Higher-Order Methods for Large Scale Optimization

Kimon Fountoulakis
supervised by: J. Gondzio and A. Grothey

Motivation

- Make 2*nd*-order methods faster
- Capture the structure of the problem such that even ill-conditioned large-scale problems can be solved fast

Solutions in my PhD

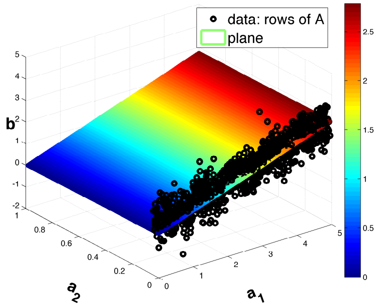
- Inexact Newton: low computational complexity per iteration (few matvecs)
    - factorization-free interior point method
    - primal-dual Newton conjugate gradients
- Provably efficient preconditioners; from $\mathcal{O}(n^3)$ to $\mathcal{O}(n)$ (in practice)
- Worst-case iteration complexity for Newton conjugate gradient methods
- A flexible problem generator and many experiments on real and synthetic data sets

Contribution I
Factorization-free primal-dual interior point method for sparse signal reconstruction

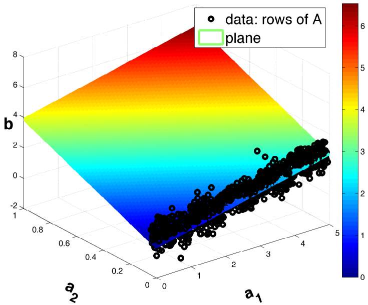
[K. F., J. Gondzio and P. Zhlobich, *Math. Prog. Computation.* 6 (1): 1-31, 2014]

Data fitting: sparse signal reconstruction

$$\text{minimize } \tau\|x\|_1 + \tfrac{1}{2}\|Ax - b\|_2^2$$



$$\text{minimize } \tfrac{1}{2}\|Ax - b\|_2^2$$



Shift to a higher dimensional space

Set $x = u - v$, where $u, v \geq 0$, which is true for

$$u_i = \max(x_i, 0), \quad v_i = \max(-x_i, 0) \quad \forall i = 1, 2, \cdots, n.$$

We thus have

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

Primal

$$\begin{array}{ll} \min\limits_{z \in \mathbb{R}^{2n}} & c^T z + \frac{1}{2} z^T F^T F z \\ \textit{subject to:} & z \geq 0 \end{array}$$

Dual

$$\begin{array}{ll} \max\limits_{z, s \in \mathbb{R}^{2n}} & -\frac{1}{2} z^T F^T F z \\ \textit{subject to:} & F^T F z - s = -c \\ & z, s \geq 0 \end{array}$$

$$F^T F = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix} \in \mathbb{R}^{2n \times 2n}, \quad c = \begin{bmatrix} \tau 1_n - A^T b \\ \tau 1_n + A^T b \end{bmatrix} \in \mathbb{R}^{2n}$$

Linear systems and properties

At every iteration we solve inexactly the following system using PCG:

$$(\underbrace{\Theta^{-1}}_{\text{diagonal}} + F^{\intercal}F) \times \Delta x = *, \quad F^{T}F = \begin{bmatrix} A^{T}A & -A^{T}A \\ -A^{T}A & A^{T}A \end{bmatrix}, \quad A \in \mathbb{R}^{m \times n}, \;\; m \ll n$$

- Rows of A are nearly orthogonal, i.e. $\|AA^{\intercal} - I_n\|_2 \leq \delta$, where δ is small
- Subsets of columns of A are nearly orthogonal

$A =$ 

- For at most q-sparse optimal solutions we have that q entries of $\Theta^{-1} \to 0$, $2n - q$ entries of $\Theta^{-1} \to \infty$

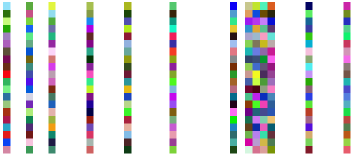
Linear systems and properties

At every iteration we solve inexactly the following system using PCG:

$$(\underbrace{\Theta^{-1}}_{\text{diagonal}} + F^\intercal F) \times \Delta x = *, \quad F^T F = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix}, \quad A \in \mathbb{R}^{m \times n}, \ m \ll n$$

- Rows of A are nearly orthogonal, i.e. $\|AA^\intercal - I_n\|_2 \leq \delta$, where δ is small
- Subsets of columns of A are nearly orthogonal

$$\bar{A} =$$



- For at most q-sparse optimal solutions we have that q entries of $\Theta^{-1} \to 0$, $2n - q$ entries of $\Theta^{-1} \to \infty$

Linear systems and properties

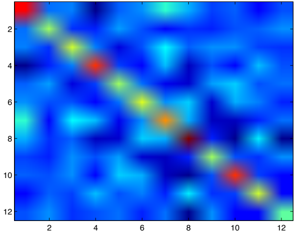
At every iteration we solve inexactly the following system using PCG:

$$(\underbrace{\Theta^{-1}}_{\text{diagonal}} + F^{\mathsf{T}}F) \times \Delta x = *, \quad F^T F = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix}, \;\; A \in \mathbb{R}^{m \times n}, \; m \ll n$$

- Rows of A are nearly orthogonal, i.e. $\|AA^{\mathsf{T}} - I_n\|_2 \leq \delta$, where δ is small

- Subsets of columns of A are nearly orthogonal

$$\bar{A}^T \bar{A} =$$



- For at most q-sparse optimal solutions we have that q entries of $\Theta^{-1} \rightarrow 0$, $2n - q$ entries of $\Theta^{-1} \rightarrow \infty$

Linear systems and properties

At every iteration we solve inexactly using PCG the system:

$$(\underbrace{\Theta^{-1}}_{\text{diagonal}} + F^\intercal F) \times \Delta x = *, \quad F^T F = \begin{bmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{bmatrix}, \;\; A \in \mathbb{R}^{m \times n}, \;\; m \ll n$$

- Rows of A are nearly orthogonal, i.e. $\|AA^\intercal - I_n\|_2 \leq \delta$, where δ is small.

- There exists $\delta_q < 1/2$ such that Restricted Isometry Property (RIP) holds:

$$(1 - \delta_q)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_q)\|x\|_2^2,$$

  for all at most q-sparse $x \in R^n$.

- For at most q-sparse optimal solutions we have that q entries of $\Theta^{-1} \to 0$, $2n - q$ entries of $\Theta^{-1} \to \infty$

Preconditioner

Approximate: $$\mathsf{M} = \Theta^{-1} + \begin{bmatrix} A^TA & -A^TA \\ -A^TA & A^TA \end{bmatrix}$$

with: $$\mathsf{P} = \Theta^{-1} + \rho \begin{bmatrix} \mathsf{I} & -\mathsf{I} \\ -\mathsf{I} & \mathsf{I} \end{bmatrix}, \quad \rho = m/n$$

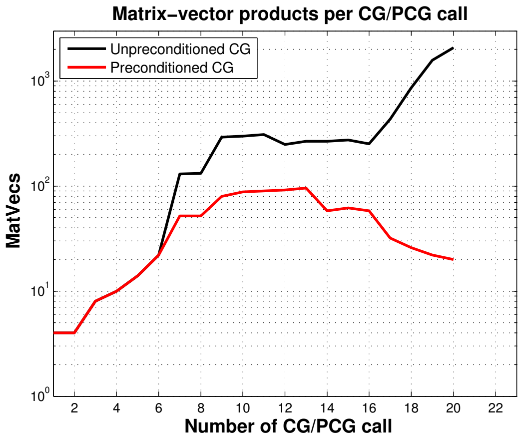
Systems can be solved with P in $\mathcal{O}(n)$ time!

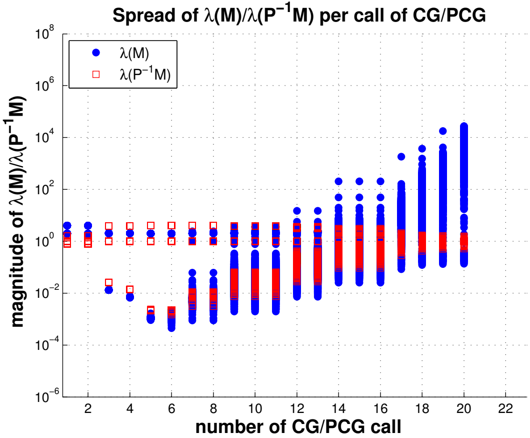
Theorem (Brief description).

- *Exactly n eigenvalues of $P^{-1}M$ are 1.*
- *The remaining n satisfy $|\lambda(P^{-1}M) - 1| \leq \delta_q + \frac{n}{m\delta_q L}$*

where $L \to \infty$ and δ_q is the RIP-constant.

Practical performance





Required accuracy: 1.0e-6

Experiments: sparco test suite by M. P. Friedlander et al.

ID	rhs	Accuracy	mfipm	ℓ_1_ℓ_s	pdco	fpc_as cg	spgl1
2	$\tilde{b}$ (noisy)	3.0e-04	61	48	687	9	40000
	b (noiseless)	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
7	$\tilde{b}$	4.0e-03	179	272	425	62	39
	b	1.0e-06	255	850	601	76	81
9	$\tilde{b}$	1.0e-03	689	1546	7065	1680	40000
	b	5.0e-12	649	1886	6845	40016	40000
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

Contribution II
Theoretical analysis of primal-dual Newton Conjugate Gradients

[K. F. and J. Gondzio, *Math. Prog. A.* DOI: 10.1007/s10107-015-0875-4]

General fidelity term

$$\text{minimize } f_\tau(x) := \tau\|x\|_1 + \varphi(x)$$

- $x \in \mathbb{R}^n$, $\varphi(x) : \mathbb{R}^n \to \mathbb{R}$, $\tau > 0$

Assumptions

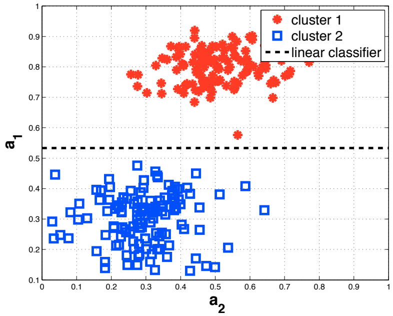
- φ is a smooth and strongly convex function

Plenty of data

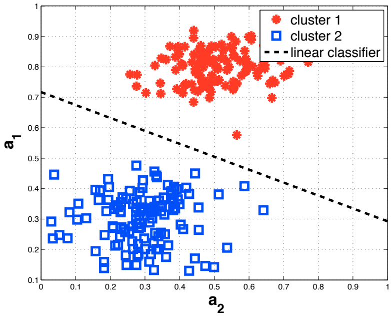
- n is very large. i.e. of order millions or billions

Binary classification

$$\text{minimize } \tau\|x\|_1 + \textstyle\sum_{i=1}^m \log(1 + e^{-b_i x^\intercal a_i})$$



$$\text{minimize } \tau\|x\|_2^2 + \textstyle\sum_{i=1}^m \log(1 + e^{-b_i x^\intercal a_i})$$



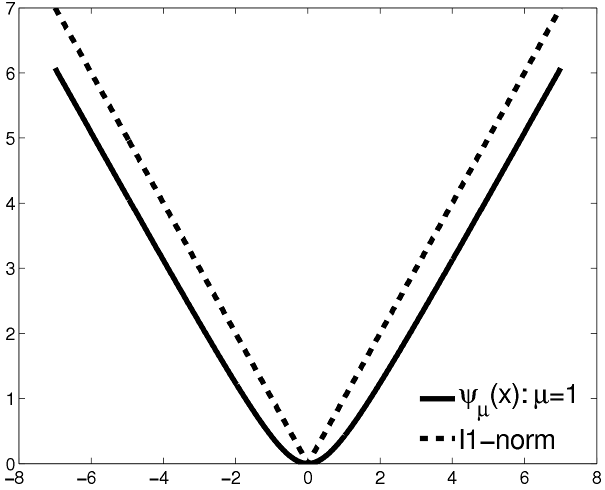
Smoothing in pdNCG: Moreau-Yosida

Replace $\|x\|_1 = \sup_{\|g\|_\infty \leq 1} g^\intercal x$

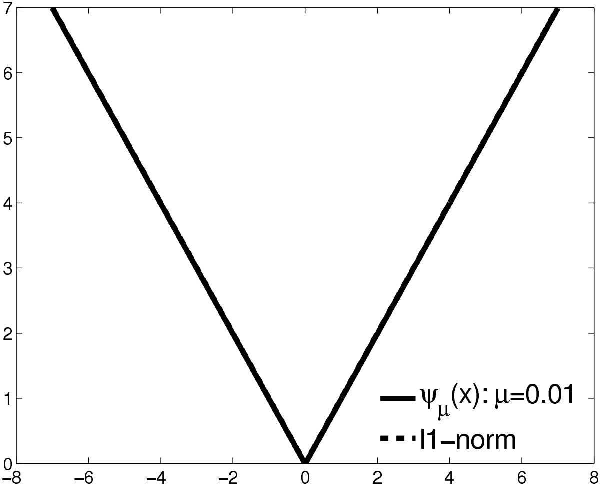
with $\psi_\mu(x) = \sup_{\|g\|_\infty \leq 1} g^\intercal x - \mu d(g)$, $\quad \underbrace{d(g) = n - \sum_{i=1}^{n} (1 - g_i^2)^{1/2}}_{\text{proximity function on } \|g\|_\infty \leq 1}$

$$\text{Pseudo-Huber: } \psi_\mu(x) = \sum_{i=1}^{n} \left(\sqrt{\mu^2 + x_i^2} - \mu \right)$$

Smoothing in pdNCG: Moreau-Yosida



Smoothing in pdNCG: Moreau-Yosida



A better linearization

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

where $D := diag(D_1, D_2, \cdots, D_n)$ with

$$D_i := (\mu^2 + x_i^2)^{-\frac{1}{2}} \quad \forall i = 1, 2, \cdots, n$$

Set $g = Dx$ and linearise the blue instead of the red equations.

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

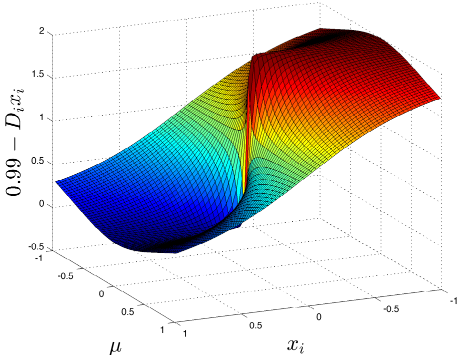
Essentially we are solving the primal-dual problem:

$$\min_x \ \frac{1}{2}\|Ax-b\|_2^2 + \tau \sup_{\|g\|_\infty \le 1} g^\intercal x - \mu d(g)$$

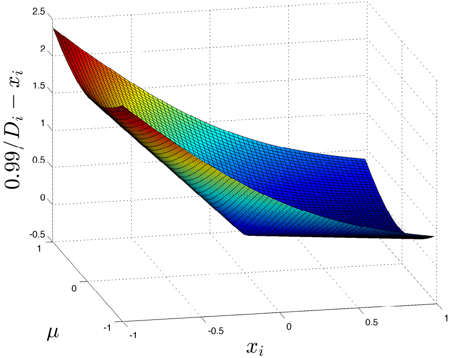
[Chan, Golub, Mulet, *SIAM. J. Sci. Comput.* 20 (6) 1999 pp. 1964-1977]

A better linearisation

Example: $g_i = 0.99$



$g_i = D_i x_i$



$D_i^{-1} g_i = x_i$

Newton-type directions

Linearisation of the optimality conditions reduces to

$$B(x,g)\Delta x = -\nabla f_\tau^\mu(x) \quad \text{where} \quad B := \tau\nabla^2\psi(x,g) + A^\mathsf{T}A. \tag{1}$$

Δg is inexpensive to calculate

- $B \succ 0$ if $\|g\|_\infty \leq 1$
- Use PCG to solve (1) approximately

Primal-dual Newton Conjugate Gradient (pdNCG)

Input: x^0, g^0, where $\|g^0\|_\infty \leq 1$

Loop: For $k = 1, 2, \ldots$, until termination criteria are met

  Calculate primal-dual directions Δx^k, Δg^k *approximately* with PCG

  $\mathbf{g^{k+1}} := \mathbf{P}_{\|\cdot\|_\infty \leq \mathbf{1}}(\mathbf{g^k} + \mathbf{\Delta g^k})$, $P_{\|\cdot\|_\infty \leq 1}(\cdot)$ is the projection on the ℓ_∞ ball

  Perform backtracking line search for the direction Δx^k

  Set $x^{k+1} := x^k + \alpha \Delta x^k$

[Chan, Golub, Mulet, *SIAM. J. Sci. Comput.* 20 (6) 1999 pp. 1964-1977]

pdNCG: convergence

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 (Dual convergence). *The sequences of dual variables generated by pdNCG satisfy* $\{g^k\}_{k=0}^{\infty} \to \nabla \psi_\mu(x_{\tau,\mu})$*.*

Lemma (Convergence of approximate Hessian). *Let the sequences* $\{x^k\}_{k=0}^{\infty}$ *and* $\{g^k\}_{k=0}^{\infty}$ *be generated by pdNCG. Then* $B(x^k, g^k) \to \nabla^2 f_\tau^\mu(x_{\tau,\mu})$*.*

[K. F. and J. Gondzio, *Math. Prog. A.* DOI: 10.1007/s10107-015-0875-4]

pdNCG: worst case iteration complexity

pdNCG needs at most

$$\mathcal{O}\left(\frac{\kappa^2}{(1-\eta^2)^2}\right) + \log_2 \log_2 \frac{\mathit{const.}}{\epsilon}$$

iterations to converge to a solution x^k of accuracy

$$f(x^k) - f^* \leq \epsilon.$$

Damped Newton, see S. Boyd and L. Vandenberghe, *Convex Optimization*

$$\mathcal{O}(\kappa^2) + \log_2 \log_2 \frac{\mathit{const.}}{\epsilon}.$$

[K. F. and J. Gondzio, *Math. Prog. A.* DOI: 10.1007/s10107-015-0875-4]

Contribution III
A preconditioner for primal-dual Newton Conjugate Gradients for sparse signal reconstruction

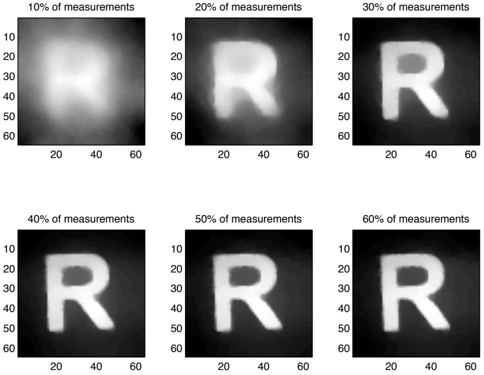
[I. Dassios, K. F. and J. Gondzio, *Technical Report ERGO-14-021*]

(2nd round of revisions SIAM Scientific Computing)

Data fitting: non-separable regularizers

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

where W^* is *tridiagonal*; a discretization of the ∇ operator for images



Preconditioner and spectral properties

Approximate: $\bar{B} := \tau/2(\tilde{B} + \tilde{B}^{\intercal}) + A^{\intercal}A$,

with: $N := \tau/2 \underbrace{(\tilde{B} + \tilde{B}^{\intercal})}_{5-diagonal} + \rho I_m$, where $\rho \in [\delta_q, 1/2]$ and $\delta_q < 1/2$.

Assumptions

- Rows of A are nearly orthogonal, i.e. $\|AA^{\intercal} - I_n\|_2 \leq \delta$, where δ is small.
- There exists $\delta_q < 1/2$ such that Restricted Isometry Property (W-RIP) holds:

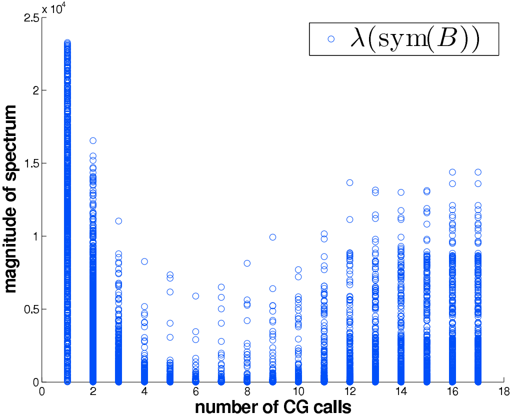
$$(1 - \delta_q)\|Wz\|_2^2 \leq \|AWz\|_2^2 \leq (1 + \delta_q)\|Wz\|_2^2,$$

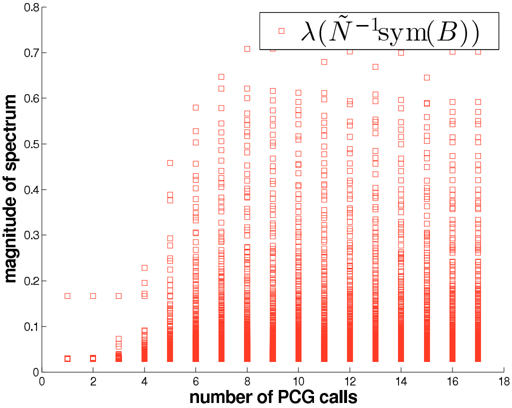
for all at most q-sparse $z \in E^l$. *In subspaces defined by any at most q columns of W matrix* $A^{\intercal}A$ behaves like a scaled identity.

Theorem (Brief description). *Let* $\lambda \in spec(N^{-1}\bar{B})$*, then close to the solution the following holds*

- $|\lambda - 1| \leq \frac{1}{2}(\chi + 1 + (5\chi^2 - 2\chi + 1)^{\frac{1}{2}})\mathcal{O}(\mu)$, *where* $\chi := 1 + \delta - \rho$, *and* $\mu \approx 0$

Spectrum in practice ($\mu = 1.0e$-5)



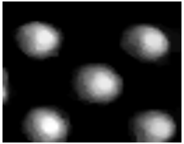


Dependence of pdNCG on problem size

The image Shepp-Logan has been used for this experiment, 25% of the measurements are used and SNR is fixed to 15 dB.

Solver	64×64	128×128	256×256	512×512	1024×1024
TFOCS (Candès et al.)	17	23	56	260	1018
TVAL3 (Wotao et al.)	5	8	37	99	365
pdNCG	**2**	**6**	**12**	**62**	**250**

Single pixel camera benchmarks



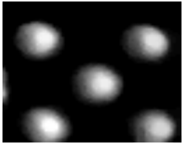
TFOCS, 25 sec.

24 sec.

37 sec.

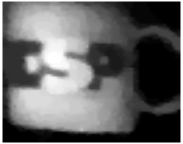
26 sec.

49 sec.



pdNCG, 7 sec.

15 sec.



15 sec.

27 sec.

33 sec.

Contribution IV
A problem generator for ℓ_1-regularized least squares

[K. F. and J. Gondzio, *Technical Report ERGO-15-005* (submitted)]

Motivation

Issue: frequently, the performance of new methods is tested on well-conditioned randomly generated problems

Need: controlled testing – a problem generator which can reveal weaknesses and strengths of new methods

A problem generator

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

$\tau > 0$, $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$

- The generator is inexpensive and has a
- low-memory-footprint

The generator allows control of the

- *dimensions m, n*
- *sparsity and the values of the optimal solution x^**
- *sparsity of A and $A^\intercal A$ (independently of A)*
- *singular value decomposition of A*

A trillion variable problem

ARCHER

- 25th fastest supercomputer worldwide out of 500 supercomputers (based on TOP500 commercial supercomputers list)
- 118,080 cores, we used 65,536

n	processors	terabytes	seconds
$2^{36} \approx 68$ billion	4096	12.288	1970
2^{38}	16384	49.152	1990
$2^{40} \approx 1$ trillion	65536	196.608	2006

All problems have been solved to a relative error of order 10^{-4} using pdNCG

Givens rotation

$G(i,j,\theta) \in \mathbb{R}^{n\times n}$, which rotates plane i-j by an angle θ:

$$G(i,j,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix},$$

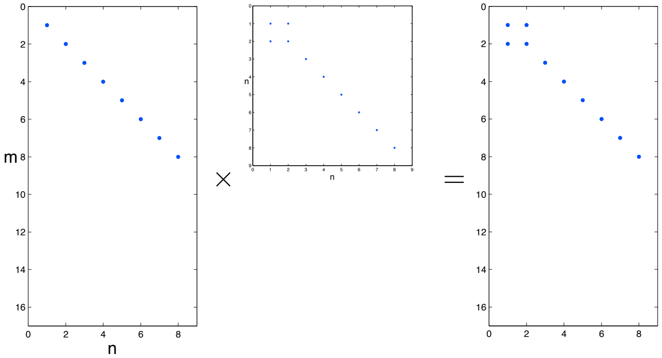
where $i,j \in \{1,2,\cdots,n\}$, $c = \cos\theta$ and $s = \sin\theta$.

- Memory requirements: coordinates i,j and a 2×2 matrix

An example: using Givens rotations

Fix an angle θ.

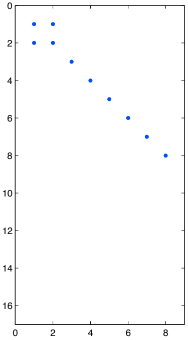
$$\Sigma G(1,2,\theta)^{\mathsf{T}} = A_1$$



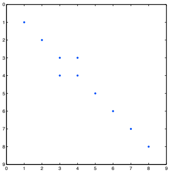
An example: using Givens rotations

Fix an angle θ.

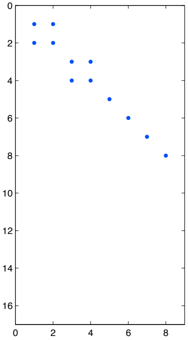
$$A_1 G(3,4,\theta)^{\mathsf{T}} = A_2$$



$\times$



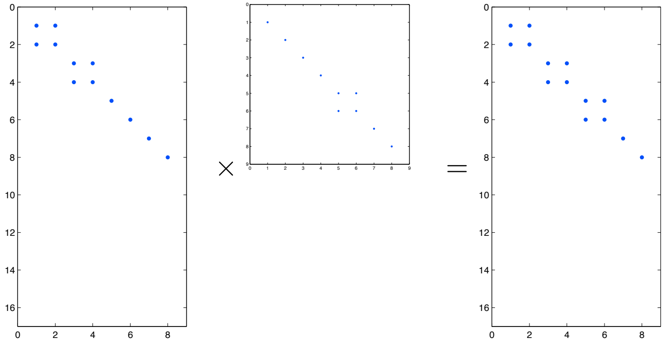
$=$



An example: using Givens rotations

Fix an angle θ.

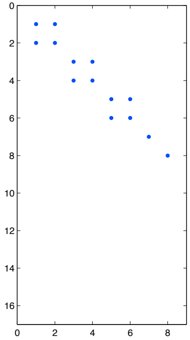
$$A_2 G(5, 6, \theta)^{\mathsf{T}} = A_3$$



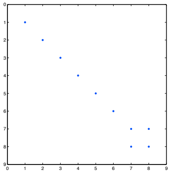
An example: using Givens rotations

Fix an angle θ.

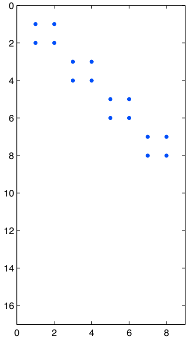
$$A_3 G(7, 8, \theta)^\mathsf{T} = A_4$$



$\times$

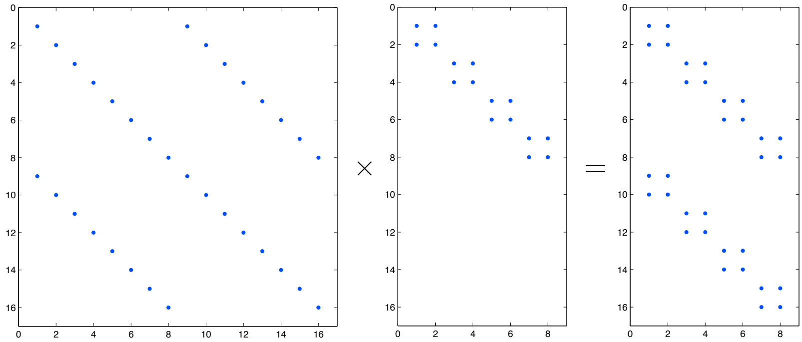


$=$



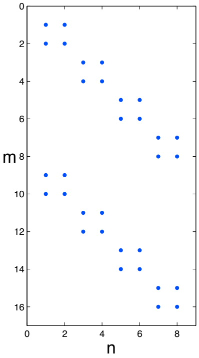
An example: using Givens rotations

$$\tilde{G}_{5:12}(\theta)A_4 = A_{12}$$

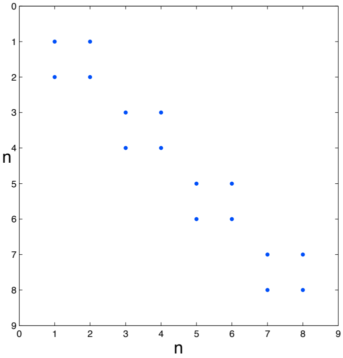


- $A_{12} = \tilde{G}_{5:12}(\theta)\Sigma(G(7,8,\theta)G(5,6,\theta)G(3,4,\theta)G(1,2,\theta))^\intercal$
- We only need to store a 2×2 matrix
- Matrix-vector products can be computed fast using a simple algorithmic process

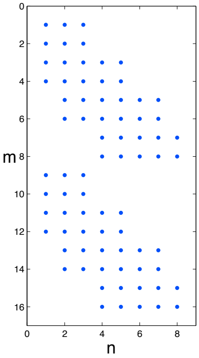
Control of sparsity using Givens rotations



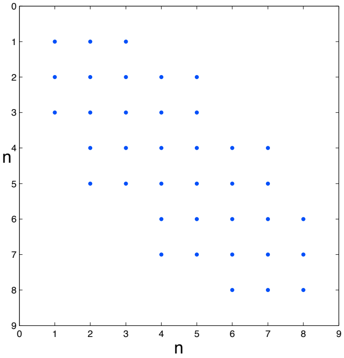
A



$A^\intercal A$



A



$A^\intercal A$

Thank you!