On the solution of nonnegative least-squares problems by interior point Newton-like methods

Stefania Bellavia, Maria Macconi, Benedetta Morini

Dipartimento di Energetica "S.Stecco", Università degli Studi di Firenze

NAday, Bologna September 18, 2006

Non-Negative Least-Squares (NNLS) problems

$$\min_{x \geq 0} q(x) = \frac{1}{2}\|Ax - b\|_2^2,$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ are given and $m \geq n$.

Assume n is large, A is sparse

- NNLS problems model:
  data fitting with variables in meaningful intervals; nonnegative image restoration problems; contact problems for mechanical systems; control problems.

- A has full column rank $\Rightarrow$ there is an unique solution x^*. Allow x^* to be degenerate:

$$g(x) = \nabla q(x), \qquad x_i^* = 0, \quad g_i(x^*) = 0 \quad \text{for some} \quad i, \quad 1 \leq i \leq n$$

Interior Point methods for NNLS problems

- IP methods for large bound-constrained quadratic programming and bound constrained least-squares problems:
  - First order optimality conditions are reformulated as a nonlinear system.
  - A sequence of strictly feasible iterates is generated by a Newton like method. Global convergence is guaranteed.

    [Coleman, Hulbert, 1993],
    [Coleman, Li, SIOPT 1996]
    [Coleman, Li, COAP 2000],
    [Portugal, Judice and Vicente, Math. Comput. 1994].
- Procedures tailored for ill-posed problems arising in image reconstruction

  e.g. [Calvetti, Landi, Reichel, Sgallari, Inv. Problems, 2004] [Rojas, Steihaugh, Inv. Problems, 2002].

To our knowledge all Interior Point methods for quadratic programming and NNLS problems do not ensure fast convergence in presence of degeneracy.

$$\Downarrow$$

We propose new Interior Point methods:

- Fast convergence in presence of degeneracy.
- The structure of the NNLS problems is exploited:

  in the linear algebra phase;

  in the globalization strategy.
- The method is tailored for general NNLS problems. It is not be suited for handling ill-posed problems.

Basic transformation of NNLS problems

Let $g(x) = \nabla q(x) = A^T(Ax - b)$. The first order optimality conditions are:

$$x^* \geq 0, \qquad g(x^*) \geq 0, \qquad g(x^*)^T x^* = 0,$$

$$\Downarrow$$

- x^* solves the system of nonlinear equations:

$$D(x)g(x) = 0,$$

where $D(x) = diag(d_1(x), \ldots, d_n(x))$, $\quad d_i(x) = \begin{cases} x_i & \text{if} \quad g_i(x) \geq 0, \\ 1 & \text{otherwise .} \end{cases}$

- $D(x)g(x)$ is continuous for $x \geq 0$.
- $D(x)$ is invertible for $x > 0$.

[Coleman and Li, SIOPT 1996]

Interior Point methods for $\min_{l \le x \le u} f(x)$.

- Given $x_k > 0$, the Newton-like system for $D(x)g(x) = D(x)(A^T(Ax-b)) = 0$ is

$$\boxed{(D_k\, A^T A + E_k)p = -D_k g_k,}$$

$g_k = g(x_k),\ D_k = D(x_k),\ E_k = E(x_k)$ and

$$E(x) = diag(e_1(x), \ldots, e_n(x)), \qquad e_i(x) = \begin{cases} g_i(x) & \text{if } g_i(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

- To remain strictly feasible, the step p is possibly truncated.

- The method is locally fast convergent if x^* is nondegenerate.
- Approaching a degenerate solution x^*, the coefficient matrix $D_k\, A^T A + E_k$ tends to become singular.

$$\text{If} \quad \lim_{k\to\infty} x_k = x^*, \quad \text{and} \quad x_i^* = 0,\ g_i(x^*) = 0 \quad \text{for some} \quad i,\ 1 \le i \le n$$

$$\Downarrow$$

$$\lim_{k\to\infty} d_i(x_k) = \lim_{k\to\infty} e_i(x_k) = 0 \quad \text{i.e.} \quad \lim_{k\to\infty} \|(D_k\, A^T A + E_k)^{-1}\| = \infty.$$

- [M. Heinkenschloss, M. Ulbrich, S.Ulbrich, Math. Program. 1999]
  New Interior Points methods for $\min_{l\le x\le u} f(x)$.

1) $W(x) = diag(w_1(x), \ldots, w_n(x)), \qquad w_i(x) = \dfrac{1}{d_i(x) + e_i(x)}, \qquad x > 0$

$$W_k(D_k\, A^T A + E_k)p = -W_k D_k g_k,$$

The matrix $W(x)(D(x)A^T A + E(x))^{-1}$ exists and is uniformly bounded for $x \geq 0$.

2) $E(x) = diag(e_1(x), .., e_n(x)), \quad e_i(x) = \begin{cases} g_i(x) & \text{if } 0 \leq g_i(x) < x_i^2 \\ & \text{or } \quad x_i < g_i(x)^2 \\ 0 & \text{otherwise} \end{cases}$

Around x^*, the modification introduced affects $E(x)$ only in the presence of degeneracy and allows to develop fast convergent methods.

3) To remain strictly feasible, the step p_k used is

$$p_k = \max\{\sigma,\ 1 - \|P(x_k + p) - x_k\|_2\}\,(P(x_k + p) - x_k), \qquad \sigma < 1,$$

where $P(x) = \max\{0, x\}$ with $\max$ meant componentwise.

- The sequence $\{x_k\}$ generated by the Newton-like method converges locally and quadratically toward x^* in the presence of degeneracy too.

- However, the straightforward application of this method to NNLS problems may be inappropriate:

  – Linear algebra issues: a proper implementation of the method.

  – Globalization strategy: convergence does not depend critically on the initial guess.

Linear algebra issues

$$W_k(D_k\, A^T A + E_k)p = -W_k D_k g_k,$$

$\nabla^2 q(x) = A^T A$ symmetric positive definite (s.d.p.);
$W_k(D_k\, A^T A + E_k)$ is nonsymmetric.

- Form $A^T A$, apply direct methods for large nonsymmetric systems.

  Prohibitively costly in terms of storage and operations if A is large and sparse and $A^T A$ is almost dense.

- Apply iterative methods, the action of A and A^T on vectors are needed.
  - Short recurrences method (QMR, BI-CG, BI-CGSTAB) are not optimal in the sense of error or residual minimization.
  - GMRES minimizes the residual norm but requires full-term recurrence; a restarted version may stagnate.

Formulation of the Newton system as a s.d.p. system

$$W_k(D_k\, A^T A + E_k)p = -W_k D_k g_k,$$

$$\Downarrow$$

$$\underbrace{W_k(D_k\, A^T A D_k + E_k D_k)W_k}_{\text{s.d.p. matrix}}\ \tilde{p} = -W_k D_k g_k, \qquad \tilde{p} = D_k^{-1} W_k^{-1} p$$

- Approaching a solution x^* where a component is active and nondegerate, the coefficient matrix $W_k(D_k\, A^T A D_k + E_k D_k)W_k$ tends to become singular:

$$\lim_{k\to\infty} \|(W_k(D_k\, A^T A D_k + E_k D_k)W_k\ )^{-1}\| = \infty.$$

$$W_k(D_k\,A^TA+E_k)p=-W_kD_kg_k \quad\Leftrightarrow\quad W_k^{\frac{1}{2}}(D_k^{\frac{1}{2}}\,A^TA+D_k^{-\frac{1}{2}}E_k)\,p=-W_k^{\frac{1}{2}}D_k^{\frac{1}{2}}\,g_k$$

$$\Downarrow$$

$$\underbrace{W_k^{\frac{1}{2}}(D_k^{\frac{1}{2}}\,A^TAD_k^{\frac{1}{2}}+E_k)W_k^{\frac{1}{2}}}_{\text{s.d.p. matrix}}\;\tilde{p}=-W_k^{\frac{1}{2}}D_k^{\frac{1}{2}}\,g_k,\qquad \tilde{p}=D_k^{-\frac{1}{2}}W_k^{-\frac{1}{2}}\,p$$

Properties of $Z(x)=W(x)^{\frac{1}{2}}(D(x)^{\frac{1}{2}}A^TAD(x)^{\frac{1}{2}}+E(x))W(x)^{\frac{1}{2}},\quad \forall x>0$:

- $Z(x)$ is s.d.p.;
- $\|Z(x)^{-1}\|\le C,\qquad$ for some $C>0$;
- $k_2(Z(x))\le k_2(W(x)(D(x)\,A^TA+E(x)))$.

Potential benefits - Direct methods

- Symmetric approximate minimum degree permutation of Z_k + Cholesky factorization + iterative refinement for

$$Z_k \tilde{p} = -W_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \, g_k$$

- Sparse methods + iterative refinement for the augmented system

$$\begin{pmatrix} W_k E_k & W_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \, A^T \\ A \, W_k^{\frac{1}{2}} D_k^{\frac{1}{2}} & -I_m \end{pmatrix} \begin{pmatrix} \tilde{p} \\ \\ q \end{pmatrix} = \begin{pmatrix} -W_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \, g_k \\ \\ 0 \end{pmatrix}$$

Routines from Harwell Subroutine Library [Duff, Erisman and Reid, 1986].

Potential benefits - Iterative solvers

- Solve the Newton equation approximately $\Rightarrow$ Inexact Newton method:

$$W_k(D_k\, A^T A + E_k)p = -W_k D_k g_k + r_k, \qquad \|r_k\|_2 \leq \eta_k \|W_k D_k g_k\|_2, \quad \eta_k \in [0, 1).$$

- Solve the s.p.d. linear system:

$$Z_k \tilde{p} = -W_k^{\frac{1}{2}} D_k^{\frac{1}{2}}\, g_k + \tilde{r}_k, \qquad \tilde{r}_k = D_k^{-\frac{1}{2}} W_k^{-\frac{1}{2}} r_k$$

$$\|\tilde{r}_k\|_2 \leq \eta_k \|W_k D_k g_k\|_2 \quad \Rightarrow \quad \|r_k\|_2 \leq \eta_k \|W_k D_k g_k\|_2$$

Apply Conjiugate Gradient (CG) methods: short recurrence, theoretical finite termination.

CG-like methods

$$Z_k \tilde{p} = -W_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \, g_k \qquad \Leftrightarrow \qquad \min_{\tilde{p} \in \mathbb{R}^n} \left\| \begin{pmatrix} A W_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \\ \\ W_k^{\frac{1}{2}} E_k^{\frac{1}{2}} \end{pmatrix} \tilde{p} + \begin{pmatrix} A x_k - b \\ \\ 0 \end{pmatrix} \right\|_2$$

- Apply CGLS, LSQR to the least-squares problem.

CGLS and LSQR are analytically equivalent to CG, [Paige, Saunders, ACM Trans. Math. Softw. 1982].

Formulation of the Newton-like method

0. Given $x_0 > 0$, $\sigma < 1$.

1. For $k = 0,\ 1, \ldots$

   1.1 Choose $\eta_k \in [0, 1)$.

   1.2 Solve $Z_k \tilde{p} = -W_k^{\frac{1}{2}} D_k^{\frac{1}{2}}\ g_k + \tilde{r}_k, \qquad \|\tilde{r}_k\|_2 \leq \eta_k \|W_k D_k g_k\|_2$

   1.3 Set $p = W_k^{\frac{1}{2}} D_k^{\frac{1}{2}} \tilde{p}$

   1.4 Set $p_k = \max\{\sigma,\ 1 - \|P(x_k + p) - x_k\|_2\}\ (P(x_k + p) - x_k)$

   1.5 Set $x_{k+1} = x_k + p_k$

Theorem. Let x_0 be sufficiently near to x^*.

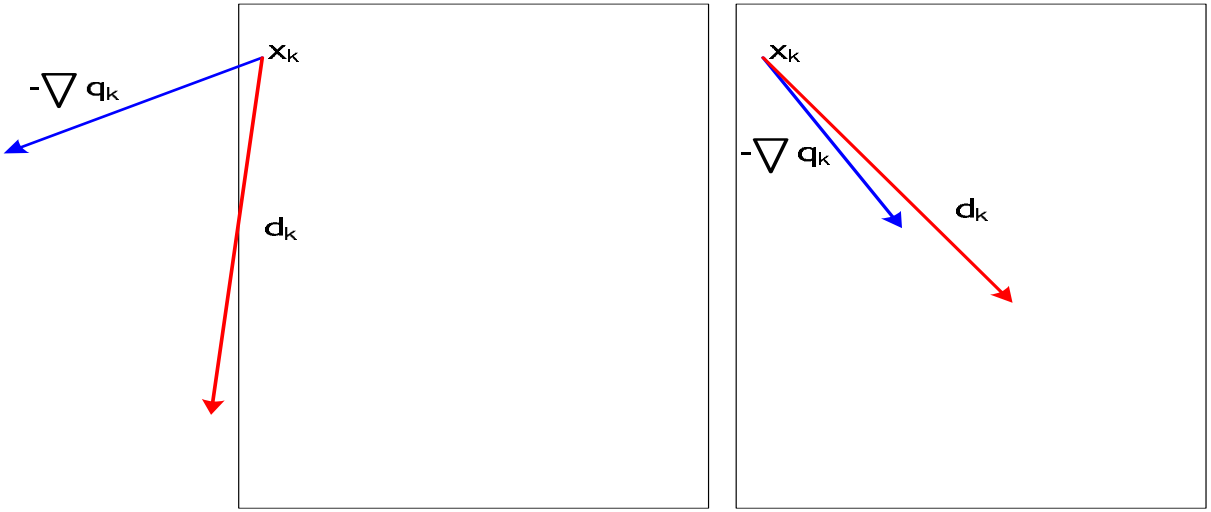
If $\eta_k = 0,\ \forall k$, then $x_k \longrightarrow x^*$ with quadratic convergence rate.

If $\eta_k = O(\|W_k D_k g_k\|)$, then $x_k \longrightarrow x^*$ with quadratic convergence rate.

A globalization strategy

Consider the well-angled scaled stepeest descent direction $d_k = -D_k\ g_k$

d_k is biased towards the interior of Ω as D_k penalizes the step $-g_k$ preventing a step directly toward a boundary point.



- Consider the quadratic model:

$$\psi_k(p) = \frac{1}{2} p^T (A^T A + D_k^{-1} E_k) p + p^T g_k.$$

  Note that the Newton step is the global minimizer of $\psi_k(p)$.

- Define the generalized Cauchy step p_k^C:
  - Let p_d be the minimizer of ψ_k along d_k $\Rightarrow$ $p_k^C = p_d$ if $x_k + p_d > 0$
  - Otherwise, let $\theta \in (0, 1)$ and λ_k be the stepsize along d_k to the boundary. $\Rightarrow$ $p_k^C = \theta \lambda_k d_k$

- The functions $\psi_k(p)$ and $q(x)$ are related so that

$$q(x_k) - q(x_k + p) = -\psi_k(p) + \frac{1}{2} p^T D_k^{-1} E_k p.$$

Since $D_k^{-1} E_k$ is positive semidefinite

$$q(x_k) - q(x_k + p) \geq -\psi_k(p).$$

- Global convergence depends on taking a step $\bar{p}_k$ satisfying

$$\frac{\psi_k(\bar{p}_k)}{\psi_k(p_k^C)} \geq \beta, \qquad \beta \in (0,\, 1),$$

that implies

$$q(x_k) - q(x_k + \bar{p}_k) \geq -\beta \psi_k(p_k^C)$$

i.e. $q(x_k + \bar{p}_k) < q(x_k)$.

Forming the new iterate x_{k+1}

Embed the Newton-like method into the globalization strategy. Let p_k be the projected (inexact) Newton step.

$$\bar{p}_k = p_k, \qquad \text{if} \quad \frac{\psi_k(p_k)}{\psi_k(p_k^C)} \geq \beta$$

$$\bar{p}_k = (1-t)p_k + tp_k^C, \quad t \in [0,1) \text{ s.t.} \quad \frac{\psi_k(\bar{p}_k)}{\psi_k(p_k^C)} = \beta, \quad \text{otherwise}$$

$$\text{Set} \qquad x_{k+1} = x_k + \bar{p}_k$$

Theorem. Let $x_0 > 0$ be an arbitrary initial point. Then

- $\lim_{k\to\infty} x_k = x^*$;
- Eventually, the projected Newton step p_k is taken $\Rightarrow \{x_k\}$ converges to x^* quadratically.

- Advantageous:
  strategy easy and cheap to implement;
  good theoretical results.

- Drawback:
  The value $\|D_k g_k\|_2 = \|D_k A^T(Ax_k - b)\|_2$ may oscillate for ill-conditioned problems and it can exhibit a large growth at some iterations (see [Nocedal, Sartenear, Zhu, COAP 2002]).

  The direction of p_k^C is $D_k g_k$. We select a point on the segment connecting the projected Newton step p_k and the Cauchy step p_k^C.

  $$\Downarrow$$

  The oscillating behaviour of $\|D_k g_k\|_2$ may slower the iterative process so that several iterations are needed to reach the vicinity of x^*.

Numerical results

We implemented the method in a `Matlab` code, $\epsilon_m = 2.\ 10^{-16}$.

- Initial guess $x_0 = (1, \ldots, 1)^T$.
- Stopping criteria:

$$\begin{cases} q_{k-1} - q_k < \tau\ (1 + q_{k-1}), \\ \|x_k - x_{k-1}\|_2 \leq \sqrt{\tau}\ (\ 1 + \|x_k\|_2\ ) \\ \|P(x_k + g_k) - x_k\|_2 < \tau^{\frac{1}{3}}\ (\ 1 + \|g_k\|_2\ ) \end{cases} \quad \text{or} \quad \|D_k\, g_k\|_2 \leq \tau$$

  with $\tau = 10^{-9}$.

- A failure is declared after 300 iterations.

Numerical solution of the Newton equation:

- direct method:
  Symmetric approximate minimum degree permutation (function symamd) + Cholesky decomposition + one step of iterative refinement.

- iterative method:
  CGLS method + preconditioner (diagonal/IC with minimum degree preordering).

  We iterate CGLS until

$$\|\tilde{r}_k\|_2 \leq \begin{cases} \min\{10^{-1},\ \|W_k D_k g_k\|_2\}\, \|W_k D_k g_k\|_2 & \text{if } \|W_k D_k g_k\|_2^2 > 500\epsilon_m \\ 500\epsilon_m & \text{otherwise} \end{cases}$$

  Eventually, $\eta_k = \|W_k D_k g_k\|_2$ i.e. quadratic convergence.

Harwell Boeing Collection tests

Well-conditioned or moderately ill-conditioned $A \in \mathbb{R}^{m \times n}$, solution may be degenerate.

Performance of the Newton-like method.

Test name	m	n	$k_2(A)$	$\|x_0 - x^*\|_2$	it
add20.rua	2395	2395	$4.9\,10^1$	$1.7\,10^4$	13
illc1033	1033	320	$1.9\,10^4$	$5.8\,10^3$	35
illc1850	1850	712	$1.4\,10^4$	$6.1\,10^3$	16
well1033	1033	320	$1.7\,10^2$	$5.8\,10^3$	14
well1850	1850	712	$1.1\,10^2$	$5.2\,10^3$	16

Effects of conditioning of A and degeneracy of x^*

- Test generator by [Portugal, Judice and Vicente, Math. Comput. 1994]: constructs NNLS problems where x^*, g^*, $k_2(A)$ are prescribed.

- We fixed

$$A \in \mathbb{R}^{5000 \times 2000}$$

$$dens(A) = \frac{nnz(A)}{m\,n} = 5\ 10^{-3}$$

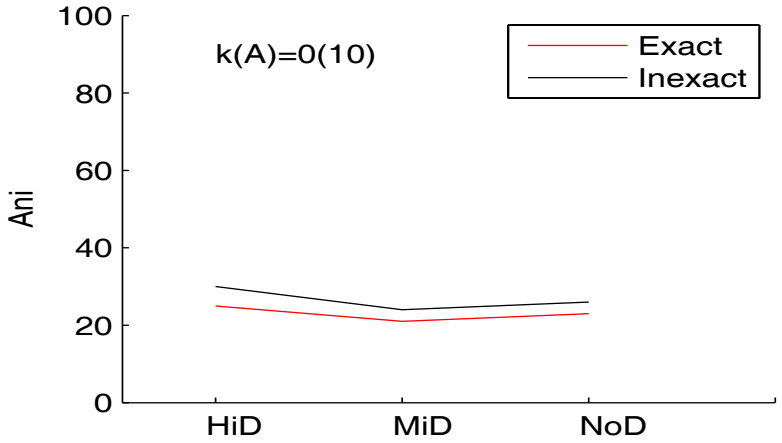
$$x^* = (\overbrace{1, 2, 3, 4, 5, \ldots}^{inactive}, \quad \overbrace{0, 0, \ldots 0}^{nondegenerate} \quad , \quad \overbrace{0, 0, \ldots 0}^{degenerate})^T$$

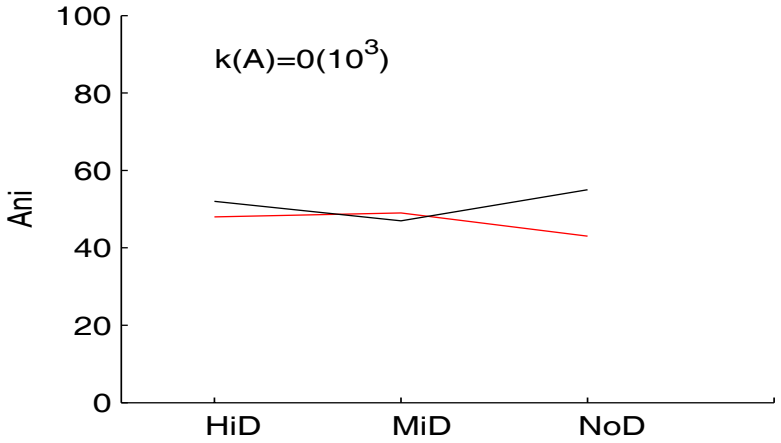
Set of tests	components of x^*			$\|x_0 - x^*\|_2$
	# inactive	# nondegenerate	# degenerate	
`HiD` - Higly degenerate	1000	900	100	$1.8\ 10^4$
`MiD` - Mildly degenerate	500	1490	10	$6.4\ 10^3$
`NoD` - Non degenerate	1500	500	0	$3.3\ 10^4$

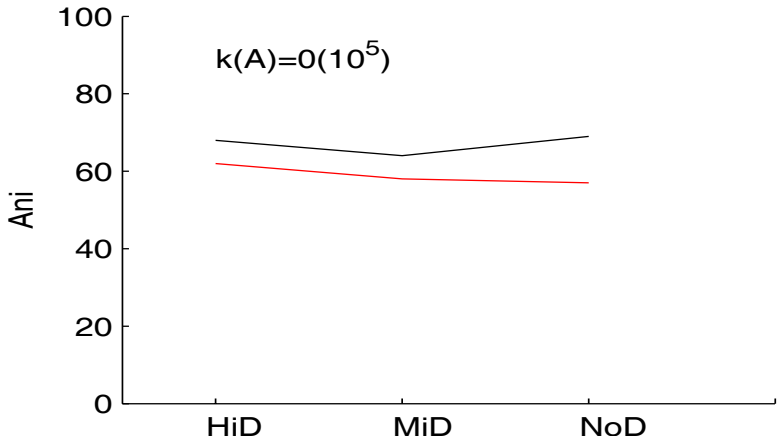
Each `Set` contains 30 tests:

$$\begin{cases} \text{10 tests where} \quad k_2(A) = 0(10), \\ \text{10 tests where} \quad k_2(A) = 0(10^3), \\ \text{10 tests where} \quad k_2(A) = 0(10^5), \end{cases}$$

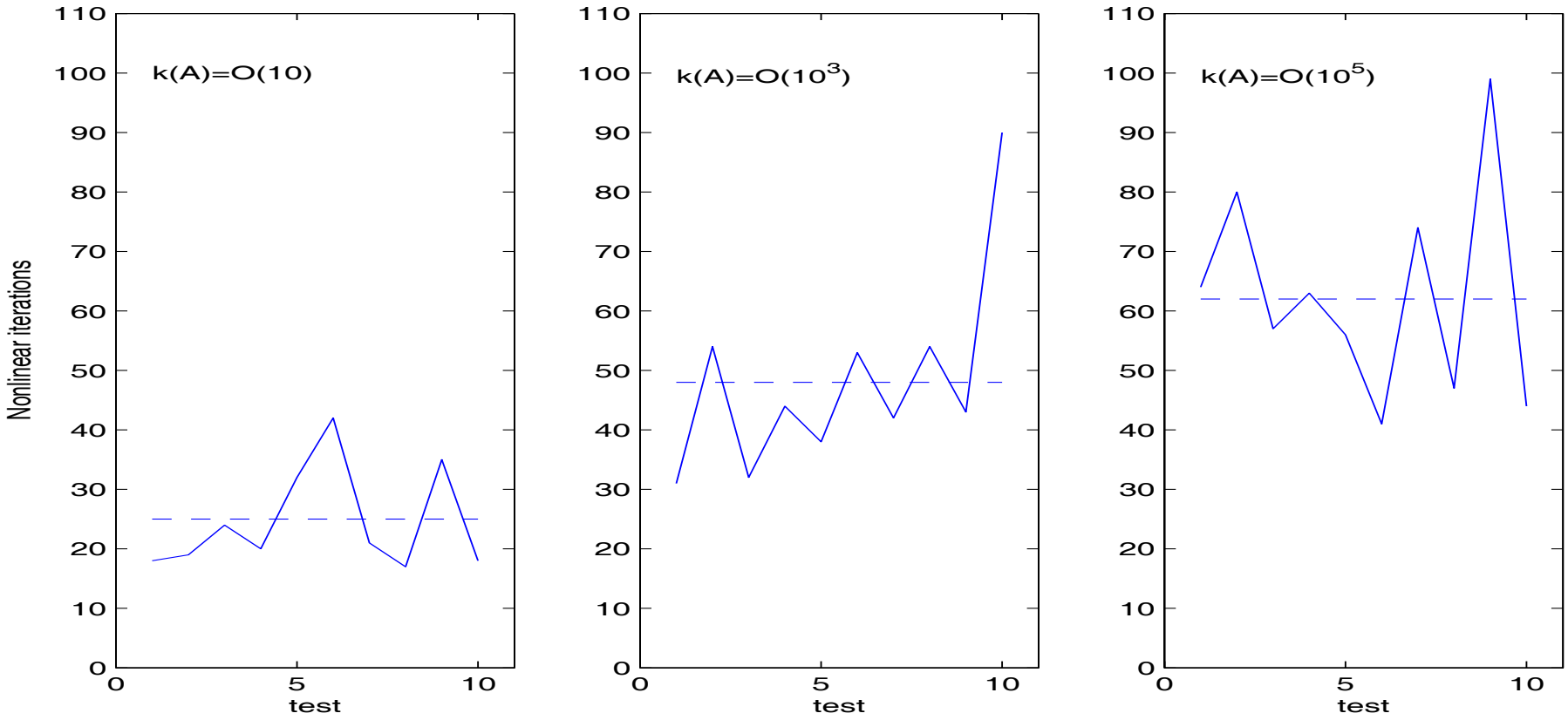
Average of nonlinear iterations







Number of nonlinear iterations - HiD Set



Few runs are very expensive.

Average number of linear iterations (Inexact method)

Set	$k_2(A) = O(10^1)$	$k_2(A) = O(10^3)$	$k_2(A) = O(10^5)$
HiD	23	32	36
MiD	22	24	55
NoD	25	38	55

Randomly generated tests

$A \in \mathbb{R}^{5000\times 5000}$

10 tests for $k_2(A) = 0(10^\gamma),\ \gamma = 1, 3, 5 \ \Rightarrow$ total of 30 tests

Performance of the Inexact Newton-like method

$k_2(A)$	Ani	Mni	Ali
$0(10)$	12	14	25
$0(10^3)$	21	28	23
$0(10^5)$	30	46	32

Ani : average number of nonlinear iteration performed over 10 tests.

Mni : maximum number of nonlinear iteration performed over 10 tests.

Ali : average number of linear iterations performed over 10 tests.

Features of the methods

- All runs reveal fast local convergence.
- The performance of the Inexact Newton-like method is comparable to that of the Newton-like method.
- The number of nonlinear iterations is insensitive to:

  problem's dimension;

  number of active constraints at the solution x^*;

  degeneracy of x^*;
- The number of nonlinear iterations increases when the conditioning deteriorate due to the globalization strategy.

Future work

- Barzilai-Borwein methods for NNLS problems.

- Scaling tecniques for the augmented system & direct solvers.