Received: 25 July 2019 | Revised: 24 April 2020 | Accepted: 11 May 2020

DOI: 10.1002/nla.2313

RESEARCH ARTICLE

WILEY

Spectral estimates for saddle point matrices arising in weak constraint four-dimensional variational data assimilation

Ieva Daužickaitė[1] | **Amos S. Lawless**[1,2] | **Jennifer A. Scott**[1,3] | **Peter Jan van Leeuwen**[1,2,4]

[1]School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, UK

[2]National Centre for Earth Observation, Reading, UK

[3]Scientific Computing Department, STFC Rutherford Appleton Laboratory, Didcot, UK

[4]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

Correspondence
Ieva Daužickaitė, Department of Mathematics and Statistics, University of Reading, PO Box 220, Reading RG6 6AX, UK.
Email: i.dauzickaite@pgr.reading.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: EP/L016613/1; European Research Council CUNDA project, Grant/Award Number: 694509; NERC National Centre for Earth Observation

Summary

We consider the large sparse symmetric linear systems of equations that arise in the solution of weak constraint four-dimensional variational data assimilation, a method of high interest for numerical weather prediction. These systems can be written as saddle point systems with a 3×3 block structure but block eliminations can be performed to reduce them to saddle point systems with a 2×2 block structure, or further to symmetric positive definite systems. In this article, we analyse how sensitive the spectra of these matrices are to the number of observations of the underlying dynamical system. We also obtain bounds on the eigenvalues of the matrices. Numerical experiments are used to confirm the theoretical analysis and bounds.

KEYWORDS

data assimilation, saddle point systems, sparse linear systems, spectral estimates, weak constraint 4D-Var

1 | INTRODUCTION

Data assimilation estimates the state of a dynamical system by combining observations of the system with a prior estimate. The latter is called a background state and it is usually an output of a numerical model that simulates the dynamics of the system. The impact that the observations and the background state have on the state estimate depends on their errors whose statistical properties we assume are known. Data assimilation is used to produce initial conditions in numerical weather prediction (NWP),[1,2] as well as other areas, for example, flood forecasting,[3] research into atmospheric composition,[4] and neuroscience.[5] In operational applications, the process is made more challenging by the size of the system, for example, the numerical model may be operating on 10^8 state variables and 10^5–10^6 observations may be incorporated.[6,7] Moreover, there is usually a constraint on the time that can be spent on calculations.

The solution, called the analysis, is obtained by combining the observations and the background state in an optimal way. One approach is to solve a weighted least-squares problem, which requires minimizing a cost function. An active

research topic in this area is the weak constraint four-dimensional variational (4D-Var) data assimilation method.[8-14] It is employed in the search for states of the system over a time period, called the assimilation window. This method uses a cost function that is formulated under the assumption that the numerical model is not perfect and penalizes the weighted discrepancy between the analysis and the observations, the analysis and the background state, and the difference between the analysis and the trajectory given by integrating the dynamical model.

Effective minimization techniques need evaluations of the cost function and its gradient that involve expensive operations with the dynamical model and its linearized variant. Such approaches are impractical in operational applications. One way to approximate the minimum of the weak constraint 4D-Var is to use an inexact Gauss-Newton method,[15] in which a series of linearized quadratic cost functions with a low resolution model are minimized,[16] and the minima are used to update the high resolution state estimate. The state estimate update is found by solving sparse symmetric linear systems using an iterative method.[17]

To increase the potential of using parallel computations when computing the state update with weak constraint 4D-Var, Fisher and Gürol[12] introduced a symmetric 3×3 block saddle point formulation. The resulting large symmetric linear systems are solved using Krylov subspace solvers.[14,17,18] One criteria that affects their convergence is the spectra of the coefficient matrices.[18] We derive bounds for the eigenvalues of the 3×3 block matrix using the work of Rusten and Winther.[19] In addition, inspired by the practice in solving saddle point systems that arise from interior point methods,[20,21] we reduce the 3×3 block system to a 2×2 block saddle point formulation and derive eigenvalue bounds for this system. We also consider a 1×1 block formulation with a positive definite coefficient matrix, which corresponds to the standard method.[8,9] Some of the blocks in the 3×3 and 2×2 block saddle point coefficient matrices, and the 1×1 block positive definite coefficient matrix depend on the available observations of the dynamical system. We present a novel examination of how adding new observations influences the spectrum of these coefficient matrices.

In Section 2, we formulate the data assimilation problem and introduce weak constraint 4D-Var with the 3×3 block and 2×2 block saddle point formulations and the 1×1 block symmetric positive definite formulation. Eigenvalue bounds for the saddle point and positive definite matrices and results on how the extreme eigenvalues and the bounds depend on the number of observations are presented in Section 3. Section 4 illustrates the theoretical considerations using numerical examples, and concluding remarks and future directions are presented in Section 5.

2 | VARIATIONAL DATA ASSIMILATION

The state of the dynamical system of interest at times $t_0 < t_1 < \cdots < t_N$ is represented by the state vectors $x_0, x_1, \ldots, x_N$ with $x_i \in \mathbb{R}^n$. A nonlinear model m_i that is assumed to have errors describes the transition from the state at time t_i to the state at time t_{i+1}, that is

$$x_{i+1} = m_i(x_i) + \eta_{i+1}, \tag{1}$$

where η_i represents the model error at time t_i. It is further assumed that the model errors are Gaussian with zero mean and covariance matrix $Q_i \in \mathbb{R}^{n \times n}$, and that they are uncorrelated in time, that is, there is no relationship between the model errors at different times. In NWP, the model comes from the discretization of the partial differential equations that describe the flow and thermodynamics of a stratified multiphase fluid in interaction with radiation.[1] It also involves parameters that characterize processes arising at spatial scales that are smaller than the distance between the grid points.[22] Errors due to the discretization of the equations, errors in the boundary conditions, inaccurate parameters, and so on are components of the model error.[23]

The background information about the state at time t_0 is denoted by $x^b \in \mathbb{R}^n$. x^b usually comes from a previous short range forecast and is chosen to be the first guess of the state. It is assumed that the background term has errors that are Gaussian with zero mean and covariance matrix $B \in \mathbb{R}^{n \times n}$.

Observations of the dynamical system at time t_i are given by $y_i \in \mathbb{R}^{p_i}$. In NWP, there are considerably fewer observations than state variables, that is, $p_i \ll n$. In addition, there may be indirect observations of the variables in the state vector and a comparison is obtained by mapping the state variables to the observation space using a nonlinear operator $\mathcal{H}_i$. For example, satellite observations used in NWP provide top of the atmosphere radiance data, whereas the model operates on different meteorological variables, for example, temperature, pressure, wind, and so on.[24] Hence, values of meteorological variables are transformed into top of the atmosphere radiances in order to compare the model output with the observations. In this case, the operator $\mathcal{H}_i$ is nonlinear and complicated. Approximations made when mapping the state variables

to the observation space, different spatial and temporal scales between the model and some observations (observations may give information at a finer resolution than the model), and preprocessing, or quality control, of the observations (see, eg, section 5.8 of Kalnay[1]) comprise the representativity error.[25] The observation error is made up of the representativity error combined with the error arising due to the limited precision of the measurements. It is assumed to be Gaussian with zero mean and covariance matrix $R_i \in \mathbb{R}^{p_i \times p_i}$. The observation errors are assumed to be uncorrelated in time.[7]

2.1 | Weak constraint 4D-Var

In weak constraint 4D-Var, the analysis $x_0^a, x_1^a, \dots, x_N^a$ is obtained by minimizing the following nonlinear cost function

$$J(x_0, x_1, \dots, x_N) = \frac{1}{2}(x_0 - x^b)^T B^{-1}(x_0 - x^b) + \frac{1}{2}\sum_{i=0}^{N}(y_i - \mathcal{H}_i(x_i))^T R_i^{-1}(y_i - \mathcal{H}_i(x_i)) + \frac{1}{2}\sum_{i=0}^{N-1}(x_{i+1} - m_i(x_i))^T Q_{i+1}^{-1}(x_{i+1} - m_i(x_i)). \tag{2}$$

This cost function is referred to as the state control variable formulation. Here, the control variables are defined as the variables with respect to which the cost function is minimized, that is, x_0,x_1,… ,x_N in (2). Choosing different control variables and obtaining different formulations of the cost function is possible.[8] If the model is assumed to have no errors (ie, $x_{i+1} = m_i(x_i)$), the cost function simplifies as the last term in (2) is removed; this is called strong constraint 4D-Var. Rejecting this perfect model assumption and using weak constraint 4D-Var may lead to a better analysis.[9]

Iterative gradient-based optimization methods are used in practical data assimilation.[7,26] These require the cost function and its gradient to be evaluated at every iteration. In operational applications, integrating the model over the assimilation window to evaluate the cost function is computationally expensive. The gradient is obtained by the adjoint method (see, eg, section 2 of Lawless[7] and section 3.2 of Talagrand[26] for an introduction), which is a few times more computationally expensive than evaluating the cost function. This makes the minimisation of the nonlinear weak constraint 4D-Var cost function impractical. Hence, approximations have to be made. We introduce such an approach in the following section.

2.2 | Incremental formulation

Minimisation of the 4D-Var cost function in an operational setting is made feasible by employing an iterative Gauss-Newton method, as first proposed by Courtier et al[16] for the strong constraint 4D-Var. In this approach, the solution of the weak constraint 4D-Var is approximated by solving a sequence of linearised problems, that is, the (*l+1*)th approximation of the state is

$$x_i^{(l+1)} = x_i^{(l)} + \delta x_i^{(l)}, \quad i \in \{0, 1, \dots, N\}, \tag{3}$$

where $\delta x_i^{(l)}$ is obtained as the minimizer of the linearised cost function

$$J^{\delta}(\delta x_0^{(l)}, \delta x_1^{(l)}, \dots, \delta x_N^{(l)}) = (\delta x_0^{(l)} - b^{(l)})^T B^{-1}(\delta x_0^{(l)} - b^{(l)}) + \frac{1}{2}\sum_{i=0}^{N}(H_i^{(l)}\delta x_i^{(l)} - d_i^{(l)})^T R_i^{-1}(H_i^{(l)}\delta x_i^{(l)} - d_i^{(l)}) + \frac{1}{2}\sum_{i=0}^{N-1}(M_i^{(l)}\delta x_i^{(l)} - \delta x_{i+1}^{(l)} - \eta_{i+1}^{(l)})^T Q_{i+1}^{-1}(M_i^{(l)}\delta x_i^{(l)} - \delta x_{i+1}^{(l)} - \eta_{i+1}^{(l)}), \tag{4}$$

where $b^{(l)} = x_0^{(l)} - x^b$, $d_i^{(l)} = y_i - \mathcal{H}_i(x_i^{(l)})$, $\eta_i^{(l)} = x_i^{(l)} - m_{i-1}(x_{i-1}^{(l)})$ (as in (1)) and $M_i^{(l)}$ and $H_i^{(l)}$ are the model m_i and the observation operator $\mathcal{H}_i$, respectively, linearised at $x_i^{(l)}$. Minimisation of (4) is called the inner loop. The lth outer loop consists of updating the approximation of the state (3), linearizing the model m_i and the observation operator $\mathcal{H}_i$, and computing the values $b^{(l)}$, $d_i^{(l)}$, and $\eta_i^{(l)}$. This cost function is quadratic, which allows the use of effective minimisation techniques,

such as conjugate gradients (see Chapter 5 of Nocedal and Wright[27]). In NWP, the computational cost of minimizing the 4D-Var cost function is reduced by using a version of the inner loop cost function that is defined for a model with lower spatial resolution, that is, on a coarser grid.[28] We do not consider such an approach in the subsequent work, because our results on the change of the spectra of the coefficient matrices and the bounds (that are introduced in the following section) hold for models with any spatial resolution.

For ease of notation, we introduce the following 4D (in the sense that they contain information in space and time) vectors and matrices. These vectors and matrices are indicated in bold.

$$\mathbf{x}^{(l)} = \begin{pmatrix} x_0^{(l)} \\ x_1^{(l)} \\ \vdots \\ x_N^{(l)} \end{pmatrix}, \ \delta\mathbf{x}^{(l)} = \begin{pmatrix} \delta x_0^{(l)} \\ \delta x_1^{(l)} \\ \vdots \\ \delta x_N^{(l)} \end{pmatrix}, \mathbf{b}^{(l)} = \begin{pmatrix} b^{(l)} \\ -\eta_1^{(l)} \\ \vdots \\ -\eta_N^{(l)} \end{pmatrix}, \ \mathbf{d}^{(l)} = \begin{pmatrix} y_0 - \mathcal{H}_0(x_0^{(l)}) \\ y_1 - \mathcal{H}_1(x_1^{(l)}) \\ \vdots \\ y_N - \mathcal{H}_N(x_N^{(l)}) \end{pmatrix},$$

where $\mathbf{x}^{(l)}, \delta\mathbf{x}^{(l)}, \mathbf{b}^{(l)} \in \mathbb{R}^{(N+1)n}$, and $\mathbf{d}^{(l)} \in \mathbb{R}^p, p = \Sigma_{i=0}^N p_i$. We also define the matrices

$$\mathbf{L}^{(l)} = \begin{pmatrix} I & & & & \\ -M_0^{(l)} & I & & & \\ & -M_1^{(l)} & I & & \\ & & \ddots & \ddots & \\ & & & -M_{N-1}^{(l)} & I \end{pmatrix}, \quad \mathbf{H}^{(l)} = \begin{pmatrix} H_0^{(l)} & & & \\ & H_1^{(l)} & & \\ & & \ddots & \\ & & & H_N^{(l)} \end{pmatrix},$$

where $I \in \mathbb{R}^{n\times n}$ is the identity matrix, $\mathbf{L}^{(l)} \in \mathbb{R}^{(N+1)n\times(N+1)n}$ and $\mathbf{H}^{(l)} \in \mathbb{R}^{p\times(N+1)n}$. We define the block diagonal covariance matrices

$$\mathbf{D} = \begin{pmatrix} B & & & \\ & Q_1 & & \\ & & \ddots & \\ & & & Q_N \end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} R_0 & & & \\ & R_1 & & \\ & & \ddots & \\ & & & R_N \end{pmatrix},$$

$\mathbf{D} \in \mathbb{R}^{(N+1)n\times(N+1)n}$ and $\mathbf{R} \in \mathbb{R}^{p\times p}$. The state update (3) may then be written as

$$\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} + \delta\mathbf{x}^{(l)},$$

and the quadratic cost function (4) becomes

$$J^\delta(\delta\mathbf{x}^{(l)}) = \frac{1}{2}||\mathbf{L}^{(l)}\delta\mathbf{x}^{(l)} - \mathbf{b}^{(l)}||^2_{\mathbf{D}^{-1}} + \frac{1}{2}||\mathbf{H}^{(l)}\delta\mathbf{x}^{(l)} - \mathbf{d}^{(l)}||^2_{\mathbf{R}^{-1}}, \tag{5}$$

where $||\mathbf{a}||^2_{\mathbf{A}^{-1}} = \mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}$. We omit the superscript (l) for the outer iteration in the subsequent discussions. The minimum of (5) can be found by solving linear systems. We discuss different formulations of these in the next three subsections.

2.2.1 | 3 × 3 block saddle point formulation

In pursuance of exploiting parallel computations in data assimilation, Fisher and Gürol[12] proposed obtaining the state increment $\delta\mathbf{x}$ by solving a saddle point system (see also Freitag and Green[14]). New variables are introduced

$$\boldsymbol{\lambda} = \mathbf{D}^{-1}(\mathbf{b} - \mathbf{L}\delta\mathbf{x}) \ \in \mathbb{R}^{(N+1)n}, \tag{6}$$

$$\boldsymbol{\mu} = \mathbf{R}^{-1}(\mathbf{d} - \mathbf{H}\delta\mathbf{x}) \ \in \mathbb{R}^p. \tag{7}$$

The gradient of the cost function (5) with respect to $\delta\mathbf{x}$ provides the optimality constraint

$$\begin{aligned} \mathbf{0} &= \mathbf{L}^T\mathbf{D}^{-1}(\mathbf{L}\delta\mathbf{x} - \mathbf{b}) + \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{H}\delta\mathbf{x} - \mathbf{d}) \\ &= -(\mathbf{L}^T\boldsymbol{\lambda} + \mathbf{H}^T\boldsymbol{\mu}). \end{aligned} \tag{8}$$

Multiplying (6) by $\mathbf{D}$ and (7) by $\mathbf{R}$ together with (8), yields a coupled linear system of equations:

$$\mathcal{A}_3 \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \\ \delta\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ \mathbf{d} \\ \mathbf{0} \end{pmatrix}, \tag{9}$$

where the coefficient matrix is given by

$$\mathcal{A}_3 = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{R} & \mathbf{H} \\ \mathbf{L}^T & \mathbf{H}^T & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{(2(N+1)n+p)\times(2(N+1)n+p)}. \tag{10}$$

$\mathcal{A}_3$ is a sparse symmetric indefinite saddle point matrix that has a 3×3 block form. Such systems are explored in the optimization literature.[20,21,29] When solving these systems iteratively, it is usually assumed that calculations involving the blocks on the diagonal are computationally expensive, while the off-diagonal blocks are cheap to apply and easily approximated. However, in our application, operations with the diagonal blocks are relatively cheap and the off-diagonal blocks are computationally expensive, particularly because of the integrations of the model and its adjoint in $\mathbf{L}$ and $\mathbf{L}^T$.

Recall that the sizes of the blocks $\mathbf{R}$, $\mathbf{H}$, and $\mathbf{H}^T$ depend on the number of observations p. Thus, the size of $\mathcal{A}_3$ and possibly some of its characteristics are also affected by p. The saddle point systems that arise in different outer loops vary in the right-hand sides and the linearization states of $\mathbf{L}$ and $\mathbf{H}$.

Because of the memory requirements of sparse direct solvers, they cannot be used to solve the 3×3 block saddle point systems that arise in an operational setting. Iterative solvers (such as MINRES,[30] SYMMLQ,[30] GMRES[31]) need to be used. These Krylov subspace methods require matrix-vector products with $\mathcal{A}_3$ at each iteration. Note that the matrix-vector product $\mathcal{A}_3\mathbf{q}$, $\mathbf{q}^T = (q_1^T, q_2^T, q_3^T)$, $q_1, q_3 \in \mathbb{R}^{(N+1)n}, q_2 \in \mathbb{R}^p$, involves multiplying $\mathbf{D}$ and $\mathbf{L}^T$ by q_1, $\mathbf{R}$ and $\mathbf{H}^T$ by q_2, and $\mathbf{L}$ and $\mathbf{H}$ by q_3. These matrix-vector products may be performed in parallel. Furthermore, multiplication of each component of each block matrix with the respective part of the vector q_i can be performed in parallel. The possibility of multiplying a vector with the blocks in $\mathbf{L}$ and $\mathbf{L}^T$ in parallel is particularly attractive, because the expensive operations involving the linearised model M_i and its adjoint M_i^T can be performed at the same time for every $i \in \{0,1,\ldots,N-1\}$.

2.2.2 | 2×2 block saddle point formulation

The saddle point systems with 3×3 block coefficient matrices that arise from interior point methods are often reduced to 2×2 block systems.[20,21] The 2×2 block formulation has not been used in data assimilation before. Because of its smaller size, it may be advantageous. Therefore, we now explore using this approach in the weak constraint 4D-Var setting.

Multiplying Equation (6) by $\mathbf{D}$ and eliminating $\boldsymbol{\mu}$ in (8) gives the following equivalent system of equations

$$\mathcal{A}_2 \begin{pmatrix} \boldsymbol{\lambda} \\ \delta\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{b} \\ -\mathbf{H}^T\mathbf{R}^{-1}\mathbf{d} \end{pmatrix}, \tag{11}$$

where

$$\mathcal{A}_2 = \begin{pmatrix} \mathbf{D} & \mathbf{L} \\ \mathbf{L}^T & -\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \end{pmatrix} \in \mathbb{R}^{2(N+1)n\times 2(N+1)n}. \tag{12}$$

The reduced matrix $\mathcal{A}_2$ is a sparse symmetric indefinite saddle point matrix with a 2×2 block form. Unlike the 3×3 block matrix $\mathcal{A}_3$, its size is independent of the number of observations. $\mathcal{A}_2$ involves the matrix $\mathbf{R}^{-1}$, which is usually available in data assimilation applications. The computationally most expensive blocks $\mathbf{L}$ and $\mathbf{L}^T$ are again the off-diagonal blocks.

Solving (11) in parallel might be less appealing compared with solving (9) in parallel: for a Krylov subspace method, the (2, 2) block $-\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ need not be formed separately, that is, only operators to perform the matrix-vector products with $\mathbf{H}^T$, $\mathbf{R}^{-1}$, and $\mathbf{H}$ need to be stored. Hence, a matrix-vector product $\mathcal{A}_2\mathbf{q}$, $\mathbf{q}^T = (q_1^T, q_3^T)$, $q_1, q_3 \in \mathbb{R}^{(N+1)n}$, requires multiplying $\mathbf{D}$ and $\mathbf{L}^T$ by q_1, $\mathbf{L}$ and $\mathbf{H}$ by q_3 (which may be done in parallel) and subsequently $\mathbf{R}^{-1}$ by $\mathbf{H}q_3$, followed by $-\mathbf{H}^T$ by $\mathbf{R}^{-1}\mathbf{H}q_3$. Hence, the cost of matrix-vector products for the 3×3 and 2×2 block formulations differs in that the former needs matrix-vector products with $\mathbf{R}$ while the latter requires products with $\mathbf{R}^{-1}$, and the 2×2 block formulation

requires some sequential calculations. However, notice that the expensive calculations that involve applying the operators $\mathbf{L}$ and $\mathbf{L}^T$ (the linearised model and its adjoint) can still be performed in parallel.

2.2.3 | 1 × 1 block formulation

The 2 × 2 block system can be further reduced to a 1 × 1 block system, that is, to the standard formulation (see, eg, Trémolet[8] and A. El-Said[10] for a more detailed consideration):

$$(\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\delta\mathbf{x} = \mathbf{L}^T\mathbf{D}^{-1}\mathbf{b} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{d}. \tag{13}$$

Observe that the coefficient matrix

$$\begin{aligned}\mathcal{A}_1 &= \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \\ &= (\mathbf{L}^T \quad \mathbf{H}^T)\begin{pmatrix}\mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{L} \\ \mathbf{H}\end{pmatrix}\end{aligned} \tag{14}$$

is the negative Schur complement of $\begin{pmatrix}\mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}\end{pmatrix}$ in $\mathcal{A}_3$. The matrix $\mathcal{A}_1$ is block tridiagonal and symmetric positive definite, hence the conjugate gradient method by Hestenes and Stiefel[32] can be used. The computations with the linearised model in $\mathbf{L}$ at every time step can again be performed in parallel. However, the adjoint of the linearised model in $\mathbf{L}^T$ can only be applied after the computations with the model are finished, thus limiting the potential for parallelism.

3 | EIGENVALUES OF THE SADDLE POINT FORMULATIONS

One factor that influences the rate of convergence of Krylov subspace iterative solvers for symmetric systems is the spectrum of the coefficient matrix (see, eg, section 9 in the survey article[18] and Lectures 35 and 38 in the textbook[33] for a discussion). Simoncini and Szyld[34] have shown that any eigenvalues of the saddle point systems that lie close to zero can cause the iterative solver MINRES to stagnate for a number of iterations while the rate of convergence can improve if the eigenvalues are clustered.

In the following, we examine how the eigenvalues of the block matrices $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$ change when new observations are added. This is done by considering the shift in the extreme eigenvalues of these matrices, that is, the smallest and largest positive and negative eigenvalues. We then present bounds for the eigenvalues of these matrices.

3.1 | Preliminaries

In order to determine how changing the number of observations influences the spectra of $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$, we explore the extreme singular values and eigenvalues of some blocks in $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$. We state two theorems that we will require. Here, we employ the notation $\lambda_k(A)$ to denote the kth largest eigenvalue of a matrix A and subscripts min and max are used to denote the smallest and largest eigenvalues, respectively.

Theorem 1 (See section 8.1.2 of Golub and Van Loan[35]). *If A and C are $n \times n$ Hermitian matrices, then*
$\lambda_k(A)+\lambda_{\min}(C) \leq \lambda_k(A+C) \leq \lambda_k(A)+\lambda_{\max}(C), k \in \{1,2,\dots,n\}$.

Theorem 2 (Cauchy's Interlace Theorem, see Theorem 4.2 in Chapter 4 of Stewart and Sun[36]). *If A is an $n \times n$ Hermitian matrix and C is a $(n-1)\times(n-1)$ principal submatrix of A (a matrix obtained by eliminating a row and a corresponding column of A), then*
$\lambda_n(A) \leq \lambda_{n-1}(C) \leq \lambda_{n-1}(A) \leq \cdots \leq \lambda_2(A) \leq \lambda_1(C) \leq \lambda_1(A)$.

In the following lemmas we describe how the smallest and largest singular values of $(\mathbf{L}^T\mathbf{H}^T)$ (here $\mathbf{L}$ and $\mathbf{H}$ are as defined in Section 2.2) and the extreme eigenvalues of the observation error covariance matrix $\mathbf{R}$ change when new observations are introduced. The same is done for the largest eigenvalues of $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ assuming that $\mathbf{R}$ is diagonal. In these lemmas the subscript $k \in \{0,1,\dots,(N+1)n-1\}$ denotes the number of available observations and the subscript $k+1$ indicates that a new observation is added to the system with k observations, that is, matrices $\mathbf{R}_k \in \mathbb{R}^{k\times k}$ and $\mathbf{H}_k \in \mathbb{R}^{k\times(N+1)n}$

correspond to a system with k observations and $\mathbf{R}_{k+1}$ and $\mathbf{H}_{k+1}$ correspond to the system with an additional observation. We write $\mathbf{R}_{k+1} = \begin{pmatrix} \mathbf{R}_k & r \\ r^T & \alpha \end{pmatrix}$ and $\mathbf{H}_{k+1} = \begin{pmatrix} \mathbf{H}_k \\ h_{k+1}^T \end{pmatrix}$, where $r \in \mathbb{R}^k$, $\alpha \in \mathbb{R}^1$, $\alpha>0$ and $h_{k+1} \in \mathbb{R}^{(N+1)n}$ correspond to the new observation.

Lemma 1. *Let $\omega_{\min}$ and $\omega_{\max}$ be the smallest and largest singular values of $(\mathbf{L}^T\ \mathbf{H}_k^T)$, and $\phi_{\min}$ and $\phi_{\max}$ be the smallest and largest singular values of $(\mathbf{L}^T\ \mathbf{H}_{k+1}^T)$. Then*

$$\omega_{\min}^2 \leq \phi_{\min}^2 \quad \text{and} \quad \omega_{\max}^2 \leq \phi_{\max}^2$$

that is, the smallest and largest singular values of $(\mathbf{L}^T\mathbf{H}^T)$ increase or are unchanged when new observations are added.

Proof. We consider the eigenvalues of $\mathbf{L}^T\mathbf{L} + \mathbf{H}_k^T\mathbf{H}_k$ and $\mathbf{L}^T\mathbf{L} + \mathbf{H}_{k+1}^T\mathbf{H}_{k+1}$, which are the squares of the singular values of $(\mathbf{L}^T\ \mathbf{H}_k^T)$ and $(\mathbf{L}^T\ \mathbf{H}_{k+1}^T)$, respectively (see section 2.4.2 of Golub and Van Loan[35]). We can write

$$\mathbf{H}_{k+1}^T\mathbf{H}_{k+1} = \begin{pmatrix} \mathbf{H}_k^T & h_{k+1} \end{pmatrix} \begin{pmatrix} \mathbf{H}_k \\ h_{k+1}^T \end{pmatrix} = \mathbf{H}_k^T\mathbf{H}_k + h_{k+1}h_{k+1}^T.$$

Then by Theorem 1,

$$\omega_{\min}^2 + \lambda_{\min}(h_{k+1}h_{k+1}^T) \leq \phi_{\min}^2, \quad k \in \{0, 1, \dots, (N+1)n-1\},$$

and since $h_{k+1}h_{k+1}^T$ is a rank 1 symmetric positive semidefinite matrix, $\lambda_{\min}(h_{k+1}h_{k+1}^T) = 0$.

The proof for the largest singular values is analogous. ∎

Lemma 2. *Consider the observation error covariance matrices $\mathbf{R}_k \in \mathbb{R}^{k\times k}$ and $\mathbf{R}_{k+1} \in \mathbb{R}^{(k+1)\times(k+1)}$. Then*

$$\lambda_{\min}(\mathbf{R}_{k+1}) \leq \lambda_{\min}(\mathbf{R}_k) \quad \text{and} \quad \lambda_{\max}(\mathbf{R}_k) \leq \lambda_{\max}(\mathbf{R}_{k+1}), \quad k \in \{0, 1, \dots, (N+1)n-1\},$$

that is, the largest (respectively, smallest) eigenvalue of $\mathbf{R}$ increases (respectively, decreases), or is unchanged when new observations are introduced.

Proof. When adding an observation, a row and a corresponding column are appended to $\mathbf{R}_k$ while the other entries of $\mathbf{R}_k$ are unchanged. The result is immediate by applying Theorem 2. ∎

Lemma 3. *If the observation errors are uncorrelated, that is, $\mathbf{R}$ is diagonal, then*

$$\lambda_{\max}(\mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k) \leq \lambda_{\max}(\mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{H}_{k+1}), \quad k \in \{0, 1, \dots, (N+1)n-1\},$$

that is, for diagonal $\mathbf{R}$, the largest eigenvalue of $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ increases or is unchanged when new observations are introduced.

Proof. The proof is similar to that of Lemma 1. For diagonal $\mathbf{R}_{k+1}$:

$$\mathbf{R}_{k+1}^{-1} = \begin{pmatrix} \mathbf{R}_k^{-1} & \\ & \alpha^{-1} \end{pmatrix}, \ \alpha > 0.$$

Then

$$\mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{H}_{k+1} = \begin{pmatrix} \mathbf{H}_k^T & h_{k+1} \end{pmatrix} \begin{pmatrix} \mathbf{R}_k^{-1} & \\ & \alpha^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{H}_k \\ h_{k+1}^T \end{pmatrix} = \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k + \alpha^{-1}h_{k+1}h_{k+1}^T.$$

Hence, by Theorem 1,

$$\lambda_{\max}(\mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k) + \alpha^{-1}\lambda_{\min}(h_{k+1}h_{k+1}^T) \leq \lambda_{\max}(\mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{H}_{k+1}), \quad k \in \{0, 1, \dots, (N+1)n-1\},$$

and since $\lambda_{\min}(h_{k+1}h_{k+1}^T) = 0$ the result is proved. ∎

Matrix	$\mathcal{A}_3$	$\mathcal{A}_2$	$\mathcal{A}_1$	$\mathbf{D}$	$\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$	$\mathbf{R}$
Eigenvalue	γ_i	ζ_i	χ_i	ψ_i	ν_i	ρ_i
Matrix	$(\mathbf{L}^T\mathbf{H}^T)$	$\mathbf{L}$				
Singular value	θ_i	σ_i				

TABLE 1 Notation for the eigenvalues and singular values

Notation

In the following, we use the notation given in Table 1 for the eigenvalues of $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$, and the eigenvalues and singular values of the blocks within them. We use subscripts min and max to denote the smallest and largest eigenvalues or singular values, respectively, and $\theta_{\min}$ denote the smallest nonzero singular value of $(\mathbf{L}^T\mathbf{H}^T)$. In addition, $||\cdot||$ denotes the L_2 norm.

We also use

$$\tau_{\min} = \min\{\psi_{\min}, \rho_{\min}\}, \tag{15}$$

$$\tau_{\max} = \max\{\psi_{\max}, \rho_{\max}\}. \tag{16}$$

3.2 | Bounds for the 3×3 block formulation

To determine the numbers of positive and negative eigenvalues of $\mathcal{A}_3$ given in (10), we write $\mathcal{A}_3$ as a congruence transformation

$$\mathcal{A}_3 = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{R} & \mathbf{H} \\ \mathbf{L}^T & \mathbf{H}^T & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} & \mathbf{0} \\ \mathbf{L}^T & \mathbf{H}^T & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L} - \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \end{pmatrix} \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{R} & \mathbf{H} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{pmatrix} = \hat{\mathbf{L}}\hat{\mathbf{B}}\hat{\mathbf{L}}^T,$$

where $\mathbf{I} \in \mathbb{R}^{(N+1)n \times (N+1)n}$ is the identity matrix. Thus, by Sylvester's law of inertia (see Section 8.1.5 of Golub and Van Loan[35]), $\mathcal{A}_3$ and $\hat{\mathbf{B}}$ have the same inertia, that is, the same number of positive, negative, and zero eigenvalues. Since the blocks $\mathbf{D}^{-1}$, $\mathbf{R}^{-1}$ and $\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} = \mathcal{A}_1$ are symmetric positive definite matrices, $\mathcal{A}_3$ has $(N+1)n+p$ positive and $(N+1)n$ negative eigenvalues. In the following theorem, we explore how the extreme eigenvalues of $\mathcal{A}_3$ change when new observations are introduced.

Theorem 3. *The smallest and largest negative eigenvalues of $\mathcal{A}_3$ either move away from zero or are unchanged when new observations are introduced. The same holds for the largest positive eigenvalue, while the smallest positive eigenvalue approaches zero or is unchanged.*

Proof. Let $\mathcal{A}_{3,k}$ denote $\mathcal{A}_3$ where $p = k$. To account for an additional observation, a row and a corresponding column is added to $\mathcal{A}_3$, hence $\mathcal{A}_{3,k}$ is a principal submatrix of $\mathcal{A}_{3,k+1}$. Let

$$\lambda_{-(N+1)n}(\mathcal{A}_{3,k}) \le \lambda_{-((N+1)n-1)}(\mathcal{A}_{3,k}) \le \cdots \le \lambda_{-1}(\mathcal{A}_{3,k}) < 0 < \lambda_1(\mathcal{A}_{3,k}) \le \cdots \le \lambda_{(N+1)n+k}(\mathcal{A}_{3,k})$$

be the eigenvalues of $\mathcal{A}_{3,k}$, and

$$\lambda_{-(N+1)n}(\mathcal{A}_{3,k+1}) \le \lambda_{-((N+1)n-1)}(\mathcal{A}_{3,k+1}) \le \cdots \le \lambda_{-1}(\mathcal{A}_{3,k+1}) < 0 < \lambda_1(\mathcal{A}_{3,k+1}) \le \cdots \le \lambda_{(N+1)n+k+1}(\mathcal{A}_{3,k+1})$$

be the eigenvalues of $\mathcal{A}_{3,k+1}$. Then by Theorem 2:

$$\begin{aligned}
&\text{smallest negative eigenvalues}: && \lambda_{-(N+1)n}(\mathcal{A}_{3,k+1}) \le \lambda_{-(N+1)n}(\mathcal{A}_{3,k}),\\
&\text{largest negative eigenvalues}: && \lambda_{-1}(\mathcal{A}_{3,k+1}) \le \lambda_{-1}(\mathcal{A}_{3,k}),\\
&\text{smallest positive eigenvalues}: && \lambda_1(\mathcal{A}_{3,k+1}) \le \lambda_1(\mathcal{A}_{3,k}),\\
&\text{largest positive eigenvalues}: && \lambda_{(N+1)n+k}(\mathcal{A}_{3,k}) \le \lambda_{(N+1)n+k+1}(\mathcal{A}_{3,k+1}).
\end{aligned}$$

■

To obtain information on not only how the eigenvalues of $\mathcal{A}_3$ change because of new observations, but also on where the eigenvalues lie when the number of observations is fixed, we formulate intervals for the negative and positive eigenvalues of $\mathcal{A}_3$.

Theorem 4. *The negative eigenvalues of $\mathcal{A}_3$ lie in the interval*

$$I_- = \left[\frac{1}{2}\left(\tau_{\min} - \sqrt{\tau_{\min}^2 + 4\theta_{\max}^2}\right), \frac{1}{2}\left(\tau_{\max} - \sqrt{\tau_{\max}^2 + 4\theta_{\min}^2}\right)\right] \tag{17}$$

and the positive eigenvalues lie in the interval

$$I_+ = \left[\tau_{\min}, \frac{1}{2}\left(\tau_{\max} + \sqrt{\tau_{\max}^2 + 4\theta_{\max}^2}\right)\right], \tag{18}$$

where $\tau_{\min}$,$\tau_{\max}$, and θ_i are defined in (15), (16), and Table 1.

Proof. Lemma 2.1 of Rusten and Winther[19] gives eigenvalue intervals for matrices of the form $A = \begin{pmatrix} C & E \\ E^T & 0 \end{pmatrix}$. Applying these intervals in the case $C = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$ and $E^T = \begin{pmatrix} \mathbf{L}^T & \mathbf{H}^T \end{pmatrix}$ yields the required results. ∎

We present two corollaries that describe how the bounds in Theorem 4 change if additional observations are introduced and conclude that the change of the bounds is consistent with the results in Theorem 3.

Corollary 1. *The interval for the positive eigenvalues of $\mathcal{A}_3$ in (18) either expands or is unchanged when new observations are added.*

Proof. First, consider the positive upper bound $\frac{1}{2}\left(\tau_{\max} + \sqrt{\tau_{\max}^2 + 4\theta_{\max}^2}\right)$. By Lemma 1, $\theta_{\max}^2$ increases or is unchanged when additional observations are included. If $\tau_{\max} = \rho_{\max}$, the same holds for $\tau_{\max}$ (by Lemma 2). If $\tau_{\max}$=$\psi_{\max}$, changing the number of observations does not affect $\tau_{\max}$. Hence, the positive upper bound increases or is unchanged.

The positive lower bound $\tau_{\min}$ is unaltered if $\tau_{\min}$=$\psi_{\min}$. If $\tau_{\min}$=$\rho_{\min}$, the bound decreases or is unchanged by Lemma 2. ∎

Corollary 2. *If $\tau_{\max}$=$\psi_{\max}$, the upper bound for the negative eigenvalues of $\mathcal{A}_3$ in (17) is either unchanged or moves away from zero when new observations are added. If $\tau_{\min}$=$\psi_{\min}$, the same holds for the lower bound for negative eigenvalues in (17).*

Proof. The results follow from the facts that $\psi_{\max}$ and $\psi_{\min}$ do not change if observations are added, whereas $\theta_{\min}$ and $\theta_{\max}$ increase or are unchanged by Lemma 1. ∎

If $\tau_{\max} = \rho_{\max}$ or $\tau_{\min} = \rho_{\min}$, it is unclear how the interval for the negative eigenvalues in (17) changes, because $\sqrt{\tau_{\min}^2 + 4\theta_{\max}^2}$ can increase, decrease or be unchanged, and both $\tau_{\max}$ and $\sqrt{\tau_{\max}^2 + 4\theta_{\min}^2}$ can increase or be unchanged.

3.3 | Bounds for the 2 × 2 block formulation

$\mathcal{A}_2$ given in (12) is equal to the following congruence transformation

$$\mathcal{A}_2 = \begin{pmatrix} \mathbf{D} & \mathbf{L} \\ \mathbf{L}^T & -\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \end{pmatrix} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{L}^T & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L} - \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \end{pmatrix} \begin{pmatrix} \mathbf{D} & \mathbf{L} \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I} \in \mathbb{R}^{(N+1)n \times (N+1)n}$ is the identity matrix. Then by Sylvester's law, $\mathcal{A}_2$ has $(N+1)n$ positive and $(N+1)n$ negative eigenvalues. The change of the extreme negative and positive eigenvalues of $\mathcal{A}_2$ due to the additional observations is analyzed in the subsequent theorem. However, the result holds only in the case of uncorrelated observation errors, unlike the general analysis for $\mathcal{A}_3$ in Theorem 3.

Theorem 5. *If the observation errors are uncorrelated, that is,* $\mathbf{R}$ *is diagonal, then the smallest and largest negative eigenvalues of* $\mathcal{A}_2$ *either move away from zero or are unchanged when new observations are added. Contrarily, the smallest and largest positive eigenvalues of* $\mathcal{A}_2$ *approach zero or are unchanged.*

Proof. Matrices $\mathbf{D}$ and $\mathbf{L}$ do not depend on the number of observations. In Lemma 3, we have shown that $\mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{H}_{k+1} = \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k + \alpha^{-1}h_{k+1}h_{k+1}^T$, $(\alpha > 0)$ for diagonal $\mathbf{R}$. Hence, when $\mathcal{A}_{2,k}$ denotes $\mathcal{A}_2$ with $p{=}k$, we can write

$$\mathcal{A}_{2,k+1} = \mathcal{A}_{2,k} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\alpha^{-1}h_{k+1}h_{k+1}^T \end{pmatrix} = \mathcal{A}_{2,k} + \mathcal{E}_2,$$

where $\mathcal{E}_2$ has negative and zero eigenvalues. Let

$$\lambda_{-(N+1)n}(\mathcal{A}_{2,k}) \le \cdots \le \lambda_{-1}(\mathcal{A}_{2,k}) < 0 < \lambda_1(\mathcal{A}_{2,k}) \le \cdots \le \lambda_{(N+1)n}(\mathcal{A}_{2,k})$$

be the eigenvalues of $\mathcal{A}_{2,k}$, and

$$\lambda_{-(N+1)n}(\mathcal{A}_{2,k+1}) \le \cdots \le \lambda_{-1}(\mathcal{A}_{2,k+1}) < 0 < \lambda_1(\mathcal{A}_{2,k+1}) \le \cdots \le \lambda_{(N+1)n}(\mathcal{A}_{2,k+1})$$

be the eigenvalues of $\mathcal{A}_{2,k+1}$. By Theorem 1,

$$\begin{aligned}
&\text{smallest negative eigenvalues}: \quad \lambda_{-(N+1)n}(\mathcal{A}_{2,k}) - \alpha^{-1}\lambda_{\max}(h_{k+1}h_{k+1}^T) \le \lambda_{-(N+1)n}(\mathcal{A}_{2,k+1}) \le \lambda_{-(N+1)n}(\mathcal{A}_{2,k}),\\
&\text{largest negative eigenvalues}: \quad \lambda_{-1}(\mathcal{A}_{2,k}) - \alpha^{-1}\lambda_{\max}(h_{k+1}h_{k+1}^T) \le \lambda_{-1}(\mathcal{A}_{2,k+1}) \le \lambda_{-1}(\mathcal{A}_{2,k}),\\
&\text{smallest positive eigenvalues}: \quad \lambda_1(\mathcal{A}_{2,k}) - \alpha^{-1}\lambda_{\max}(h_{k+1}h_{k+1}^T) \le \lambda_1(\mathcal{A}_{2,k+1}) \le \lambda_1(\mathcal{A}_{2,k}),\\
&\text{largest positive eigenvalues}: \quad \lambda_{(N+1)n}(\mathcal{A}_{2,k}) - \alpha^{-1}\lambda_{\max}(h_{k+1}h_{k+1}^T) \le \lambda_{(N+1)n}(\mathcal{A}_{2,k+1}) \le \lambda_{(N+1)n}(\mathcal{A}_{2,k}).
\end{aligned}$$

■

We further search for the intervals in which the negative and positive eigenvalues of $\mathcal{A}_2$ lie. We follow a similar line of thought as in Silvester and Wathen,[37] with the energy arguments for any nonzero vector $\mathbf{w} \in \mathbb{R}^{(N+1)n}$

$$\psi_{\min}||\mathbf{w}||^2 \le \mathbf{w}^T\mathbf{D}\mathbf{w} \le \psi_{\max}||\mathbf{w}||^2, \tag{19}$$

$$-\nu_{\max}||\mathbf{w}||^2 \le -\mathbf{w}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{w} \le -\nu_{\min}||\mathbf{w}||^2, \tag{20}$$

$$\sigma_{\min}||\mathbf{w}|| \le ||\mathbf{L}^T\mathbf{w}|| \le \sigma_{\max}||\mathbf{w}||, \tag{21}$$

$$\theta_{\min}||\mathbf{w}|| \le ||(\mathbf{L}^T\ \mathbf{H}^T)^T\mathbf{w}|| \le \theta_{\max}||\mathbf{w}||. \tag{22}$$

Theorem 6. *The negative eigenvalues of* $\mathcal{A}_2$ *lie in the interval*

$$I_- = \left[\frac{1}{2}\left(\psi_{\min} - \nu_{\max} - \sqrt{(\psi_{\min} + \nu_{\max})^2 + 4\sigma_{\max}^2}\right), \min\{\beta_1, \max\{\beta_2, \beta_3\}\}\right], \tag{23}$$

where

$$\beta_1 = \frac{1}{2}\left(\psi_{\max} - \nu_{\min} - \sqrt{(\psi_{\max} + \nu_{\min})^2 + 4\sigma_{\min}^2}\right), \tag{24}$$

$$\beta_2 = -\rho_{\max}^{-1}\theta_{\min}^2, \tag{25}$$

$$\beta_3 = \frac{1}{2}\left(\psi_{\max} - \sqrt{\psi_{\max}^2 + 4\theta_{\min}^2}\right), \tag{26}$$

and the positive ones lie in the interval

$$I_+ = \left[\frac{1}{2}\left(\psi_{\min} - \nu_{\max} + \sqrt{(\psi_{\min} + \nu_{\max})^2 + 4\sigma_{\min}^2}\right), \frac{1}{2}\left(\psi_{\max} - \nu_{\min} + \sqrt{(\psi_{\max} + \nu_{\min})^2 + 4\sigma_{\max}^2}\right)\right]. \tag{27}$$

Proof. Assume that $(\mathbf{u}^T, \mathbf{v}^T)^T$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{(N+1)n}$ is an eigenvector of $\mathcal{A}_2$ with an eigenvalue ζ. Then the eigenvalue equations are

$$\mathbf{D}\mathbf{u} + \mathbf{L}\mathbf{v} = \zeta\mathbf{u}, \tag{28}$$

$$\mathbf{L}^T\mathbf{u} - \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{v} = \zeta\mathbf{v}. \tag{29}$$

We note that if $\mathbf{u} = \mathbf{0}$ then $\mathbf{v} = \mathbf{0}$ by (28) and if $\mathbf{v} = \mathbf{0}$ then $\mathbf{u} = \mathbf{0}$ by (29). Hence, $\mathbf{u}, \mathbf{v} \neq \mathbf{0}$.

First, we consider $\zeta > 0$. Equation (29) gives $\mathbf{v} = (\mathbf{I}\zeta + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{L}^T\mathbf{u}$, where $\mathbf{I} \in \mathbb{R}^{(N+1)n \times (N+1)n}$. The matrix $\mathbf{I}\zeta + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ is positive definite, hence nonsingular. We multiply (28) by $\mathbf{u}^T$ and use the previous expression for $\mathbf{v}$ to get

$$\mathbf{u}^T\mathbf{D}\mathbf{u} + \mathbf{u}^T\mathbf{L}(\mathbf{I}\zeta + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{L}^T\mathbf{u} = \zeta||\mathbf{u}||^2. \tag{30}$$

The eigenvalues of $(\mathbf{I}\zeta + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}$ in increasing order are $(\zeta + \nu_{\max})^{-1}, \ldots, (\zeta + \nu_{\min})^{-1}$. Then

$$\begin{aligned}\mathbf{u}^T\mathbf{L}(\mathbf{I}\zeta + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{L}^T\mathbf{u} &\geq \frac{1}{\zeta + \nu_{\max}}||\mathbf{L}^T\mathbf{u}||^2 \\ &\geq \frac{1}{\zeta + \nu_{\max}}\sigma_{\min}^2||\mathbf{u}||^2 \quad \text{[by (21)]}.\end{aligned}$$

Hence, this inequality together with (19) and (30) gives

$$\zeta||\mathbf{u}||^2 \geq \psi_{\min}||\mathbf{u}||^2 + \frac{1}{\zeta + \nu_{\max}}\sigma_{\min}^2||\mathbf{u}||^2$$

and solving

$$\zeta^2 + (\nu_{\max} - \psi_{\min})\zeta - \psi_{\min}\nu_{\max} - \sigma_{\min}^2 \geq 0$$

results in

$$\zeta \geq \frac{1}{2}\left(\psi_{\min} - \nu_{\max} + \sqrt{(\psi_{\min} + \nu_{\max})^2 + 4\sigma_{\min}^2}\right).$$

Similarly, using the upper bound from (19) and employing (30) yields the upper bound

$$\zeta \leq \frac{1}{2}\left(\psi_{\max} - \nu_{\min} + \sqrt{(\psi_{\max} + \nu_{\min})^2 + 4\sigma_{\max}^2}\right).$$

Now consider the case $\zeta < 0$. Since $\mathbf{D} - \zeta\mathbf{I}$ is positive definite, from (28) $\mathbf{u} = -(\mathbf{D} - \zeta\mathbf{I})^{-1}\mathbf{L}\mathbf{v}$. Using this expression and multiplying (29) by $\mathbf{v}^T$ gives

$$-\zeta||\mathbf{v}||^2 = \mathbf{v}^T\mathbf{L}^T(\mathbf{D} - \zeta\mathbf{I}_{(N+1)n})^{-1}\mathbf{L}\mathbf{v} + \mathbf{v}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{v}. \tag{31}$$

Then using (20), (21) and the fact that the smallest eigenvalue of $(\mathbf{D} - \zeta\mathbf{I})^{-1}$ is $(\psi_{\max} - \zeta)^{-1}$ results in the inequality

$$-\zeta||\mathbf{v}||^2 \geq \sigma_{\min}^2||\mathbf{v}||^2\frac{1}{\psi_{\max} - \zeta} + \nu_{\min}||\mathbf{v}||^2,$$

which can be expressed as

$$\zeta^2 - (\psi_{\max} - \nu_{\min})\zeta - \nu_{\min}\psi_{\max} - \sigma_{\min}^2 \geq 0,$$

and its solution gives the upper bound

$$\zeta \leq \frac{1}{2}\left(\psi_{\max} - \nu_{\min} - \sqrt{(\psi_{\max} + \nu_{\min})^2 + 4\sigma_{\min}^2}\right) = \beta_1. \tag{32}$$

Notice that the bound (32) takes into account information on observations only if the system is fully observed. Otherwise, $p<(N+1)n$ and $\nu_{\min}=0$.

We obtain an alternative upper bound for the negative eigenvalues that depends on the observational information and might be useful for the fully observed case, too. Equation (31) may be written as

$$-\zeta||\mathbf{v}||^2 = \mathbf{v}^T(\mathbf{L}^T\ \mathbf{H}^T)\begin{pmatrix}(\mathbf{D}-\zeta\mathbf{I})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{L}\\ \mathbf{H}\end{pmatrix}\mathbf{v}.$$

Eigenvalues of the 2×2 block matrix in the previous equation are the eigenvalues of $(\mathbf{D}-\zeta\mathbf{I})^{-1}$ and $\mathbf{R}^{-1}$. Thus, by an energy argument (19),

$$\begin{aligned}-\zeta||\mathbf{v}||^2 &\geq \min\{\rho_{\max}^{-1}, (-\zeta+\psi_{\max})^{-1}\}||(\mathbf{L}^T\ \mathbf{H}^T)^T\mathbf{v}||^2\\ &\geq \min\{\rho_{\max}^{-1}, (-\zeta+\psi_{\max})^{-1}\}\theta_{\min}^2||\mathbf{v}||^2 \quad \text{[by (22)]}.\end{aligned}$$

Hence,

$$\zeta \leq -\theta_{\min}^2\iota,$$

where $\iota = \min\{\rho_{\max}^{-1}, (-\zeta+\psi_{\max})^{-1}\}$. If $\iota = \rho_{\max}^{-1}$, the upper bound is

$$\zeta \leq -\rho_{\max}^{-1}\theta_{\min}^2 = \beta_2.$$

If $\iota = (-\zeta+\psi_{\max})^{-1}$, the following inequality

$$\zeta^2 - \psi_{\max}\zeta - \theta_{\min}^2 \geq 0$$

gives the bound

$$\zeta \leq \frac{1}{2}\left(\psi_{\max} - \sqrt{\psi_{\max}^2 + 4\theta_{\min}^2}\right) = \beta_3.$$

Hence,

$$\zeta \leq \max\{\beta_2, \beta_3\}. \tag{33}$$

The required upper bound follows from (32) and (33).

Next, we obtain the lower bound for the negative eigenvalues. Using Equation (31) with the largest eigenvalue of $(\mathbf{D}-\zeta\mathbf{I})^{-1}$ and other parts of (20) and (21) yields

$$-\zeta||\mathbf{v}||^2 \leq \sigma_{\max}^2||\mathbf{v}||^2\frac{1}{\psi_{\min}-\zeta} + \nu_{\max}||\mathbf{v}||^2.$$

Solving

$$\zeta^2 - (\psi_{\min}-\nu_{\max})\zeta - \nu_{\max}\psi_{\min} - \sigma_{\max}^2 \leq 0$$

results in

$$\zeta \geq \frac{1}{2}\left(\psi_{\min} - \nu_{\max} - \sqrt{(\psi_{\min}+\nu_{\max})^2 + 4\sigma_{\max}^2}\right).$$

■

We observe that if the system is not fully observed, then $p<(N+1)n$ and $\nu_{\min}=0$, and the upper bound for the positive eigenvalues and the upper bound for the negative eigenvalues (24) in Theorem 6 reduce to (2.11) and (2.13) of Silvester and Wathen.[37]

We are interested in how the bounds in Theorem 6 change if additional observations are introduced. The change to the upper negative bound in (23) depends on which of (24), (25), or (26) gives the bound. Hence, in Corollary 3 we comment on when (26) is larger than (25) and Corollary 4 describes a setting when the negative upper bound is given by (26).

Corollary 3.

$$\max\{\beta_2, \beta_3\} = \beta_3 \quad \Leftrightarrow \frac{1}{2}(\psi_{\max} + \sqrt{\psi_{\max}^2 + \theta_{\min}^2}) \geq \rho_{\max}.$$

Proof. $\max\{\beta_2,\beta_3\} = \beta_3$ if and only if

$$\frac{1}{2}\left(\psi_{\max} - \sqrt{\psi_{\max}^2 + 4\theta_{\min}^2}\right) \geq -\rho_{\max}^{-1}\theta_{\min}^2.$$

Rearranging this inequality gives

$$\psi_{\max} + 2\rho_{\max}^{-1}\theta_{\min}^2 \geq \sqrt{\psi_{\max}^2 + 4\theta_{\min}^2}.$$

Squaring both sides with further rearrangement results in

$$\theta_{\min}^2(\rho_{\max}^{-1}\psi_{\max} + \rho_{\max}^{-2}\theta_{\min}^2 - 1) \geq 0.$$

Since $\theta_{\min}^2 > 0$, this is equivalent to

$$\rho_{\max}^2 - \rho_{\max}\psi_{\max} - \theta_{\min}^2 \leq 0,$$

from which it follows that

$$\rho_{\max} \leq \frac{1}{2}\left(\psi_{\max} + \sqrt{\psi_{\max}^2 + 4\theta_{\min}^2}\right).$$

■

Corollary 3 can be used to check if the assumption in the following corollary holds.

Corollary 4. *If the system is not fully observed and* $\max\{\beta_2,\beta_3\} = \beta_3$*, then the upper bound for the negative eigenvalues of* $\mathcal{A}_2$ *is given by (26).*

Proof. The singular values of $\mathbf{L}$ and $(\mathbf{L}^T\mathbf{H}^T)$ are the square roots of the eigenvalues of $\mathbf{L}^T\mathbf{L}$ and $\mathbf{L}^T\mathbf{L}+\mathbf{H}^T\mathbf{H}$, respectively. Hence, by Theorem 1,

$$\sigma_{\min}^2 + \lambda_{\min}(\mathbf{H}^T\mathbf{H}) \leq \theta_{\min}^2,$$

where $\lambda_{\min}(\mathbf{H}^T\mathbf{H}) \geq 0$, since $\mathbf{H}^T\mathbf{H}$ is symmetric positive semidefinite. In addition, if $p<(N+1)n$, then $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ is singular, that is, $\nu_{\min}=0$, and from (24) and (26)

$$\beta_1 = \frac{1}{2}\left(\psi_{\max} - \sqrt{\psi_{\max}^2 + 4\sigma_{\min}^2}\right) \geq \frac{1}{2}\left(\psi_{\max} - \sqrt{\psi_{\max}^2 + 4\theta_{\min}^2}\right) = \beta_3 = \max\{\beta_2, \beta_3\}.$$

■

We further describe how the negative upper bound changes if it is given by (24) or (26), including the case described in Corollary 4.

Corollary 5. *If the upper bound for the negative eigenvalues of* $\mathcal{A}_2$ *in (23) is given by* β_1 *or* β_3*, then the bound moves away from zero or stays the same when new observations are added.*

Proof. β_1 does not change while the system is not fully observed. When the system becomes fully observed, $\nu_{\min}>0$ and β_1 decreases. β_3 decreases or stays the same by Lemma 1. ■

Note that if the negative upper bound in (23) is given by β_2, it is unclear how the bound changes with the number of observations, since both $\rho_{\max}$ and $\theta^2_{\min}$ increase or stay the same. The same is true for the positive bounds in (27). Only $\nu_{\max}$ and $\nu_{\min}$ depend on the available observations and they are contained in elements with positive and negative signs.

The result in Corollary 5 that applies for $\mathcal{A}_2$ with a general $\mathbf{R}$ is consistent with the result in Theorem 5 that considers $\mathcal{A}_2$ with a diagonal $\mathbf{R}$. The same holds for the result in the following corollary, that determines how the lower bound for the negative eigenvalues of $\mathcal{A}_2$ changes in the special case of uncorrelated observational errors.

Corollary 6. *If the observation error covariance matrix* $\mathbf{R}$ *is diagonal, the negative lower bound in (23) moves away from zero or stays the same when additional observations are introduced.*

Proof. The result follows by applying Lemma 3 to see how $\nu_{\max}$ changes. ∎

In the following corollary, we consider the intervals for the positive eigenvalues of $\mathcal{A}_3$ and $\mathcal{A}_2$ with a fixed number of observations. It suggests that we may expect the positive eigenvalues of $\mathcal{A}_2$ to be more clustered than those of $\mathcal{A}_3$.

Corollary 7. *The interval for the positive eigenvalues of* $\mathcal{A}_2$ *is contained in the interval for the positive eigenvalues of* $\mathcal{A}_3$*, that is,*

$$\left[\frac{1}{2}\left(\psi_{\min} - \nu_{\max} + \sqrt{(\psi_{\min} + \nu_{\max})^2 + 4\sigma^2_{\min}}\right), \frac{1}{2}\left(\psi_{\max} - \nu_{\min} + \sqrt{(\psi_{\max} + \nu_{\min})^2 + 4\sigma^2_{\max}}\right)\right] \subseteq \left[\tau_{\min}, \frac{1}{2}\left(\tau_{\max} + \sqrt{\tau^2_{\max} + 4\theta^2_{\max}}\right)\right].$$

Proof. As observed in Corollary 4,

$$\sigma^2_{\max} + \lambda_{\min}(\mathbf{H}^T\mathbf{H}) \leq \theta^2_{\max},$$

with $\lambda_{\min}(\mathbf{H}^T\mathbf{H}) \geq 0$. In addition, by definition $\tau_{\max} \geq \psi_{\max}$ and the following inequality for the upper bound for the positive eigenvalues of $\mathcal{A}_3$ holds

$$\frac{1}{2}\left(\tau_{\max} + \sqrt{\tau^2_{\max} + 4\theta^2_{\max}}\right) \geq \frac{1}{2}\left(\psi_{\max} + \sqrt{\psi^2_{\max} + 4\theta^2_{\max}}\right).$$

Thus, we show that the upper bound for positive eigenvalues of $\mathcal{A}_3$ is larger than the upper bound for positive eigenvalues of $\mathcal{A}_2$:

$$\begin{aligned}
&\frac{1}{2}\left(\psi_{\max} + \sqrt{\psi^2_{\max} + 4\theta^2_{\max}}\right) \geq \frac{1}{2}\left(\psi_{\max} - \nu_{\min} + \sqrt{(\psi_{\max} + \nu_{\min})^2 + 4\sigma^2_{\max}}\right)\\
&\quad \Leftrightarrow\ \nu_{\min} + \sqrt{\psi^2_{\max} + 4\theta^2_{\max}} \geq \sqrt{(\psi_{\max} + \nu_{\min})^2 + 4\sigma^2_{\max}}\\
&\quad \text{(squaring both sides and simplifying)}\ \Leftrightarrow\ 2\theta^2_{\max} + \nu_{\min}\sqrt{\psi^2_{\max} + 4\theta^2_{\max}} \geq \psi_{\max}\nu_{\min} + 2\sigma^2_{\max}\\
&\quad \text{(rearranging)}\quad \Leftrightarrow\ 2(\theta^2_{\max} - \sigma^2_{\max}) \geq \nu_{\min}(\psi_{\max} - \sqrt{\psi^2_{\max} + 4\theta^2_{\max}}).
\end{aligned} \tag{34}$$

Inequality (34) always holds because the left-hand side is positive and the right-hand side is negative.

We also show that the lower bound for the positive eigenvalues of $\mathcal{A}_3$ is smaller than the lower bound for the positive eigenvalues of $\mathcal{A}_2$:

$$\tau_{\min} \leq \frac{1}{2}\left(\psi_{\min} - \nu_{\max} + \sqrt{(\psi_{\min} + \nu_{\max})^2 + 4\sigma^2_{\min}}\right).$$

Note that by definition $\tau_{\min} \leq \psi_{\min}$ and the following inequality always holds

$$\psi_{\min} \leq \frac{1}{2}\left(\psi_{\min} - \nu_{\max} + \sqrt{(\psi_{\min} + \nu_{\max})^2 + 4\sigma^2_{\min}}\right),$$

because it can be simplified to

$$\begin{aligned} & \psi_{\min} + \nu_{\max} \leq \sqrt{(\psi_{\min} + \nu_{\max})^2 + 4\sigma_{\min}^2} \\ \text{(squaring both sides)} \quad & \Leftrightarrow \ (\psi_{\min} + \nu_{\max})^2 \leq (\psi_{\min} + \nu_{\max})^2 + 4\sigma_{\min}^2 \\ & \Leftrightarrow \ 0 \leq 4\sigma_{\min}^2. \end{aligned}$$

■

3.4 | Bounds for the 1×1 block formulation

The system matrix $\mathcal{A}_1$ given by (14) is symmetric positive definite and so its eigenvalues are positive. We determine how these change due to additional observations when the observation errors are uncorrelated (as for the extreme eigenvalues of $\mathcal{A}_2$ in Theorem 5).

Theorem 7. *If the observation errors are uncorrelated, that is,* $\mathbf{R}$ *is diagonal, then the eigenvalues of* $\mathcal{A}_1$ *move away from zero or are unchanged when new observations are added.*

Proof. Let $\mathcal{A}_{1,k}$ denote $\mathcal{A}_1$ where $p = k$. Then $\mathcal{A}_{1,k+1} = \mathbf{L}^T\mathbf{D}^{-1}\mathbf{L} + \mathbf{H}_{k+1}^T\mathbf{R}_{k+1}^{-1}\mathbf{H}_{k+1} = \mathcal{A}_{1,k} + \alpha^{-1}h_{k+1}h_{k+1}^T$. The result follows by applying Theorem 1. ■

We formulate spectral bounds for $\mathcal{A}_1$ that depend on the largest and smallest eigenvalues of $\mathbf{D}$ and $\mathbf{R}$, and the largest and smallest singular values of $(\mathbf{L}^T\mathbf{H}^T)$.

Theorem 8. *The eigenvalues of* $\mathcal{A}_1$ *lie in the interval*

$$I_+ = \left[\theta_{\min}^2/\tau_{\max}, \theta_{\max}^2/\tau_{\min}\right],$$

where θ_i *and* τ_i *are defined in Table 1, and (15) and (16).*

Proof. Assume that $\mathbf{u} \in \mathbb{R}^{(N+1)n}$ is an eigenvector of $\mathcal{A}_1$. Then the eigenvalue equation premultiplied by $\mathbf{u}^T$ can be written as

$$\chi||\mathbf{u}||^2 = \mathbf{u}^T(\mathbf{L}^T\ \mathbf{H}^T)\begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{L} \\ \mathbf{H} \end{pmatrix}\mathbf{u},$$

where χ is an eigenvalue of $\mathcal{A}_1$. The smallest and largest eigenvalues of $\begin{pmatrix} \mathbf{D}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1} \end{pmatrix}$ are $\tau_{\max}^{-1}$ and $\tau_{\min}^{-1}$, respectively. The bounds follow from the following inequalities that are obtained using (22):

$$\begin{aligned} \chi||\mathbf{u}||^2 &\geq \tau_{\max}^{-1}\mathbf{u}^T(\mathbf{L}^T\ \mathbf{H}^T)\begin{pmatrix} \mathbf{L} \\ \mathbf{H} \end{pmatrix}\mathbf{u} \geq \tau_{\max}^{-1}\theta_{\min}^2||\mathbf{u}||^2, \\ \chi||\mathbf{u}||^2 &\leq \tau_{\min}^{-1}\mathbf{u}^T(\mathbf{L}^T\ \mathbf{H}^T)\begin{pmatrix} \mathbf{L} \\ \mathbf{H} \end{pmatrix}\mathbf{u} \leq \tau_{\min}^{-1}\theta_{\max}^2||\mathbf{u}||^2. \end{aligned}$$

■

The following corollary explains how the upper bound for the eigenvalues of $\mathcal{A}_1$ changes with the addition of new observations. This result that applies for $\mathcal{A}_1$ with a general $\mathbf{R}$ is consistent with Theorem 7 that considers $\mathcal{A}_1$ with a diagonal $\mathbf{R}$.

Corollary 8. *The upper bound in Theorem 8 moves away from zero or is unchanged when new observations are added.*

Proof. If $\tau_{\min} = \rho_{\min}$, $\tau_{\min}$ decreases by Lemma 2. Otherwise $\tau_{\min}$ does not change. The result follows by applying Lemma 1 to determine the change to $\theta_{\max}$. ■

It is unclear how the lower bound in Theorem 8 changes with respect to the number of observations, because both the numerator and denominator grow or stay unchanged by Lemmas 1 and 2, respectively.

3.5 | Alternative bounds

Alternative eigenvalue bounds for symmetric saddle point matrices have been formulated by Axelsson and Neytcheva.[38] These depend on the eigenvalues of the matrices $\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$, $\mathbf{R}$, $\mathbf{D}$, and $\mathcal{A}_1$, and $\xi = \max\{|\lambda_i(\mathcal{A}_1^{-1/2}\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}\mathcal{A}_1^{-1/2})|, i = 1, \dots, (N+1)n\}$.

Theorem 9 (From Theorem 1 (c) of Axelsson and Neytcheva[38]). *The negative eigenvalues of $\mathcal{A}_3$ lie in the interval*

$$I_- = \left[\frac{1}{2}\left(\tau_{\max} - \sqrt{\tau_{\max}^2 + 4\tau_{\max}\lambda_{\max}(\mathcal{A}_1)}\right), \frac{1}{2}\left(\tau_{\min} - \sqrt{\tau_{\min}^2 + 4\tau_{\min}\lambda_{\min}(\mathcal{A}_1)}\right)\right]$$

and the positive ones lie in the interval

$$I_+ = \left[\tau_{\min}, \frac{1}{2}\left(\tau_{\max} + \sqrt{\tau_{\max}^2 + 4\tau_{\max}\lambda_{\max}(\mathcal{A}_1)}\right)\right].$$

Note that the lower bound for the positive eigenvalues in Theorem 9 is the same as in Theorem 4.

Theorem 10 (From Theorem 1 (a) and (b) of Axelsson and Neytcheva[38]). *The negative eigenvalues of $\mathcal{A}_2$ lie in the interval*

$$I_- = \left[-\lambda_{\max}(\mathcal{A}_1), \frac{-\lambda_{\min}(\mathcal{A}_1)}{1 + \frac{\xi\lambda_{\min}(\mathcal{A}_1)}{\psi_{\min}}}\right],$$

and the positive ones lie in the interval

$$I_+ = \left[\psi_{\min}, \frac{1}{2}\left(\psi_{\max} + \sqrt{\psi_{\max}^2 + 4\psi_{\max}\lambda_{\max}(\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L})}\right)\right]. \tag{35}$$

We observe that the bound (35) for the positive eigenvalues, unlike our bound in Theorem 6, is independent of the number of observations. In addition, in practical applications it may not be possible to compute the upper bound for the negative eigenvalues because of the ξ term.

4 | NUMERICAL EXPERIMENTS

4.1 | System setup

We present results of numerical experiments using the Lorenz 96 model,[39] where the state of the system at time t_i is $x_i = (X_i^1, X_i^2, \dots, X_i^n)^T$ and the evolution of x_i components X^j,$j\in\{1,2,\dots,n\}$, is governed by a set of n coupled ODEs:

$$\frac{dX^j}{dt} = -X^{j-2}X^{j-1} + X^{j-1}X^{j+1} - X^j + F,$$

where $X^{-1} = X^{n-1}$,$X^0 = X^n$ and $X^{n+1} = X^1$. This model is continuous in time and discrete in space. We assume that X^1,$X^2\dots$,X^n are equally spaced on a periodic domain of length one and take the space increment to be $\Delta X = 1/n$. We require the linearization of this model $M_i^{(l)}$, $i\in\{0,\dots,N-1\}$ to define $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$. In our experiments, we set n=40 and $F = 8$, since the system shows chaotic behavior with the latter value. The equations are integrated using a fourth-order Runge-Kutta scheme.[40] The time step is set to $\Delta t = 2.5\times10^{-2}$ and the system is run for $N = 15$ time steps.

The assimilation system is set up for so-called identical twin experiments, by which synthetic data are generated using the same model as is used in the assimilation. We generate a reference, or "true," model trajectory $\mathbf{x}^t$ by running the Lorenz 96 model over the time window from prescribed initial conditions and with prescribed Gaussian model errors η_i. An initial background state x^b and observations y_i at each time t_i are then generated by adding Gaussian noise to $\mathbf{x}^t$. Assimilation experiments are run using this background state and observations, assuming that the true state is unknown.

TABLE 2 Computed spectral intervals and extreme eigenvalues of $\mathcal{A}_3$ from Theorem 4 for different observation networks (O.n.)

O.n.	I_-	Eigenvalues	I_+	Eigenvalues
a	$[-2.193,-2.66\times10^{-2}]$	$[-2.192,-2.99\times10^{-2}]$	$[5.93\times10^{-4},2.198]$	$[3.56\times10^{-3},2.195]$
c	$[-2.249,-5.88\times10^{-2}]$	$[-2.247,-6.18\times10^{-2}]$	$[5.93\times10^{-4},2.254]$	$[1.70\times10^{-3},2.251]$
d	$[-2.360,-1.28\times10^{-1}]$	$[-2.358,-1.31\times10^{-1}]$	$[5.93\times10^{-4},2.365]$	$[1.13\times10^{-3},2.362]$
f	$[-2.410,-9.96\times10^{-1}]$	$[-2.408,-9.96\times10^{-1}]$	$[5.93\times10^{-4},2.416]$	$[9.14\times10^{-4},2.413]$

The error covariance matrices that are used to generate the model error in $\mathbf{x}^t$ and the observation error in y_i are also used for the assimilation, that is, in the 3×3 block, 2×2 block, and 1×1 block matrices. These error covariance matrices do not change over time. The observation error covariance matrix is $R_i=\sigma_o^2 I_{p_i}$, where p_i is the number of observations at time t_i, (diagonal R_i is a common choice in data assimilation experiments[13,14]) and the model error covariance matrix is equal to the background error covariance matrix $Q_i=B=\sigma_b^2 C_b$, where C_b is a second-order auto-regressive correlation matrix[41] with correlation length scale 1.5×10^{-2}. We have also performed numerical experiments with $Q_i=\sigma_q^2 C_q\neq B$, where C_q is a Laplacian correlation matrix,[42] and σ_q and σ_b vary by a factor of two. We observed similar results to those presented here. In our experiments, the parameters are chosen so that the observations are close to the real values of the variables, and the background and the model errors are low, in particular, we set $\sigma_o=10^{-1}$, which is about 5% of the mean of the values in $\mathbf{x}^t$, and $\sigma_b=5\times10^{-2}$. y_i consists of direct observations of the variables $X^j j\in\{1,2,\dots,n\}$ at time t_i, hence the observation operator $\mathcal{H}_i$ is linear.

All computations are performed using Matlab R2016b. In particular, the eigenvalues are computed using the Matlab function *eig*. If only extreme eigenvalues are needed, *eigs* is used, and the extreme singular values are given by *svds*.

4.2 | Eigenvalue bounds

We present numerically calculated eigenvalue bounds and eigenvalues of $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$ and illustrate their change with the number of observations and the quality of the spectral estimates, presented in Section 3. We consider the following observation networks that have different numbers of observations ($p=\sum_{i=0}^{N} p_i$):

a. 1 observation at the final time t_{15},
b. 20 observations, observing every eighth model variable at every fourth time step (at times t_3,t_7,t_{11},t_{15}),
c. 80 observations, observing every fourth model variable at every second time step (at times $t_1,t_3,t_5,t_7,t_9,t_{11},t_{13},t_{15}$),
d. 160 observations, observing every second model variable at every second time step (at the same times as in observation network c),
e. 320 observations, observing every second model variable at every time step,
f. 640 observations, fully observed system.

In Figure 1, we plot the eigenvalues of the matrices $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$ together with the bounds from Theorems 4,6, and 8, respectively, for each of the observation networks a-f. In these experiments, as expected from Theorem 3, as the number of observations increases, the smallest and largest negative and the largest positive eigenvalues of $\mathcal{A}_3$ move away from zero and the smallest positive eigenvalue approaches zero. In addition, as determined in Corollary 1, the upper bound for the positive eigenvalues of $\mathcal{A}_3$ presented in Figure 1(I) grows and the lower bound stays the same (because the eigenvalues of $\mathbf{R}$ do not change) when more observations are added. The change is too small to observe in the plots, hence we report the extreme eigenvalues of $\mathcal{A}_3$ and the intervals from Theorem 4 for the networks a, c, e, and f in Table 2. Moreover, the negative bounds for the eigenvalues of $\mathcal{A}_3$ in Figure 1(II) move away from zero. This agrees with Corollary 2, because here $\tau_{\min}=\psi_{\min}$. However, in this setting $\tau_{\max}=\rho_{\max}$ and the same Corollary cannot be used to explain the change to the upper bound. In general, the outer bounds (the largest positive and the smallest negative) for the eigenvalues of $\mathcal{A}_3$ are tight and the inner bounds (the smallest positive and the largest negative) get tighter as the number of observations increases.

The positive eigenvalues of $\mathcal{A}_2$ displayed in Figure 1(III) approach zero as observations are added, whereas the negative eigenvalues in Figure 1(IV) move away from it. This is consistent with Theorem 5, which holds for this experiment

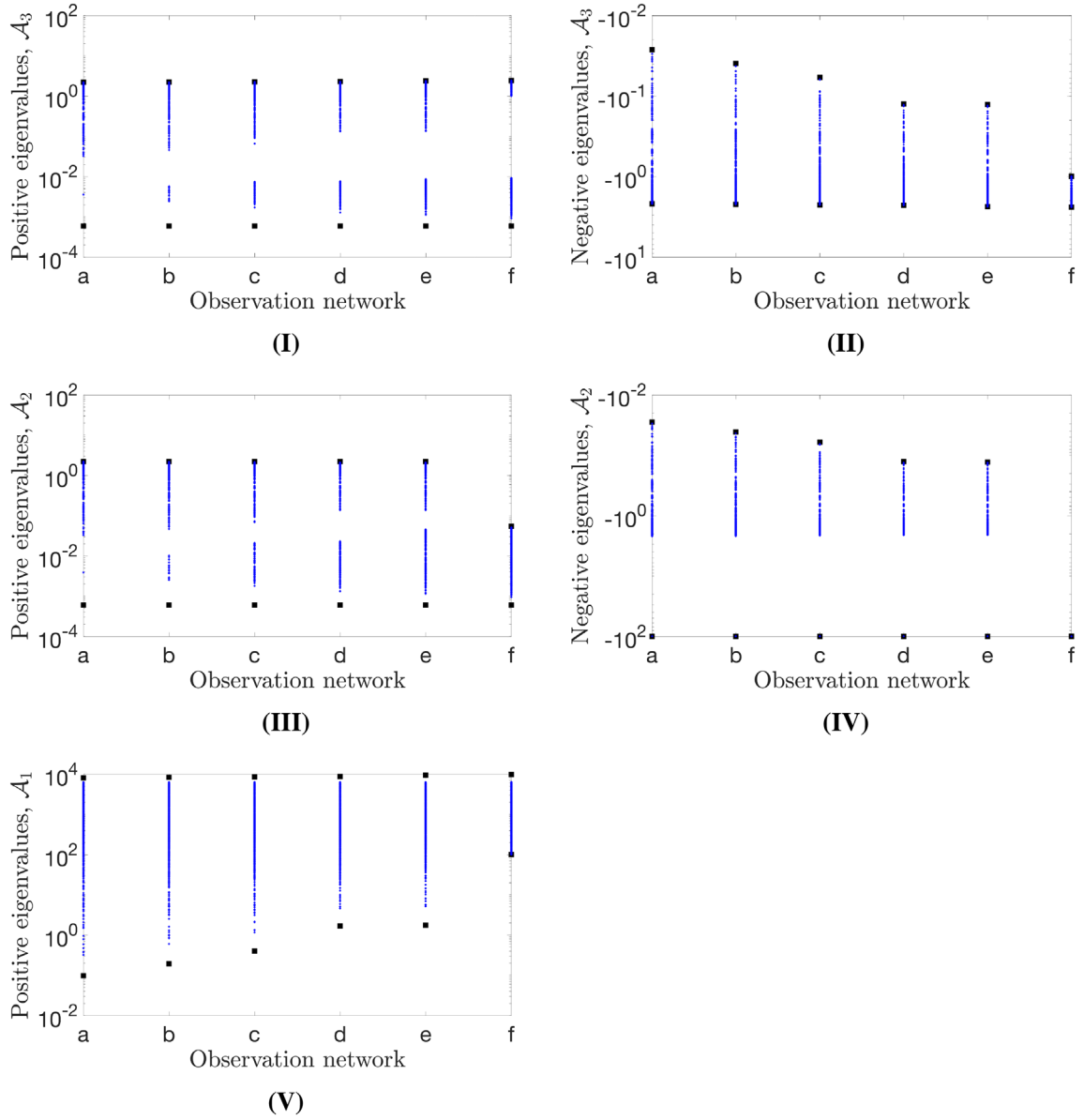

FIGURE 1 Semilogarithmic plots of the positive and negative eigenvalues of the matrices $\mathcal{A}_3$ (I and II) and $\mathcal{A}_2$ (III and IV), and the positive eigenvalues of $\mathcal{A}_1$ in V for the different observation networks (a-f). Eigenvalues are denoted with merged blue dots. The filled black squares mark the bounds for eigenvalues of $\mathcal{A}_3$ in Theorem 4, $\mathcal{A}_2$ in Theorem 6, and $\mathcal{A}_1$ in Theorem 8. Note that the smallest negative eigenvalues of $\mathcal{A}_2$ coincide with the bounds

because we have chosen diagonal $\mathbf{R}$. The lower bounds for the positive and negative eigenvalues of $\mathcal{A}_2$ stay the same when the observation network is changed. In these bounds only $\nu_{\max}$ (the largest eigenvalue of $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$) depends on the observations. In our experiments, $\nu_{\max}$ does not change because of our choice of $\mathbf{H}$ and $\mathbf{R}$. The constant negative lower bound is consistent with Corollary 6. The numerical values of the intervals from Theorem 6 and of the extreme eigenvalues of $\mathcal{A}_2$ for the networks a, c, d, and f are presented in Table 3. The upper positive bound moves toward zero when the system becomes fully observed and is constant for the other networks, because the smallest eigenvalue $\nu_{\min}$ of $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ is nonzero only for the fully observed system. The negative upper bound for the spectrum of $\mathcal{A}_2$ is given by β_1 in (24) for the fully observed system and β_3 in (26) otherwise, and moves away from zero, in agreement with Corollary 5. Notice that the eigenvalue bounds are tight. In addition, the numerical results confirm the statement of Corollary 7 that the interval for the positive eigenvalues of $\mathcal{A}_3$ contains the bounds for positive eigenvalues of $\mathcal{A}_2$.

Note that $\mathcal{A}_2$ has p distinct eigenvalues that coincide with the negative lower bound in the plots. The distinct eigenvalues are explained by the bounds for individual eigenvalues in Corollary 9 in Appendix A, because in our experiments

TABLE 3 Computed spectral intervals and extreme eigenvalues of $\mathcal{A}_2$ from Theorem 6 for different observation networks (O.n.)

O.n.	I_-	Eigenvalues	I_+	Eigenvalues
a	$[-1.0005\times10^{2},-2.83\times10^{-2}]$	$[-1.0001\times10^{2},-2.99\times10^{-2}]$	$[6.03\times10^{-4},2.196]$	$[3.91\times10^{-3},2.195]$
c	$[-1.0005\times10^{2},-6.07\times10^{-2}]$	$[-1.0002\times10^{2},-6.50\times10^{-2}]$	$[6.03\times10^{-4},2.196]$	$[1.78\times10^{-3},2.148]$
d	$[-1.0005\times10^{2},-1.29\times10^{-1}]$	$[-1.0004\times10^{2},-1.33\times10^{-1}]$	$[6.03\times10^{-4},2.196]$	$[1.15\times10^{-3},2.101]$
f	$[-1.0005\times10^{2},-1.00\times10^{2}]$	$[-1.0005\times10^{2},-1.00\times10^{2}]$	$[6.03\times10^{-4},5.42\times10^{-2}]$	$[9.35\times10^{-4},5.15\times10^{-2}]$

TABLE 4 Computed spectral intervals and extreme eigenvalues of $\mathcal{A}_1$ from Theorem 8 with different observation networks (O.n.)

O.n.	I_+	Eigenvalues
a	$[9.72\times10^{-2},8.11\times10^{3}]$	$[3.23\times10^{-1},6.30\times10^{3}]$
c	$[4.05\times10^{-1},8.53\times10^{3}]$	$[1.16,6.32\times10^{3}]$
d	$[1.75,9.40\times10^{3}]$	$[5.21,6.35\times10^{3}]$
f	$[1.00\times10^{2},9.80\times10^{3}]$	$[1.00\times10^{2},6.40\times10^{3}]$

TABLE 5 Computed spectral intervals and extreme eigenvalues of $\mathcal{A}_3$ from Theorems 4 and 9 for observation network d with σ_o=1.5 and σ_b=1

Eigenvalues of $\mathcal{A}_3$	Bounds from Th. 4	Bounds from Th. 9
$[-1.93,-1.38\times10^{-2}]$	$[-2.17,-5.83\times10^{-3}]$	$[-5.10,-1.33\times10^{-2}]$
$[2.98\times10^{-1},3.59]$	$[2.37\times10^{-1},3.81]$	$[2.37\times10^{-1},7.53]$

TABLE 6 Computed spectral intervals and extreme eigenvalues of $\mathcal{A}_2$ from Theorems 6 and 10 for observation network d with σ_o=1.5 and σ_b=1

Eigenvalues of $\mathcal{A}_2$	Bounds from Th. 6	Bounds from Th. 10
$[-1.97,-1.39\times10^{-2}]$	$[-2.33,-5.83\times10^{-3}]$	$[-15.79,-1.33\times10^{-2}]$
$[3.00\times10^{-1},3.51]$	$[2.38\times10^{-1},3.74]$	$[2.37\times10^{-1},7.51]$

$\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ has eigenvalues that are equal to $\sigma_o^{-2}=10^2$ and the largest singular value $\sigma_{\max}$ of $\mathbf{L}$ is less than 10. Hence, there are p eigenvalues of $\mathcal{A}_2$ in the interval $[-110,-90]$ and $(N+1)n-p$ eigenvalues no further than 10 from zero.

The eigenvalues of $\mathcal{A}_1$ and their bounds presented in Figure 1(V) move away from zero when more observations are used. This is as expected, because Theorem 7 holds for our choice of diagonal $\mathbf{R}$. The variation of the bounds is explained by the fact that with our choice of $\mathbf{R}$ values of $\tau_{\min}$ and $\tau_{\max}$ do not change, and $\theta_{\min}$ and $\theta_{\max}$ grow. Such behavior of the upper bound agrees with Corollary 8. However, as can be seen in Table 4 the upper value of the intervals in Theorem 8 are too pessimistic.

Better eigenvalue clustering away from zero when more observations are used can speed up the convergence of iterative solvers when solving the 1×1 block formulation. However, nothing definite can be said about the 3×3 block and 2×2 block formulations: the negative eigenvalues become more clustered, but the smallest positive eigenvalues approach zero when new observations are introduced.

We also calculate the alternative eigenvalue bounds given in Theorems 9 and 10. With the choice of parameters and observations considered in this section, the bounds given in these theorems are not as sharp as those in Theorems 4 and 6. However, this is not always the case, as is illustrated in Tables 5 and 6. Here, σ_o=1.5, σ_b=1 and the observation network d is used.

4.3 | Solving the systems

We solve the 3×3 block, 2×2 block, and 1×1 block systems with the coefficient matrices discussed in the previous subsection, and the right-hand sides defined in (9), (11), and (13), respectively. The saddle point systems are solved with MINRES and the symmetric positive definite systems are solved with CG. The relative residual at the jth iteration of the iterative method is defined as $||\mathbf{r}_j||/||\mathbf{r}_0||$, where $||\cdot||$ is the L_2 norm and $\mathbf{r}_j$ is the residual on iteration j. The iterative method terminates after 400 iterations or when the relative residual reaches 10^{-4}. The initial guess is taken to be the zero vector.

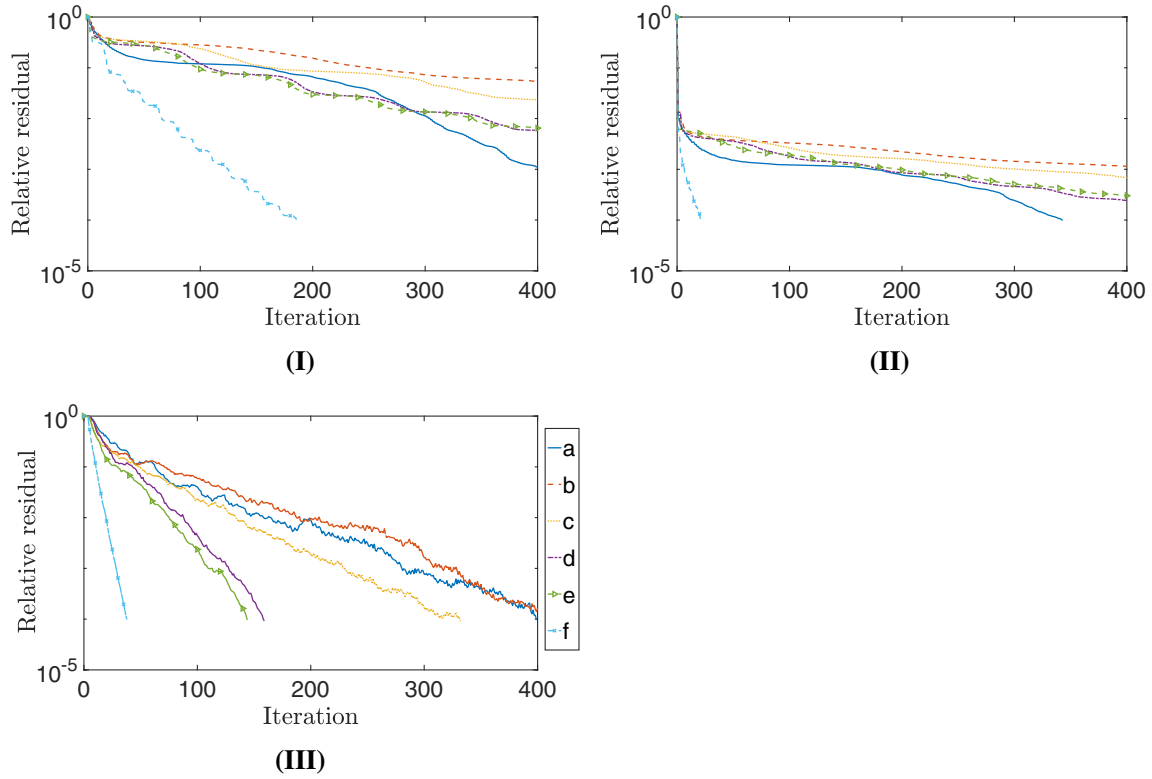

FIGURE 2 Semilogarithmic plots of the relative residual of MINRES when solving the 3×3 block (I) and 2×2 block (II) systems, and the relative residual of CG when solving the 1×1 block (III) system for different observation networks (a-f)

In Figure 2, we plot the relative residuals. Note that the residual reaches 10^{-4} in the fully observed case (observation network f) when solving each of the systems and convergence is most rapid in this case. This is expected because of the clustering of the eigenvalues. The convergence rates are similar for networks d and e, which is consistent with Figure 1. The convergence of MINRES for the observation network a with a single observation is not explained by the spectra of $\mathcal{A}_3$ and $\mathcal{A}_2$. However, the convergence in other cases agrees with our eigenvalue analysis.

5 | CONCLUSIONS

Weak constraint 4D-Var data assimilation requires the minimisation of a cost function in order to obtain an estimate of the state of a dynamical system. Its solution can be approximated by solving a series of linear systems. We have analyzed three different formulations of these systems, namely, the standard system with 1×1 block symmetric positive definite coefficient matrix $\mathcal{A}_1$, a new system with a 2×2 block saddle point coefficient matrix $\mathcal{A}_2$, and the version with 3×3 block saddle point coefficient matrix $\mathcal{A}_3$ of Fisher and Gürol.[12] We have focused on the dependency of the coefficient matrices on the number of observations.

We have found that the spectra of $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$ are sensitive to the number of observations and examined how they change when new observations are added. The results hold with any choice of the blocks in $\mathcal{A}_3$, whereas we can only make inference about the change of the spectra of $\mathcal{A}_2$ and $\mathcal{A}_1$ for uncorrelated observation errors (diagonal $\mathbf{R}$). We have shown that the negative eigenvalues of both $\mathcal{A}_3$ and $\mathcal{A}_2$ move away from zero or are unchanged when observations are added. The smallest and largest positive eigenvalues of $\mathcal{A}_2$, as well as the smallest positive eigenvalue of $\mathcal{A}_3$, approach zero or are unchanged, whereas the largest positive eigenvalue of $\mathcal{A}_3$ moves away from zero or is unchanged. The smallest and largest eigenvalues of $\mathcal{A}_1$ move away from zero or are unchanged. The extreme eigenvalues may cause convergence problems for Krylov subspace solvers, hence we may expect the small positive eigenvalues of $\mathcal{A}_2$ and $\mathcal{A}_3$ to cause these issues when new observations are added. We summarise these results together with the properties of the three systems in Table 7.

TABLE 7 A summary of the properties of the 3×3 block, 2×2 block, and 1×1 systems

	$\mathcal{A}_3$	$\mathcal{A}_2$	$\mathcal{A}_1$
Type	Symmetric indefinite	Symmetric indefinite	Symmetric positive definite
Iterative solver	MINRES/SYMMLQ	MINRES/SYMMLQ	CG
Order	$2(N+1)n+p$	$2(N+1)n$	$(N+1)n$
$\mathbf{D}^{-1}$ needed	No	No	Yes
$\mathbf{R}^{-1}$ needed	No	Yes	Yes
Sequential matrix products	None	$\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$	$\mathbf{L}^T\mathbf{D}^{-1}\mathbf{L}$, $\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$
Eigenvalues that may move toward zero with new observations	Smallest positive	Positive*	None*
Eigenvalues that may move away from zero with new observations	Largest positive, negative	Negative*	All*

Note: *Applies to systems with diagonal $\mathbf{R}$.

We have used the work of Rusten and Winther[19] to determine the bounds for the spectrum of $\mathcal{A}_3$ and derived novel bounds for the spectral intervals of the saddle point matrix $\mathcal{A}_2$ and the positive definite matrix $\mathcal{A}_1$. We have observed that the change to the intervals due to new observations is consistent with the change of the extreme eigenvalues of the matrices. Our numerical experiments agree with these findings. In general, the bounds for the saddle point matrices are tight whereas the upper bounds for the positive definite matrix are too pessimistic.

Our numerical experiments show slow convergence, particularly with a few observations, and the need for preconditioning is evident. Previous work on the 3×3 block saddle point system considered iteratively solving the problem when inexact constraint preconditioners of Bergamaschi et al[43] are used (see, Fisher and Gürol,[12] Freitag and Green,[14] Gratton et al[13]). It was shown that such a preconditioning approach is not optimal and further research into effective preconditioning is still an open question. Preconditioning may transform the coefficient matrix into a nonnormal one with GMRES as an iterative solver of choice. Although the spectrum of a nonnormal matrix may not be enough to describe the convergence of GMRES,[44] Benzi et al[18] claim that fast convergence often appears if the spectrum is clustered away from the origin. Hence, a better understanding of the spectrum of $\mathcal{A}_3$, $\mathcal{A}_2$, and $\mathcal{A}_1$ may help in finding a suitable preconditioner for these matrices. We suggest that including the information on observations coming from the observation error covariance matrix $\mathbf{R}$ and the linearised observation operator $\mathbf{H}$ could be beneficial for preconditioning, given that the spectra of all the considered matrices depend on the observations. The design of such preconditioners that are cheap to construct and apply is an interesting area for future research.

ACKNOWLEDGEMENTS

We would like to kindly thank Dr. Adam El-Said for his code for the weak constraint 4D-Var assimilation system. We are also grateful to two anonymous reviewers for their constructive comments that have led to improvements to the article. This work does not have any conflicts of interest.

ORCID

Ieva Daužickaitė https://orcid.org/0000-0002-1285-1764
Jennifer A. Scott https://orcid.org/0000-0003-2130-1091

REFERENCES

1. Kalnay E. Atmospheric modeling data assimilation and predictability. Cambridge, MA: Cambridge University Press, 2002.
2. Swinbank R. Numerical weather prediction. In: Lahoz W, Khattatov B, Menard R, editors. Data assimilation: Making sense of observations. Berlin, Heidelberg/Germany: Springer-Verlag, 2010; p. 381–406.
3. Chen H, Yang D, Hong Y, Gourley JJ, Zhang Y. Hydrological data assimilation with the ensemble square-root-filter: Use of streamflow observations to update model states for real-time flash flood forecasting. Adv Water Res. 2013;59:209–220.
4. Elbern H, Schmidt H, Ebel A. Variational data assimilation for tropospheric chemistry modeling. J Geophys Res: Atmosph. 1997;102(D13):15967–15985.
5. Moye MJ, Diekman CO. Data assimilation methods for neuronal state and parameter estimation. J Math Neurosci. 2018;8(11):1–38.

6. Nichols NK. Mathematical concepts of data assimilation. In: Lahoz W, Khattatov B, Menard R, editors. Data assimilation: Making sense of observations. Berlin, Heidelberg/Germany: Springer-Verlag, 2010; p. 13–39.
7. Lawless AS. Variational data assimilation for very large environmental problems. In: Cullen MJP, Freitag MA, Kindermann S, Scheichl R, editors. Large scale inverse problems: Computational methods and applications in the earth sciences. Radon series on computational and applied mathematics. Volume 13, Berlin, Boston: De Gruyter, 2013; p. 55-90.
8. Trémolet Y. Accounting for an imperfect model in 4D-Var. Quart J Royal Meteorol Soc. 2006;132(621):2483–2504.
9. Trémolet Y. Model-error estimation in 4D-Var. Quart J Royal Meteorol Soc. 2007;133(626):1267–1280.
10. El-Said A. Conditioning of the weak-constraint variational data assimilation problem for numerical weather prediction (PhD thesis). Department of Mathematics and Statistics, University of Reading; 2015.
11. Bonavita M, Trémolet Y, Hólm E, et al. A strategy for data assimilation. ECMWF Technical Memorandum no. 800. 2017;1-42.
12. Fisher M, Gürol S. Parallelisation in the time dimension of four-dimensional variational data assimilation. Quart J Royal Meteorol Soc. 2017;143(703):1136–1147.
13. Gratton S, Gürol S, Simon E, Toint PL. Guaranteeing the convergence of the saddle formulation for weakly-constrained 4D-Var data assimilation. Quart J Royal Meteorol Soc. 2018;144(717):2592–2602.
14. Freitag MA, Green DLH. A low-rank approach to the solution of weak constraint variational data assimilation problems. J Comput Phys. 2018;357:263–281.
15. Gratton S, Lawless AS, Nichols NK. Approximate Gauss-Newton methods for nonlinear least squares problems. SIAM J Optim. 2007;18(1):106–132.
16. Courtier P, Thépaut JN, Hollingsworth A. A strategy for operational implementation of 4D-Var, using an incremental approach. Quart J Royal Meteorol Soc. 1994;120(519):1367–1387.
17. Saad Y. Iterative methods for sparse linear systems. 2nd ed., Philadelphia: SIAM, 2003.
18. Benzi M, Golub GH, Liesen J. Numerical solution of saddle point problems. Acta Numerica. 2005;14:1–137.
19. Rusten T, Winther R. A preconditioned iterative method for saddle point problems. SIAM J Matrix Anal Appl. 1992;13(3):887–904.
20. Greif C, Moulding E, Orban D. Bounds on eigenvalues of matrices arising from interior-point methods. SIAM J Optim. 2014;24(1):49–83.
21. Morini B, Simoncini V, Tani M. A comparison of reduced and unreduced KKT systems arising from interior point methods. Comput Optim Appl. 2017;68(1):1–27.
22. Rood RB. The role of the model in the data assimilation system. In: Lahoz W, Khattatov B, Menard R, editors. Data assimilation: Making sense of observations. Berlin, Heidelberg/Germany: Springer-Verlag, 2010; p. 351–379.
23. Griffith A, Nichols NK. Adjoint methods in data assimilation for estimating model error. Flow Turbul Combust. 2000;65(3 - 4):469–488.
24. Andersson E, Thépaut JN. Assimilation of operational data. In: Lahoz W, Khattatov B, Menard R, editors. Data assimilation: Making sense of observations. Berlin, Heidelberg/Germany: Springer-Verlag, 2010; p. 283–299.
25. Janjić T, Bormann N, Bocquet M, et al. On the representation error in data assimilation. Quart J Royal Meteorol Soc. 2018;144(713):1257–1278.
26. Talagrand O. Variational assimilation. In: Lahoz W, Khattatov B, Menard R, editors. Data assimilation: Making sense of observations. Berlin, Heidelberg/Germany: Springer-Verlag, 2010; p. 41–68.
27. Nocedal J, Wright SJ. Numerical optimization. 2nd ed. New York: World Scientific, 2006.
28. Fisher M. Minimization algorithms for variational data assimilation. Paper presented at: Proceedings of the Seminar on Recent Developments in Numerical Methods for Atmospheric Modelling. European Centre for Medium Range Weather Forecasts; 1998. p. 364 – 385.
29. Morini B, Simoncini V, Tani M. Spectral estimates for unreduced symmetric KKT systems arising from interior point methods. Numer Linear Algebr Appl. 2016;23(5):776–800.
30. Paige CC, Saunders MA. Solution of sparse indefinite systems of linear equations. SIAM J Numer Anal. 1975;12(4):617–629.
31. Saad Y, Schultz MH. GMRES: A generalized minimal residual algorithm for solving non-symmetric linear systems. SIAM J Sci Stat Comput. 1986;7(3):856–869.
32. Hestenes MR, Stiefel E. Methods of conjugate gradients for solving linear systems. J Res Nat Bureau Stand. 1952;49(6):409–436.
33. Trefethen LN, Bau D. Numerical linear algebra. Philadelphia: SIAM, 1997.
34. Simoncini V, Szyld DB. On the superlinear convergence of MINRES. In: Cangiani A, Davidchack R, Georgoulis E, Gorban A, Levesley J, Tretyakov M, editors. Numerical mathematics and advanced applications. New York, NY: Springer, 2013; p. 733–740.
35. Golub GH, Van Loan CF. Matrix computations. 4th ed. Baltimore: JHU Press, 2013.
36. Stewart GW, Sun JG. Matrix perturbation theory. Computer science and scientific computing. Boston: Academic Press, 1990.
37. Silvester D, Wathen A. Fast iterative solution of stabilised Stokes systems. Part II: Using general block preconditioners. SIAM J Numer Anal. 1994;31(5):1352–1367.
38. Axelsson O, Neytcheva M. Eigenvalue estimates for preconditioned saddle point matrices. Numer Linear Algebr Appl. 2006;13(4):339–360.
39. Lorenz E. Predictability - A problem partly solved. Paper presented at: Proceedings of the Seminar on Predictability. European Centre for Medium Range Weather Forecasts; vol. 1. 1996. p. 1 – 18.
40. Butcher JC. The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods. Hoboken, NJ: Wiley-Interscience, 1987.
41. Daley R. Atmospheric data analysis. Vol 2. Cambridge, MA: Cambridge University Press, 1993.
42. Johnson C, Hoskins BJ, Nichols NK. A singular vector perspective of 4D-Var: Filtering and interpolation. Quart J Royal Meteorol Soc. 2005;131(605):1–19.

43. Bergamaschi L, Gondzio J, Venturin M, Zilli G. Inexact constraint preconditioners for linear systems arising in interior point methods. Comput Optim Appl. 2007;36(2):137–147.
44. Greenbaum A, Pták V, Strakoš Z. Any nonincreasing convergence curve is possible for GMRES. SIAM J Matrix Anal Appl. 1996;17(3):465–469.
45. Silvester J. Determinants of block matrices. Math Gazette. 2000;84(501):460–467.

How to cite this article: Daužickaitė I, Lawless AS, Scott JA, van Leeuwen PJ. Spectral estimates for saddle point matrices arising in weak constraint four-dimensional variational data assimilation. *Numer Linear Algebra Appl*. 2020;27:e2313. https://doi.org/10.1002/nla.2313

APPENDIX A. Bounds for individual eigenvalues of $\mathcal{A}_3$ and $\mathcal{A}_2$

We derive bounds for the individual eigenvalues of $\mathcal{A}_3$ and $\mathcal{A}_2$ (Theorems 13 and 14, respectively). First, we state two theorems that are used in deriving these bounds. The notation of Table 1 is used.

Theorem 11 (See Theorem 3 in Silvester[45]). *If* $A = \begin{pmatrix} C & E \\ F & G \end{pmatrix}$, $C, E, F, G \in \mathbb{R}^{n\times n}$, *and* $FG = GF$, *then* $det(A) = det(CG-EF)$.

Theorem 12 (Jordan-Wielandt Theorem, see Theorem 4.2 in Chapter 1 of Stewart and Sun[36]). *Let* $U^H AV = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$, $\Sigma = \mathrm{diag}(\sigma_1,\dots,\sigma_n)$ *be the singular value decomposition of* $A \in \mathbb{C}^{m\times n}$, $m \geq n$. *Then the eigenvalues of the matrix*

$$C = \begin{pmatrix} 0 & A \\ A^H & 0 \end{pmatrix}$$

are $\pm\sigma_1,\dots,\pm\sigma_n$, *corresponding to the eigenvectors* $\begin{pmatrix} u_i \\ \pm v_i \end{pmatrix}$, $i = 1,\dots,n$, *where* u_i *and* v_i *are the ith columns of U and V, respectively. C also has* $m-n$ *zero eigenvalues with eigenvectors* $\begin{pmatrix} u_i \\ 0 \end{pmatrix}$, $i = n+1,\dots,m$.

Theorem 13. *Let* ω_i, $i = 1,\dots,(N+1)n+p$ *be the ith value in* $\{\psi_k,\rho_j | k = 1,\dots,(N+1)n, j = 1,\dots,p\}$ *(the set of eigenvalues of* $\mathbf{D}$ *and* $\mathbf{R}$*). Then the kth eigenvalue of* $\mathcal{A}_3$ *is bounded by*

$$\begin{aligned} &\text{positive eigenvalues:} && \omega_k - \theta_{\max} \leq \gamma_k \leq \omega_k + \theta_{\max}, && k = 1,\dots,(N+1)n+p, \\ &\text{negative eigenvalues:} && -\theta_{\max} \leq \gamma_{k+(N+1)n+p} < 0, && k = 1,\dots,(N+1)n. \end{aligned}$$

Proof. We can write $\mathcal{A}_3$ as a sum of two symmetric matrices:

$$\mathcal{A}_3 = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{R} & \mathbf{H} \\ \mathbf{L}^T & \mathbf{H}^T & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{D} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{L} \\ \mathbf{0} & \mathbf{0} & \mathbf{H} \\ \mathbf{L}^T & \mathbf{H}^T & \mathbf{0} \end{pmatrix} = \mathbf{S}_D^{3x3} + \mathbf{S}_L^{3x3}.$$

The spectrum of $\mathbf{S}_D^{3x3}$ is the union of the eigenvalues of $\mathbf{D}$, $\mathbf{R}$ and zeros. By Theorem 12, the eigenvalues λ of the indefinite matrix $\mathbf{S}_L^{3x3}$ are the singular values of $(\mathbf{L}^T\mathbf{H}^T)$ with plus and minus signs, thus $\lambda_{\min} = -\theta_{\max}$ and $\lambda_{\max} = \theta_{\max}$.

The result follows from applying Theorem 1 to the matrices $\mathbf{S}_D^{3x3}$ and $\mathbf{S}_L^{3x3}$. ∎

Theorem 14. *The eigenvalues of $\mathcal{A}_2$ are bounded by*

$$\begin{aligned} &\text{positive eigenvalues:} && \psi_k - \sigma_{\max} \le \zeta_k \le \psi_k + \sigma_{\max}, && k = 1, \dots, (N+1)n. \\ &\text{negative eigenvalues:} && -\nu_k - \sigma_{\max} \le \zeta_{k+(N+1)n} \le -\nu_k + \sigma_{\max}, && k = 1, \dots, (N+1)n, \end{aligned} \tag{A1}$$

Proof. As in Theorem 13, we express $\mathcal{A}_2$ as a sum of two symmetric matrices

$$\mathcal{A}_2 = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & -\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{L} \\ \mathbf{L}^T & \mathbf{0} \end{pmatrix} = \mathbf{S}_D^{2x2} + \mathbf{S}_L^{2x2}.$$

The rest of the proof is analogous to that of Theorem 13. ∎

Corollary 9. *If there are $p < (N+1)n$ observations, (A1) in Theorem 14 becomes*

$$\begin{aligned} -\sigma_{\max} \le &\zeta_{k+(N+1)n} \le 0, && k = 1, \dots, (N+1)n - p, \\ -\nu_k - \sigma_{\max} \le &\zeta_{k+2(N+1)n-p} < -\nu_k + \sigma_{\max}, && k = 1, \dots, p. \end{aligned}$$

Proof. The result follows from noticing that $-\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}$ has $(N+1)n-p$ zero eigenvalues. ∎