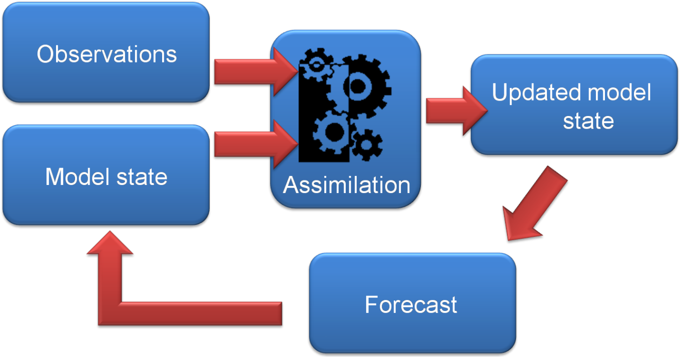
Preconditioners for saddle point weak-constraint 4D-Var with correlated observation errors

Jemima M. Tabeart
TU Eindhoven
j.m.tabeart@tue.nl

With John Pearson
University of Edinburgh

Data assimilation: observation + prior info = ???

Weighted combination of observation and prior information (typically from numerical model)



Areas of recent research interest: engineering design, COVID prediction, economics, renewable energy sector, ecology, personalised medicine...

Data assimilation for numerical weather prediction presents challenges and opportunities

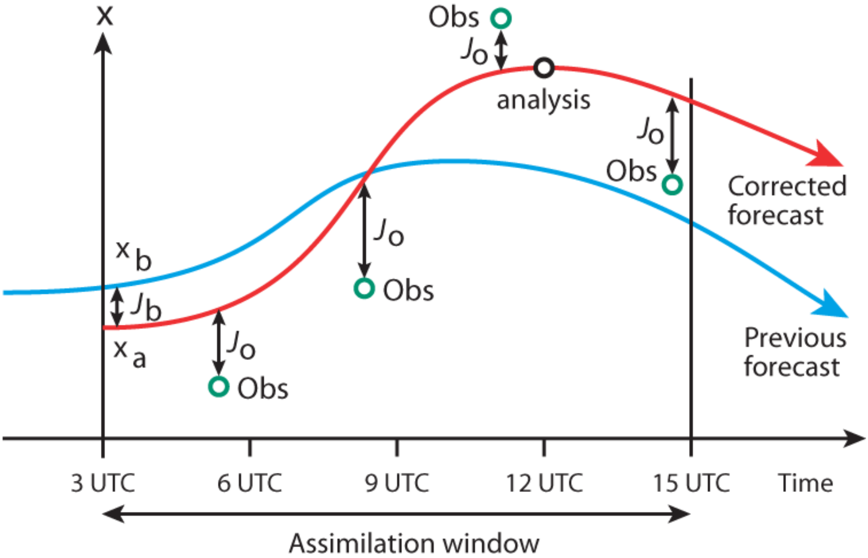
- Very high dimensional systems (10^9 state variables and 10^6 observations)
- Extreme time constraints: e.g. 30 minutes for DA component of a traditional 6 hour forecast cycles, JMA: update forecasts every 10 minutes.
- Noisy data with gaps

Data assimilation for numerical weather prediction presents challenges and opportunities

- Very high dimensional systems (10^9 state variables and 10^6 observations)
- Extreme time constraints: e.g. 30 minutes for DA component of a traditional 6 hour forecast cycles, JMA: update forecasts every 10 minutes.
- Noisy data with gaps

+ Data/linear systems possess lots of inherent structure
+ Mature applications: exploit expert knowledge of physics/instruments when designing new approaches
+ Large amount of data/community models for testing

DA applied to numerical weather prediction



Variational DA can be viewed as a minimization problem

Need to solve

$$\min_{x \in \mathbb{R}^{s(N+1)}} J(x), \quad x = \text{vec}([x_0, \dots, x_N]) = (x_0^T, \dots, x_N^T)^T$$

where

$$J(x) = \frac{1}{2}\|x_0 - x_0^B\|^2_{B^{-1}} + \frac{1}{2}\sum_{i=0}^{N}\|y_i - \mathcal{H}_i(x_i)\|^2_{R_i^{-1}} + \frac{1}{2}\sum_{i=0}^{N-1}\|x_{i+1} - \mathcal{M}_i(x_i)\|^2_{Q_{i+1}^{-1}}$$

- $x_i \in \mathbb{R}^s$ model state at time t_i
- $y_i \in \mathbb{R}^p$ observation at time t_i
- $\mathcal{H}_i$ new observation operator, $y_i = \mathcal{H}_i(x_i^t) + \epsilon_i$, x_i^t true state, $\epsilon_i \sim \mathcal{N}(0, R_i)$
- $\mathcal{M}_i$ (inexact) forecast model, $x_{i+1} = \mathcal{M}_i(x_i) + \epsilon_i^M$, $\epsilon_i^M \sim \mathcal{N}(0, Q_i)$
- $x_0^B = x_0^t + \epsilon^B$, $\epsilon^B \sim \mathcal{N}(0, B)$

Inner loop we solve a SPD linear system

$$\underbrace{(\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}+\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})}_{\mathcal{S}}\delta x=\mathbf{D}^{-1}b+\mathbf{L}^T\mathbf{H}^T\mathbf{R}^{-1}d$$

$$\mathbf{D}=\begin{pmatrix} B & & & \\ & Q_1 & & \\ & & \ddots & \\ & & & Q_N \end{pmatrix}, \qquad \mathbf{L}=\begin{pmatrix} I & & & \\ -M_1 & I & & \\ & \ddots & \ddots & \\ & & -M_N & I \end{pmatrix},$$

$$\mathbf{R}=\begin{pmatrix} R_0 & & & \\ & R_1 & & \\ & & \ddots & \\ & & & R_N \end{pmatrix}, \qquad \mathbf{H}=\begin{pmatrix} H_0 & & & \\ & H_1 & & \\ & & \ddots & \\ & & & H_N \end{pmatrix}.$$

Saddle point formulation of weak-constraint data assimilation

Re-write linearised objective function in saddle point form

$$\begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{R} & \mathbf{H} \\ \mathbf{L}^\top & \mathbf{H}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \delta\eta \\ \delta\nu \\ \delta\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{d} \\ \mathbf{0} \end{pmatrix}. \quad (1)$$

$$\mathbf{D} = \texttt{blkdiag}\left(\mathbf{B}, \mathbf{Q}_1, \mathbf{Q}_2, ..., \mathbf{Q}_N\right) \in \mathbb{R}^{(N+1)s\times(N+1)s},$$

$$\mathbf{R} = \texttt{blkdiag}\left(\mathbf{R}_0, \mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N\right) \in \mathbb{R}^{(N+1)p\times(N+1)p},$$

$$\mathbf{H} = \texttt{blkdiag}\left(\mathbf{H}_0^{(l)}, \mathbf{H}_1^{(l)}, \mathbf{H}_2^{(l)}, ..., \mathbf{H}_N^{(l)}\right) \in \mathbb{R}^{(N+1)s\times(N+1)s},$$

$$\mathbf{L} = \begin{pmatrix} \mathbf{I} & & & & \\ -\mathbf{M}_1^{(l)} & \mathbf{I} & & & \\ & -\mathbf{M}_2^{(l)} & \mathbf{I} & & \\ & & \ddots & \ddots & \\ & & & -\mathbf{M}_N^{(l)} & \mathbf{I} \end{pmatrix}. \quad (2)$$

Why the saddle point formulation

- Saddle point systems well-studied in numerical linear algebra
  - Standard preconditioning approaches
  - Eigenvalue bounds - guarantee good performance of MINRES
- Reveal structure that is obscured in objective function form
  - Block-diagonal structure means we can immediately parallelise multiplication with saddle matrix (typical DA motivation)

Much more varied options for preconditioners than the primal form

Some preconditioners for saddle point problems

$$\mathcal{P}_D = \begin{bmatrix} \widehat{\mathbf{D}} & & \\ & \widehat{\mathbf{R}} & \\ & & \widehat{\mathbf{S}} \end{bmatrix}, \quad \mathcal{P}_T = \begin{bmatrix} \widehat{\mathbf{D}} & 0 & \mathbf{L} \\ & \widehat{\mathbf{R}} & \mathbf{H} \\ & & \widehat{\mathbf{S}} \end{bmatrix}, \quad \mathcal{P}_{\mathcal{C}} := \begin{bmatrix} \widehat{\mathbf{D}} & 0 & \widehat{\mathbf{L}} \\ 0 & \widehat{\mathbf{R}} & 0 \\ \widehat{\mathbf{L}}^T & 0 & 0 \end{bmatrix}$$

$$\mathbf{S} = \mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L} + \mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}$$

Some preconditioners for saddle point problems

$$\mathcal{P}_D = \begin{bmatrix} \widehat{\mathbf{D}} & & \\ & \widehat{\mathbf{R}} & \\ & & \widehat{\mathbf{S}} \end{bmatrix}, \quad \mathcal{P}_T = \begin{bmatrix} \widehat{\mathbf{D}} & 0 & \mathbf{L} \\ & \widehat{\mathbf{R}} & \mathbf{H} \\ & & \widehat{\mathbf{S}} \end{bmatrix}, \quad \mathcal{P}_\mathcal{C} := \begin{bmatrix} \widehat{\mathbf{D}} & 0 & \widehat{\mathbf{L}} \\ 0 & \widehat{\mathbf{R}} & 0 \\ \widehat{\mathbf{L}}^T & 0 & 0 \end{bmatrix}$$

$$\mathbf{S} = \mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L} + \mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}$$

$$\mathcal{P}_D^{-1} = \begin{bmatrix} \widehat{\mathbf{D}}^{-1} & & \\ & \widehat{\mathbf{R}}^{-1} & \\ & & \widehat{\mathbf{S}}^{-1} \end{bmatrix}, \mathcal{P}_T = \begin{bmatrix} \widehat{\mathbf{D}}^{-1} & 0 & -\widehat{\mathbf{D}}^{-1}\mathbf{L}\widehat{\mathbf{S}}^{-1} \\ & \widehat{\mathbf{R}}^{-1} & -\widehat{\mathbf{R}}\mathbf{H}\widehat{\mathbf{S}}^{-1} \\ & & \widehat{\mathbf{S}}^{-1} \end{bmatrix}$$

$$\mathcal{P}_\mathcal{C}^{-1} := \begin{bmatrix} 0 & 0 & \widehat{\mathbf{L}}^{-T} \\ 0 & \widehat{\mathbf{R}}^{-1} & 0 \\ \widehat{\mathbf{L}}^{-1} & 0 & -\widehat{\mathbf{S}}_0^{-1} \end{bmatrix}$$

$$\mathbf{S}_0 = \mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L}$$

Bounds on the preconditioned spectrum (block diagonal)

$$\mathcal{P}_{\mathcal{D}} := \begin{pmatrix} \widehat{\mathbf{D}} & & \\ & \widehat{\mathbf{R}} & \\ & & \widehat{\mathbf{S}} \end{pmatrix},$$

$$\lambda(\widehat{\mathbf{D}}^{-1}\mathbf{D}) \in [\lambda_{\mathbf{D}}, \Lambda_{\mathbf{D}}], \qquad \lambda(\widehat{\mathbf{R}}^{-1}\mathbf{R}) \in [\lambda_{\mathbf{R}}, \Lambda_{\mathbf{R}}], \qquad \lambda(\widehat{\mathbf{S}}^{-1}\mathbf{S}) \in [\delta, \Delta],$$

Theorem ([JMT and Pearson 2023a])

The eigenvalues of $\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}}$ are real, and satisfy:

$$\lambda(\boldsymbol{\mathcal{P}}_D^{-1}\boldsymbol{\mathcal{A}}) \in \left[\frac{\phi - \sqrt{\phi^2 + 4\Phi\Delta}}{2}, \frac{\Phi - \sqrt{\Phi^2 + 4\phi\delta}}{2}\right] \cup [\phi, \Phi] \cup \left[\frac{\phi + \sqrt{\phi^2 + 4\phi\delta}}{2}, \frac{\Phi + \sqrt{\Phi^2 + 4\Phi\Delta}}{2}\right],$$

where $\phi = \min\{\lambda_{\mathbf{D}}, \lambda_{\mathbf{R}}\}$, $\Phi = \max\{\Lambda_{\mathbf{D}}, \Lambda_{\mathbf{R}}\}$.

Standard preconditioning neglects observation term of Schur complement

One popular choice of preconditioner is given by

$$\widehat{\mathbf{S}} = \widehat{\mathbf{L}}^\top \mathbf{D}^{-1} \widehat{\mathbf{L}}. \tag{3}$$

- Neglect observation term completely
- Approximate $\mathbf{L}$

Standard preconditioning neglects observation term of Schur complement

One popular choice of preconditioner is given by

$$\widehat{\mathbf{S}} = \widehat{\mathbf{L}}^\top \mathbf{D}^{-1} \widehat{\mathbf{L}}. \tag{3}$$

- Neglect observation term completely
- Approximate $\mathbf{L}$

$$\widehat{\mathbf{S}}^{-1} = \widehat{\mathbf{L}}^{-1} \mathbf{D} \widehat{\mathbf{L}}^{-\top}$$

1. What are some good choices for $\widehat{\mathbf{L}}$?
2. Is including observation information in $\widehat{\mathbf{S}}$ a good idea:
   - when $\widehat{\mathbf{L}} = \mathbf{L}$?
   - when $\widehat{\mathbf{L}} \neq \mathbf{L}$?

Why do we need to approximate L in a preconditioner?

$$\mathbf{L} = \begin{pmatrix} \mathbf{I} & & & & \\ -\mathbf{M}_1^{(l)} & \mathbf{I} & & & \\ & -\mathbf{M}_2^{(l)} & \mathbf{I} & & \\ & & \ddots & \ddots & \\ & & & -\mathbf{M}_N^{(l)} & \mathbf{I} \end{pmatrix}.$$

$$\mathbf{L}^{-1} = \begin{pmatrix} \mathbf{I} & & & & \\ \mathbf{M}_{1,1} & \mathbf{I} & & & \\ \mathbf{M}_{1,2} & \mathbf{M}_{2,2} & \mathbf{I} & & \\ \vdots & \vdots & \ddots & \ddots & \\ \mathbf{M}_{1,N} & \mathbf{M}_{2,N} & \cdots & \mathbf{M}_{N,N} & \mathbf{I} \end{pmatrix}$$

where $\mathbf{M}_{i,j} = \mathbf{M}_i^{(l)}\mathbf{M}_{i+1}^{(l)}\dots\mathbf{M}_j^{(l)}$.

Standard approximations to $\mathbf{L}$ don't include model information

$$\mathbf{L}_0 = \begin{pmatrix} \mathbf{I} & & & & \\ & \mathbf{I} & & & \\ & & \mathbf{I} & & \\ & & & \ddots & \\ & & & & \mathbf{I} \end{pmatrix} = \mathbf{I}_{(N+1)s}$$

$$\mathbf{L}_I = \begin{pmatrix} \mathbf{I}_s & & & & \\ -\mathbf{I}_s & \mathbf{I}_s & & & \\ & -\mathbf{I}_s & \mathbf{I}_s & & \\ & & \ddots & \ddots & \\ & & & -\mathbf{I}_s & \mathbf{I}_s \end{pmatrix}$$

[Fisher et al 2018, Gratton et al 2018]

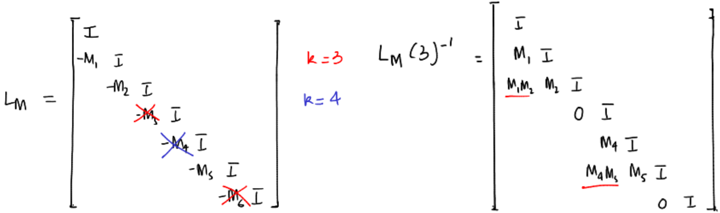
Proposed **L**: contains model info and is parallelisable

$$L_M = \begin{bmatrix} I & & & & & & \\ -M_1 & I & & & & & \\ & -M_2 & I & & & & \\ & & \cancel{-M_3} & I & & & \\ & & & \cancel{-M_4} & I & & \\ & & & & -M_5 & I & \\ & & & & & \cancel{-M_6} & I \end{bmatrix}$$

$k=3$

$k=4$

Proposed L: contains model info and is parallelisable



Parameter k controls the dimensions of the block diagonals.

- Highly parallelisable (block diagonal structure).
- $k = 1$ yields $\mathbf{L}_0$.
- $k = N + 1$ yields $\mathbf{L}$.
- Expect best performance/parallellisation trade-off for small k.

Eigenvalues of $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$

Theorem

We can write $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1} = \mathbf{I} + \mathbf{A}(\mathbf{M})$ *where the block entries of* $\mathbf{A}(\mathbf{M}) \in \mathbb{R}^{s(N+1)\times s(N+1)}$ *are defined as follows. For* $n = 1, \ldots, \left\lfloor \frac{N}{k} \right\rfloor$,

$$[\mathbf{A}(\mathbf{M})]_{i,j} = \begin{cases} (\prod_{t=i}^{nk} \mathbf{M}_t^\top)(\prod_{q=j}^{nk} \mathbf{M}_{nk-q+j}) & \text{for } (n-1)k+1 \leq i,j \leq nk, \\ -\prod_{t=j}^{nk} \mathbf{M}_{nk-t+j} & \text{for } i = nk+1, (n-1)k+1 \leq j \leq nk, \\ -\prod_{t=i}^{nk} \mathbf{M}_t^\top & \text{for } j = nk+1, (n-1)k+1 \leq i \leq nk, \\ \mathbf{0} & \text{otherwise}, \end{cases}$$

where $[\mathbf{A}(\mathbf{M})]_{i,j}$ *denotes the* (i,j)*th block of* $\mathbf{A}(\mathbf{M})$.

Theorem

Let $\mathbf{L}$ *be defined as in* (2) *and* $\mathbf{L}_M$ *as in Lemma 2. For* $2 \leq k \leq N+1$, $\mathbf{L}_M^{-\top}\mathbf{L}^\top\mathbf{L}\mathbf{L}_M^{-1}$ *has* rs *unit eigenvalues where* $r = N+1-2\left\lfloor \frac{N}{k} \right\rfloor$.

Theoretical conclusions

- Using model information we obtain more unit eigenvalues for the preconditioned $\mathbf{L}$ term than using $\mathbf{L}_0$.
- r is not strictly monotonic - increasing k increases/maintains the number of unit eigenvalues of the preconditioned system.
- Under additional assumptions on the $\mathbf{M}_i$s we can bound the remaining eigenvalues above

Lorenz 96 experiments

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + 8 \qquad (4)$$

where we have periodic boundary conditions ($x_{-1} = x_{s-1}$ and $x_0 = x_s$ and $x_{s+1} = x_1$). $F = 8$ gives us chaotic behaviour.

- $s = 2500, 1250, N = 15$
- $\mathbf{B}, \mathbf{Q}$ - truncated spatial (SOAR)
- $\mathbf{H}$ randomly selected direct/averaged observations
- $\mathbf{R}$ noisy block structure

Performance with changing k

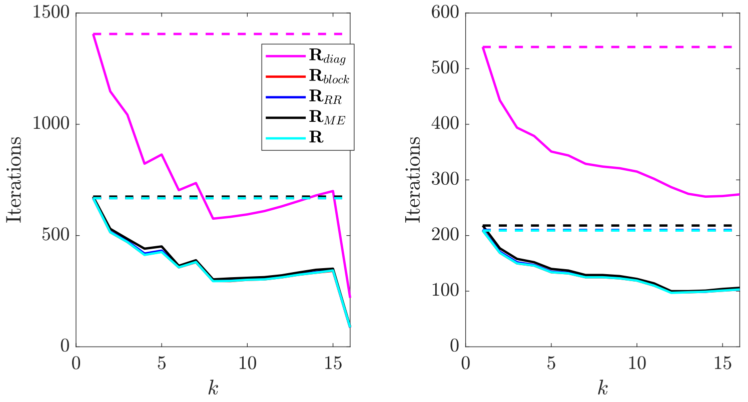


Figure: Performance of inexact constraint preconditioner for Lorenz 96 problem for changing values of k. Dimension of problem is $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{100,000 \times 100,0000}$.

Computational cost – matrix-vector products

k	$\mathbf{R}_i$	$\mathbf{D}_i$	$\widehat{\mathbf{D}}_i^{-1}$	$\mathbf{M}_i/\mathbf{M}_i^\top$	$\mathbf{R}_i$	$\mathbf{R}_{block}^{-1}$	$\mathbf{D}_i$	$\widehat{\mathbf{D}}_i^{-1}$	$\mathbf{M}_i/\mathbf{M}_i^\top$
1	22496	44992	22496	42180	10704	10704	21408	10704	20070
3	16688	33376	16688	47978	7536	7536	15072	7536	21666
4	13168	26336	13168	41150	6624	6624	13248	6624	20700
7	11776	23552	11776	41216	6080	6080	12160	6080	21280
10	9520	19040	9520	34510	4816	4816	9632	4816	17458
16	3520	7040	3520	12760	1376	1376	2752	1376	4988

Table: $\mathcal{P}_D$ for increasing k for $\mathbf{R}_{diag}$ (left) and $\mathbf{R}_{block}$ (right).

k	$\mathbf{R}_i$	$\mathbf{D}_i$	$\mathbf{M}_i/\mathbf{M}_i^\top$	$\mathbf{R}_i$	$\mathbf{R}_{block}^{-1}$	$\mathbf{D}_i$	$\mathbf{M}_i/\mathbf{M}_i^\top$
1	8624	17248	16170	3344	3344	6688	6270
3	6304	12608	18124	2400	2400	4800	6900
4	6064	12128	18950	2336	2336	4672	7300
7	5264	10528	18424	2000	2000	4000	7000
10	5040	10080	18270	1904	1904	3808	6902
16	4384	8768	15892	1648	1648	3296	5974

Table: $\mathcal{P}_I$ for increasing k for $\mathbf{R}_{diag}$ (left) and $\mathbf{R}_{block}$ (right).

Convergence for a large dimensional example

	$\mathbf{R}_{block}$	$\mathbf{R}_{RR}$	$\mathbf{R}$	$\mathbf{R}_{block}$	$\mathbf{R}_{RR}$	$\mathbf{R}$
$\mathbf{L}_0$	759	822	822	359	275	275
$\mathbf{L}_M$, $k = 3$	433	466	467	244	205	205
$\mathbf{L}_M$, $k = 4$	348	335	336	228	200	200
$\mathbf{L}_M$, $k = 5$	367	354	355	206	182	182

Table: Experiment A: Number of iterations required for convergence of MINRES with the block diagonal preconditioner $\mathcal{P}_D$ (left) and $\mathcal{P}_I$ (right) applied to the Lorenz 96 problem, using $\mathbf{R}_{block}$, $\mathbf{R}_{RR}$, $\mathbf{R}$ in combination with $\mathbf{L}_0$, $\mathbf{L}_M$ ($k = 3, 4, 5$). Here, $\mathcal{A} \in \mathbb{R}^{1,600,000 \times 1,600,000}$.

Conclusions of [JMT and Pearson 2023a]

- Better approximations to $\mathbf{L}$ improve convergence in terms of iterations
- Smaller values of k allow us to reduce/maintain the number of matrix-vector products with $\mathbf{M}_i$ and decrease the number of matrix-vector products with covariance matrices.
- Using a correlated choice of $\widehat{\mathbf{R}}$ compared to $\mathbf{R}_{diag}$ leads to large reduction in iterations and matrix-vector products.

Conclusions of [JMT and Pearson 2023a]

- Better approximations to $\mathbf{L}$ improve convergence in terms of iterations
- Smaller values of k allow us to reduce/maintain the number of matrix-vector products with $\mathbf{M}_i$ and decrease the number of matrix-vector products with covariance matrices.
- Using a correlated choice of $\widehat{\mathbf{R}}$ compared to $\mathbf{R}_{diag}$ leads to large reduction in iterations and matrix-vector products.

Will we see improvements when accounting for the observation term in $\mathbf{S}$?

Accounting for the observation term in $\widehat{\mathbf{S}}$

- Will only work for $\mathcal{P}_D, \mathcal{P}_T$ (recall the Schur complement for $\mathcal{P}_I^{-1}$ has no observation term)
- Can also be used within the primal formulation (where we solve a system of the form $\mathbf{S}\delta x = b$)

Start by considering the case $\widehat{\mathbf{L}} = \mathbf{L}$ and then extend our approach to the case of approximate $\mathbf{L}$.

$\lambda_{min}(\mathbf{R})$ is still important if $\widehat{\mathbf{S}} = \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$

If we precondition with the exact first term, we can bound the eigenvalues

Theorem ([JMT et al. 2021])

Let $\hat{\mathbf{S}}^{-1}\mathbf{S} = \mathbf{I} + \mathbf{D}^{1/2}\mathbf{L}^{-T}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{L}^{-1}\mathbf{D}^{1/2}$ *be the Hessian of the preconditioned data assimilation problem. Then we can bound the condition number of the preconditioned system above by:*

$$\kappa(\hat{\mathbf{S}}^{-1}\mathbf{S}) \leq 1 + \frac{\lambda_{max}^{LDL}}{\lambda_{min}(\mathbf{R})}\lambda_{max}(\mathbf{H}\mathbf{H}^T)$$

where $\lambda_{min}^{LDL} = \lambda_{min}(\mathbf{L}^{-1}\mathbf{D}\mathbf{L}^{-T})$, $\lambda_{max}^{LDL} = \lambda_{max}(\mathbf{L}^{-1}\mathbf{D}\mathbf{L}^{-T})$.

- Preconditioned system is identity plus low rank – smallest eigenvalue is 1.
- How tight/pessimistic is this bound?

R is still causing us problems

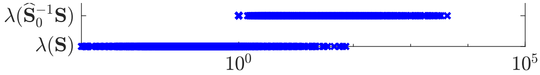


Figure: Eigenvalues of unpreconditioned and preconditioned system, using the level-1 preconditioner $\widehat{\mathbf{S}}_0^{-1}\mathbf{S}$

It is possible to end up with a worse condition number than you started with due to very large eigenvalues!

R is still causing us problems

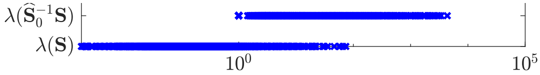


Figure: Eigenvalues of unpreconditioned and preconditioned system, using the level-1 preconditioner $\widehat{\mathbf{S}}_0^{-1}\mathbf{S}$

It is possible to end up with a worse condition number than you started with due to very large eigenvalues!

Can we mitigate the impact of some of these very large eigenvalues in a computationally efficient way?

Existing approaches

Limited memory preconditioner approach
[Daužickaitė et al 2021, Fisher et al 2018]

1. Precondition symmetrically with exact first term: $\mathbf{P}_1 = \mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L}$

$$\mathbf{P}_1^{-1}\mathbf{S} = \mathbf{I} + \mathbf{D}^{1/2}\mathbf{L}^{-\top}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{L}^{-1}\mathbf{D}^{1/2}$$

2. Estimate k leading terms of $\mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^\top \approx \mathbf{D}^{1/2}\mathbf{L}^{-\top}\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H}\mathbf{L}^{-1}\mathbf{D}^{1/2}$
3. $\mathbf{P}_2^{-1} = \mathbf{I} - \mathbf{U}\widetilde{\boldsymbol{\Gamma}}\mathbf{U}^\top$ where $\widetilde{\boldsymbol{\Gamma}}_{ii} = 1 - \frac{1}{\gamma_i}$ for $i = 1, \ldots, k$.

Challenges:

- We have to sketch this term
- Preconditioning with $\mathbf{P}_1^{-1}$ is done via a transformation in the primal form, but not so straightforward in saddle point form
- Restricted to using exact $\mathbf{L}$

Observation low-rank correction (OLC) approach

Propose a preconditioner of the form

$$\mathbf{S}_r = \mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L} + \mathbf{K}_r^\top \mathbf{K}_r, \tag{5}$$

where and $\mathbf{K}_r = \mathbf{\Lambda}_r^{1/2} \mathbf{V}_r^\top \in \mathbb{R}^{r \times s(N+1)}$ defines a rank-r approximation to $\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}$ such that

$$\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H} = \mathbf{V}_r \mathbf{\Lambda}_r \mathbf{V}_r^\top + \widetilde{\mathbf{V}} \widetilde{\mathbf{\Lambda}} \widetilde{\mathbf{V}}^\top.$$

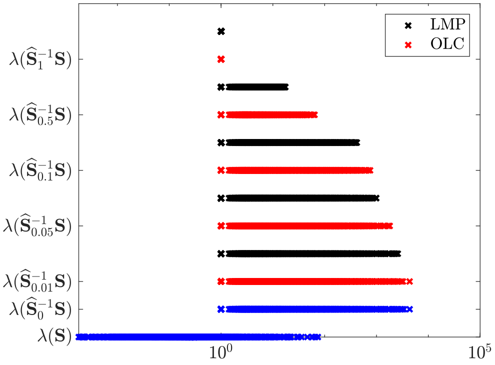
Here, $\mathbf{\Lambda}_r \in \mathbb{R}^{r \times r}$ contains the r leading eigenvalues of $\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}$ (with $r < s(N+1)$), and $\mathbf{V}_r$ the corresponding eigenvectors.
Properties:

- Applied additively rather than multiplicatively – we automatically get symmetry of the updated preconditioner
- No requirement for a square root decomposition of $\mathbf{D}$
- We can exploit the block structure of $\mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}$ – much cheaper to obtain eigenvalue/vector information

Comparison of LMP vs OLC

Both methods:

- Preserve the minimum eigenvalue
- Increase the number of unit eigenvalues by r.
- Can be extended to the case of approximate **L**



Applying OLC efficiently

We may apply the inverse operation of the matrix (5) using the Sherman–Morrison–Woodbury identity via

$$\widehat{\mathbf{S}}^{-1} = \widehat{\mathbf{L}}^{-1}\mathbf{D}\widehat{\mathbf{L}}^{-\top}\left(\mathbf{I}_{s(N+1)} - \mathbf{K}_r^{\top}(\mathbf{I}_r + \mathbf{K}_r\widehat{\mathbf{L}}^{-1}\mathbf{D}\widehat{\mathbf{L}}^{-\top}\mathbf{K}_r^{\top})^{-1}\mathbf{K}_r\widehat{\mathbf{L}}^{-1}\mathbf{D}\widehat{\mathbf{L}}^{-\top}\right).$$

Retain beneficial properties of $\widehat{\mathbf{S}}$

- Re-use approximations/implementations of $\widehat{\mathbf{L}}$ [JMT and Pearson 2023a].
- Inverse is small dimension so can be computed explicitly.
- $\mathbf{K}_r$ also has a block structure.

Zoom in on largest eigenvalues

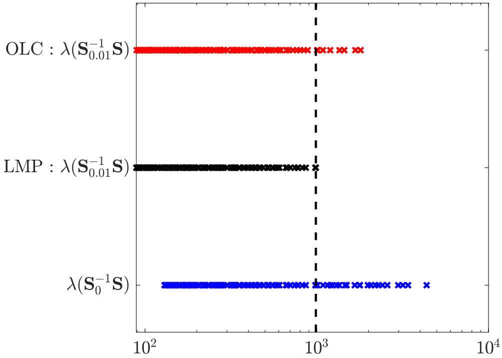


Figure: Zoom in on largest eigenvalues. Dashed line represents 23rd eigenvalues of the first level preconditioned system, $r = 22$.

Including low-rank information improves convergence

r	0	5	10	20	30	40	50
OLC	70	55	44	33	28	24	20
LMP	70	40	34	27	22	19	17
$\mathcal{P}_D$, OLC	67	65	55	43	37	31	27
$\mathcal{P}_D$, LMP	67	37	29	21	17	13	9
$\mathcal{P}_T$, OLC	39	34	29	23	20	18	16
$\mathcal{P}_T$, LMP	39	22	17	12	10	8	6

Table: Convergence for Lorenz 96 problem with p=100, s=400, N=7 using $\widehat{\mathbf{D}} = \mathbf{D}, \widehat{\mathbf{R}} = \mathbf{R}$.

Randomised approach performs similarly in terms of iterations and better in terms of speed.

Extending OLC/LMP to the case of approximate **L**: motivation

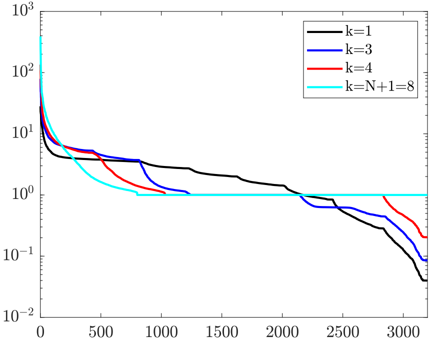


Figure: Spectrum of $\widehat{\mathbf{S}}^{-1}\mathbf{S}$ for $\mathbf{S} = \mathbf{L}_M^T \mathbf{D}^{-1} \mathbf{L}_M$ for different values of k

Extending OLC/LMP to the case of approximate $\mathbf{L}$

Theorem

If $\widehat{\mathbf{L}} \neq \mathbf{L}$ is given as in [JMT and Pearson 2023a] for $k < N+1$ then we can re-write the first-level preconditioned system as

$$\widehat{\mathbf{S}}_0 = \mathbf{I}_{s(N+1)} + \mathbf{D}^{1/2}\widehat{\mathbf{L}}^{-T}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{C})\widehat{\mathbf{L}}^{-1}\mathbf{D}^{1/2}$$

where

$$[\mathbf{C}]_{ij} = \begin{cases} M_i^T Q_i^{-1} M_i & \text{if } i = j \text{ and } k\lfloor \frac{i}{k} \rfloor = i, 1 \leq i,j \leq N-1 \\ -Q_j^{-1} M_j & \text{if } k\lfloor \frac{j}{k} \rfloor = j \text{ and } i = j+1, 1 \leq j \leq N \\ -M_i^T Q_i^{-1} & \text{if } k\lfloor \frac{i}{k} \rfloor = i \text{ and } j = i+1, 1 \leq i \leq N \\ 0 & \text{otherwise} \end{cases}$$

- Each non-zero block of $\mathbf{C}$ has s positive eigenvalues and s negative eigenvalues
- $rank(\mathbf{C}) = 2s\lfloor \frac{N-1}{k} \rfloor$ for $k \geq 2$
- We can prove (pessimistic) upper bounds on the number of observations required for $\mathbf{C} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ to be symmetric indefinite

Extending LMP/OLC

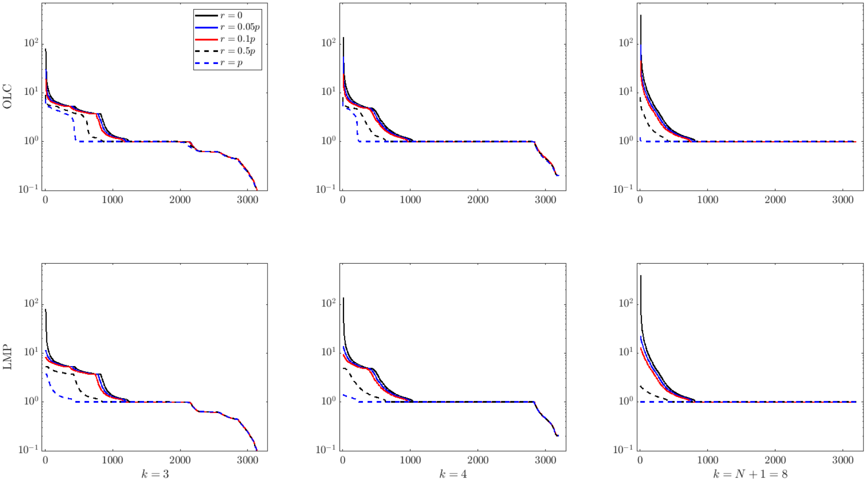
Apply both methods to $\mathbf{C} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$.

- **LMP**: for $k = N + 1$ the second term is SPSD, so we can sketch this using e.g. Nyström and then add 1 to the eigenvalues. Here, we want to sketch the full first-level preconditioned term (as the second term may be indefinite and this is hard to determine a priori)
- ¨: $\mathbf{C} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ has a block diagonal structure - distinguish between blocks
  - $[\mathbf{C}]_{i,j} = 0$ compute/approximate eigendecomposition of $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \in \mathbb{R}^{s\times s}$
  - $[\mathbf{C}]_{i,j} \neq 0$ compute/approximate eigendecomposition of $[\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{C}]_{i:i+1,j+j+1} \in \mathbb{R}^{2s\times 2s}$

Properties:

- r additional unit eigenvalues when using second-level preconditioning
- Small eigenvalues unchanged (smaller than 1) unless r very large

Numerical experiments



Improvement to iterations - $\mathcal{A} \in \mathbb{R}^{7200\times 7200}$

$\mathcal{P}_D$	r	0	5	10	20	30	40	50
OLC	1	113	93	85	81	79	77	77
	3	105	97	83	73	69	67	65
	4	85	79	67	55	49	47	45
	$N+1=8$	67	65	55	43	37	31	27
LMP	1	113	79	71	65	63	63	63
	3	105	69	61	55	53	51	49
	4	85	52	45	37	35	33	33
	$N+1=8$	67	37	29	21	17	13	9

$\mathcal{P}_T$	r	0	5	10	20	30	40	50
OLC	1	70	58	54	50	49	48	47
	3	64	56	50	44	42	41	39
	4	52	44	39	32	29	27	26
	$N+1=8$	39	34	29	23	20	18	16
LMP	1	70	50	44	40	39	39	38
	3	64	45	39	34	33	31	31
	4	52	33	27	23	21	20	19
	$N+1=8$	39	22	17	12	10	8	6

Experimental results

- Still benefit to including the observation term in the Schur complement in the case of approximate $\mathbf{L}$
- Including small amounts of observation information results in fewer iterations than increasing k with $r = 0$ (and might be more computationally affordable) – this could be problem specific
- For same choice of r LMP leads to bigger reduction in iterations
    - Potentially can afford to use larger r for OLC than LMP
    - OLC can be used in the case where $\mathbf{D}^{1/2}$ unavailable/with MINRES for $\mathcal{P}_D$ – not the case for LMP

Conclusions

- New preconditioners for the saddle point formulation of weak-constraint 4D-Var
- Including model information in the preconditioner can reduce iterations, careful parameter choice ensures control over computational cost in terms of matrix-vector products.
- Low-rank correction methods allow us to include some observation information in the Schur complement term
- Presented a new method (OLC) and extended this and LMP to the case of an approximate first term

Outlook and related work

- Alternative choices for $\widehat{\mathbf{L}}$:
  - Replace $\mathbf{M}_i$ with average value $\widehat{\mathbf{M}}$ and exploit Toeplitz structure via solution of matrix equations [Palitta and JMT 2023]
  - Similar to above but using a block-circulant preconditioner $\mathbf{L}$.
- Other preconditioners that avoid the application of $\widehat{\mathbf{D}}^{-1}$ but allow observation information in the Schur complement term

References: *Saddle point preconditioners for weak-constraint 4D-Var*
J. M. Tabeart and J. W. Pearson
https://arXiv:2105.06975
Stein-based Preconditioners for Weak-constraint 4D-Var
D. Palitta and J. M. Tabeart
https://doi.org/10.1016/j.jcp.2023.112068

References I

J. M. Tabeart, S. L. Dance, S. A. Haben, A. S. Lawless, N. K. Nichols, and J. A. Waller (2018)

The conditioning of least squares problems in variational data assimilation.

Numerical Linear Algebra with Applications http://dx.doi.org/10.1002/nla.2165

J. M. Tabeart, S. L. Dance, A. S. Lawless, N. K. Nichols, and J. A. Waller. (2021)

New bounds on the condition number of the Hessian of the preconditioned variational data assimilation problem

Numerical Linear Algebra with Applications,(2021) e2405. https://doi.org/10.1002/nla.2405

J. M. Tabeart and D. Palitta

Stein-based Preconditioners for Weak-constraint 4D-Var

Journal of Computational Physics, 482, (2023) 112068.

J. M. Tabeart, S. L. Dance, A. S. Lawless, N. K. Nichols, and J. A. Waller.

Improving the conditioning of estimated covariance matrices

Tellus A: Dynamic Meteorology and Oceanography, 72, 1 (2020), 1 – 19

References II

L. Stewart (2010)
Correlated observation errors in data assimilation
PhD thesis University of Reading

P. Weston, W. Bell and J. R. Eyre (2014)
Accounting for correlated error in the assimilation of high-resolution sounder data
Q. J. R Met Soc 140, 2420 – 2429.

J. M. Tabeart and J. W. Pearson
Using low-rank observation information to precondition weak-constraint 4D-Var,
in preparation. (2023)

J. M. Tabeart and J. W. Pearson
Saddle point preconditioners for weak-constraint 4D-Var
arXiv preprint arXiv:2105.06975 (2023).

References III

S. Gratton, S. Gürol, E. Simon and P. L. Toint (2018)
Guaranteeing the convergence of the saddle formulation for weakly constrained 4D-Var data assimilation
Quarterly Journal of the Royal Meteorological Society (144), 713, pp 2592–2602

M. Fisher, S. Gratton, S. Gürol, Y. Trémolet and X. Vasseur (2018)
Low rank updates in preconditioning the saddle point systems arising from data assimilation problems
Optimization Methods and Software, 33(1), 45-69.

I. Daužickaitė, A. S.Lawless, J. A. Scott, and P. J. Van Leeuwen (2021)
Randomised preconditioning for the forcing formulation of weak-constraint 4D-Var
Quarterly Journal of the Royal Meteorological Society 147(740), 3719-3734