Krylov methods and test functions for detecting bifurcations in one parameter-dependent partial differential equations

Bosco García-Archilla

Departamento de Matemática Aplicada II, Universidad de Sevilla

Escuela Superior de Ingenieros, Camino de los Descubrimientos, s/n, 41092 Sevilla (Spain)

e-mail: bosco.garcia@esi.us.es

Juan Sánchez

Departament de Física Aplicada, Universitat Politècnica de Catalunya

Jordi Girona, 1–3, mòdul B4–B5, Campus Nord, 08034 Barcelona (Spain)

e-mail: sanchez@fa.upc.es

Carles Simó

Departament de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona

Gran Via de les Corts Catalanes, 585, 08071 Barcelona (Spain)

e-mail: carles@maia.ub.es

Abstract

In this paper, we study the computation of the sign of the determinant of a large matrix as a byproduct of the preconditioned GMRES method when applied to solve the linear systems arising in the discretization of partial differential equations (PDEs). Convergence is proved using not the eigenvalues but the singular values of the PDE operator when premultiplied by a preconditioner. Numerical experiments are presented where the technique is applied to detect and locate pitchfork and transcritical bifurcation points on a one parameter dependent system. With an appropriate selection of the initial guess in the GMRES method, the technique is shown to accurately locate bifurcation points.

Mathematics Subject Classification: 65F40, 15A60, 65P30, 35B60, 65N35, 65F10.

Keywords and Phrases: Determinants, Arnoldi decomposition, compact operators in Hilbert spaces, spectral methods for PDEs, continuation methods, bifurcation location.

1 Introduction

Much of the well-developed numerical techniques to study the bifurcations of dynamical systems in $\mathbb{R}^m$ (see e.g., [10], [25], [35], [36]) are inapplicable or have their computational

efficiency greatly diminished when the systems arise from the discretization of partial differential equations (PDE). One of the reasons for this is the impossibility of storing and factoring $m \times m$ matrices for m large. Although discretizations by finite-element or finite-difference methods usually give rise to sparse matrices, this is not usually the case of spectral methods when applied to nonlinear equations. In these cases, for the Linear Algebra problems to be solved in the study of the dynamical systems (solution of systems of linear equations, eigenvalue computation, etc) one must resort to matrix-free methods based only on matrix vector products (plus preconditioning) without explicitly building the matrices involved. Iterative matrix-free methods for solving linear systems or computing eigenvalues are well-established today [16], [26], and are regularly used in the studies of large dynamical systems (see e.g. [11], [14], [13], [28], [29]). However, to our knowledge, no matrix-free method has been studied for the computation of determinants, which, in the study of one-parameter dependent dynamical systems, can be used for detection and location of pitchfork and transcritical bifurcations.

In this paper, we study the computation of the sign of the determinant of a large matrix as a byproduct of the GMRES method [33] for solving linear systems. The matrices we deal with arise from the discretization of one parameter-dependent partial differential equations (PDE) problems. The PDE problems we consider are of the form

$$\text{find} \quad u \in \mathcal{H}, \quad \mu \in \mathbb{R}, \quad \text{such that} \qquad f(u,\mu) \equiv \mathcal{A}u + R(u,\mu) = 0, \tag{1}$$

where $\mathcal{H}$ is a Hilbert space (typically $L^2(\Omega)$ for some domain $\Omega \subset \mathbb{R}^d$) $\mathcal{A}$ is an operator such as the Laplacian, the biharmonic operator or some other elliptic operator, subject to appropriate boundary conditions, and $R(\cdot,\lambda)$ is a nonlinear differential operator with $D(R(\cdot,\mu)) \subset D(\mathcal{A})$, so that the Fréchet differential of $\mathcal{A}^{-1}R(\cdot,\mu)$ is compact. This is typically the case when the derivatives featuring in R are of lower order than those in $\mathcal{A}$. The solutions of (1) will be assumed to be sufficiently smooth so that standard discretizations by spectral, finite-element or finite-difference methods converge.

We will also assume that a fast solver is available for $\mathcal{A}$ or its discretization, as it is usually the case of spectral methods, and thus, $\mathcal{A}$ can be used as a preconditioner. This assumption is for simplicity, since the results here apply also when more sophisticated preconditioners are used.

Although in practice the GMRES method is applied to solve linear systems in $\mathbb{R}^m$, following [9], [13], we will analyze first the case where the coefficient matrix is replaced by compact perturbation of the identity in a Hilbert space, in order to deduce results in $\mathbb{R}^m$ independent of the value of m.

In Section 2, besides preliminary material, we introduce an approximation to the determinant of a matrix by means of the Arnoldi decomposition. In Section 3 we analyze how correctly the Arnoldi decomposition reproduces part of the spectrum of an operator. Borrowing from [31], singular values rather than eigenvalues will be a key element in the convergence result in Theorem 4. In Section 4 we deal with issues of practical implementation. Section 5 contains numerical experiments where the previous material is applied to locate pitchfork and transcritical bifurcations. The last section is devoted to conclusions and further remarks.

2 Preliminaries

2.1 Discretization and Continuation

In the discretization of (1) by spectral methods, a complete orthonormal set $p_1, p_2, \ldots$ of $\mathcal{H}$ is given. Typically, the p_j are the Fourier modes, their real or imaginary parts, Legendre or Chebyshev polynomials, eigenfunctions of some boundary value problem or tensor products of these functions. Let us denote $\mathcal{H}_m = \text{span}(p_1, \ldots, p_{m-1})$, and let $P_m : \mathcal{H} \to \mathcal{H}_m$ be the orthogonal projection onto $\mathcal{H}_m$. Although it is not always the case, quite frequently $P_m\mathcal{A} = \mathcal{A}P_m$. Thus, for simplicity, we make this assumption, the reader bearing in mind that only minor modifications of what follows apply to a more general case. Problem (1) is then replaced by the family of problems

$$\text{find} \quad \mathbf{u}_m = \begin{bmatrix} u_m \\ \mu \end{bmatrix} \in \mathcal{H}_m \times \mathbb{R} \quad \text{s.t.} \qquad f^{(m)}(\mathbf{u}_m) \equiv \mathcal{A}u_m + P_m R(\mathbf{u}_m) = 0. \tag{2}$$

Here and in the sequel, we reserve boldface characters for elements of

$$\hat{\mathcal{H}} = \mathcal{H} \times \mathbb{R};$$

similarly, for simplicity we will drop the subindex m and the superindex (m) when no confusion arises.

As mentioned in the introduction, we assume that the Jacobian matrix $f^{(m)}_{u_m}$ is full but the mapping

$$v \mapsto f^{(m)}_{u_m} v$$

can be computed at a cost of $O(m \log(m))$ flops with Fast Fourier Transform (FFT) techniques, and without explicitly building $f^{(m)}_{u_m}$.

Solutions of (2), when not isolated, appear in branches $s \mapsto \mathbf{u}(s) \in \hat{\mathcal{H}}$ for certain parameter s such as the arclength. A discrete set of points is computed sequentially along the branch. A much-used technique is Keller's pseudo-arclength continuation [22], [23], [25], which requires at every step the solution of the system

$$F^{(m)}(u, \mu) = \begin{bmatrix} f^{(m)}(u, \mu) \\ (\mathbf{t}_r, (\mathbf{u} - \mathbf{u}_r)) - \delta_r \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}, \tag{3}$$

where $\mathbf{u}_r$ is a reference point, typically the most recently available solution of (2), $\mathbf{t}_r$ is an approximation to the tangent $\mathbf{t}(s) = d\mathbf{u}(s)/ds$ of the branch at $\mathbf{u}_r$, δ_r is the corresponding increment along this tangent, and $(\cdot, \cdot)$ denotes the inner product in the Hilbert space $\hat{\mathcal{H}}$

$$(\mathbf{x}_1, \mathbf{x}_2) = \langle u_1, u_2 \rangle + \mu_1 \mu_2,$$

$\langle \cdot, \rangle$ being the inner product in $\mathcal{H}$.

System (3) is typically solved by Newton's method. This requires to solve systems of the form

$$F_{\mathbf{u}} \mathbf{d} = -F. \tag{4}$$

Of the various methods to solve (4) based on $\mathbf{v} \mapsto F_{\mathbf{u}}\mathbf{v}$ operations (see e.g. [16]) we focus on the GMRES method [33], although our discussion covers other methods based on the Arnoldi decomposition of a matrix such as ORTHORES [20].

Note however that the systems in (4) arise from the discretization of PDEs, and thus, are typically ill-conditioned. This ruins the computational efficiency of the GMRES method. For this reason, preconditioning is used in practice. Here, we consider left-preconditioning, which amounts to replace the systems in (4) by $\mathcal{P}^{-1}F_{\mathbf{u}}\mathbf{d} = -\mathcal{P}^{-1}F$, where the preconditioner $\mathcal{P}$ is a matrix (hopefully) close to $F_{\mathbf{u}}$ and easily invertible. For simplicity, we take as preconditioner the operator $\mathcal{P}$ defined by

$$\mathcal{P}\begin{bmatrix} u \\ \mu \end{bmatrix} = \begin{bmatrix} \mathcal{A}u \\ \mu \end{bmatrix}. \tag{5}$$

Typically, when using spectral methods, fast solvers are available for $\mathcal{P}$. In fact, in Fourier spectral methods, $\mathcal{P}$ is a diagonal matrix.

2.2 Arnoldi decomposition

In an Arnoldi decomposition of an operator A such as $\mathcal{P}^{-1}F_{\mathbf{u}}$, starting from an initial vector $\mathbf{v}_1$ (for a linear system $A\mathbf{x} = \mathbf{b}$, $\mathbf{v}_1 = (\mathbf{b} - A\mathbf{x}_0)/\left\|\mathbf{b} - A\mathbf{x}_0\right\|$, for an initial approximation $\mathbf{x}_0$ to $\mathbf{x}$), a sequence of orthonormal vectors, the *Arnoldi vectors*, is obtained recursively by

$$\mathbf{v}_{k+1} = \frac{(I - P_{\mathcal{V}_k})A\mathbf{v}_k}{\|(I - P_{\mathcal{V}_k})A\mathbf{v}_k\|}, \qquad k = 1, 2, \ldots, L-1, \tag{6}$$

where $\mathcal{V}_k$ is the k-dimensional Krylov subspace

$$\mathcal{V}_k = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k), \tag{7}$$

and where here and in the sequel $P_{\mathcal{V}}$ denotes the orthogonal projection onto the subspace $\mathcal{V}$ of $\hat{\mathcal{H}}$. The integer L in (6) is such that $(I - P_{\mathcal{V}_L})A\mathbf{v}_L = 0$. In fact, it is easy to check that $\mathcal{V}_L$ is the minimal subspace invariant by A containing $\mathbf{v}_1$.

Denoting by V_k the row vector $V_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ with entries in $\hat{\mathcal{H}}$, from (6) follows the Arnoldi decomposition of the operator A,

$$AV_k = V_{k+1}\tilde{H}_k = V_k H_k + [\mathbf{0}, \ldots, \mathbf{0}, h_{k+1,k}\mathbf{v}_{k+1}], \tag{8}$$

where $\tilde{H}_k$ is a $(k+1) \times k$ matrix, and H_k is the $k \times k$ matrix obtained from $\tilde{H}_k$ by deleting the last row. The entries $h_{i,j}$ of $\tilde{H}_k$ are given by

$$h_{i,j} = (\mathbf{v}_i, A\mathbf{v}_j), \qquad 1 \le i \le k+1, \quad 1 \le j \le k.$$

Observe that H_k is the matrix of the operator $P_{\mathcal{V}_k}A_{|_{\mathcal{V}_k}}$, given by the restriction of $P_{\mathcal{V}_k}A$ to the subspace $\mathcal{V}_k$, expressed in the basis of the Arnoldi vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

For solving a linear system $A\mathbf{x} = \mathbf{b}$, the GMRES method approximates $\mathbf{x}$ by $\mathbf{x}_k$, found by solving the least-squares problem $\min_{\mathbf{x}_k \in \mathcal{V}_k} \|\mathbf{b} - A\mathbf{x}_k\|$, and the ORTHORES [20] method approximates $\mathbf{x}$ by $\mathbf{x}_k \in \mathcal{V}_k$ such that the residual $\mathbf{b} - A\mathbf{x}_k$ is orthogonal to $\mathcal{V}_k$. In practical implementations, the matrices $\tilde{H}_k$ are explicitly built in a sequential process, starting from $k = 1$ and stopping when the residual $\|\mathbf{b} - A\mathbf{x}_k\| / \|\mathbf{b}\|$ satisfies a prescribed tolerance test.

2.3 Bifurcation detection via determinants

Typically, solutions of (1) (or, in practice, (2)) are the equilibria or steady-states of evolution equations of the form $u_t = f(u, \mu)$, for a given value of μ. It is of interest then to know how many such steady-states exist for a given value of μ, their asymptotic stability, the appearance of periodic orbits (Hopf bifurcations) etc. A steady-state u is asymptotically stable if all the eigenvalues of the Fréchet derivative $f_u(u, \mu)$ have negative real part. Thus, if when approximating branches of equilibria by solving (3), one is also interested in their stability, the spectrum of f_u or its most relevant part [26], [27], must be computed. This not only provides information on the stability of equilibria, but also informs about saddle–node bifurcations, the appearance of periodic orbits in Hopf bifurcations as well as crossings of two branches of equilibria at branching points in pitchfork or transcritical bifurcations. In branching points, not only f_u has a zero eigenvalue, but also one eigenvalue of $F_{\mathbf{u}}$ changes sign along any of the two crossing branches. The bifurcation point can then be located by finding the zero of the smallest eigenvalue or, as in [1], [2], the smallest singular value of $F_{\mathbf{u}}$, or even by solving some generalized eigenvalue problem as in [13]. Also, due to the ill-conditioning of $F_{\mathbf{u}}$, shift-invert techniques [30] (i.e. solving linear systems) are usually required when more sophisticated techniques [26] are used to locate the smallest eigenvalue of $F_{\mathbf{u}}$.

The same approach can be used to study bifurcations of some solutions different from steady-states, like periodic solutions and invariant tori. There are different approaches to reduce these problems to fixed point problems, $u = f(u, \mu)$, in suitable spaces (see [21], [34], [37], and references therein).

However, since eigenvalue computations are in general much more costly than solving linear systems, specially when the dimension $m - 1$ in (2) is large, it is useful in many instances to compute first all branches of equilibria so that the study of stability can be focused on places of most interest at a later stage (although it is advisable to do some eigenvalue computations, at least at bifurcation points, since besides providing important geometric information, they help to better focus the stability computations). In order to do this, it is necessary to be able to detect and locate branching points without computing the most relevant part of the spectrum of f_u or $F_{\mathbf{u}}$.

Detection of branching points can be done by monitoring $\mathrm{sign}(\det(F_{\mathbf{u}}))$ and, for their location, there are robust techniques based on augmented systems of larger dimension than (3), [3], [18], [42], although they require more than providing the computation of f and the action of f_u. Location of branching points can also be done by by finding the zero of $\mathrm{sign}(\det(F_{\mathbf{u}}))$ with the help of the bisection method (to avoid possible overflow or underflow, it is advisable to compute $\mathrm{sign}(\det(F_{\mathbf{u}}))$ rather than $\det(F_{\mathbf{u}})$ [10], despite computing the logarithm of $|\det(F_{\mathbf{u}})|$ can reduce the number of iterations when locating a bifurcation point as a zero of $\det(F_{\mathbf{u}})$). This is particularly appealing if $\mathrm{sign}(\det(F_{\mathbf{u}}))$ can be obtained at little cost as a byproduct of Newton's method when solving (3), as it is suggested in [10]. Such is the case if m is small and Gaussian elimination is used to solve the linear systems (4).

When an iterative method based on an Arnoldi decomposition of $A = \mathcal{P}^{-1}F_{\mathbf{u}}$ is used to solve (4), the following argument suggests how to approximate $\mathrm{sign}(\det(F_{\mathbf{u}}))$. Since H_k is a matrix representation of $P_{\mathcal{V}_k} A_{|\mathcal{V}_k}$, and $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \ldots$ with increasing dimensions, for the larger k reached in the iterative solution of (4), we may take

$$\chi_k = \mathrm{sign}(\det(\mathcal{P}))\mathrm{sign}(\det(H_k)) \tag{9}$$

as an approximation to $\text{sign}(\det(F_{\mathbf{u}}))$. Next section will show that under certain hypothesis this approximation is correct. We end this section commenting on how readily available is this approximation in the standard implementations of the GMRES method.

In practice, the matrices $\tilde{H}_k$ are reduced to upper triangular form by means of orthogonal transformations,

$$Q_k \tilde{H}_k = \tilde{R}_k = \begin{bmatrix} R_k \\ 0^T \end{bmatrix},$$

where R_k is $k \times k$ and upper triangular. The transformation Q_k is built recursively as a product of Givens rotations,

$$Q_k = G_k \begin{bmatrix} Q_{k-1} & 0 \\ 0^T & 1 \end{bmatrix}, \qquad k = 1, \ldots, L-1,$$

where if we express

$$\begin{bmatrix} Q_{k-1} & 0 \\ 0^T & 1 \end{bmatrix} \tilde{H}_k = \begin{bmatrix} R_{k-1} & r_k \\ 0^T & \eta_k \\ 0^T & h_{k+1,k} \end{bmatrix}, \tag{10}$$

then

$$G_k = \begin{bmatrix} I_{k-1} & 0 & 0 \\ 0^T & c_k & s_k \\ 0^T & -s_k & c_k \end{bmatrix}, \qquad c_k = \frac{\eta_k}{\sqrt{\eta_k^2 + h_{k,k+1}^2}}, \quad s_k = \frac{h_{k,k+1}}{\sqrt{\eta_k^2 + h_{k+1,k}^2}}. \tag{11}$$

In view of (11), we observe that for $l = 1, \ldots, L-1$, $\det(G_l) = 1$, and, consequently, $\det(Q_l) = 1$. Thus, in view of (10), we have that $\det(H_k) = \eta_k \det(R_{l-1})$. But, due to the way the matrices G_l, $l = 1, \ldots, L-1$ are built, we deduce that all diagonal elements of R_{k-1} are positive. Consequently,

$$\text{sign}(\det(H_k)) = \text{sign}(\eta_k) = \text{sign}(c_k), \qquad k = 1, \ldots, L-1. \tag{12}$$

Generally, routines for the GMRES method provide as optional output the sequence of values c_l and s_l, so that no extra computation is needed to estimate $\text{sign}(\det(F_{\mathbf{u}}))$.

3 Convergence Analysis

Although in practice, computations are carried out in finite-dimensional spaces, we will gain the right perspective of what happens in practice by first analyzing the estimate $\text{sign}(\det(H_k)) \approx \text{sign}(\det(\mathcal{P}^{-1}F_{\mathbf{u}}))$ in the Hilbert space $\hat{\mathcal{H}}$, that is, the case $m = \infty$ in (2), where $\mathcal{H}_\infty = \mathcal{H}$ and $P_\infty = I$ (further below we explain how to get rid of determinants, which, in general, are not defined for operators in infinite-dimensional spaces). This is not unnecessary artificial since, as discussed in [12], any possible pathological behaviour of Krylov methods is attainable in finite-dimensional spaces. The following matrix, used in [19] to show that the GMRES method may achieve the worst possible convergence sequence, also shows that the approximation $\text{sign}(\det(H_k)) \approx \text{sign}(\det(A))$ is completely wrong until the final step $k = m$. Consider the matrix,

$$A = \begin{bmatrix} 0 & 0 & \ldots & 0 & -\alpha_0 \\ 1 & 0 & \ldots & 0 & -\alpha_1 \\ 0 & 1 & & 0 & -\alpha_2 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & & 1 & -\alpha_{m-1} \end{bmatrix}.$$

Taking as v_1 the first coordinate vector in $\mathbb{R}^m$, the resulting matrices H_k are the upper-left $k \times k$ submatrices of A. Thus, for $k = 1, \ldots, m-1$, $\det(H_k) = 0 \neq (-1)^m \text{sign}(\alpha_0) = \text{sign}(\det(A))$. We will see that this situation cannot happen in the infinite-dimensional case, and then thanks to the convergence properties of the discretization, it will be excluded also for $m < \infty$.

3.1 The infinite-dimensional case

The mappings $f^{(m)}$ in (2) are approximations to the mapping in $\mathcal{H}$ given by $f(u, \mu) = \mathcal{A}u + R(u, \mu)$. For a given solution (u, μ) of (1), let us denote by $R_u(u, \mu)$ and $R_\mu(u, \mu)$ the Fréchet derivatives of R with respect to u and μ, respectively, and let us consider the operator in $\hat{\mathcal{H}}$ given by

$$A\mathbf{v} = A \begin{bmatrix} v \\ \delta \end{bmatrix} = \begin{bmatrix} v + \mathcal{A}^{-1} R_u(u, \mu)v + \delta \mathcal{A}^{-1} R_\mu(u, \mu) \\ (\mathbf{t}, \mathbf{v}) \end{bmatrix},$$

where

$$\mathbf{t} = \begin{bmatrix} t \\ \tau \end{bmatrix}$$

is a unit vector such that $f_u(u, \mu)t + f_\mu(u, \mu)\tau = 0$. Observe then that, since the solutions (u_m, μ_m) of (2) converge to (u, μ) as $m \to \infty$, the operators $\mathcal{P}^{-1} F^{(m)}_{\mathbf{u}}$ converge to A.

As mentioned in the introduction, we assume that the operator in $\mathcal{H}$ given by $\mathcal{A}^{-1} R_u(u, \mu)$ is compact. Thus, it is clear that the operator A can be written as

$$A = I + T,$$

where T is a compact operator in $\hat{\mathcal{H}}$.

For a compact operator K let us denote by $\sigma(I + K)$ the spectrum of $(I + K)$. It is well-known that $\sigma(I + K)$ is composed of a sequence of isolated eigenvalues converging to 1, together with $\lambda^\infty = 1$. Let us also denote

$$\sigma_-(I + K) = \sigma(A) \cap \{\text{Re}(z) < 0\}, \qquad \sigma_+(I + K) = \sigma(A) \cap \{\text{Re}(z) \geq 0\},$$

the set of eigenvalues of $I + K$ with negative and non negative real part respectively. Observe that $\sigma_-(I + K)$ is a finite set. For an isolated eigenvalue $\lambda \in \sigma(I + K)$, the invariant subspace E_λ associated with λ is the range of P_λ

$$E_\lambda = \mathcal{R}(P_\lambda), \qquad P_\lambda = \frac{1}{2\pi} \int_{|z-\lambda|=\delta} (zI - (I + K))^{-1} \, \mathrm{d}z,$$

for $\delta > 0$ such that $\sigma(I + K) \cap \{z \in \mathbb{C} \mid |z - \lambda| \leq \delta\} = \{\lambda\}$. Since K is compact, $\dim(E_\lambda) < \infty$, and it is known as the *algebraic multiplicity* of λ (see e. g. [24]).

For operators in infinite-dimensional spaces, it is not generally possible to extend the idea of determinant [24], § X.1.4. Thus, we consider instead the orientation. If $(I + K)$ is invertible, we define the *orientation* of $I + K$ as

$$\text{or}(I + K) = (-1)^{d(K)}, \qquad d(K) = \sum_{\lambda \in \sigma_-(I+K)} \dim(E_\lambda). \tag{13}$$

Observe that if K is an operator in a finite-dimensional space (and, hence, compact) $\text{or}(I + K) = \text{sign}(\det(I + K))$.

The following result is a consequence of Theorem 3.11 and Remark 3.2 in [24], § IV.5, and [24], § I.4.6.

Theorem 1 *Let K be a compact operator such that $-1 \notin \sigma(K)$, and Γ a closed curve in $\{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0\}$ such that is interior encloses $\sigma_-(I+K)$ and no other eigenvalue $I+K$. If B is a bounded operator such that*

$$\|B\| \leq \frac{1}{2\min_{\xi\in\Gamma} \|(\xi I - (I+K))^{-1}\|}, \tag{14}$$

then

$$\mathrm{or}(I+K+B) = \mathrm{or}(I+K).$$

Let us consider now $(\mathbf{v}_k)_{k=1}^{\infty}$, the sequence of Arnoldi vectors of $I+T$, that is the vectors obtained by the process in (6–8). Recall then that H_k is the matrix of $P_{\mathcal{V}_k}(I+T)_{|\mathcal{V}_k}$. Let us denote by $\mathcal{V}$ the minimal closed invariant subspace containing the first Arnoldi vector $\mathbf{v}_1$, that is

$$\mathcal{V} = \overline{\cap_{\mathbf{v_1}\in\mathcal{X}}\mathcal{X}}, \qquad \mathcal{X} \text{ subspace of } \hat{\mathcal{H}}, \quad \text{and } T\mathcal{X} \subset \mathcal{X}, \tag{15}$$

where $\overline{X}$ denotes the closure of the set X. If the subspace $\mathcal{V}$ is finite-dimensional, then the Arnoldi process finishes in a finite number of steps.

We consider the operators

$$T_k = P_{\mathcal{V}_k} T P_{\mathcal{V}_k}, \qquad k = 1, 2, \ldots .$$

It is clear then that

$$\mathrm{sign}(\det(H_k)) = \mathrm{or}(I+T_k).$$

Observe also that, since $(\mathbf{v}_k)_{k=1}^{\infty}$ is a complete orthonormal set in $\mathcal{V}$, we have that

$$\lim_{k\to\infty}(I+T_k) = (I+T)_{|\mathcal{V}},$$

Thus, taking $K = T_{|\mathcal{V}}$ in Theorem 1, we have the following result.

Theorem 2 *If*

$$\bigcup_{\lambda\in\sigma_-(I+T)\cap\mathbb{R}} E_\lambda \subset \mathcal{V}, \tag{16}$$

then $\lim_{k\to\infty}\mathrm{sign}(\det(H_k)) = \mathrm{or}(I+T)$.

Remark 1 Observe that since $\mathcal{V}$ is an invariant subspace of $I+T$, if (16) does not hold and, then $\lim_{k\to\infty}\mathrm{sign}(\det(H_k)) = \mathrm{or}\big((I+T)_{|\mathcal{V}}\big)$.

3.2 Results for the discretization

The same arguments that lead to Theorem 2 also show that $\mathrm{sign}(\det(\mathcal{P}^{-1}F^{(m)}_{\mathbf{u}_m})) \to \mathrm{or}(T)$, as $m \to \infty$. However notice that this only guarantees that the approximation (9) is correct when $k = m$. In practice, though, this is an undesirable situation since one usually aims at $k \ll m$ in the GMRES method from which we obtain the approximation (9). The purpose of this section is to show that the correct value of $\mathrm{sign}(\det(\mathcal{P}^{-1}F^{(m)}_{\mathbf{u}_m}))$ is obtained for $k \ll m$.

As in the previous section we denote as $(\mathbf{v}_k)_{k=1}^{\infty}$ the Arnoldi vectors of $(I+T)$, $\mathcal{V}_k$ their Krylov subspaces and $\tilde{H}_k$ and H_k the corresponding matrices of the Arnoldi

decomposition. Besides, in this section, for every $m = 1, 2, \ldots$, we denote $(\mathbf{v}_k^{(m)})_{k=1}^m$ the Arnoldi vectors of $((I + P_m T)_{|\mathcal{H}_m})$, $\mathcal{V}_k^{(m)}$ their Krylov subspaces and $\tilde{H}_k^{(m)}$ and $H_k^{(m)}$ the corresponding matrices of the Arnoldi decomposition. Notice that $(I + P_m T)_{|\mathcal{H}_m}$ does not coincide with $\mathcal{P}^{-1} F_{\mathbf{u}_m}^{(m)}$ since the former has the term $P_m \mathcal{A}^{-1} R_u(u, \mu)$ and the later $P_m \mathcal{A}^{-1} R_u(u_m, \mu)$. However, since $u_m \to u$ as $m \to \infty$, all results about $(I + P_m T)_{|\mathcal{H}_m}$ apply also to $\mathcal{P}^{-1} F_{\mathbf{u}_m}^{(m)}$. Similarly, for simplicity, we will assume that $\mathbf{v}_1^{(m)} = P_m \mathbf{v_1}$, although any other values satisfying $\left\| \mathbf{v}_1 - \mathbf{v}_1^{(m)} \right\| \to 0$ as $m \to \infty$ may be taken.

The main result of this Section is Theorem 4 below. Its proof will be a consequence of preliminary results which are presented next.

Lemma 1 *Let $\mathcal{X} \subset \mathcal{Y}$ and $\mathcal{U}$ be closed subspaces of the Hilbert space $\hat{\mathcal{H}}$ satisfying that for some positive $\epsilon_s < 1$,*

$$\|P_{\mathcal{X}} u\| \geq \sqrt{1 - \epsilon_s^2}\, \|P_{\mathcal{Y}} u\|, \qquad \forall u \in \mathcal{U}. \tag{17}$$

Then, for every $y \in \mathcal{Y}$ such that $P_{\mathcal{X}} y = 0$ the following bound holds:

$$\|P_{\mathcal{U}} y\| \leq \epsilon_s \|y\|. \tag{18}$$

Proof. We write

$$\begin{aligned} y &= (I - P_{\mathcal{U}}) y + P_{\mathcal{U}} y & (19)\\ &= (I - P_{\mathcal{U}}) y + (I - P_{\mathcal{Y}}) P_{\mathcal{U}} y + P_{\mathcal{Y}} P_{\mathcal{U}} y \\ &= (I - P_{\mathcal{U}}) y + (I - P_{\mathcal{Y}}) P_{\mathcal{U}} y + (I - P_{\mathcal{X}}) P_{\mathcal{Y}} P_{\mathcal{U}} y + P_{\mathcal{X}} P_{\mathcal{Y}} P_{\mathcal{U}} y. & (20) \end{aligned}$$

Taking inner product with y we have

$$\|y\|^2 = \|(I - P_{\mathcal{U}}) y\|^2 + (y, (I - P_{\mathcal{X}}) P_{\mathcal{Y}} P_{\mathcal{U}} y). \tag{21}$$

Since $P_{\mathcal{X}} P_{\mathcal{Y}} P_{\mathcal{U}} y = P_{\mathcal{X}} P_{\mathcal{U}} y$ and $P_{\mathcal{U}} y \in \mathcal{U}$, (17) implies that $\|(I - P_{\mathcal{X}}) P_{\mathcal{Y}} P_{\mathcal{U}} y\| \leq \epsilon_s \|P_{\mathcal{Y}} P_{\mathcal{U}} y\| \leq \epsilon_s \|P_{\mathcal{U}} y\|$, so that from (21) it follows that $\|y\|^2 \leq \|(I - P_{\mathcal{U}}) y\|^2 + \epsilon_s \|y\| \, \|P_{\mathcal{U}} y\|$, and hence,

$$\|(I - P_{\mathcal{U}}) y\|^2 \geq \|y\|^2 - \epsilon_s \|y\| \, \|P_{\mathcal{U}} y\|. \tag{22}$$

Going back to (19), by Pythagoras Theorem we have $\|y\|^2 = \|(I - P_{\mathcal{U}}) y\|^2 + \|P_{\mathcal{U}} y\|^2$, and, taking into account (22), it follows that

$$0 \geq \|P_{\mathcal{U}} y\|^2 + \epsilon_s \|y\| \, \|P_{\mathcal{U}} y\|,$$

from where (18) follows. □

We remark that for (17) to hold, it must be $\dim(P_{\mathcal{Y}} \mathcal{U}) \leq \dim(\mathcal{X})$. In fact, if $\dim(\mathcal{U}) < +\infty$ (which will be the case in the analysis that follows), $\epsilon_s = \sin(\theta)$, θ being the smallest principal angle between $\mathcal{U}$ and $\mathcal{X}$ (see e.g. [15], § 12.4.3).

We need the singular value decomposition of an operator. Let K a compact operator in a Hilbert space and let $\sigma_1^2 \geq \sigma_2^2 \ldots$, the eigenvalues (counted with their multiplicities) of $K^* K$. Let $u_1, u_2, \ldots$, be the complete orthonormal set of the associated eigenvectors and, for $j = 1, 2, \ldots$, let us denote $w_j = \sigma_j^{-1} K u_j$, and u_j^* the functional given by the

inner product $u^*x = (u, x)$. The *singular value decomposition* (SVD) of K is the one given by

$$K = \sum_{j=1}^{\infty} \sigma_j w_j u_j^*.$$

Notice that $\|K\| = \sigma_1$. For a given $\epsilon > 0$ let

$$\mathcal{U}_\epsilon = \operatorname{span}\{u_j \mid \sigma_j \geq \epsilon\}. \tag{23}$$

Observe that relevant information about an operator is provided by the singular vectors. However, in practice, we have the Arnoldi vectors instead. The following two results aim at establishing when the properties of the singular vectors are enjoyed in some sense by the Arnoldi vectors.

Lemma 2 *Let $\mathcal{X} \subset \mathcal{Y}$ be closed subspaces of $\hat{\mathcal{H}}$ and $0 < \epsilon_s < 1$ such that $\|P_{\mathcal{X}} u\| \geq \sqrt{1-\epsilon_s^2}\,\|P_{\mathcal{Y}} u\|$, for all $u \in \mathcal{U}_\epsilon$. Then, for every $y \in \mathcal{Y}$ such that $P_{\mathcal{X}} y = 0$,*

$$\|Ky\| \leq \left(\epsilon_s^2 \|K\|^2 + \epsilon^2\right)^{1/2} \|y\|. \tag{24}$$

Proof. Let y be such that $P_{\mathcal{X}} y = 0$ and let us denote $u = P_{\mathcal{U}_\epsilon} y$ and $z = (I - P_{\mathcal{U}_\epsilon})y$ so that $y = u + z$. We have that

$$\|Ky\|^2 = \|Ku\|^2 + \|Kz\|^2 \leq \sigma_1^2 \|u\|^2 + \epsilon^2 \|z\|^2 = \|K\|^2 \|u\|^2 + \epsilon^2 \|z\|^2.$$

Applying Lemma 1 and taking into account that $\|z\| \leq \|y\|$, (24) follows. □

Theorem 3 *For every $0 < \epsilon \leq 1$ there exist positive integers k_1 and m_1 given by (27), (28) below, such that the Arnoldi vectors $(\mathbf{v}_k)_{k=1}^{\infty}$ of $I + T$ satisfy*

$$\|T\mathbf{v}_k\| \leq \epsilon, \quad \text{for } k \geq k_1, \tag{25}$$

and for $m \geq m_1$, the Arnoldi vectors $(\mathbf{v}_k^{(m)})_{k=1}^{\infty}$ of $I + P_m T$ satisfy

$$\left\|P_m T \mathbf{v}_k^{(m)}\right\| \leq \epsilon, \quad \text{for } k_1 \leq k \leq m. \tag{26}$$

Proof. If $\|T\| < \epsilon$ both (25) and (26) follow trivially. Hence, we can assume $\|T\| \geq \epsilon$. We take

$$\epsilon_1 = \frac{\epsilon}{\sqrt{2}}, \qquad \epsilon_s = \frac{\epsilon}{\sqrt{2}\,\|T\|}.$$

Let $T = \Sigma_{j=1}^{\infty} \sigma_j w_j u_j^*$ be a SVD of T and let us consider $\mathcal{U}_{\epsilon_1} = \operatorname{span}\{u_j \mid \sigma_j \geq \epsilon_1\}$.

Since in the invariant subspace $\mathcal{V}$ defined in (15), $P_{\mathcal{V}_k} \to I$ as $k \to \infty$, there exists an integer k_1 such that

$$\left\|P_{\mathcal{V}_{k_1}} u\right\| \geq \sqrt{1 - \frac{\epsilon_s^2}{2}}\,\|P_{\mathcal{V}} u\| = \sqrt{1 - \frac{\epsilon^2}{4\|T\|^2}}\,\|P_{\mathcal{V}} u\|, \qquad \text{for all } u \in \mathcal{U}_{\epsilon_1}. \tag{27}$$

Now, we apply Lemma 2 with $K = T$, $\mathcal{X} = \mathcal{V}_{k_1}$ and $\mathcal{Y} = \mathcal{V}$, to get that,

$$\|Ty\| \leq \left(\frac{\epsilon^2}{4} + \frac{\epsilon^2}{2}\right)^{1/2} \|y\| \leq \epsilon \|y\|, \qquad \text{for} \quad y \in (I - P_{\mathcal{V}_{k_1}})\mathcal{V},$$

which implies (25).

We now prove (26). Since the generation of the Arnoldi basis $(\mathbf{v}_k)_{k=1}^{\infty}$ is, for a fixed k, a continuous mapping of $\mathbf{v}_1$, and $P_m \to I$ as $m \to \infty$, there exists a m_1 such that

$$\left\|P_{\mathcal{V}_{k_1}^{(m)}} - P_{\mathcal{V}_{k_1}}\right\| \le \frac{\epsilon_s^2}{8} = \frac{\epsilon^2}{16\left\|T\right\|^2}, \qquad \text{for all } m \ge m_1. \tag{28}$$

Let us consider now $m \ge m_1$, and let us denote by $\mathcal{V}^{(m)}$ the minimal T-invariant subspace containing the first Arnoldi vector $\mathbf{v}_1^{(m)}$. For $u \in \mathcal{U}_{\epsilon_1}$, let us denote $v = P_{\mathcal{V}^{(m)}}u$. In view of (27–28), we have

$$\left\|P_{\mathcal{V}_{k_1}^{(m)}}u\right\| = \left\|P_{\mathcal{V}_{k_1}^{(m)}}v\right\| \ge \left\|P_{\mathcal{V}_{k_1}}v\right\| - \left\|(P_{\mathcal{V}_{k_1}^{(m)}} - P_{\mathcal{V}_{k_1}})v\right\| \ge \sqrt{1-\frac{\epsilon_s^2}{2}}\left\|v\right\| - \frac{\epsilon_s^2}{8}\left\|v\right\|.$$

A simple calculation shows that $\sqrt{1-\epsilon_s^2/2} - \epsilon_s^2/8 \ge \sqrt{1-\epsilon_s^2}$, so that

$$\left\|P_{\mathcal{V}_{k_1}^{(m)}}u\right\| \ge \sqrt{1-\epsilon_s^2}\left\|P_{\mathcal{V}^{(m)}}u\right\|, \qquad u \in \mathcal{U}_{\epsilon_1}, \quad m \ge m_1. \tag{29}$$

For $k_1 \le k \le m$, we have $\mathcal{V}_{k_1}^{(m)} \subset \mathcal{V}_k^{(m)}$, so that $\|P_{\mathcal{V}_k^{(m)}}u\| \ge \|P_{\mathcal{V}_{k_1}^{(m)}}u\|$, and, hence, (29) holds with k_1 replaced by k. Thus, applying Lemma 2 with $K = T$, $\mathcal{X} = \mathcal{V}_k^{(m)}$ and $\mathcal{Y} = \mathcal{V}^{(m)}$, for $k_1 \le k \le m$ we have that

$$\left\|Ty\right\| \le \left(\frac{\epsilon^2}{2} + \frac{\epsilon^2}{2}\right)^{1/2}\left\|y\right\| = \epsilon\left\|y\right\|, \qquad \text{for} \quad y \in (I - P_{\mathcal{V}_k^{(m)}})\mathcal{V}^{(m)}, \quad m \ge m_1,$$

which, taking into account that $\|P_m Ty\| \le \|Ty\|$, implies (26). ∎

The following result states sufficient conditions for the sequence $\mathrm{sign}(\det(H_k))$, $k = 1, 2, \ldots,$ to become stationary. This has to be so if the sequence $\mathrm{sign}(\det(H_k))_{k=1}^{\infty}$ is to converge to the orientation $\mathrm{or}(I+T)$.

Lemma 3 *Assume that the value of s_k in (11) satisfies that $|s_k| < 1/\sqrt{2}$. If $\|T\mathbf{v}_{k+1}\| < 1/\sqrt{2}$, then*

$$\mathrm{sign}(\det(H_{k+1})) = \mathrm{sign}(\det(H_k)). \tag{30}$$

Proof. Since in (12) we saw that $\mathrm{sign}(\det(H_k)) = \mathrm{sign}(\eta_k) = \mathrm{sign}(c_k)$, we will show that $\mathrm{sign}(\eta_{k+1}) = \mathrm{sign}(c_k)$. Recall (10–12), and let us write

$$\begin{bmatrix} Q_{k-1} & 0 & 0 \\ 0^T & 1 & 0 \\ 0^T & 0 & 1 \end{bmatrix} \tilde{H}_{k+1} = \begin{bmatrix} R_{k-1} & r_k & f \\ 0^T & \eta_k & \gamma \\ 0^T & h_{k+1,k} & h_{k+1,k+1} \\ 0^T & 0 & h_{k+2,k+1} \end{bmatrix}.$$

Observe that

$$\begin{bmatrix} f \\ \gamma \end{bmatrix} = \begin{bmatrix} Q_{k-1} & 0 \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} h_{1,k+1} \\ \vdots \\ h_{k,k+1} \end{bmatrix},$$

and thus

$$(\left\|f\right\|^2 + |\gamma|^2)^{1/2} = \left\|P_{\mathcal{V}_k}(I+T)\mathbf{v}_{k+1}\right\| = \left\|P_{\mathcal{V}_k}T\mathbf{v}_{k+1}\right\|, \tag{31}$$

so that

$$(\|f\|^2 + |\gamma|^2)^{1/2} = \Big(\|(I+T)\mathbf{v}_{k+1}\|^2 - |1 + (\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|^2 - |h_{k+2,k+1})|^2)\Big)^{1/2}.$$

On the other hand,

$$\eta_{k+1} = c_k h_{k+1,k+1} - s_k\gamma = c_k + \Big(c_k(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1}) - s_k\gamma\Big). \tag{32}$$

Thus, η_{k+1} and c_k will have the same sign if the last term on the right-hand side above is sufficiently small. Applying Holder's inequality and recalling (31) we have that

$$\begin{aligned} |c_k(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1}) - s_k\gamma| &\leq (c_k^2 + s_k^2)\Big(|(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|^2 + |\gamma|^2\Big)^{1/2} \\ &\leq \Big(|(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|^2 + \|P_{\mathcal{V}_k} T\mathbf{v}_{k+1}\|^2\Big)^{1/2} \\ &\leq \Big(\|T\mathbf{v}_{k+1}\|^2 - |h_{k+2,k+1}|\Big)^{1/2} < \frac{1}{\sqrt{2}} < |c_k|\,, \end{aligned} \tag{33}$$

and thus, $\operatorname{sign}(\eta_{k+1}) = \operatorname{sign}(c_k)$. □

Lemma 3 guarantees the same sign for the determinants of H_k and H_{k+1} provided $\|T\mathbf{v}_{k+1}\|$ and s_k are small. Theorem 3 gives conditions to ensure that $\|T\mathbf{v}_{k+1}\|$ is small. For the smallness of s_k, Lemma 6.11 in [12] states that

$$|s_k| \leq \big\|(I+T)^{-1}\big\|\,|(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|\,, \qquad k = 1, 2, \ldots\,. \tag{34}$$

Thus, choosing

$$\epsilon = \frac{1}{\sqrt{2}\max(1, \|(I+T)^{-1}\|)}, \tag{35}$$

Theorem 3 and (34) allow us to conclude the following result.

Theorem 4 *Assume that (16) holds and let $T = \sum_{j=1}^{\infty} \sigma_j w_j P_{u_j}$ be a SVD of T. For ϵ defined in (35) set $\mathcal{U}_{\epsilon/\sqrt{2}} = \operatorname{span}\{u_j \mid \sigma_j \geq \epsilon/\sqrt{2}\}$. Choose k_1 such that*

$$\big\|P_{\mathcal{V}_{k_1}} u\big\| \geq \Big(1 - \frac{\epsilon}{4\,\|T\|^2}\Big)^{1/2}\|u\|\,, \qquad \text{for all } u \in \mathcal{U}_{\epsilon/\sqrt{2}}.$$

Then, $\operatorname{sign}(\det(H_k)) = \operatorname{sign}(\det(H_{k_1})) = \operatorname{or}(T)$*, for all $k \geq k_1$.*

Furthermore, choose m_1 such that

$$\big\|P_{\mathcal{V}_{k_1}^{(m)}} - P_{\mathcal{V}_{k_1}}\big\| \leq \frac{\epsilon^2}{16\,\|T\|^2}, \qquad \text{for all } m \geq m_1.$$

Then, $\operatorname{sign}(\det(H_k^{(m)})) = \operatorname{sign}(\det(H_{k_1}^{(m)})) = \operatorname{or}(T)$*, for all $k_1 \leq k \leq m$, and $m \geq m_1$.*

Notice that, it is also possible to find $m_2 \geq m_1$ such that for $m \geq m_2$, $\operatorname{sign}(\det(H_k^{(m)}))$ $= \operatorname{sign}(\det(H_k))$, for $k = 1, \ldots, m$. This value will obviously depend on factors such as the rate of decay of $\|(I - P_m)T\|$.

Remark 2 As argued in (34–35), the fact that $\det(H_k^{(m)})$ has the correct sign relies in $\|T\mathbf{v}_{k+1}\|$ being sufficiently small, which, according to the proof of Theorem 3, is attained for k sufficiently large. Observe then that, in view of (34), in order to guarantee the hypothesis $|s_k| < 1/\sqrt{2}$ in Lemma 3, k may have to be quite large if $I + T$ is close to singular. We will refer to this remark when commenting our numerical tests.

4 Practical implementation

In practice, $\chi_k = \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H_k^{(m)}))$ may easily fail to change sign at a branching point $\mathbf{u}(s_0)$ in a branch of equilibria $s \mapsto \mathbf{u}(s)$ whenever this branch is in an invariant subspace shared by both F and the preconditioner $\mathcal{P}$. This scenario is typical when f is equivariant by an appropriate group Γ of symmetries (see e.g., [6], [7], [17]), that is

$$f(Su, \mu) = Sf(u, \mu), \qquad S \in \Gamma, \quad \mu \in \mathbb{R}, \tag{36}$$

where, for the purposes of the present section, we may assume that Γ is a group of isometric operators in $\mathcal{H}$. The fixed space of Γ,

$$\mathcal{E} = \{u \in \mathcal{H} \mid Su = u, \quad \forall S \in \Gamma\},$$

is then invariant by f and its differential f_u.

For a branch $s \mapsto \mathbf{u}(s) \in \hat{\mathcal{E}} = \mathcal{E} \times \mathbb{R}$, branching takes place at $\mathbf{u}_0 = \mathbf{u}(s_0)$ if, for instance, besides some further assumptions [6], § 2.3.2, a single eigenvalue of f_u changes sign at s_0 and

$$\text{Ker}(f_u(u_0, \mu_0)) = \text{span}(v), \qquad v \notin \mathcal{E}. \tag{37}$$

For example, a symmetry-breaking pitchfork bifurcation happens if $Sv = -v$ for some $S \in \Gamma$ with $S^2 = I$.

Notice then that, in spite of $\mathbf{u}_0$ being a singular point of $f : \mathcal{H} \times \mathbb{R} \to \mathcal{H}$, however, since $\mathcal{E}$ is an invariant subspace of f, (37) implies that $\mathbf{u}_0$ is a regular point of f as a mapping from $\mathcal{E} \times \mathbb{R}$ onto $\mathcal{E}$.

Now recall how the mapping F in Keller's pseudo-arclength continuation technique is built in (2–3), and then it is a simple exercise to check that from (36) it follows that, in the case $m = \infty$, $\hat{\mathcal{E}}$ is invariant by F and $F_{\mathbf{u}}$. If, in addition, $\hat{\mathcal{E}}$ is also invariant by the preconditioner $\mathcal{P}$, we also have

$$-\mathcal{P}^{-1}F(\mathbf{u}) \in \hat{\mathcal{E}}, \qquad \forall \mathbf{u} \in \hat{\mathcal{E}}, \qquad\qquad \mathcal{P}^{-1}F_{\mathbf{u}}(\mathbf{u})\mathbf{x} \in \hat{\mathcal{E}}, \qquad \forall \mathbf{u} \in \hat{\mathcal{E}}, \quad \forall \mathbf{x} \in \hat{\mathcal{E}}. \tag{38}$$

Furthermore, let us denote

$$\mathbf{v} = \begin{bmatrix} v \\ 0 \end{bmatrix}, \qquad \mathbf{t}_0 = \frac{d\mathbf{u}}{ds}(s_0),$$

where v is the same as in (37). Ovserve that $\mathbf{t}_0 \in \hat{\mathcal{E}}$ and $\mathbf{v} \notin \hat{\mathcal{E}}$. On the other hand, a simple calculation shows that for some $\alpha \in \mathbb{R}$,

$$\text{Ker}(F_{\mathbf{u}}(\mathbf{u}_0)) = \text{Ker}(\mathcal{P}^{-1}F_{\mathbf{u}}(\mathbf{u}_0)) = \text{span}(\mathbf{v} + \alpha\mathbf{t}_0), \qquad \mathbf{v} + \alpha\mathbf{t}_0 \notin \hat{\mathcal{E}}. \tag{39}$$

Since $\mathbf{t}_0 \in \hat{\mathcal{E}}$, it follows that $\mathbf{v} + \alpha\mathbf{t}_0 \notin \hat{\mathcal{E}}$, that is, no eigenvalue of the restriction to $\hat{\mathcal{E}}$ of $F_{\mathbf{u}}$ chages sign at the bifurcation point $\mathbf{u}_0$.

Now, when using Newton's method to solve equations (2–3) (for $m = \infty$) along the branch $s \mapsto \mathbf{u}(s) \in \hat{\mathcal{E}}$, in the corresponding linear systems (4), if the initial guess $\mathbf{u}^{[0]}$ for Newton's method satisfies that $\mathbf{u}^{[0]} \in \hat{\mathcal{E}}$ (which will hold if $\mathbf{u}^{[0]}$ is computed by any of the standard extrapolation procedures used in practice), and if the initial guess $\mathbf{x}_0$ in the GMRES method is also taken in $\hat{\mathcal{E}}$ (which will be so if $\mathbf{x}_0 = \mathbf{0}$) then, the first residual

$\mathbf{r}_0 = -\mathcal{P}^{-1}(F + F_{\mathbf{u}}\mathbf{x}_0)$ is also in $\hat{\mathcal{E}}$. Since the first Arnoldi vector $\mathbf{v}_1$ is proportional to $\mathbf{r}_0$, it follows that the minimal invariant subspace $\mathcal{V}$ containing $\mathbf{v}_1$ satisfies that

$$\mathcal{V} \subset \hat{\mathcal{E}}. \tag{40}$$

Thus, according to Remark 1, the value of $\text{sign}(\det(H_k))$ for k sufficiently large, will not be that of the orientation of $\mathcal{P}^{-1}F_{\mathbf{u}}$, but that of its restriction to $\mathcal{V} \subset \hat{\mathcal{E}}$, where, as shown in (39), no eigenvalue changes sign at $\mathbf{u}_0$. In other words, since all computations are carried out in the invariant subspace $\hat{\mathcal{E}}$, these show that $\mathbf{u}_0$ is a regular point of $f : \mathcal{E} \times \mathbb{R} \to \mathcal{E}$, but not a singular point of f as a mapping from $\mathcal{H} \times \mathbb{R}$ onto $\mathcal{H}$. This is also the case of a symmetry-breaking double turning point, where, instead of (37), we have $\text{Ker}(f_u(u_0, \mu_0)) = \text{span}(v, (du/ds(s_0))$, $v \notin \mathcal{E}$ and f_μ not in the range of $f_u(u_0, \mu_0)$.

In practical computations, the discretizations by spectral methods usually inherit the symmetries of f, that is,

$$f^{(m)}(Su_m, \mu) = Sf^{(m)}(u_m, \mu), \qquad S \in \Gamma, \quad \mu \in \mathbb{R}, \tag{41}$$

so that (38) holds with F, $\mathbf{u}$ and $\hat{\mathcal{E}}$ replaced $F^{(m)}$, $\mathbf{u}_m$ and $\hat{\mathcal{E}} \cap \mathcal{H}_m$ respectively. Thus, the arguments above also show that $\chi_k = \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H_k^{(m)}))$ will fail to change sign when crossing a branching point. Indeed, as shown in Section 5 below, it may be even worse, since round-off and discretization errors may induce χ_k to suffer spurious change of signs.

A simple remedy is to "pollute" the initial guess $\mathbf{x}_0$ in the GMRES method so that $\mathbf{x}_0 \notin \hat{\mathcal{E}}$. Now observe that $\hat{\mathcal{E}}$ is a proper subspace of $\hat{\mathcal{H}}$. Thus the probability of any given vector to be in $\hat{\mathcal{E}}$ is 0. Consequently, a random initial guess suffices to take $\mathbf{x}_0$ out of $\hat{\mathcal{E}}$ (see (47) in next section).

An alternative to taking an initial random guess $\mathbf{x}_0$ in the GMRES method may be implemented if, a priori, it is known in which fixed subspace $\mathcal{E}$ is the branch $s \mapsto u(s)$. In the case of a $\mathbb{Z}_2$ symmetry, that is, when $\Gamma = \{I, S\}$, so that $S^2 = I$, it is observed in [42] that, in many practical instances, it is straightforward to code $f^{(m)}$ as a mapping $f^{(m)} : \mathcal{E}^{(m)} \times \mathbb{R} \to \mathcal{E} \cap \mathcal{H}_m$, where $\mathcal{E}^{(m)} = \mathcal{E} \cap \mathcal{H}_m$. Besides, in this case, $S^{(m)}$ can be naturally decomposed as

$$\mathcal{H}_m = \mathcal{E}^{(m)} \oplus \mathcal{E}_a^{(m)}, \qquad \mathcal{E}_a^{(m)} = \{u \in \mathcal{H}_m \mid Su = -u\},$$

In [42] it is shown that if the system $f(u, \mu)$ is augmented with a new variable $p \in \mathcal{E}_a$, and with the equations $f_u(u, \mu)p = 0$, and $(l, p) = 1$, for a suitable $l \in \mathcal{E}_a$, the resulting system $G(u, p, \mu)$ has an isolated solution in a symmetry-breaking bifurcation point. Coding $f_{u_m}^{(m)}$ as an operator in $\mathcal{E}_a^{(m)}$ brings a substantial reduction in computational cost when solving the discretization of $G(u, p, \mu) = 0$. Similar strategies are followed in [8] in the case of a $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry, that is when Γ is the group generated by S_1 and S_2 such that $S_1^2 = S_2^2 = I$.

Substantial computational costs can be obtained also in pseudo-arclenth continuation by coding $F^{(m)}$ as a (nonlinear) operator in $\mathcal{E}^{(m)} \times \mathbb{R}$. In order to detect a symmetry breaking bifurcation point, we may use the following strategy. After computing a point $\mathbf{u}_m$ on the branch, we can compute the sign of the determinant of $\mathcal{P}^{-1}F_{\mathbf{u}_m}^{(m)}(\mathbf{u}_m)$. This can be done as described in Section 2.3, if we solve by the GMRES method the system $\mathcal{P}^{-1}F_{\mathbf{u}_m}^{(m)}\mathbf{x}_0 = \mathbf{b}$, where $b = \mathcal{P}^{-1}F_{\mathbf{u}_m}^{(m)}\mathbf{x}_0$, for a random vector $\mathbf{x}_0 \in \mathcal{E}_a^{(m)} \times \mathbb{R}$, and if

the action of $\mathcal{P}^{-1}F^{(m)}_{\mathbf{u}_m}$ is coded as that of an operator in $\mathcal{E}^{(m)}_a \times \mathbb{R}$. In the case of a $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry, to such systems must be solved, with the action of $F^{(m)}_{\mathbf{u}_m}$ coded as an operator in $\mathcal{E}^{(m)}_{a,j} \times \mathbb{R}$, $j = 1, 2$, where $\mathcal{E}^{(m)}_{a,1} = \{u \in \mathcal{H} \mid S_1 x = -x, \quad S_2 x = x\}$ and $\mathcal{E}^{(m)}_{a,2} = \{u \in \mathcal{H} \mid S_1 x = x, \quad S_2 x = -x\}$.

5 Numerical experiments

In this section we test numerically how correctly $\chi_k = \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H^{(m)}_k))$ approximates $\text{sign}(\det(F^{(m)}_{\mathbf{u}_m}))$ and the capabilities of this approximation to locate pitchfork and transcritical bifurcation points. The value of k is that produced by the GMRES method for the system $A\mathbf{x} = \mathbf{b}$, under the stopping criterion

$$\|\mathbf{b} - A\mathbf{x}\| / \|\mathbf{b}\| \leq TOL_{MR}. \tag{42}$$

We initially set the value of the prescribed tolerance TOL_{MR} to $TOL_{MR} = 5 \times 10^{-4}$. We will duly point out when and why this value is altered.

For numerical experiments, we consider the following Bérnard convection problem. Let $\Omega = [-1/2, 1/2] \times [0, 1]$, and consider the following system of partial differential equations in the variables $u = u(y, z)$, $v = [v_1(y, z), v_2(y, z)]^T$, $p = p(y, z)$.

$$\begin{array}{rcll} u_t + \sqrt{\mu} v(\nabla u - e_3) & = & \Delta u, & \\ -\nabla p - v + \sqrt{\mu} u e_3 & = & 0, & \\ \nabla \cdot v & = & 0, & \\ -\nabla u \cdot n & = & 0, & y = \pm\frac{1}{2}, \quad z \in [0, 1], \\ u & = & 0, & y \in [-1/2, 1/2], \quad z = 0, 1, \\ v \cdot n & = & 0, & (y, z) \in \partial\Omega, \end{array} \tag{43}$$

where n represents the outward normal vector, $e_3 = [0, 1]^T$ is the vertical unit vector, and μ is a scalar parameter.

This systems models two-dimensional flow in a closed box, filled with fluid-saturated porous material, heated from below and cooled from above and subject to gravity (see e.g. [8], [17], [32], [38], [39] and the references cited therein). The variable u represents the deflection from the linear temperature distribution $T_l(y, z) = T_0(1 - z)$, T_0 being the temperature jump between top and bottom walls. The variables v and p represent, respectively, the velocity and pressure of the fluid, and the parameter $\mu = Ra$ stands for the Rayleigh number.

A simple application of the divergence theorem shows that the velocity v and the pressure ∇p are orthogonal in $L^2(\Omega)$, so that v can be expressed as $v = v(u)$. Thus, the system can be entirely written as an equation in terms of u, and steady-state solutions satisfy,

$$f(u, \mu) \equiv -\Delta u + \sqrt{\mu} v(u) \cdot (\nabla u - \mathbf{e}_3) = 0. \tag{44}$$

Observe that the eigenfunctions of the Laplacian operator subject to the boundary conditions imposed on u in (43) are

$$p_{j,k} = \cos(\pi j (y + \frac{1}{2})) \sin(\pi k z), \qquad j = 0, 1, \ldots, \quad k = 1, 2, \ldots .$$

Thus, it is easy to check that the left hand side of (44) is equivariant by the group of symmetries generated by S_y and S_z

$$S_y u(y,z) = u(-y,z), \qquad S_z u(y,z) = -u(y,1-z). \tag{45}$$

In addition, for $p = 2, 3, \ldots$, the subspaces

$$\mathcal{Y}_p = \text{span}\{p_{pl,k}, \quad l = 0, 1, \ldots, \quad k = 1, 2, \ldots\}$$

are invariant by f and enjoy the translational invariance $T_p^s u(y,z) = u(y + \frac{2s}{p}, z)$, for $s = 1, \ldots, p-1$ (see e.g. [32]).

Fig. 1 shows a bifurcation diagram of the equation (44), that is a representation of the branches of its steady-state solutions parameter μ for $0 \leq \mu \leq 325$ (See Remark 3 below for details of the computation of this diagram). A two-dimensional version of this diagram can be found for example in [32]. The vertical axis represents the L^2 norm of the solution u and the transversal axis the value of u at the left mid wall, that is $u(-1/2, 1/2)$. For the values of the parameter μ considered, there are bifurcations enough to test the proposed technique. Here, we have represented only the solutions emanating from the first three bifurcations of the trivial solution $u = 0$. Branches of solutions were followed until $\mu = 325$ was reached or the trivial solution was returned to. Branches of stable and unstable solutions are represented by continuous and discontinuous lines respectively.

For the discretization of (44) we use a standard pseudospectral Fourier method [4]. That is, following the notation of Section 2.1,

$$\mathcal{H}_m = \text{span}\left\{p_{j,k}, \quad 0 \leq j \leq N, \ 1 \leq k \leq N-1\right\}, \quad m = N^2.$$

Nonlinear terms are approximated by standard trigonometric interpolation and evaluated by Fast Fourier Transform (FFT) techniques. In our tests, we set $N = 48$, which makes a total number of $m = 2304$ degrees of freedom. With this value of m, solutions of (44) can be computed with a relative accuracy close to machine precision. This can be seen in Fig. 2, where we show the first nonzero Fourier coefficients of bifurcation point labelled 7 in Fig. 1, which corresponds to a value of the parameter $Ra = 243.06$ Let us mention that for smaller values of Ra a smaller number of Fourier modes suffices to obtain similar levels of accuracy. For simplicity in our computations, though, we have not implemented an adaptive value of m.

This high accuracy should not be surprising since spectral methods are reputed by their fast convergence. For this reason, in the experiments we show, we set a general tolerance TOL to the values $TOL = 5 \times 10^{-5}$, 5×10^{-7}, 5×10^{-9}, and for computation of "exact" reference solutions, $TOL = 10^{-9}$ was used. In the iterative processes such as Newton's method or the bisection method, given an estimation $\mathbf{e}$ of the error of an approximation $\mathbf{u}$, the iteration was stopped whenever

$$\frac{\|\mathbf{e}\|}{\alpha * TOL(1 + \|\mathbf{u}\|)} \leq 1,$$

where α is a factor which for Newton's method was set to $\alpha = 1$ and for the bisection method in locating bifurcations was set to $\alpha = 10$. Notice that the accuracy demanded greatly exceeds what is generally demanded in PDE applications. We use such stringent tolerances in order to severely test the proposed technique.

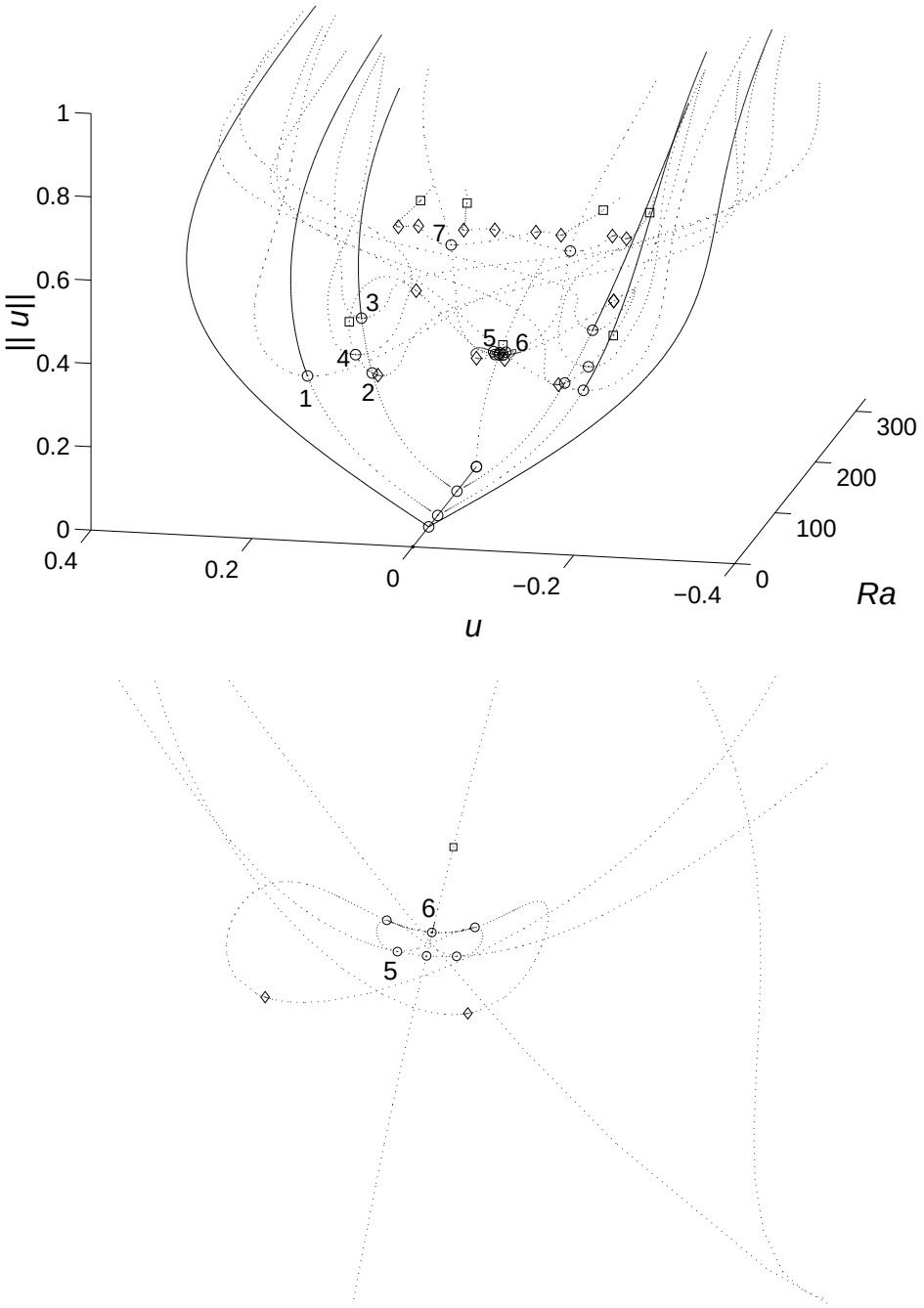

Figure 1: Bifurcation diagram of system (44), $\mu = Ra$ the Rayleigh number. Bifurcation points: ∘, pitchfork or transcritical; ◇, saddle-node; □, Hopf bifurcation. Top: Full diagram. Bottom: detail of bifurcation points 5 and 6.

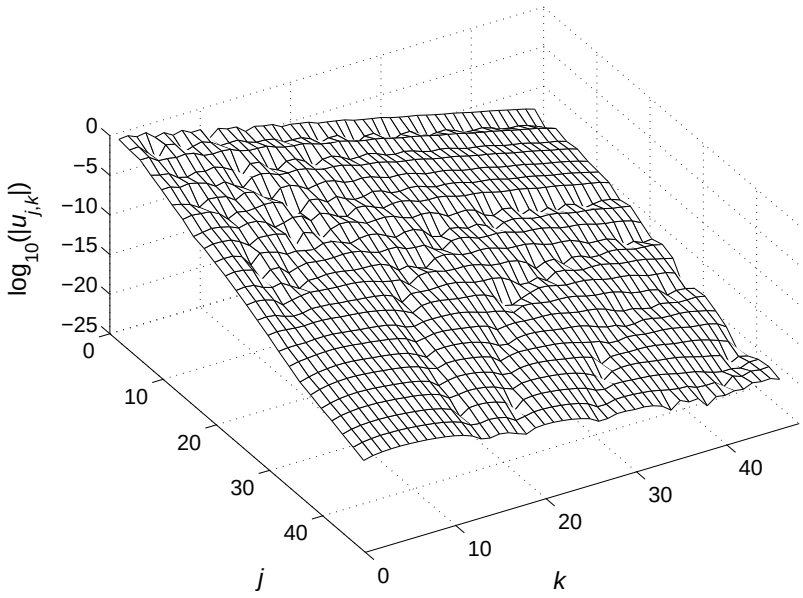

Figure 2: Nonzero Fourier coefficients of bifurcation point 7.

As preconditioner $\mathcal{P}$ in (5), we took

$$\mathcal{A} = -\Delta, \tag{46}$$

subject to the boundary conditions specified in (43), which, in terms of the Fourier coefficients, is a diagonal operator.

Despite being a very simple preconditioner, its usage resulted in a small number of GMRES iterations (see Figs. 6 and 5 below) even up to $Ra = 500$. More sophisticated preconditioners can be used in practice, as for example banded approximations to the Jacobian. In this case, for a branch of solutions $s \mapsto \mathbf{u}(s)$, we would have $\mathcal{P} = \mathcal{P}(s)$, Observe that with more sophisticated preconditioners, it is possible to have the case when $\mathcal{P}^{-1}(s)F_{\mathbf{u}} = I + T$ with $\|T\| < 1$, so that the sign of $\chi_k = \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H_k^{(m)}))$ is given by the sign of $\mathcal{P}$, independently of the value of k. For the testing purposes here, however, it is more advisable to have a simpler preconditioner and let the sign of χ_k depend entirely on $H_k^{(m)}$.

Following the analysis in Section 4, we take as initial guess $\mathbf{x}_0$ for the GMRES method when applied to solve the systems (4) in Newton's method,

$$\mathbf{x}_0 = \alpha_r \|F\| \begin{bmatrix} v_r \\ 0 \end{bmatrix}, \tag{47}$$

with v_r a randomly generated vector with uniform distribution in $(-1, 1)$, and α_r is a scaling parameter. The value of α_r was set to $\alpha_r = 10$. Other values of α_r that we tried did not essentially alter the results we present below.

It is interesting, though, to see what may happen in practice if the initial guess $\mathbf{x}_0$ in the GMRES method is taken $\mathbf{x}_0 = \mathbf{0}$. In Fig. 3, where we show χ_k as a function of $\mu = Ra$ in the branch between bifurcation points marked 2 and 3 in Fig. 1, taking $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{x}_0$ given by (47). When $\mathbf{x}_0 = \mathbf{0}$, besides failing to change sign at the bifurcation points,

the determinant changes sign four times in the middle of the branch. The points where it changes sign are not bifurcation points as it can be seen from the three eigenvalues of f_u closest to zero, represented on the lower plot. On the other hand, taking $\mathbf{x_0}$ as a

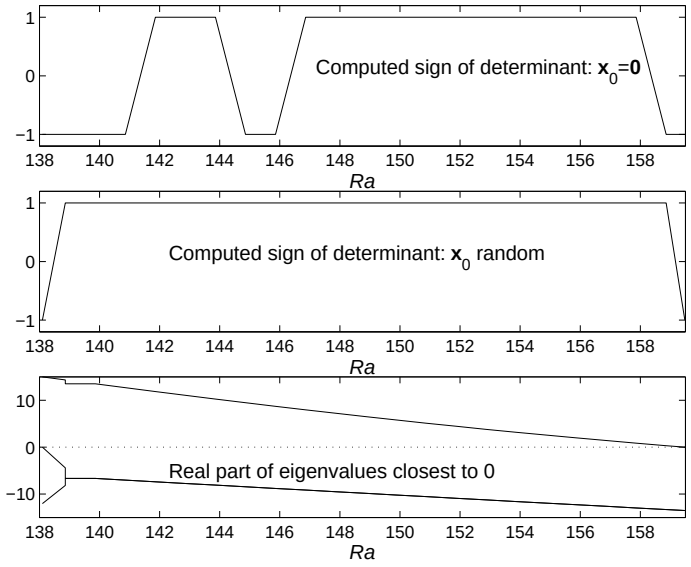

Figure 3: Top: evolution of $\chi_k = \text{sign}(\det(\mathcal{P})\text{sign}(\det(H_k^{(m)}))$ on branch between bifurcation points 2 and 3. Bottom: the three eigenvalues of f_u closest to zero are shown.

random vector (formula (47)), as argued in Section 4, χ_k behaves as the theory predicts.

The spurious changes of sign of χ_k in Fig. 3 are due to the following reasons. This branch is in the invariant subspace $\mathcal{E} \times \mathbb{R}$ where $\mathcal{E} = \mathcal{Y}_3 \cap \{x \in \mathcal{H} \mid S_y S_z x = x\}$, and there is just one eigenvalue of $\mathcal{P}^{-1} F_{\mathbf{u}}$ with negative real part whose associated eigenvector $\mathbf{v}$ satisfies that $\mathbf{v} \in (I - P_{\mathcal{E}})\mathcal{H} \times \{\mathbf{0}\}$. Factors such us round-off in FFT routines, and, in Newtons' method, extrapolation for the first iterate and the number of iterations, make that the linear systems solved with the GMRES method are of the form $A\mathbf{x} = \mathbf{b} + \delta\mathbf{b}$, where $\mathbf{b} \in \mathcal{E} \times \mathbb{R}$ and $\delta\mathbf{b} \in (I - P_{\mathcal{E}})\mathcal{H} \times \{\mathbf{0}\}$. Now recall that the first Arnoldi vector $\mathbf{v}_1$ in the GRMRES method is proportional to the first residual $\mathbf{r_0} = \mathbf{b} + \delta\mathbf{b} - A\mathbf{x}_0$. When $\mathbf{x}_0 = 0$, then $r_0 = \mathbf{b} + \delta\mathbf{b}$, and wether computaions are carried out in the invariant subspace $\mathcal{E} \times \mathbb{R}$ or not (and χ_k is spuriously altered or not) depends on the ratio ratio $\|\delta\mathbf{b}\| \,/\, \|\mathbf{b}\|$, which in practice varies from values close to round-off errors to values closer to 1 depending on the above-mentioned factors. Conversely, if $\mathbf{x_0}$ is given by (47), the ratio $\|(I - P_{\mathcal{E}})\mathbf{r_0}\| / \|\mathbf{r}_0\|$ is always close to one, so that computations are always taken out of $\mathcal{E} \times \mathbb{R}$ and χ_k comes out with right sign.

Of all the branching points shown in Fig. 1 (marked with $\circ$ marks), some of them were reached through the transversal branch ($d\mu(s)/ds = 0$), and were more effectively computed as extrema in the parameter μ. the rest of them, those labelled with a number ranging from 1 to 7 and their images by S_z, were correctly detected by a change of sign in χ_k (with $\mathbf{x}_0$ given by (47)), and furthermore, accurately located as zeros of χ_k by the bisection method. The (approximate) values of the Rayleigh nunber for these points are, respectively, 80.95, 138.01, 159.48, 146.74, 214.23, 216.04 and 243.06.

We remark that system (44) has a $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry, and, as argued in [8] double symmetry-breaking bifurcations occur generically. In these, two eigenvalues of $F_{\mathbf{u}}$ change sign, so that they cannot by detected by a change of sign of χ_k. However, as shown in [8],

[32], those double symmetry-braking bifurcation points are not to be found in the range of the Rayleigh number considered here, so that χ_k changes sign at all branching points in Fig. 1.

We now study how accurate the location of bifurcation points a zeros of χ_k can be. In Fig. 4 we show the relative errors $\|\mathbf{u}_a - \mathbf{u}_e\| / \|\mathbf{u}_e\|$ where $\mathbf{u}_e$ is the bifurcation point and $\mathbf{u}_a$ is the approximation obtained by finding a change of sign of (9) by the bisection method taking (47) as initial approximation in the GMRES method. Results are shown for general tolerance $TOL = 5 \times 10^{-5}$, 5×10^{-7}, 5×10^{-9}. The relative errors committed with the three tolerances are marked, respectively with $+$, $*$ and $\circ$, and, for each tolerance, the errors committed in the seven bifurcation points are joined by a discontinuous line. The corresponding precisions demanded in the bifurcation points, $10TOL$, are marked in the plot with dotted lines. It can be seen that the errors are

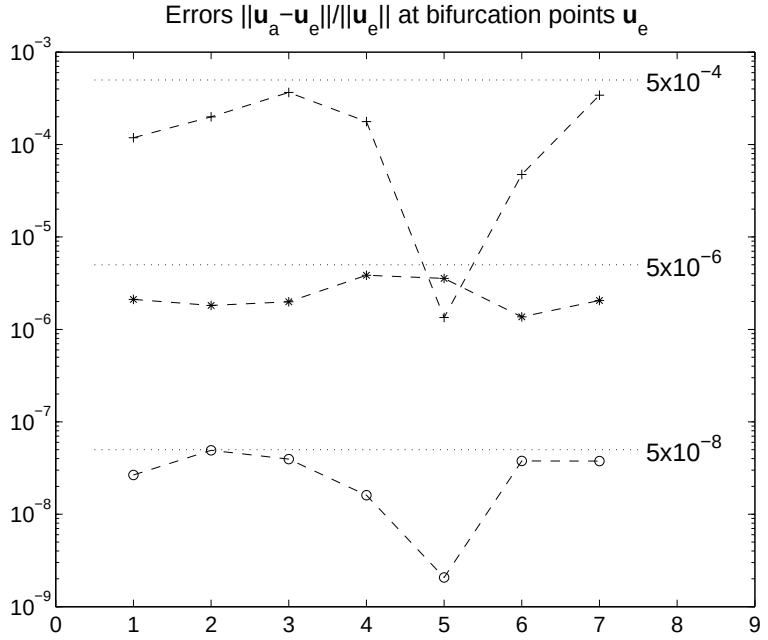

Figure 4: Errors in bifurcation location by zeroing χ_k: $+$, $TOL = 5 \times 10^{-5}$; $*$, $TOL = 5 \times 10^{-7}$; $\circ$, $TOL = 5 \times 10^{-7}$. Accuracy demanded is $10TOL$.

below the precision demanded.

To achieve these errors, it was also necessary to reduce the tolerance TOL_{MR} of the stopping criterion (42) in the GMRES method. Thus, although for general $TOL = 5 \times 10^{-5}$ we kept $TOL_{MR} = 5 \times 10^{-4}$, for $TOL = 5 \times 10^{-7}$ we set $TOL_{MR} = 5 \times 10^{-5}$, and further, for $TOL = 5 \times 10^{-9}$ we took $TOL_{MR} = 5 \times 10^{-6}$. Failing to do this resulted in some of the computations not reaching the accuracy levels demanded. For example, when using $TOL_{MR} = 5 \times 10^{-4}$ with $TOL = 5 \times 10^{-7}$, the sixth bifurcation point failed to achieve the accuracy of 5×10^{-6} (the rest of the points achieved the same errors as in Fig. 4); if $TOL_{MR} = 5 \times 10^{-5}$ is used with $TOL = 5 \times 10^{-9}$, then it was the second bifurcation point who could not be computed with errors below 5×10^{-8}.

We believe that this need to reduce TOL_{MR} when more precision is demanded in the computed solutions (i.e., when the general tolerance TOL decreases) is related to the comments made in Remark 2. Notice that, the more stringent the precision demanded,

the closer the computed points $\mathbf{u}_a$ will be to the true bifurcation point $\mathbf{u}_e$, and, consequently, the closer to singular becomes $\mathcal{P}^{-1}F_{\mathbf{u}} = I + T$. Then, reducing the stopping criterion tolerance TOL_{MR} in the GMRES method increases the number k of iterations, which is the key to guarantee $\det(H_k)$ having the correct sign.

So far, computing the correct signs has required taking the initial approximation $\mathbf{x}_0$ in the GMRES method as indicated (47), and increasing precision in the GMRES method when close to a bifurcation point. It is natural to ask how all this affects to cost. In Fig. 5 we show the numbers of GMRES iterations corresponding to the computation

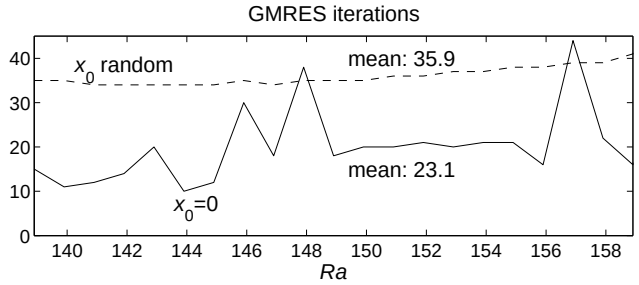

Figure 5: Iterations k in GMRES method between bifurcation points 2 and 3.

of regular points between bifurcation points 2 and 3, that is, corresponding to results shown in Fig. 3. Here TOL_{MR} is $TOL_{MR} = 5 \times 10^{-4}$. We also show the average number of iterations for the two options of choosing $\mathbf{x_0}$. We see that, on average, (47) results in a 50% increase in cost.

In Fig. 6, we show the maximum number of iterations in the GMRES method in the computation of each of the seven bifurcation points, with the same convention as in Fig. 4. Also, joined by a continuous line, we show the results corresponding to taking $\mathbf{x_0} = \mathbf{0}$ and

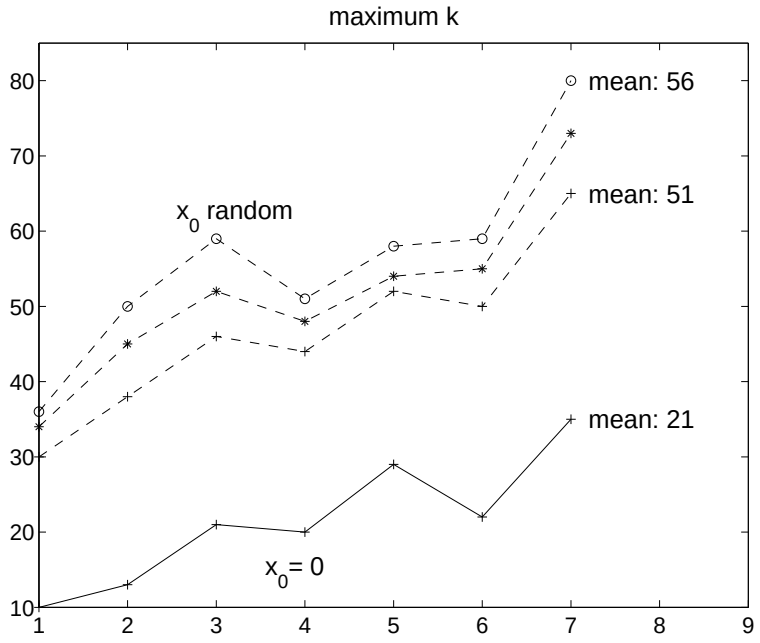

Figure 6: Maximum number of iterations k in GMRES method.

$TOL_{MR} = 5 \times 10^{-4}$. Comparing the results marking with + signs, we see that now, when computing bifurcation points, using (47) on average doubles the cost of using $\mathbf{x_0} = \mathbf{0}$, and that in some points the cost is three times larger. Comparing the results joined by discontinuous lines, we see that reducing the stopping criterion tolerance TOL_{MR} in the GMRES method from 5×10^{-4} to 5×10^{-6} results in a less than a 10% increase in cost.

In Figs. 5 and 6 we notice that the major increase in cost is due to using (47) instead of taking $\mathbf{x_0} = \mathbf{0}$. However, once that cost is paid so that $\det(H_k)$ has the correct sign, increasing the precision in the GMRES method (reducing TOL_{MR}) adds little to the cost. Since, as seen in Fig. 4, reducing TOL_{MR} has played its role in accurately computing bifurcation points, we may draw the conclusion that, in order to play on the safer side, it is advisable to increase the precision demanded to GMRES method when accurately locating bifurcations. Observe also that the number of iterations is low when compared with the dimension $m = 2304$ of the problem.

In spite of the increase in cost caused by (47), we must point out that computing bifurcation points as zeros of χ_k compares favourably with computing them as zeros of an eigenvalue of f_u. In our tests, typically, around 300 matrix-vector products were required to compute the relevant rightmost part of the spectrum of f_u, a quantity that greatly exceeds the numbers of iterations shown in previous figures.

Remark 3 All numerical experiments were carried out in a SUN Ultra 60 workstation running Solaris 8. All programs were written in FORTRAN, and compiled with the Workshop 5 compiler. Saddle-node bifurcations were located as extrema of the parameter μ as well as pitchfork bifurcations whenever possible. Hopf bifurcations were computed as zeros of the real part of one eigenvalue of f_u. Pitchfork and transcritical bifurcations, besides being computed with the techniques described in this paper, were also computed (for the purpose of having more accurate reference solutions) as zeros of eigenvalues of f_u. These were computed both by implicitly restarted Arnoldi iteration from the ARPACK package [26] as well as, for double checking, with standard QR iteration from LAPACK routines. Reference solutions were computed with a general tolerance of $TOL = 10^{-9}$. The diagram was first computed with $N = 24$ ($m = 576$). Branches where selected bifurcation points where located, were repeated with $N = 48$ ($m = 2304$).

6 Conclusions and remarks

We have seen, both theoretically and practically that the matrix H_k of the Arnoldi decomposition (8) adequately reproduces the orientation of certain operators. These are of the form $I + T$ with T compact but not necessarily self-adjoint, a form adopted by many differential operators when preconditioned by a fast solver of their higher derivative terms. From the theoretical point of view, an important role in the convergence of the Krylov sequence is played by the singular values of T. This has allowed to prove convergence of $\text{sign}(\det(H_k))$, not only for the operator $I + T$, but also for its discretization by spectral methods.

As a consequence, $\text{sign}(\det(H_k))$ is a useful tool for detecting and locating bifurcations in branches of equilibria, and convenient from the practical point of view since it is obtained as a byproduct of the preconditioned GMRES method when applied to solve the linear systems that arise in the equilibrium computations via Keller's pseudo-arclength

continuation technique. This is so provided that a random initial guess is taken in the GMRES method, so that the Arnoldi decomposition in the GMRES method is not limited to invariant subspaces which do not contain the relevant eigendirections at bifurcation points.

As argued in Section 2.3, the technique studied here is an alternative to other more sophisticated techniques, which may prove useful for preliminary computations, and whose main advantage is its simplicity. Notice that existing techniques for branching point location require either costly eigenvalue computations or providing codes for f and the action of its differential f_u as mappings from certain invariant subspaces onto themselves, or, to provide a code for the adjoint operator (recall we are dealing with problems who require matrix-free methods), which, in view of systems like (44), may discourage some potential users.

Acknowledgements. The research of B. G.-A. has been partially supported by DGICYT project PB98-0072. The research of J. S. has been partially supported by DGICYT project BFM2001-2336. The research of C. S. has been supported by grants DGICYT BFM2000-805 and CIRIT 2000SGR-27.

References

[1] R. E. Bank and T. F. Chan, PLTMGC: A multigrid continuation program for parametrized nonlinear elliptic systems, *SIAM J. Sci. Stat. Comput.*, **7** (1986), pp. 540–559.

[2] R. E. Bank *PLTMG: A Software Package for Solving Elliptic Partial Differential equations. Users Guide 8.0* SIAM, Philadelphia, 1998.

[3] W. J. Beyn, A. Champneys, E. Doedel, W. Govaerts, Y. A. Kuznetsov and B. Sandstede, Numerical Continuation and Computation of normal forms, in B. Fiedler, G. Iooss and N. Kopell (eds.), *Handbook of Dynamical Systems: Vol. 2*, Elsevier, 2002, pp. 149-164.

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer Series in Computational Physics, Springer-Verlag, Berlin, 1988.

[5] T. F. Chan, Deflation techniques and block-elimination algorithms for solving bordered singular systems, *SIAM J. Sci. Stat. Comput.*, **5** (1984), pp. 121–134.

[6] P. Chossat, and R. Lauterbach, *Methods in equivariant bifurcations and dynamical systems*, World Scientific Publishing Co., Inc., River Edge, NJ, 2000.

[7] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

[8] K. A. Cliffe and K. H. Winters, The use of symmetry in bifurcation calculations and its application to the Bérnad Problem, *J. Comput. Phys.*, **67** (1986), pp. 310–326.

[9] B. D. Davidson, Large-scale continuation and numerical bifurcation for partial differential equations, *SIAM J. Numer. Anal.*, **34** (1997), pp. 2008–2027.

[10] E. J. Doedel, *Lecture Notes on Numerical Analysis of Bifurcation Problems*, Lecture notes from Sommerschule über Nichtlineare Gleichungssysteme, Hamburg, Germany, March 17–21, 1997. (Available by anonymous ftp to ftp.cs.concordia.ca in put/doedel/doc/hamburg.ps.Z

[11] W. S. Edwards, L. S. Tuckerman, R. A. Friesner and D. C. Sorensen, Krylov methods for the incompressible Navier-Stokes equations, *J. Comput. Phys.*, **110** (1994), pp. 82–102.

[12] M. Eiermann and O. G. Ernst, Geometric aspects of Krylov subspace methods, *Acta Numerica 2001*, Cambridge University Press, Cambridge, 2000, pp. 251–312.

[13] W. R. Ferng and C. T. Kelley, Mesh independence of matrix-free methods for path following, *SIAM J. Sci. Comput.*, **21** (2000), pp. 1835–1850.

[14] I. Goldhirsch, S.A. Orszag and B.K. Maulik, An efficient method for computing leading eigenvalues and eigenvectors of large asymmetric matrices. *J. Sci. Comp.*, **2** (1987), pp. 33-58.

[15] G. H. Golub & C. F. Van Loan, *Matrix Computations (2nd. Ed.)*, The John Hopkins University Press, London, 1989.

[16] G. H. Golub and H. A. van der Vorst, Closer to solution: Iterative linear solvers, in *State of the Art in Numerical Analysis* I. S. Duff and G. A. Watson Eds. Clarendon Press, Oxford, 1997, pp. 63–92.

[17] M. Golubitsky, I. Stewart and D. G. Schaeffer, *Singularities and groups in bifurcation theory. Vol. II.*, Springer-Verlag, New York, 1988.

[18] W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, SIAM, Philadelphia, 2000.

[19] A. Greenbaum, V. Ptak and Z. Strakoš, Any nonincreasing convergence curve is possible for GMRES, *SIAM J. Matrix Anal. Appl.*, **17** (1996), pp. 465–469.

[20] K. G. Jea and D. M. Young, Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods, *Lin. Alg. Appl.*, **34** (1980), pp. 159–194.

[21] A. Jorba, Numerical computation of the normal behaviour of invariant curves of n-dimensional maps, *Nonlinearity*, **14** (2001), pp. 943–976.

[22] H. B. Keller, Constructive methods for bifurcation and nonlinear eigenvalue problems, in *Computing Methods in Applied Science and Engineering*, Lecture Notes in Math. 704, R. Glowinski and J. Lions, eds., Springer Verlag, New York, 1977, pp. 241–251.

[23] H. B. Keller, Numerical solution of bifurcation and nonlinear eigenvalue problems, in *Applications of Bifurcation Theory*, P. H. Rabinowitz, ed., Academic Press, New York, 1977, pp. 359–384.

[24] T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1995.

[25] Y. A. Kuznetsov, *Elements of applied bifurcation theory (2nd Ed.)*, Springer, New York, 1998.

[26] R. B. Lehoucq and D.C. Sorensen, Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM J. Matrix Anal. Appl.*, **17** (1996), pp. 789–821.

[27] R. B. Lehoucq and D.C. Sorensen and C. Yang, *ARPACK users' guide. Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM, Philadelphia, 1998.

[28] J.M. López, F. Marqués and J. Sánchez, Oscillatory modes in enclosed swirling flow, *J. Fluid Mech.*, **439** (2001), pp. 109-129.

[29] C. K. Mamun and L. S. Tuckerman, Asymmetry and Hopf bifurcation in spherical Couette flow, *Phys. of Fluids*, **7** (1995), pp. 80-91.

[30] K. Meerbergen and D. Roose, Matrix transformations for computing rightmost eigenvalues of large sparse non-symmetric eigenvalue problems. *IMA J. Numer. Anal.*, **16** (1996), pp. 297–346.

[31] I. Moret, A note on the superlinear convergence of GMRES, *SIAM J. Numer. Anal.*, **34** (1997), pp. 513–516

[32] R. S. Riley and K. H. Winters, Modal exchange mechanism in Lapwood convection, *J. Fluid. Mech.*, **204** (1989), pp. 325–358.

[33] Y. Saad and M. H. Schultz, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Stat. Comput.*, **7** (1986), pp. 856–869.

[34] J. Sánchez, M. Net, G. García-Archilla and C. Simó, Newton-Krylov continuation of periodic orbits for Navier-Stokes Flows, *J. Comput. Phys.* (to appear).

[35] R. Seidel, *Practical Bifurcation and Stability Analysis. From Equilibrium to Chaos. (2nd Ed)*, Springer, New York, 1994.

[36] C. Simó, Analytical and numerical computation of invariant manifolds. In *Modern methods in celestial mechanics*, Ed. D. Benest and C. Froeschlé, pp. 285-330, Editions Frontières, 1990.

[37] C. Simó, Effective Computations in Celestial Mechanics and Astrodynamics. In *Modern Methods of Analytical Mechanics and their Applications*, Ed. V. V. Rumyantsev and A. V. Karapetyan, CISM Courses and Lectures **387**, pp. 55–102, Springer, 1998.

[38] P.H. Steen, Pattern selection for finite-amplitude convection states in boxes of porous media. *J. Fluid Mech.*, **136** (1983), pp. 219–241.

[39] P.H. Steen, Container geometry and the transition to unsteady Bérnard convection in porous media, *Phys. Fluids*, **29(4)** (1964), pp. 925–933.

[40] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Applied Mathematical Sciences, **68**, Springer-Verlag, Berlin, 1988.

[41] L. N. Trefethen, D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[42] B. Werner and A. Spence, The computation of symmetry-breaking bifurcation points, *SIAM J. Numer. Anal.*, **21** (1984), pp. 388–399.