On Some Preconditioning Techniques for Linear Least-Squares Problems with Multiple Right-Hand Sides

Jean **TSHIMANGA** Ilunga(FUNDP)
joint work with
Serge Gratton (CERFACS) and Annick Sartenaer (FUNDP)

Sparse Days, Cerfacs, Toulouse, June 15-16, 2006

Outline

- General framework
- Preconditioning techniques considered
- Some properties
- Numerical experiments (data assimilation)
- Conclusions and perspectives

The goal of data assimilation

The goal of data assimilation is to find an initial state vector for which the trajector of the forecast model best fits an initial background and some observations. The vector found is the used to start a new simulation which gives an improved forecast.

Data assimilation problem formulation

A variational formulation: nonlinear least-squares problem

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}||x - x_b||^2_{B^{-1}} + \frac{1}{2}\sum_{j=0}^{N} ||\mathcal{H}_j(\mathcal{M}_j(x)) - y_j||^2_{R_j^{-1}}$$

- Size of real (operational) problems : $x, x_b \in \mathbb{R}^{10^6}$, $y_j \in \mathbb{R}^{10^5}$.
- The observations y_j and the background x_b are noisy.
- $\mathcal{M}$ is the model operator (nonlinear)
- $\mathcal{H}$ is the observation operator (nonlinear)

Solution strategy

- Typical approach : Incremental 4DVAR (i.e. inexact/truncated Gauss-Newton algorithm (GN)).
- GN leads to sequence of linear least-squares problems
- (Equivalently to) Sequence of linear symmetric positive definite systems to solve normal equations :

$$J_i^T J_i x = J_i^T r_i$$

- Whose matrix $A_i \equiv J_i^T J_i$ varies.

The key Idea (in our study)

- We consider a symmetric and positive definite (spd) matrix A.
- Solve systems $Ax = b_1$, $Ax = b_2$, ..., $Ax = b_r$ with RHS in sequence, by iterative methods: Conjugate Gradient (CG) or variants.
- Precondition the CG using informations obtained when solving the previous system.
- Extension of the idea to nonlinear process such as Gauss-Newton (GN) method. The matrix varies along the process.

The CG algorithm (A is spd and large !)

- CG is an iterative method for solving

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^T Ax - b^T x \quad \Leftrightarrow \quad Ax = b \ (\textit{stationary eq.})$$

- Iterations: Given $x_0 \in \mathbb{R}^n$; $A \in \mathbb{R}^{n\times n}$; $b \in \mathbb{R}^n$
  - Set $r_0 \leftarrow Ax_0 - b_0$; $p_0 \leftarrow -r_0$; $i \leftarrow 0$
  - Loop on i

$$\begin{aligned}
\alpha_i &\leftarrow (r_i^T r_i)/(p_i^T A p_i) \\
x_{i+1} &\leftarrow x_i + \alpha_i p_i \\
r_{i+1} &\leftarrow r_i + \alpha_i A p_i \\
\beta_{i+1} &\leftarrow (r_{i+1}^T r_{i+1})/(r_i^T r_i) \\
p_{i+1} &\leftarrow -r_{i+1} + \beta_{i+1} p_i
\end{aligned}$$

- r_i are residuals; p_i are descent directions.

The CG properties (in exact arithmetic !)

- Orthogonality of the residuals: $r_i^T r_j = 0$ if $i \neq j$.
- A-conjugacy of the descent directions: $p_i^T A p_j = 0$ if $i \neq j$.
- The distance of the iterate x_i to the solution x^* is related to the condition number of A, denoted by $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ (≥ 1):

$$||x_i - x^*||_A \leq \eta_i \; ||x_0 - x^*||_A \quad \text{with} \quad \eta_i = 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i$$

  $\Rightarrow$ The smaller $cond(A) \equiv \kappa$ is, the faster the convergence.
- Exact solution found exactly in r iterations, where $r \leq n$ is the number of distinct eigenvalues of $A \in \mathbb{R}^{n \times n}$.
  $\Rightarrow$ The more clustered the eigenvalues are, the faster the convergence.

Why to precondition ?

- Transform $Ax = b$ in an equivalent system having a more favorable eigenvalues distribution for faster convergence.
- Use a preconditioning matrix H (which must be cheap to apply).
- Ideas to design preconditioner. H would :
  - approximates A^{-1}.
  - make $cond(HA) < cond(A)$.
  - make eigenvalues of HA more clustered than those of A.
- Note: when a preconditioning is used, residuals are:
  - Orthogonal if H is factored in LL^T.
  - Conjugate w.r.t. H if H is not factored.

Preconditioning techniques considered (I)

- We consider (second level preconditioning) techniques :
  - Solve $Ax = b_1$ and extract information $info_1$.
  - Use $info_1$ to solve $Ax = b_2$ and extract information $info_2$.
  - Use $info_2$ (and possibly $info_1$) to solve $Ax = b_3$ and ...
  - ...
- $Info_k$ will be:
  - descent directions;
  - or other vectors such as eigenvectors of A ...

Preconditioning techniques considered (II)

We study and compare two approaches:

- Deflation [Frank, Vuik, 2001].
- Limited Memory Preconditioners (LMP): Preconditioners based on a set of A-conjugate directions. Generalization of known preconditioners:
  - spectral [Fisher, 1998],
  - L-BFGS [Nocedal, Morales, 2000], warm start [Gilbert, Lemaréchal, 1989].

We cover:

- Theoretical properties.
- Numerical experiments (data assimilation).

Deflation Techniques

- Given $W \in \mathbb{R}^{n \times k}$ $(k \ll n)$ formed with appropriate information obtained when solving the previous system.
- Consider the oblique projector $P = I - AW(W^T AW)^{-1} W^T$.
- Split the solution vector as follows $x^* = \underbrace{(I - P^T)x^*}_{direct} + \underbrace{P^T x^*}_{iterative}$.
- Compute $(I - P^T)x^*$ with a direct method.
- Compute $P^T x^*$ with an iterative method.

Some Properties for Deflation

- Computation of $(I - P^T)x^*$:
  - $(I - P^T)x^* = W(W^TAW)^{-1}W^TAx^* = W(W^TAW)^{-1}W^Tb$.
  - Note: $W^TAW \in \mathbb{R}^{k\times k}$ and $k \ll n$.
- Computation of P^Tx^*:
  - Consider the compatible singular (but still symmetric) system $PAy = Pb$.
  - Use CG with $y_0 = 0$ to solve $PAy = Pb$.
  - Any solution y of $PAy = Pb$ satisfies $P^Tx^* = P^Ty$.
  - Note: $cond(PA) \leq cond(A)$ and $PA = (PA)^T$.

Limited Memory Preconditioners (LMP)

- General reformulation of LMP:

$$H_{k+1} = [I - \sum_{i=0}^{k} \frac{A w_i w_i^T}{w_i^T A w_i}]^T [I - \sum_{i=0}^{k} \frac{A w_i w_i^T}{w_i^T A w_i}] + \sum_{i=0}^{k} \frac{w_i w_i^T}{w_i^T A w_i},$$

$$\text{with } w_i^T A w_j \begin{cases} = 0 & \text{if } i \neq j \\ > 0 & \text{if } i = j \end{cases}$$

- Particular forms
  - The w_i's are the descent directions obtained from CG: $w_i = p_i$ $\Rightarrow$ L-BFGS preconditioner.
  - The w_i's are eigenvectors of A: $w_i = v_i$ $\Rightarrow$ spectral preconditioner.

Spectral Properties for LMP (I)

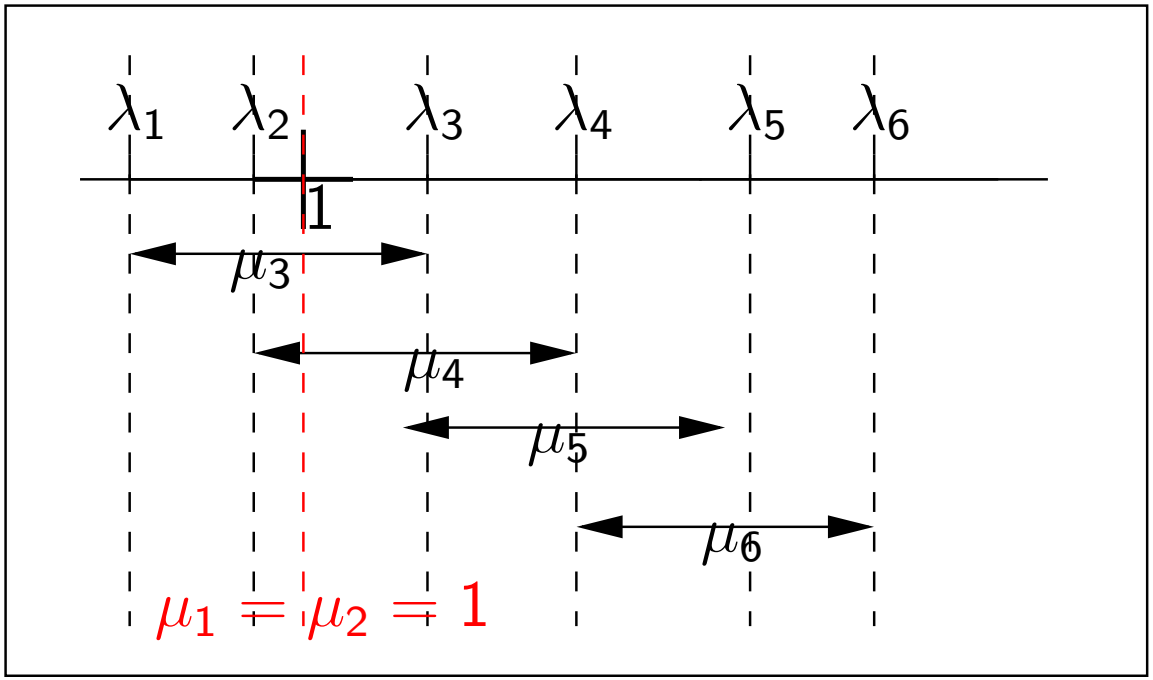
. The matrix A is fixed.

- The matrix A to precondition is the same as the previous one (only the RHS changes).
- Theorem : Assume that $\lambda_1, \ldots, \lambda_n$ is the spectrum of A, then the spectrum $\mu_1, \ldots, \mu_n$ of the preconditioned matrix $H_{k+1}A$ satisfies:

$$\begin{cases} \mu_j = 1, & \text{for } j = 1, \ldots, k \\ \lambda_{j-k}(A) \leq \mu_j \leq \lambda_j(A), & \text{for } j = k+1, \ldots, n, \end{cases}$$

where $\lambda_j(A)$ is the $j-$th eigenvalue of A (increasing order assumed).

Spectral properties for LMP (II)



- Eigenvalues translated to 1.
- The rest of the spectrum is not expanded compared to the spectrum of A.

Spectral Properties for LMP (III)

The matrix A varies.

- $\tilde{A} = A + \tau E$ with $||E||_F = 1$ and $\tau \in R^+$ (and small enougth : first order perturbation)
- Theorem : Assume that τ is small, then the $k+1$ first eigenvalues of the preconditioned matrix $H_{k+1}\tilde{A}$ satisfies:

$$\tilde{\mu}_j \leq 1 + \lambda_{max}(W^T E W) * \tau + o(\tau), \quad \text{for } j = 1, \ldots, k+1$$

- $\lambda_{max}(W^T E W) * \tau$ is a measure of the default of conjugacy of columns vectors of W w.r.t. the new matrice $\tilde{A}$

Existence of a factored form for the LMP (not the Cholesky factor and obtained by construction!)

- L-BFGS:
  - A possible factored form is $H_{k+1} = L_{k+1} L_{k+1}^T$ where:

$$L_{k+1} = \prod_{i=0}^{k} \left(I - \frac{s_i {y_i}^T}{{y_i}^T s_i} + \frac{s_i}{\sqrt{{y_i}^T s_i}} \frac{{r_i}^T}{\|r_i\|} \right),$$

    with $s_i = x_{i+1} - x_i$ and $y_i = r_{i+1} - r_i$.
  - Same cost in memory and CPU as the unfactored form.
- Spectral:
  - A possible factored form is $H_{k+1} = L_{k+1}^2$ where:

$$L_{k+1} = I + \sum_{i=1}^{k+1} \left(\frac{1}{\sqrt{\lambda_i}} - 1 \right) \frac{v_i {v_i}^T}{{v_i}^T v_i}.$$

  - Same cost in memory as the unfactored form.

Why looking for a factored form $H = LL^T$?

- With a non factored form, we use CG preconditioned by H.
- With a factored form, we solve $L^T A L u = L^T b$; $x = Lu$.

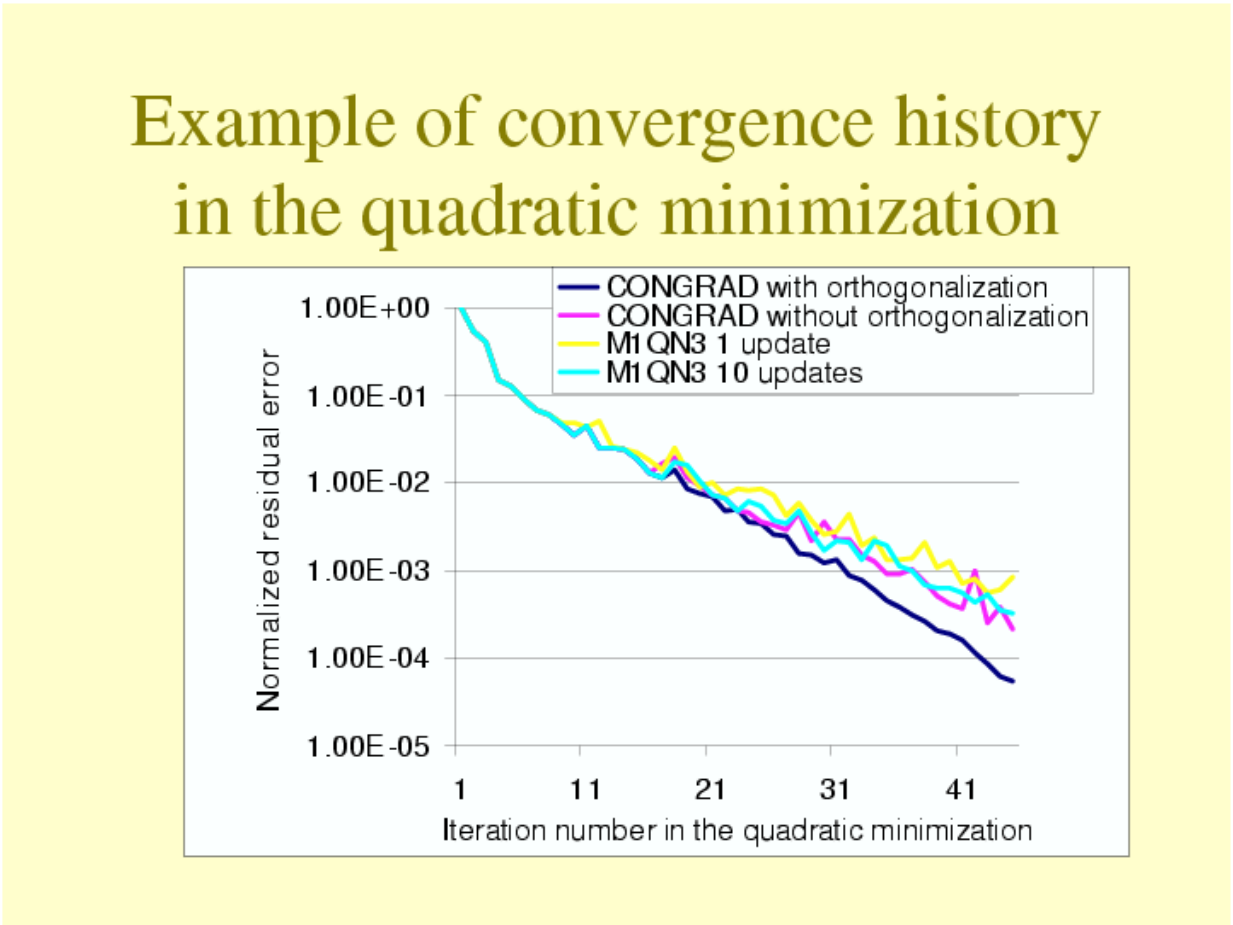
  Advantages:
  - More appropriate if reorthogonalization of the residuals is used.
  - Least-squares $min_x \|Ax - b\|$ or $AA^T x = A^T b$: LSQR (or CGLS) is more accurate than CG in presence of rounding errors but works with (A, A^T, L, L^T, b) instead of $(A^T A, A^T b, H)$.
  - When accumulating preconditioners, symmetry and positiveness are still maintained:

$$L_1^T A L_1 y_1 = L_1^T b_1, \quad L_2^T (L_1^T A L_1) L_2 y_2 = L_2^T L_1^T b_2, \quad \ldots$$

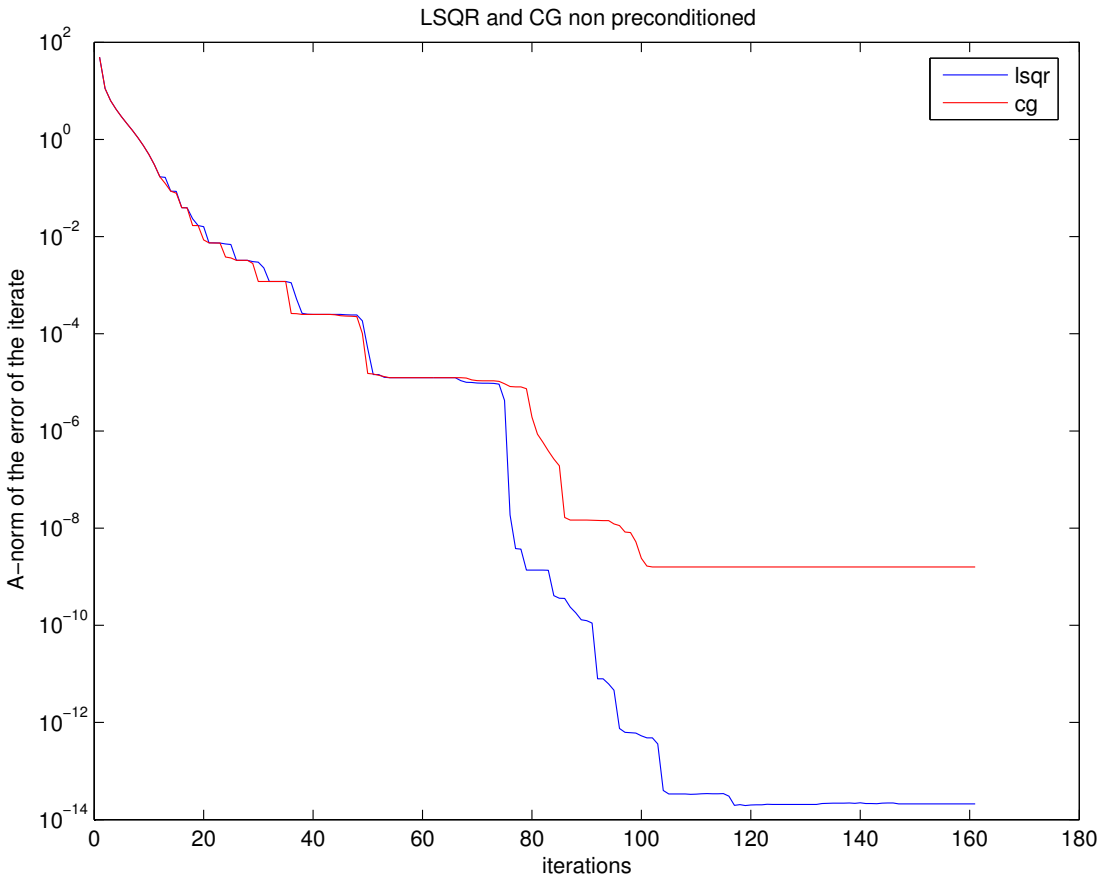
Why to reorthogonalize the residuals ?

- In finite precision, residuals often loose orthogonality (or conjugacy) and theoretical convergence is then slowed down.
- Reorthogonalization of residuals in CG is terribly successful when matrix-vector product is very expensive compared to other computations in CG (see example in the next slide).
- Note: to restore orthogonality or conjugacy, working with $L^T AL$ and the canonical inner-product is better (memory, CPU, error propagation) than working on A preconditioned by the no factored matrix H.

Example of reorthogonalisation effect : CERFACS data assimilation system (1 000 000 unknowns)

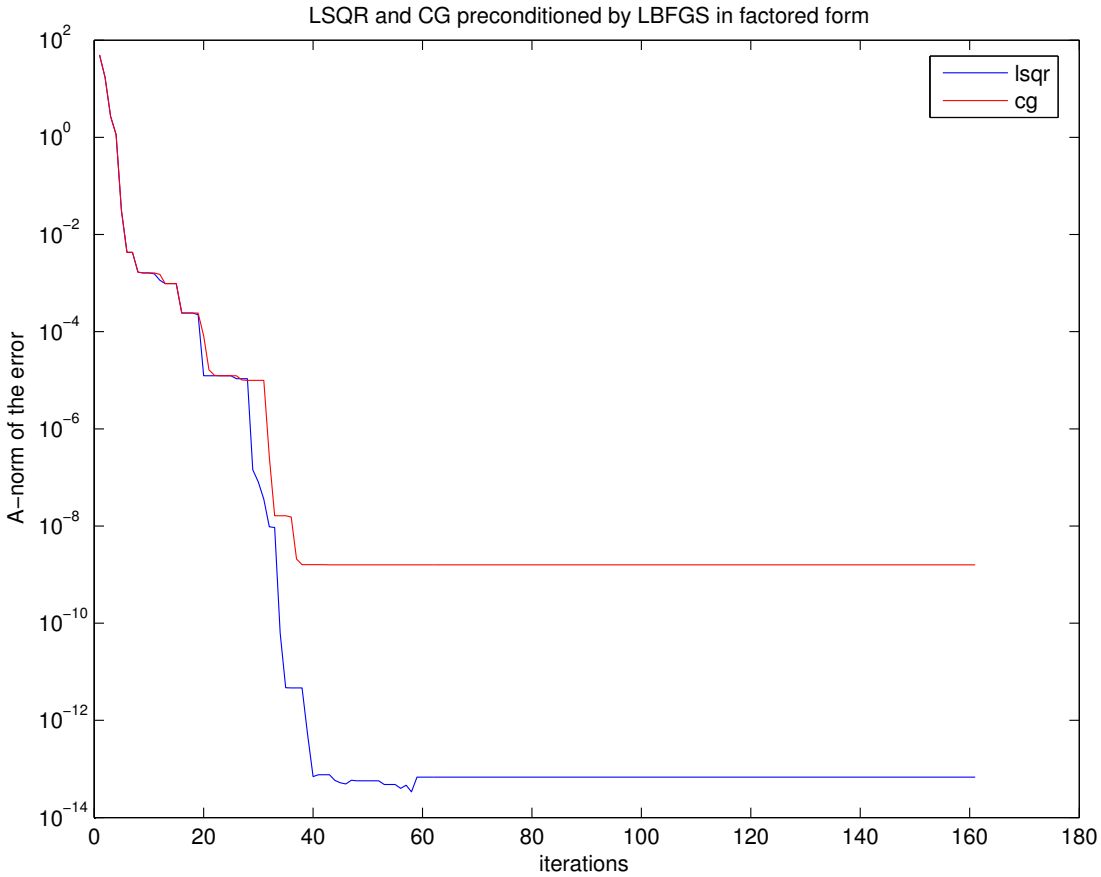


Experiments with unpreconditioned LSQR



LSQR is better than CG !

Experiments with LSQR preconditioned with factored L-BFGS



LSQR is again better than CG !

Experiments description

Algorithmic variants tested:

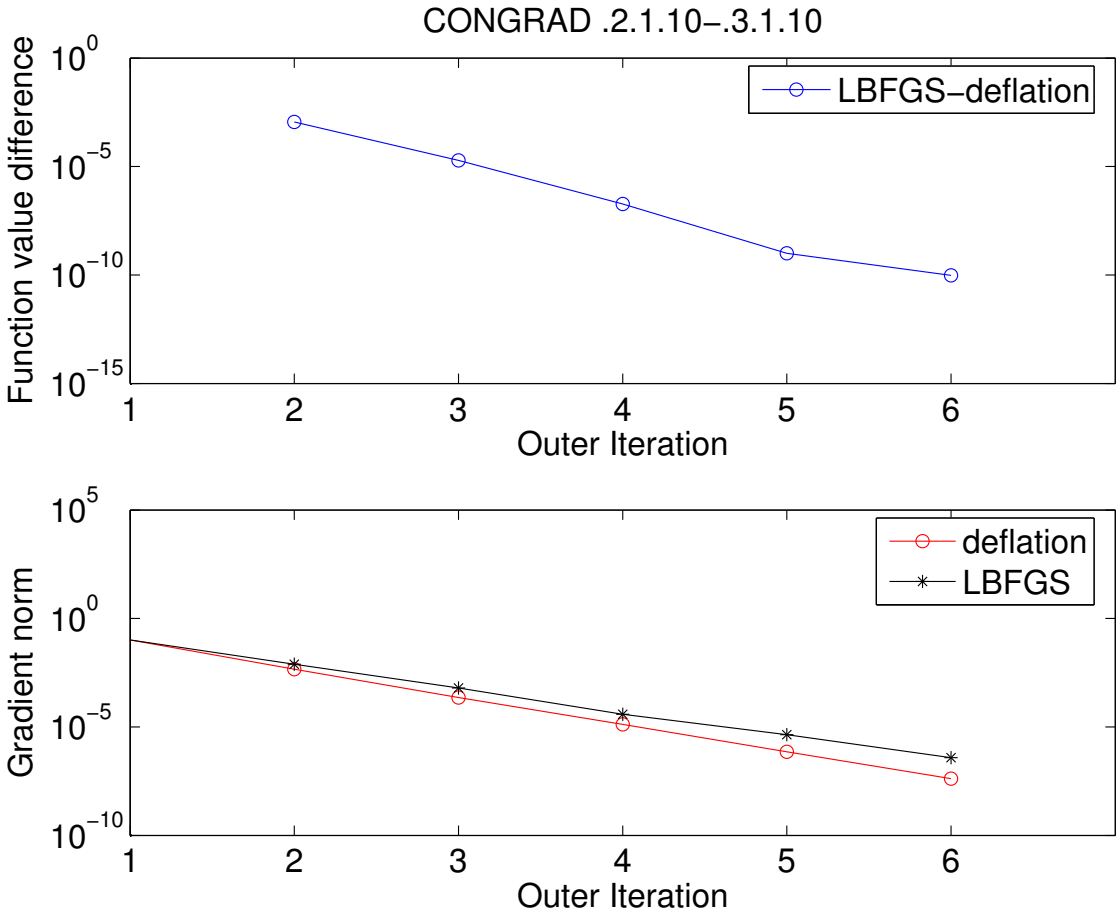
- Use CG to solve the normal equations.
- Compare 3 preconditioning techniques:
  - Deflation technique (using spectral information).
  - Spectral preconditioner (using spectral info. but differently).
  - L-BFGS preconditioner (using descent directions).
- Where spectral information is needed, use Ritz (vectors) as approximations of the eigenvectors.
- Ritz vectors are obtained by mean of a variant of CG: the Lanczos algorithm which combines linear and eigen solvers.

Experiment on a small system (I) [A. Lawless, N. Nichols, 2001, University of Reading]

Ranking of the preconditioners using the basic strategies.

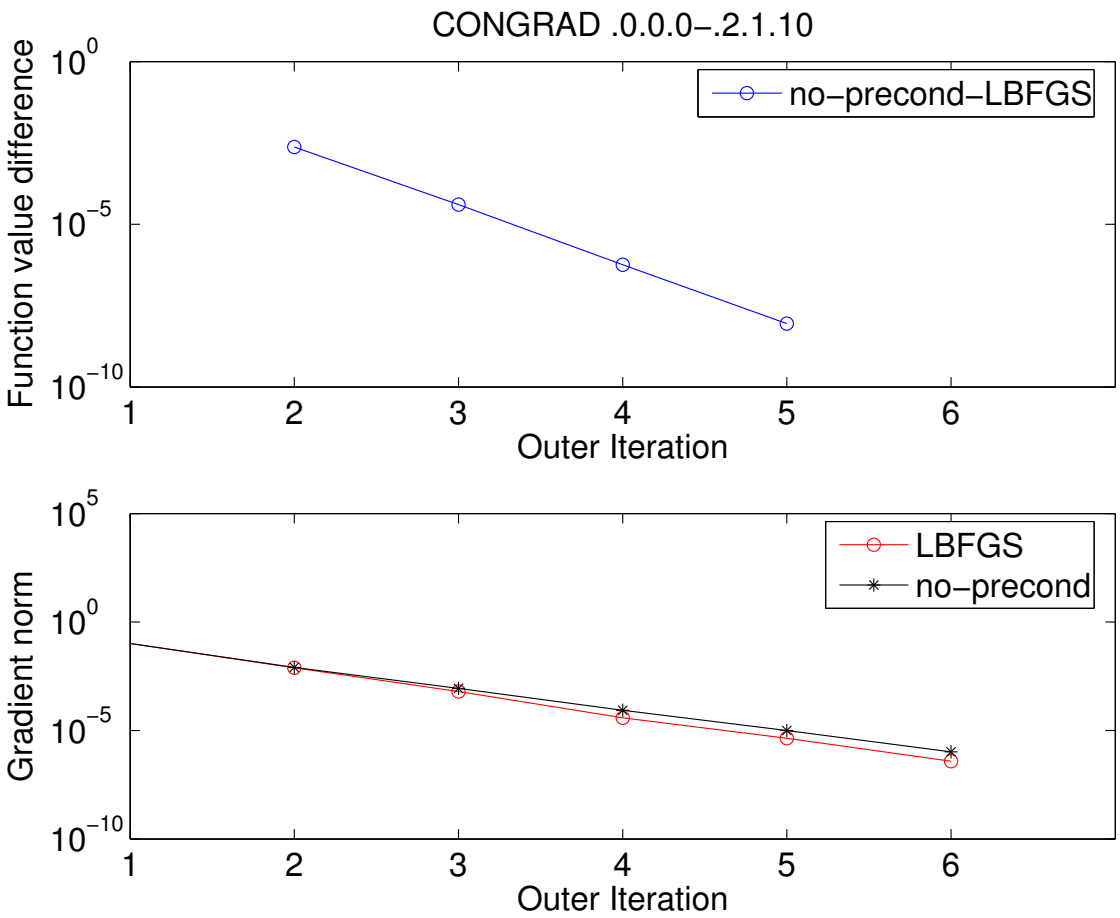
- Diagnostics are made on the non linear optimisation problem
- The number of CG iterations is fixed.

Results L-BFGS - Deflation



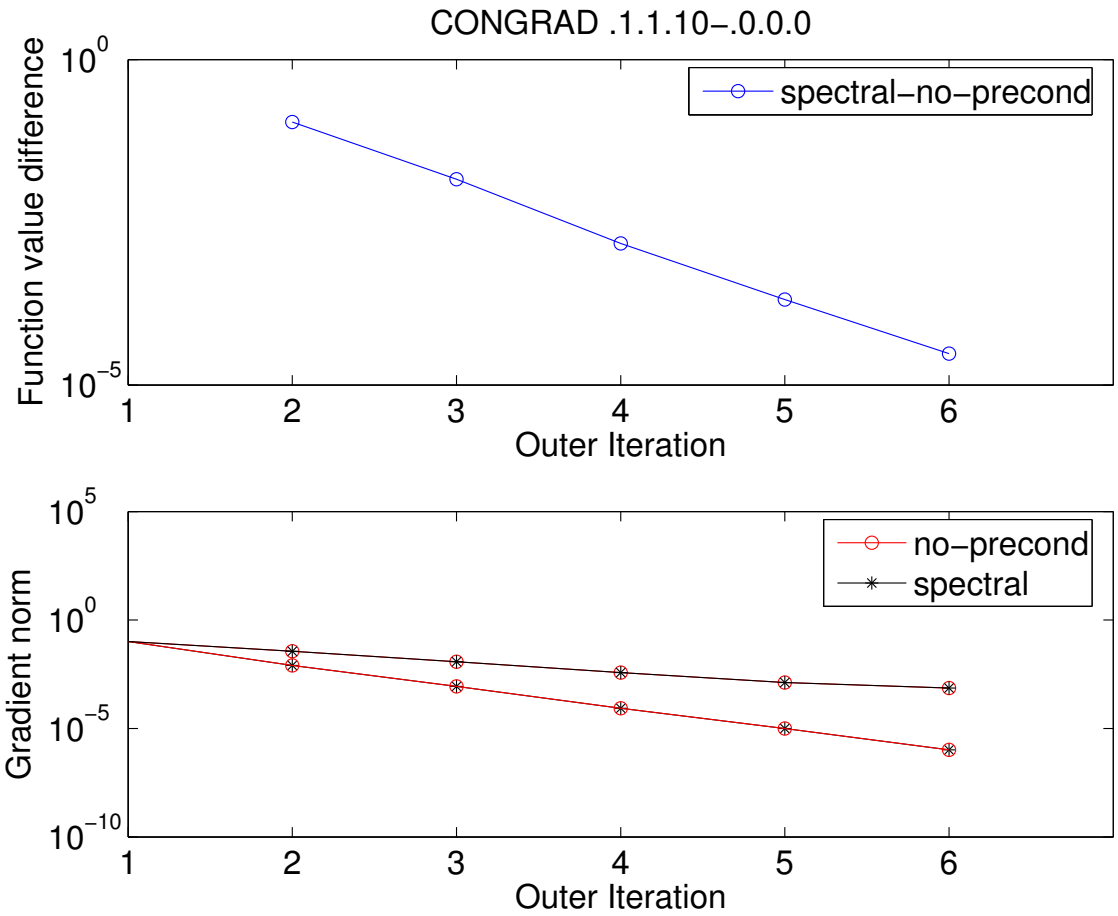
Deflation is better than L-BFGS !

Results Noprecond - L-BFGS



L-BFGS is better than Noprecond !

Results Spectral - Noprecond



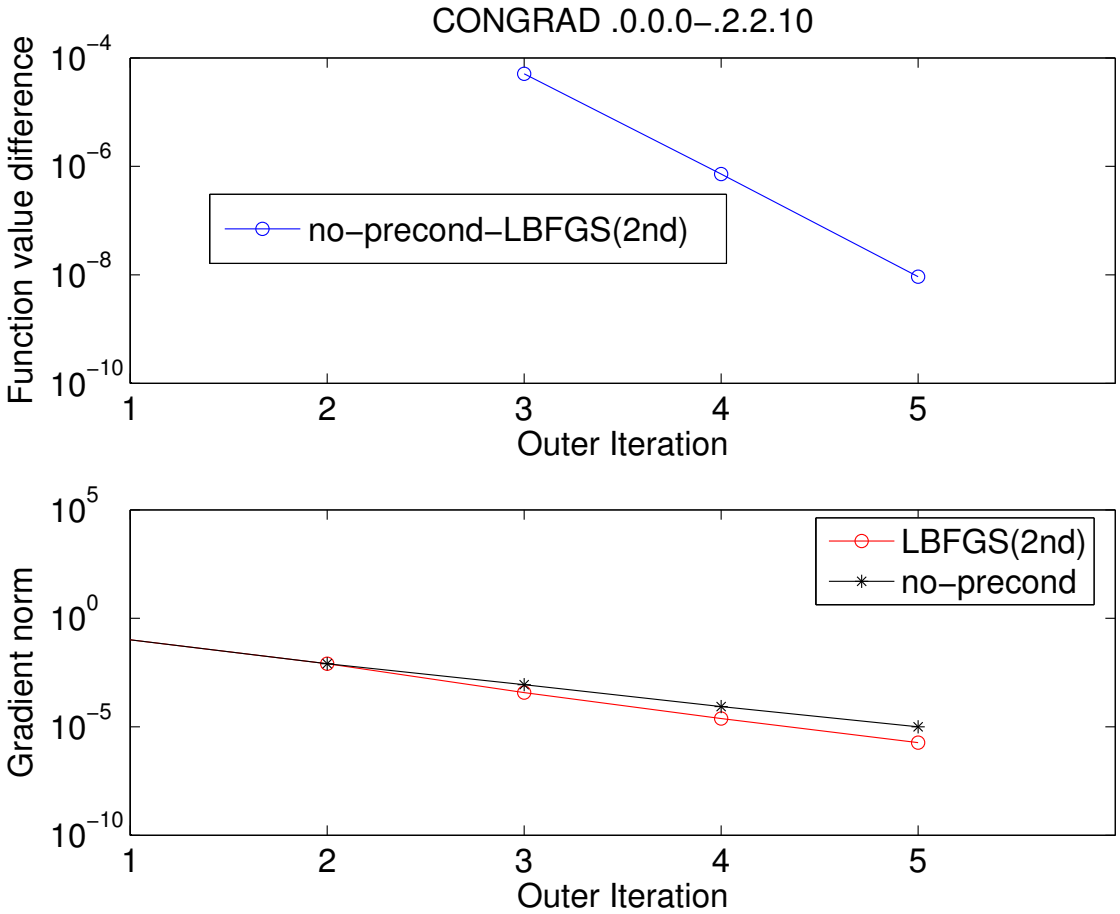
Noprecond is better than Spectral !

Experiment on a small system (II) [A. Lawless, N. Nichols, 2001, University of Reading]

Is it better to wait until the system does not vary much for the preconditioner to be efficient ?

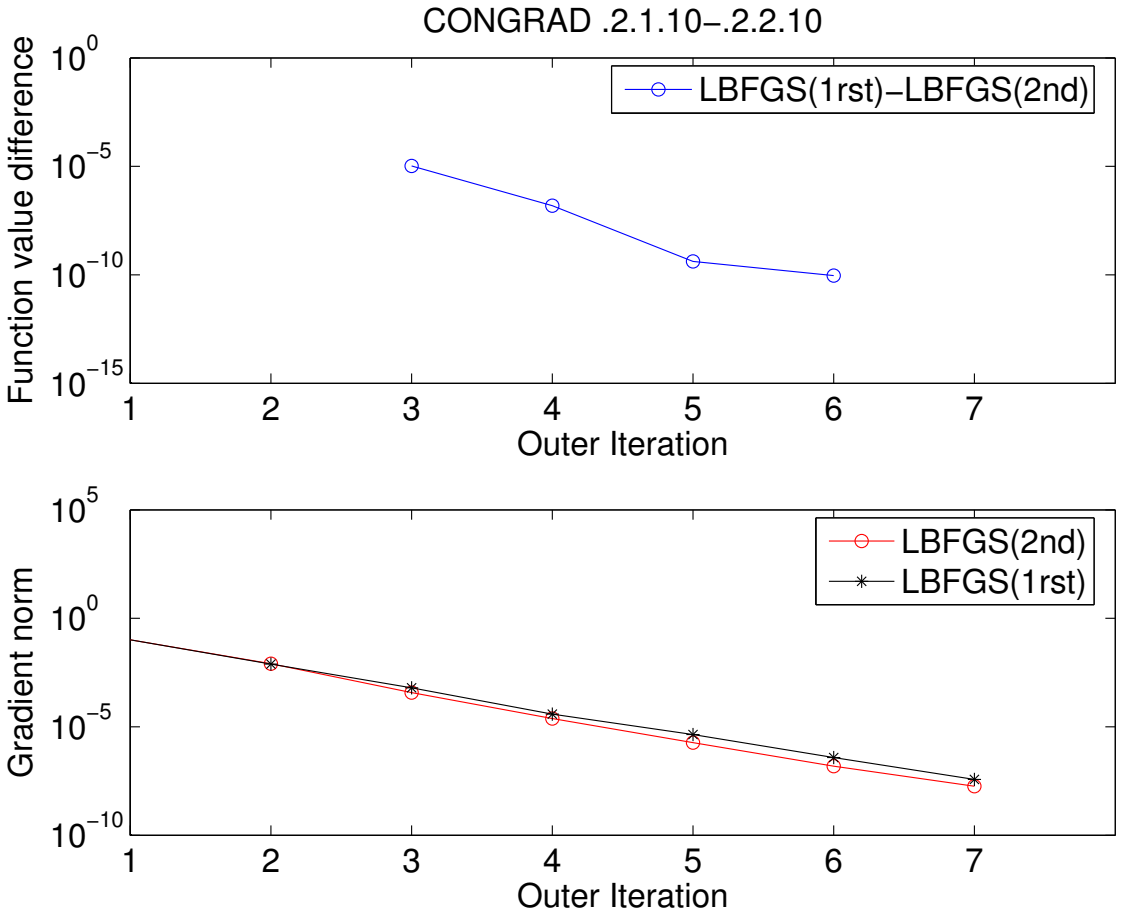
- Our main theory assumes that the matrix A is fixed. In practice (data assimilation) the matrix varies.
- It is Known that the GN method may not be locally convergent on some problems. But in our experiments the GN process converges and the steps thus become smaller.
- This means that the matrix does not change so much, when approaching the solution.

Results Noprecond - L-BFGS(2nd)



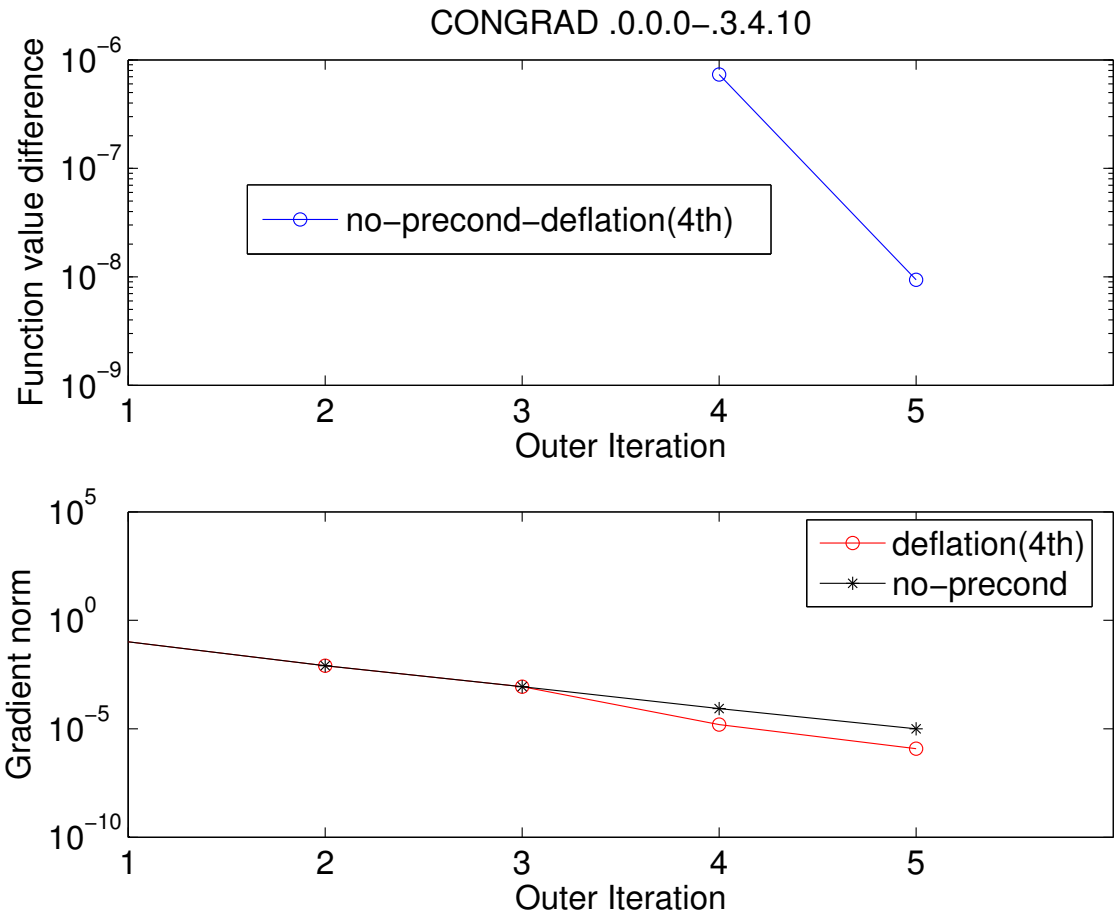
L-BFGS(2nd) is better than Noprecond !

Results L-BFGS(1rst) - L-BFGS(2nd)



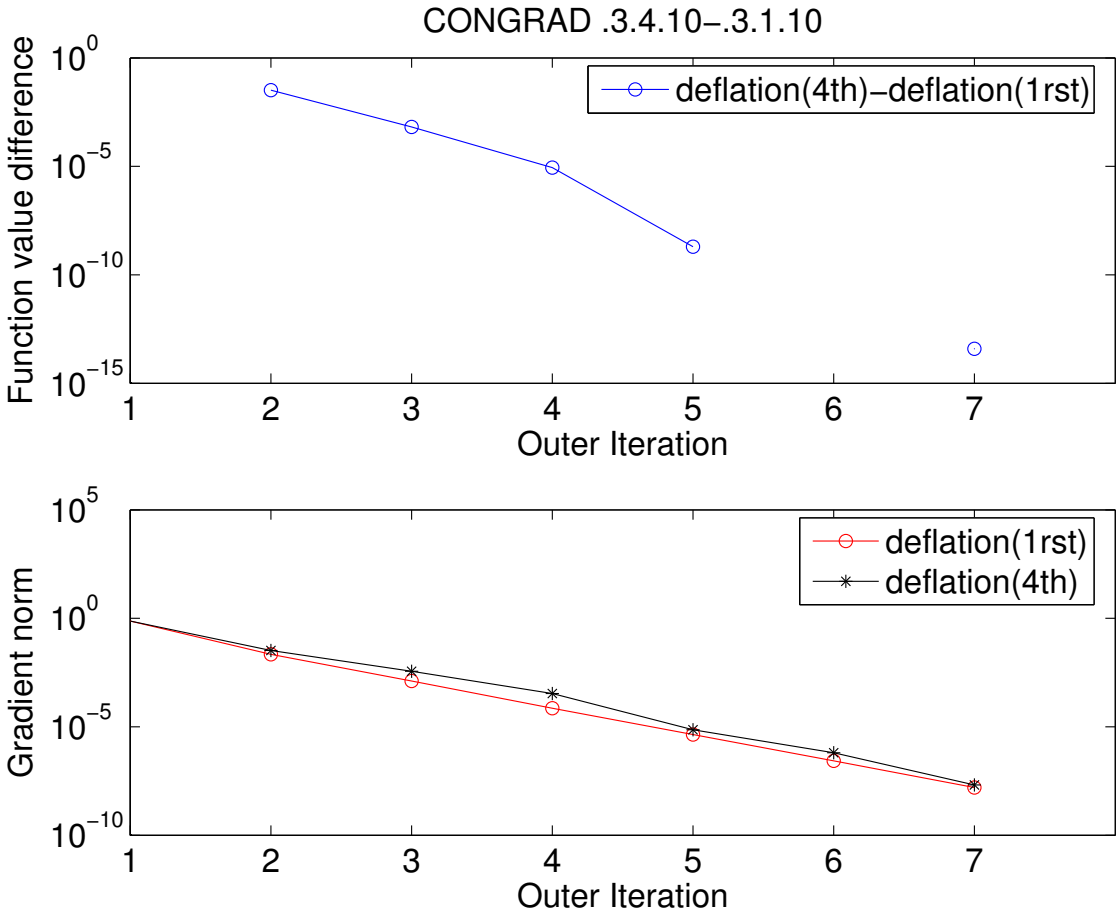
L-BFGS(2nd) is better than L-BFGS(1rst) !

Results Noprecond - deflation(4th)



Deflation(4th) is better than Noprecond !

Results deflation(4th) - deflation(1rst)



Deflation(1rst) is a little better than Deflation(4th) !

Remarks on our system !

- Spectral preconditioner:
  - Does not work in our case.
- L-BFGS preconditioner:
  - Requires no large changes in the matrix.
  - Based on by-products of CG.
  - More efficient than the spectral preconditioner or than no preconditioner.
- Deflation:
  - Is stable even when the matrix changes.
  - May be expensive ($W^T AW$) in CPU time.
  - More efficient than the other techniques.

Conclusions

- Properties of LMP preconditioners understood (when the matrix A is fixed or changes a "bit").
- Existence of factored forms for particular instances of LMP.
- Preliminary tests show weakness of spectral compared to deflation and L-BFGS in a data assimilation experiment.
- One technique (deflation) not yet used before in data assimilation has been tested in this field.

Perpectives and work in progress

- Experiment sampling techniques to select information:
  - Which information to capture ?
  - Purge the preconditioner or not ?
  - When and how to decide to apply the preconditioner ?
- Make tests in a more realistic data assimilation environment:
  - CERFACS 2D Shallow Water or any.
  - CERFACS operational system or any.