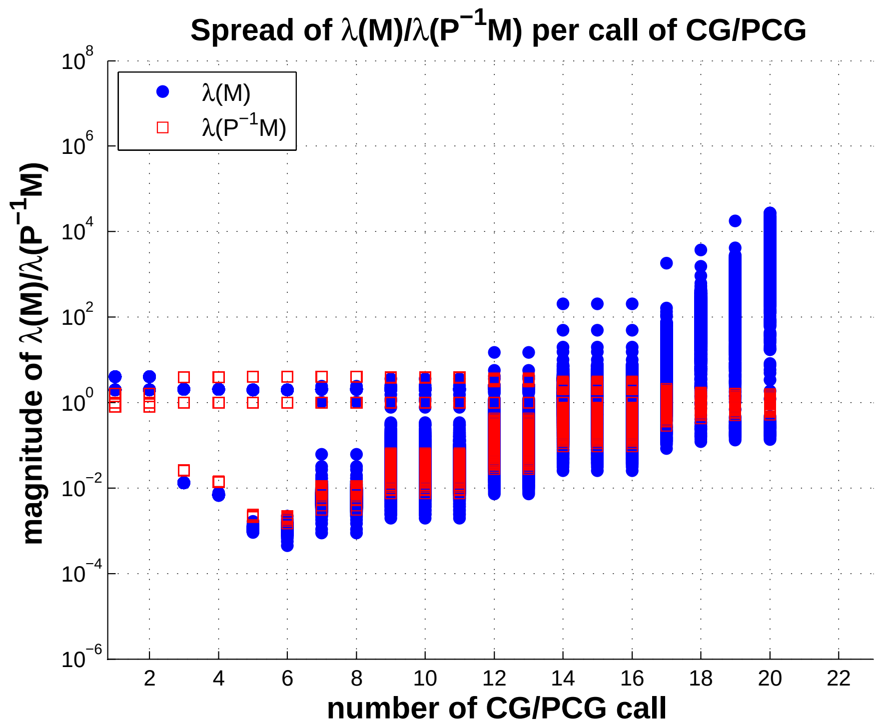
$$|\lambda(\mathcal{P}^{-1}\mathcal{M}) - 1| \leq \delta_k + \frac{n}{m\delta_k L},$$

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

Preconditioning





$\longrightarrow$ good clustering of eigenvalues

`mf-IPM` compares favourably with `NestA` on easy probs (`NestA`: Becker, Bobin and Candés).

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,
$\ell_1 - \ell_s$, *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.

ID	rhs	Accuracy	mfIPM	$\ell_1\_\ell_s$	pdco	fpc_as_cg	spgl1
2	$\tilde{b}$	3.0e-04	61	48	687	9	40000
	b	1.0e-11	65	98	40007	40002	22
3	$\tilde{b}$	7.0e-04	241	462	4941	106	40000
	b	1.0e-08	415	1612	40157	212	148
5	$\tilde{b}$	2.0e-03	5991	9842	28203	521	40000
	b	2.0e-05	7953	19684	41283	874	2567
10	$\tilde{b}$	1.0e-03	4775	8529	6203	40002	40000
	b	9.0e-10	4567	8192	41227	40161	40000
701	$\tilde{b}$	2.0e-02	947	1794	5967	1049	40000
	b	7.0e-09	1341	2656	42041	40017	15239
702	$\tilde{b}$	4.0e-03	809	1574	3341	40001	40000
	b	1.0e-07	1123	3030	49563	40157	11089

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 Preconditioner for a Primal-Dual Newton Conjugate Gradient Method for Compressed Sensing Problems, *SIAM J on Sci. Comput.* 37 (2015) A2783–A2812.

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

Theory for Continuation:

Fountoulakis and G.
A Second-order Method for Strongly Convex ℓ_1-regularization Problems, *Mathematical Programming* 156 (2016), pp. 189-219.

Computational practice:

Primal-Dual Newton Conjugate Gradients Method (pdNCG) outperforms the first-order methods.
It needs:

- **few** iterations
- with $\boldsymbol{\mathcal{O}(nz(A))}$ cost per iteration.

A better linearization

$$\tau \underbrace{Dx}_{\nabla\psi_\mu(x)} + A^T(Ax - b) = 0,$$

where $D{:=}diag(D_1,\ldots,D_n)$ with $D_i{:=}(\mu^2{+}x_i^2)^{-\frac{1}{2}} \quad \forall i{=}1,\ldots,n$

Set $g = Dx$. Use the easier form of the equations.

Difficult:

$$\tau g + A^T(Ax - b) = 0, \qquad g = Dx.$$

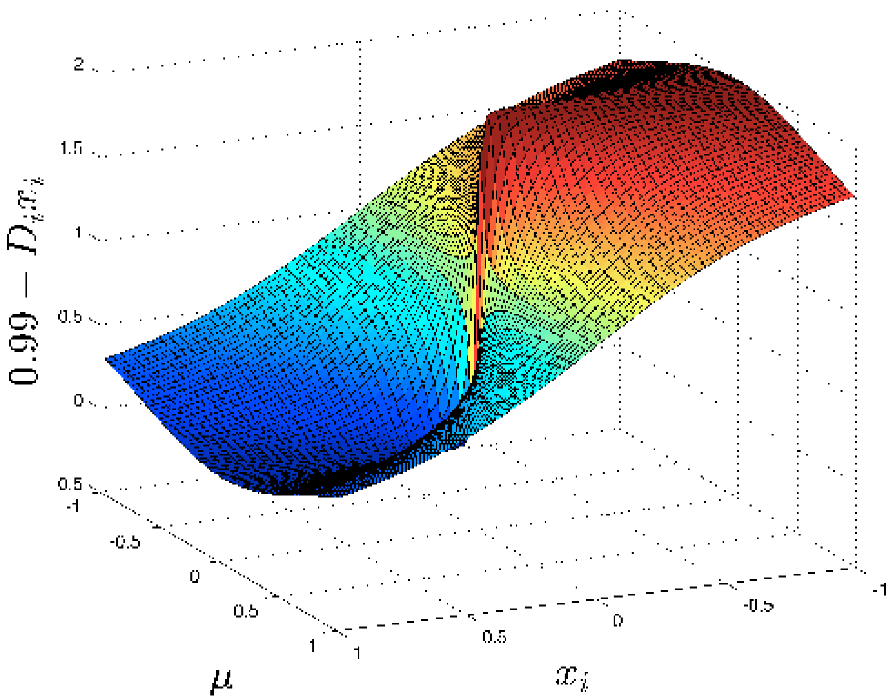
Easy:

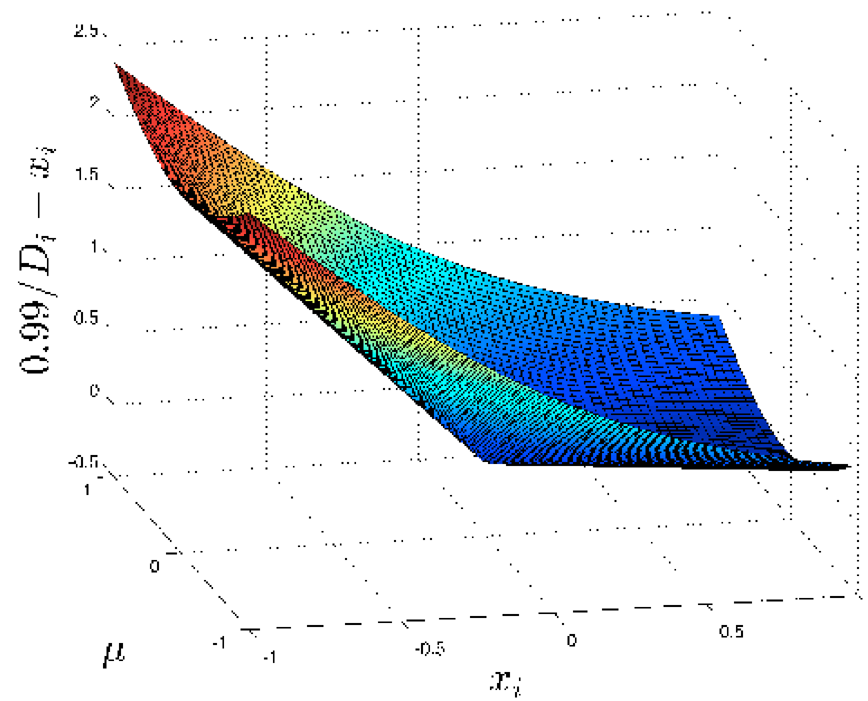
$$\tau g + A^T(Ax - b) = 0, \qquad D^{-1}g = x.$$

Chan, Golub, Mulet,
SIAM J. on Sci. Comput. 20 (1999) 1964–1977.

A better linearization

Example: $g_i = 0.99$





bad: $g_i = D_i x_i$

good: $D_i^{-1} g_i = x_i$

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 & Needell,
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

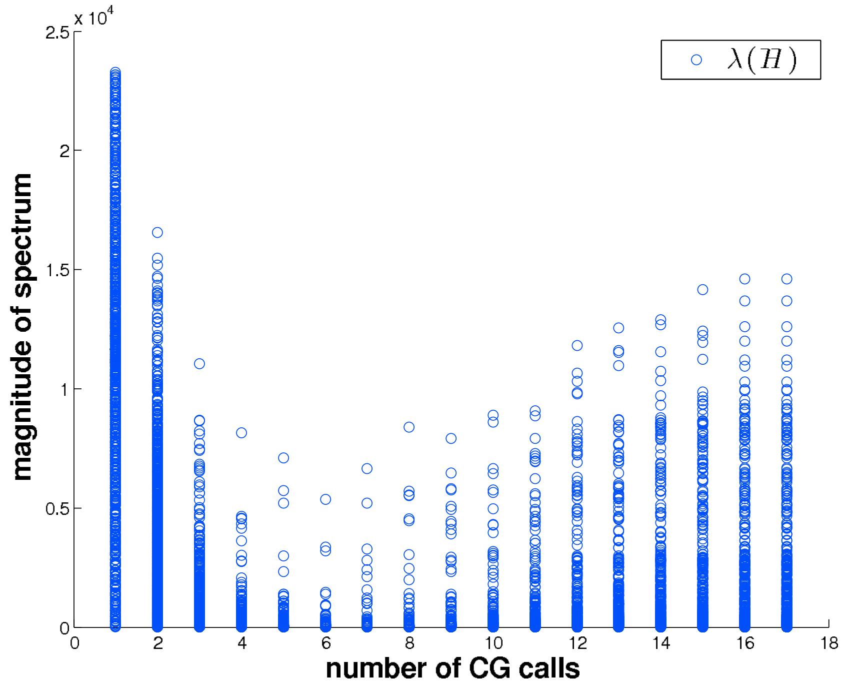
- The eigenvalues of $\mathcal{P}^{-1}\mathcal{H}$ satisfy

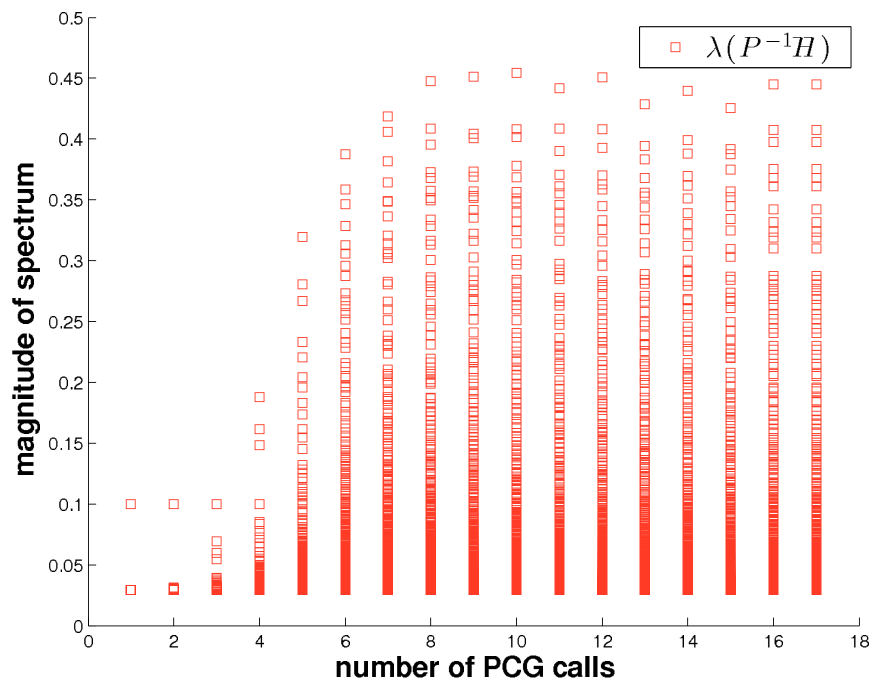
$$|\lambda(\mathcal{P}^{-1}\mathcal{H}) - 1| \leq \frac{\eta(\delta, \delta_q, \rho)}{\rho},$$

  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 Preconditioner for a Primal-Dual Newton Conjugate Gradient Method for Compressed Sensing Problems, *SIAM J on Sci. Comput.* 37 (2015) A2783–A2812.

CS: Coherent and Redundant Dictionaries





$\longrightarrow$ good clustering of eigenvalues

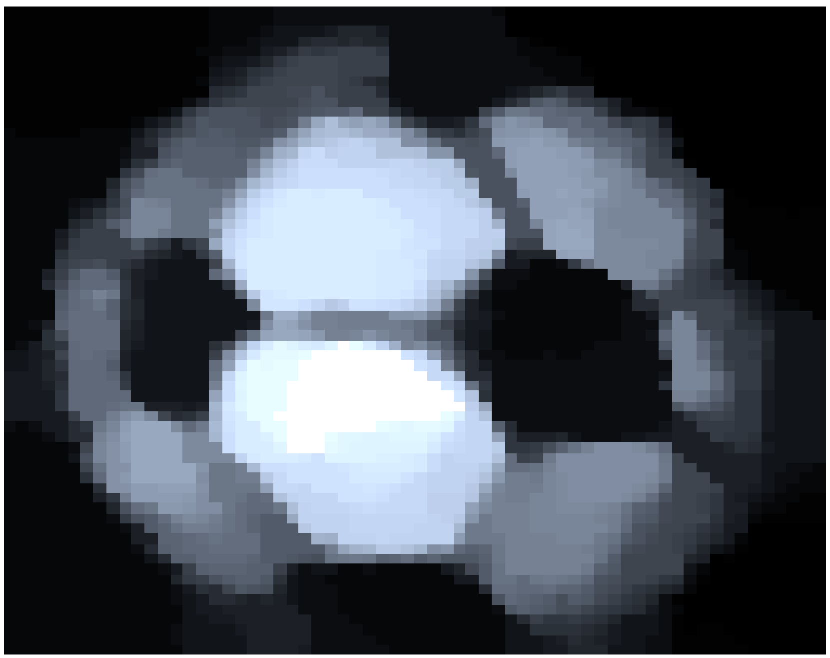
`pdNCG` outperforms `TFOCS` on several examples
(`TFOCS`: Becker, Candés and Grant).

EURO 2016 is coming!

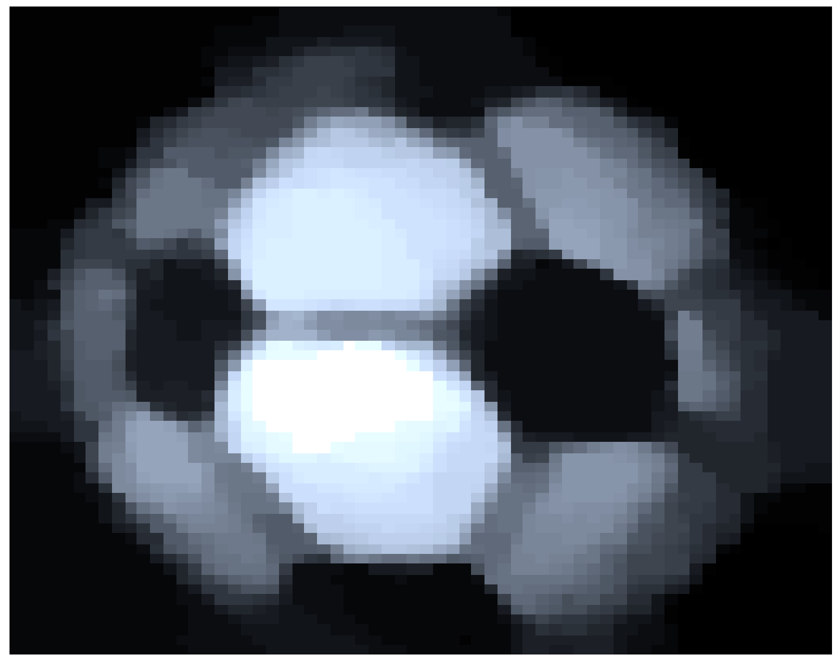
Football example: A 64×64 resolution example

Single pixel camera problem set:

`http://dsp.rice.edu/cscamera`



TFOCS, 24 sec.



pdNCG, 15 sec.

Example 3: Big Data and Optimization

with **K. Fountoulakis**.

Large dense quadratic optimization problem:

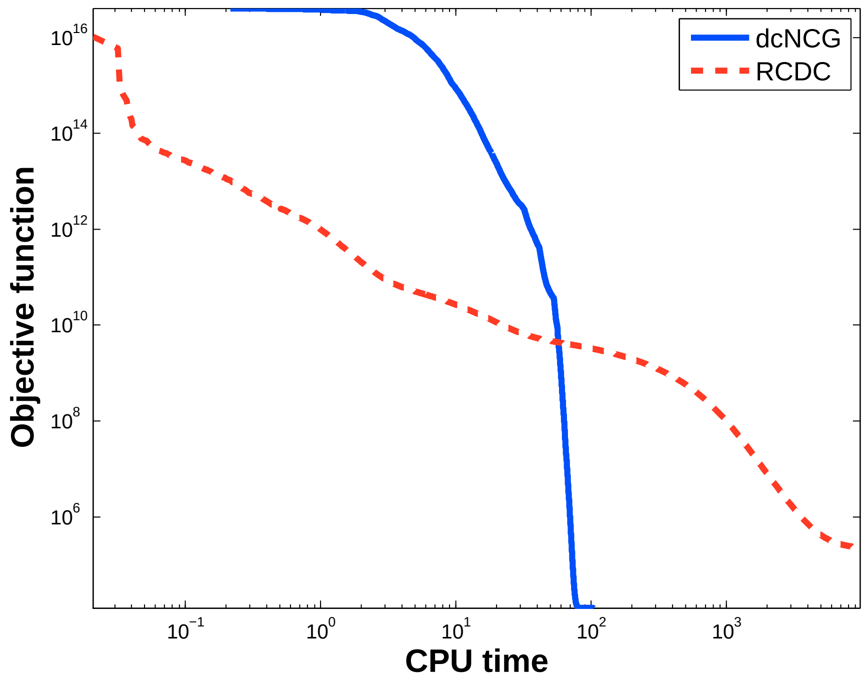
$$\min_x \ \tau\|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2,$$

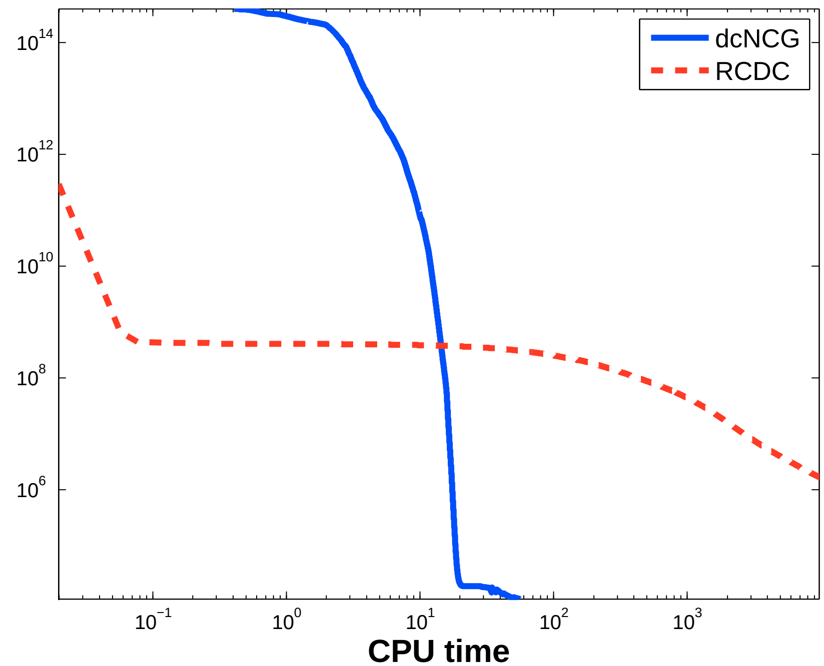
where $A \in \mathcal{R}^{m \times n}$.

Fountoulakis and G.
Performance of First- and Second-Order Methods for Big Data Optimization, *ERGO-15-005*, March 2015.

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

Baby test example: RCDC vs pdNCG



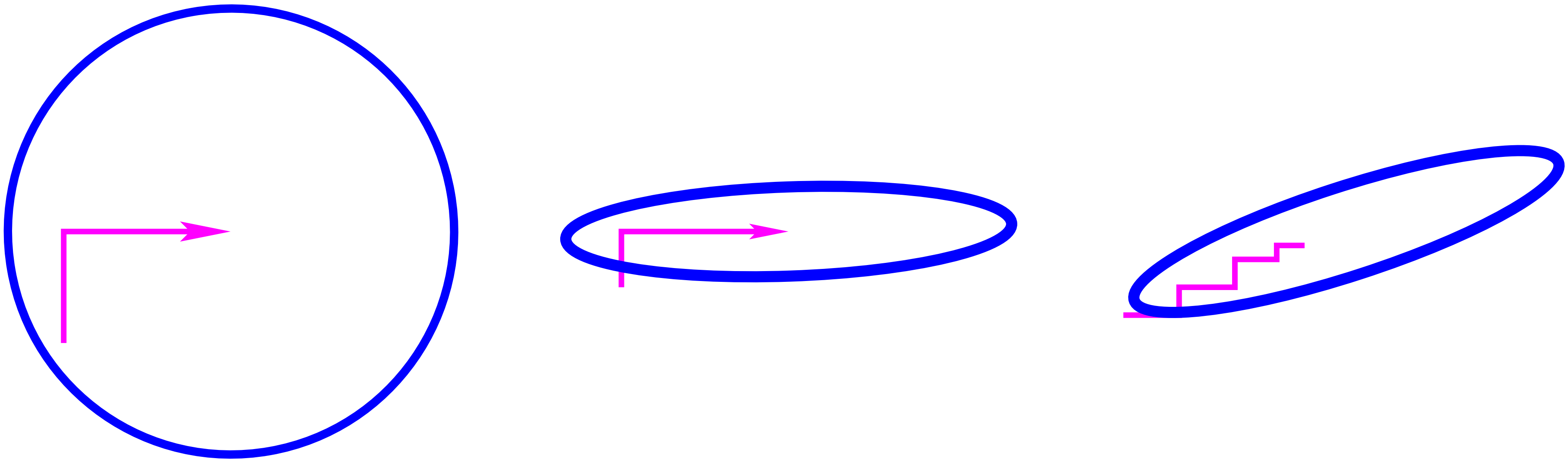


Dimensions: $\boldsymbol{m = 4 \times 10^3}$, $\boldsymbol{n = 2 \times 10^3}$.

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

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

Example: Weakness of coordinate descent



- good for well-conditioned problems with well-aligned directions,
- otherwise may be very inefficient.

The 2nd-order information is essential!

Simple example for ℓ_1-regularization

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

Special matrix given in SVD form $A = Q\Sigma G^T$.

Matlab generator:

http://www.maths.ed.ac.uk/ERGO/trillion/

The user controls:

- the condition number $\kappa(A)$,
- the sparsity of matrix A.

Simple example for ℓ_1-regularization

Suppose $m \geq n$. Write A in the SVD form:

$$A = Q \left[\frac{diag\{\sigma_1, \sigma_2, \ldots, \sigma_n\}}{0} \right] G^T,$$

where Q is an $m \times m$ orthogonal matrix,
$\sigma_1, \sigma_2, \ldots, \sigma_n$ are the singular values of A,
G is a product of Givens rotations

$$G = G(i_1, j_1, \theta_1) G(i_2, j_2, \theta_2) \ldots G(i_K, j_K, \theta_K),$$

hence it is an orthonormal matrix.
Observe that

$$A^T A = G \left[diag\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\} \right] G^T.$$

Let us go big: a trillion (2^{40}) variables

n (billions)	Processors	Memory (TB)	time (s)
1	64	0.192	1923
4	256	0.768	1968
16	1024	3.072	1986
64	4096	12.288	1970
256	16384	49.152	1990
1,024	65536	196.608	2006

ARCHER (ranked 25 on `top500.com`, 11 March 2015)
Linpack Performance (Rmax) 1,642.54 TFlop/s
Theoretical Peak (Rpeak) 2,550.53 TFlop/s

Fountoulakis and G.
Performance of First- and Second-Order Methods for Big Data Optimization, *ERGO-15-005*, March 2015.

Conclusions

2nd-order methods for optimization:

- employ **inexact Newton method**
- rely on **preconditioners**
- enjoy **matrix-free** implementation

Using the 2nd-order information:

- does not penalize the efficiency, and
- improves robustness

Simple, reliable test example for ℓ_1-regularization:
`http://www.maths.ed.ac.uk/ERGO/trillion/`

After Conclusions

What would **Shakespeare** say about Big Data?

"Much Ado About Nothing"
"Beaucoup De Bruit Pour Rien"

Convex Optimization:

- **Size** does not matter!
- **Curvature** does.