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 preconditioned by a fast solver. Numerical experiments are presented where the technique is applied to locate pitchfork and transcritical bifurcations on a one parameter dependent system. Experiments reveal that some extra precautions may have to be taken in the presence of symmetries.

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 [29] for instance) are inapplicable 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 with spectral methods and nonlinear equations, despite it is true in general that spectral methods allow to reduce the dimensionality of the problem. 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 [13], [22], and are regularly used in the studies of large dynamical systems (see e.g. [7], [10], [9], [24], [25]). 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, are used for detection and location of pitchfork and transcritical bifurcations, a particular case of bifurcations of codimension one.

In this paper, we study the computation of the sign of the determinant of a large matrix as a byproduct of the GMRES method [28] 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 operator with $D(R(\cdot,\mu)) \subset D(A)$, so that the Fréchet differential of $\mathcal{A}^{-1}R(\cdot,\mu)$ is compact. 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 and thus, $\mathcal{A}$ can be used as a preconditioner.

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. This is done first in a Hilbert space, for operators of the the form $I+T$ with T compact but not necessarily self-adjoint, and secondly for the discretizations by spectral methods of such operators. Singular values, rather than eigenvalues, will be a key element in the convergence result in Theorem 4, the main result of this paper. The last section contains numerical experiments where the previous material is applied to locate pitchfork and transcritical bifurcations. The difficulties encountered by the symmetries of (1) which are preserved by the preconditioner $\mathcal{A}$ are analyzed and, as a result, a simple remedy is proposed and tested.

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 $S_m = \mathrm{span}(p_1, \dots, p_{m-1})$, and let $P_m : \mathcal{H} \to S_m$ be the orthogonal projection onto S_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

$$\text{find} \quad \mathbf{u}_m = \left[\begin{array}{c} u_m \\ \mu \end{array} \right] \in S_m \times \mathbb{R} \quad \text{s.t.} \qquad f^{(m)}(\mathbf{u}_m) \equiv \mathcal{A} u_m + P_m R(\mathbf{u}_m) = 0. \qquad (2)$$

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 [17], [18], [21] [29], which requires at every step the solution of the system

$$F^{(m)}(u, \mu) = \left[\begin{array}{c} f^{(m)}(u, \mu) \\ (\mathbf{t}_r, (\mathbf{u} - \mathbf{u}_r)) - \delta_r \end{array} \right] = \left[\begin{array}{c} 0 \\ 0 \end{array} \right] = \mathbf{0}, \qquad (3)$$

where $\mathbf{u}_r$ is a reference point, typically the most recently available solution of (2), $\mathbf{t}_r$ is an approximation to the the tangent $d\mathbf{u}(s)/ds$ 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

$$F_{\mathbf{u}} \mathbf{d} = -F. \qquad (4)$$

Of the various methods to solve (4) based on $\mathbf{v} \mapsto F_{\mathbf{u}} \mathbf{v}$ operations (see e.g. [13]) we focus on the GMRES method [28], although our discussion covers other methods based on the Arnoldi decomposition of a matrix such as ORTHORES [15].

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, and, hence, easily invertible.

2.2 Arnoldi decomposition

In an Arnoldi decomposition of an operator A such as $\mathcal{P}^{-1}F_{\mathbf{u}}$, starting from an initial iterant $\mathbf{v}_1$ (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

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

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

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

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

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

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 $\mathbf{v}_1, \dots, \mathbf{v}_k$.

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 [15] 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 build in a sequential process, starting from $k = 1$ and stopping when the residual $\|\mathbf{b} - A\mathbf{x}_k\|$ 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

$$u_t = f(u, \mu) \tag{9}$$

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 the stability of equilibria, the spectra of $f_{u_m}^{(m)}$ or its most relevant part [22], [23], 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 in pitchfork or transcritical bifurcations.

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 [30], [16] and references therein).

However, since eigenvalue computations are in general much more costly than the solution of (3) by Newton's method, one would like to do eigenvalue-free computations. Much of these techniques are not easily implementable as matrix-free methods, specially those for Hopf bifurcations [11]. Thus, in the case we are concerned with, one must monitor the right-most part of the spectrum of f_u to detect bifurcations, although, as we explain below, once a branch crossing is detected some eigenvalue-free computations can be carried out to find its exact location.

In branch crossings, one eigenvalue of $F_{\mathbf{u}}^{(m)}$ changes sign along any of the two crossing branches in both pitchfork or transcritical bifurcations. 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 $F_{\mathbf{u}}^{(m)}$, or even by solving some generalized eigenvalue problem as in [9]. Since this must be done by inverse iteration or the power method applied to the inverse of a matrix (see e.g.[12], § 7.3.1) some extra linear systems must be solved. Also, due to the ill-conditioning of $F_{\mathbf{u}}^{(m)}$, shift-invert techniques [26] (i.e. solving linear systems) are usually required when more sophisticated techniques [22] are used to locate the smallest eigenvalue of $F_{\mathbf{u}}^{(m)}$. No extra linear system is required, though, if the bifurcation is located by finding a zero of $\mathrm{sign}(\det(F_{\mathbf{u}}))$ by the bisection method (to avoid possible overflow or underflow, it is advisable to compute $\mathrm{sign}(\det(F_{\mathbf{u}}))$ rather than $\det(F_{\mathbf{u}})$ [6], despite computing the logarithm of $|\det(F_{\mathbf{u}})|$ can reduce the number of iterations). 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 [6]. 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 make take

$$\chi_k = \mathrm{sign}(\det(\mathcal{P}))\mathrm{sign}(\det(H_k)) \tag{10}$$

as an approximation to $\mathrm{sign}(\det(F_{\mathbf{u}}))$. 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 build recursively as a product of Givens rotations,

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

where if we express

$$\begin{bmatrix} Q_{k-1} & 0 \\ 0^T & 1 \end{bmatrix} \tilde{H}_k = \begin{bmatrix} R_{k-1} & r_k \\ 0^T & \eta_k \\ 0^T & h_{k+1,k} \end{bmatrix} \tag{11}$$

then

$$G_k = \begin{bmatrix} I_{k-1} & 0 & 0 \\ 0^T & c_k & s_k \\ 0^T & -s_k & c_k \end{bmatrix}, \qquad c_k = \frac{\eta_k}{\sqrt{\eta_k^2 + h_{k,k+1}^2}}, \quad s_k = \frac{h_{k,k+1}}{\sqrt{\eta_k^2 + h_{k+1,k}^2}}. \tag{12}$$

In view of (12), we observe that for $l = 1, \dots, L-1$, $\det(G_l) = 1$, and, consequently, $\det(Q_l) = 1$. Thus, in view of (11), we have that $\det(H_k) = \eta_k \det(R_{l-1})$. But, due to the way the matrices G_l, $l = 1, \dots, L-1$ are build, we deduce that all diagonal elements of R_{k-1} are positive. Consequently,

$$\mathrm{sign}(\det(H_k)) = \mathrm{sign}(\eta_k) = \mathrm{sign}(c_k), \qquad k = 1, \dots, L-1. \tag{13}$$

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 $\mathrm{sign}(\det(F_{\mathbf{u}}))$.

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 $\mathrm{sign}(\det(H_k)) \approx \mathrm{sign}(\det(\mathcal{P}^{-1} F_{\mathbf{u}}))$ in the Hilbert space $\hat{\mathcal{H}}$, that is, the case $m = \infty$ in (2), where $S_\infty = \mathcal{H}$ 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 [8], any possible pathological behaviour of Krylov methods is attainable in finite-dimensional spaces. The following matrix, used in [14] to show that the GMRES method may achieve the worst possible convergence sequence, also shows that the approximation $\mathrm{sign}(\det(H_k)) \approx \mathrm{sign}(\det(A))$ is completely wrong until the final step $k = m$. Consider the matrix,

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

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 $k = 1, \ldots, m-1$, $\det(H_k) = 0 \neq \operatorname{sign}(\alpha_0) = \operatorname{sign}(\det(A))$. To exclude this and similar situations, we need to understand first the infinite-dimensional case.

3.1 The infinite-dimensional case

The mappings $f^{(m)}$ in (2) are approximation 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, 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 $\mathcal{P}^{-1}F_{\mathbf{u}}^{(m)}$ 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

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

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. [19]).

For operators in infinite-dimensional spaces, it is not generally possible to extend the idea of determinant [19], § X.1.4. Thus, we consider instead the orientation. If $(I+K)$ is invertible, we define the *orientation* of $I+K$ as

$$\operatorname{or}(I+K) = (-1)^{d(K)}, \qquad d(K) = \sum_{\lambda \in \sigma_-(I+K)} \dim(E_\lambda). \tag{14}$$

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

The following result is a consequence of Theorem 3.11 and Remark 3.2 in [19], § IV.5, and [19], § I.4.6.

Theorem 1 *Let K 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*

$$\|B\| \leq \frac{1}{2\min_{\xi\in\Gamma}\|(\xi I-(I+K))^{-1}\|}, \tag{15}$$

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

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

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

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

Throughout this section we will assume that the minimal invariant subspace $\mathcal{V}$ containing the first Arnoldi vector $\mathbf{v}_1$ is infinite-dimensional. Otherwise, everything could be computed exactly in a finite number of steps and there would be nothing to understand.

Also, we will assume that $\mathcal{V} = \hat{\mathcal{H}}$. Otherwise, all that is said in this section would apply with $\hat{\mathcal{H}}$ replaced by $\mathcal{V}$.

The main result of this Section is Theorem 4 below. Its proof will be a consequence of previous results which we now state and prove.

Lemma 1 *Let $\mathcal{X}$ and $\mathcal{U}$ be closed subspaces of the Hilbert space $\hat{\mathcal{H}}$ satisfying that for some positive $\epsilon_s < 1/\sqrt{2}$,*

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

Then, for every $y \in \hat{\mathcal{H}}$ such that $P_{\mathcal{X}} y = 0$ the following bounds hold:

$$\|P_{\mathcal{U}} y\| \leq \epsilon_s \|y\|, \qquad \|(I - P_{\mathcal{U}}) y\| \geq \sqrt{1-\epsilon_s^2}\, \|y\|.$$

Proof. There is no loss of generality in assuming that $\|y\| = 1$. Let us denote $c = \sqrt{1-\epsilon_s^2}$, and let us write

$$y = u + z, \qquad u = P_{\mathcal{U}} y, \qquad z = (I - P_{\mathcal{U}}) y, \tag{17}$$

$$u = x + r, \qquad x = P_{\mathcal{X}} u, \quad \text{and} \quad r = (I - P_{\mathcal{X}}) u. \tag{18}$$

By hypothesis we have that $\|x\| \geq c\|u\|$. From (18) follows that $\|u\|^2 = \|x\|^2 + \|r\|^2 \geq c^2 \|u\|^2 + \|r\|^2$ and thus

$$\|r\|^2 \leq \epsilon_s^2 \|u\|^2. \tag{19}$$

On the other hand, multiplying by y in (17) we have that

$$1 = (y, x+r) + (y, z) = (y, x+r) + \|z\|^2 = (y, r) + \|z\|^2,$$

so that

$$\|z\|^2 = 1 - (y, r) \geq 1 - |(y, r)| \geq 1 - \|y\| \|r\| = 1 - \|r\|,$$

which in view of (19) implies

$$\|z\|^2 \geq 1 - \epsilon_s \|u\|, \tag{20}$$

Since, by Pythagoras theorem,

$$1 = \|u\|^2 + \|z\|^2 \geq \|u\|^2 + (1 - \epsilon_s \|u\|) = 1 + \|u\| (\|u\| - \epsilon_s),$$

which implies that

$$\|u\| \left(\|u\| - \epsilon_s\right) \leq 0,$$

and $\|u\| \leq \epsilon_s$ follows. This and (20) imply $\|z\|^2 \geq 1 - \epsilon_s^2$. ◻

We now study the behaviour of a compact operator in a subspace which is nearly orthogonal to the subspace of the vectors that make most of its norm. Let K a compact operator in a Hilbert space and let $\sigma_1 \geq \sigma_2 \ldots$, the eigenvalues (counted with their multiplicities) of K^*K. Let $u_1, u_2, \ldots$ the complete orthonormal set of the associated eigenvectors and let us define $w_j = Ku_j$, for $j = 1, 2, \ldots$. The *singular value decomposition* (SVD) of K is the given by

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

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

$$\mathcal{U}_\epsilon = \operatorname{span}\{u_j \mid \sigma_j \geq \epsilon\}. \tag{21}$$

We have the following result.

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

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

Proof. Let y such that $\|y\| = 1$ and $P_{\mathcal{X}} y = 0$ and let us write

$$y = u + z, \qquad u = P_{\mathcal{U}_\epsilon} y, \qquad z = (I - P_{\mathcal{U}_\epsilon}) y.$$

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 $\|z\| \leq \|y\| = 1$, it follows that

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

Theorem 3 *For every $0 < \epsilon \leq 1/\sqrt{2}$ there exist positive integers k_1 and m_1 given by (24), (25) below, such that the Arnoldi vectors $(\mathbf{v}_k)_{k=1}^{\infty}$ of $I + T$ satisfy that*

$$\|T\mathbf{v}_k\| \leq \epsilon, \quad \text{for } k \geq k_1, \tag{22}$$

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

$$\left\|P_m T \mathbf{v}_k^{(m)}\right\| \leq \epsilon, \quad \text{for } k_1 \leq k \leq m. \tag{23}$$

Proof. We take
$$\epsilon_1 = \frac{\epsilon}{\sqrt{2}}, \qquad \epsilon_s = \frac{\epsilon}{\sqrt{2}\,\|T\|}.$$
Let $T = \Sigma_{j=1}^{\infty} \sigma_j w_j P_{u_j}$ a SVD of T and let us consider $\mathcal{U}_{\epsilon_1} = \operatorname{span}\{u_j \mid \sigma_j \geq \epsilon_1\}$.

Since $P_{\mathcal{V}_k} \to I$ as $k \to \infty$, there exists an integer k_1 such that
$$\left\| P_{\mathcal{V}_{k_1}} u \right\| \geq \sqrt{1 - \frac{\epsilon_s^2}{2}}\,\|u\| = \sqrt{1 - \frac{\epsilon^2}{4\,\|T\|^2}}\,\|u\|, \qquad \text{for all } u \in \mathcal{U}_{\epsilon_1}. \tag{24}$$
Now, we apply Lemma 2 with $K = T$, and $\mathcal{X} = \mathcal{V}_{k_1}$ to get that,
$$\|Ty\| \leq \left(\frac{\epsilon^2}{4} + \frac{\epsilon^2}{2} \right)^{1/2} \|y\| \leq \epsilon \|y\|, \qquad \text{for} \quad y \in (I - P_{\mathcal{V}_{k_1}})H,$$
which implies (22).

We now prove (23). Since the generation of the Arnoldi basis $(\mathbf{v}_k)_{k=1}^{\infty}$ is, for a fixed k, a continuous mapping of $\mathbf{v}_1$, there exists a m_1 such that
$$\left\| P_{\mathcal{V}_{k_1}^{(m)}} - P_{\mathcal{V}_{k_1}} \right\| \leq \frac{\epsilon_s^2}{8} = \frac{\epsilon^2}{16\,\|T\|^2}, \qquad \text{for all } m \geq m_1. \tag{25}$$
Let as consider now $m \geq m_1$. In view of (24–25), for every $u \in \mathcal{U}_{\epsilon_1}$ we have that
$$\left\| P_{\mathcal{V}_{k_1}^{(m)}} u \right\| \geq \left\| P_{\mathcal{V}_{k_1}} u \right\| - \left\| (P_{\mathcal{V}_{k_1}^{(m)}} - P_{\mathcal{V}_{k_1}}) u \right\| \geq \sqrt{1 - \frac{\epsilon_s^2}{2}}\,\|u\| - \frac{\epsilon_s^2}{8}\,\|u\|.$$
A simple calculation shows that $\sqrt{1 - \epsilon_s^2/2} - \epsilon_s^2/8 \geq \sqrt{1 - \epsilon_s^2}$, so that
$$\left\| P_{\mathcal{V}_{k_1}^{(m)}} u \right\| \geq \sqrt{1 - \epsilon_s^2}\,\|u\|, \qquad u \in \mathcal{U}_{\epsilon_1}, \quad m \geq m_1. \tag{26}$$
Thus, applying Lemma 2 with $K = T$ and $\mathcal{X} = \mathcal{V}_{k_1}^{(m)}$ we have that
$$\|Ty\| \leq \left(\frac{\epsilon^2}{2} + \frac{\epsilon^2}{2} \right)^{1/2} \|y\| = \epsilon \|y\|, \qquad \text{for} \quad y \in (I - P_{\mathcal{V}_{k_1}^{(m)}})H, \quad m \geq m_1,$$
which, taking into account that $\|P_m T y\| \leq \|Ty\|$, implies (23). □

Lemma 3 *Assume that the value of s_k in (12) satisfies that $|s_k| < 1/\sqrt{2}$. If $\|T(\mathbf{v}_{k+1})\| < 1/\sqrt{2}$, then*
$$\operatorname{sign}(\det(H_{k+1})) = \operatorname{sign}(\det(H_k)). \tag{27}$$
Proof. Since in (13) we saw that $\operatorname{sign}(\det(H_k)) = \operatorname{sign}(\eta_k) = \operatorname{sign}(c_k)$, we will show that $\operatorname{sign}(\eta_{k+1}) = \operatorname{sign}(c_k)$. Recall (11–13), 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

$$(\|f\|^2 + |\gamma|^2)^{1/2} = \left\| P_{\mathcal{V}_k}(I+T)\mathbf{v}_{k+1} \right\| = \left\| P_{\mathcal{V}_k} T\mathbf{v}_{k+1} \right\|, \tag{28}$$

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,

$$\eta_{k+1} = c_k h_{k+1,k+1} - s_k\gamma = c_k + \Big(c_k(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1}) - s_k\gamma \Big). \tag{29}$$

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 (28) we have that

$$\begin{aligned} |c_k(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1}) - s_k\gamma| &\leq (c_k^2 + s_k^2)\Big(|(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|^2 + |\gamma|^2 \Big)^{1/2} \\ &\leq \Big(|(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|^2 + \|P_{\mathcal{V}_k} T\mathbf{v}_{k+1}\|^2 \Big)^{1/2} \\ &\leq \Big(\|T\mathbf{v}_{k+1}\|^2 - |h_{k+2,k+1}| \Big)^{1/2} < \frac{1}{\sqrt{2}} < |c_k|, \end{aligned} \tag{30}$$

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

Lemma 3 guarantees the same sign for the determinants of H_k and H_{k+1} provided $\|Tv_{k+1}\|$ 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 [8] states that

$$|s_k| \leq \left\| (I+T)^{-1} \right\| |(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|, \qquad k = 1, 2, \ldots. \tag{31}$$

Thus, choosing

$$\epsilon = \frac{1}{\sqrt{2}\max(1, \|(I+T)^{-1}\|)}, \tag{32}$$

Theorem 3 and (31) 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 (32) set $\mathcal{U}_{\epsilon/\sqrt{2}} = \mathrm{span}\{u_j \mid \sigma_j \geq \epsilon/\sqrt{2}\}$. Choose k_1 such that*

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

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

Furthermore, choose m_1 such that

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

Then, $\mathrm{sign}(\det(H_k^{(m)})) = \mathrm{sign}(\det(H_{k_1}^{(m)})) = \mathrm{or}(T)$*, 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 $m \geq m_2$, $\mathrm{sign}(\det(H_k^{(m)}))$ $= \mathrm{sign}(\det(H_k))$, for $k = 1, \ldots, m$. This value will obviously depend on factors such as the rate of decay of $\|(I - P_m)T\|$.

4 Numerical experiments

In this section we test numerically how correctly $\chi_k = \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H_k^{(m)}))$ approximates $\text{sign}(\det(F_{\mathbf{u}_m}^{(m)}))$ 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 $\|\mathbf{b} - A\mathbf{x}\| / \|\mathbf{b}\| \leq TOL_{MR}$. We set the value of the prescribed tolerance TOL_{MR} to $TOL_{MR} = 5 \times 10^{-4}$.

For numerical experiments, we consider the steady-state Kuramoto-Sivashinsky (KS) equation,

$$4u_{xxxx} + \mu(u_{xx} + uu_x) = 0, \qquad u \in L^2_{per,odd} \tag{33}$$

where the solutions are demanded to 2π-periodic and odd, that is

$$L^2_{per,odd} = \{u \in L^2(-\pi, \pi) \mid u(-x) = -u(x), \quad \forall x \in (-\pi, \pi)\}.$$

The KS equation is a model equation for phenomena in Combustion, Hydrodynamics, Plasma Physics, etc (see e.g. [20], [31], [32]) and has received much attention in recent years for its rich dynamics.

Fig. 1 shows a bifurcation diagram of the KS equation, that is a representation of the branches of solutions of (33) against the parameter μ for $0 \leq \mu \leq 110$ (See Remark 1 below for details of the computation of this diagram). A two-dimensional version of this diagram can be found for example in [27]. The vertical axis represents the L^2 norm of the solution and the transversal axis the sum of the first five Fourier coefficients of the solutions, a quantity that we have denoted $\alpha(u)$. For the values of the parameter μ considered, there are bifurcations enough to test the proposed technique.

For the discretization of (33) we use a standard pseudospectral Fourier method [3]. That is, following the notation of Section 2.1, $S_m = \text{span}(p_1 = \sin(x), \ldots, p_{m-1} = \sin((m-1)x))$. Nonlinear terms are approximated by standard trigonometric interpolation and evaluated by Fast Fourier Transform (FFT) techniques. In our tests, we set $m = 32$. With this value of m, the solutions of the KS equations for the values of μ considered here, can be computed with a relative accuracy of 10^{-7}, although larger values of μ may require larger values of m to maintain that error level. This high accuracy should not be surprising since spectral methods are reputed by their fast convergence. For this reason, we set a general tolerance TOL to $TOL = 10^{-7}$. 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 $\alpha = 50$. 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.

To build the preconditioner $\mathcal{P}$ in (5), we took $\mathcal{A} = 4\partial_{xxxx} + \mu\partial_{xx}$, which, in terms of the Fourier coefficients, is a diagonal operator. The operator $\mathcal{A}$ is not invertible for the values of μ given by $\mu_j = 4j^2$, $j = 1, \ldots$. Thus, in our computations, whenever

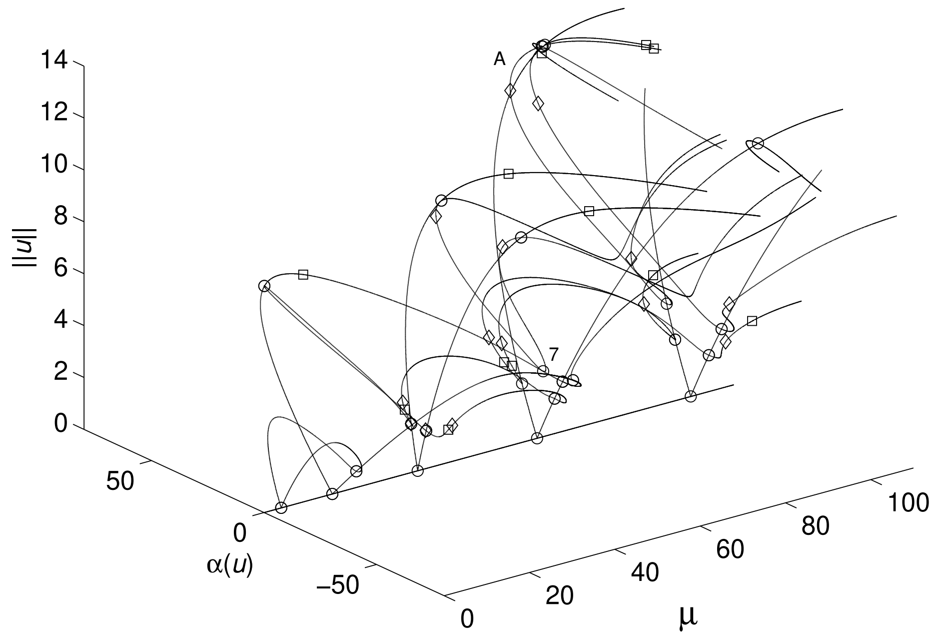

Figure 1: Bifurcation diagram of the KS equation. $\alpha(u)$ is the sum of the first five coefficients of the sine expansion of u. Bifurcation points: $\circ$, pitchfork or transcritical; $\diamond$, saddle-node; $\square$, Hopf bifurcation.

$|\mu - \mu_j| \leq 10^{-12}$, we use instead $4\partial_{xxxx} + \mu\partial_{xx} + (10^{-12} - (4j^4 - \mu j^2))P_{p_j}$, where we remind that $p_j(x) = \sin(jx)$. This procedure may be represented by setting

$$\mathcal{A} = 4\partial_{xxxx} + \mu\partial_{xx} + \sum_{j=1}^{\infty}(10^{-12} - (4j^4 - \mu j^2))\phi(\mu - \mu_j)P_{p_j}, \tag{34}$$

for some smooth function ϕ with sufficiently small compact support, and taking the value 1 in a neighbourhood of 0. Observe that this modified operator is invertible for all $\mu \in \mathbb{R}$ and the corresponding inverses are compact operators.

We start by monitoring $\chi_k = \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H_k^{(m)}))$ along the different branches and bifurcation points. And here comes a surprise: whereas in some cases χ_k neatly changes sign when crossing a bifurcation point, we find that it may also change sign in the middle of a branch far away from any bifurcation point. This is illustrated in Fig. 2, where we show χ_k as a function of μ in the branch marked A in Fig. 1. The determinant changes sign twice. 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.

If we rely on checking the eigenvalues of f_u to detect bifurcations and use the changes of sign of χ_k to locate them by means of the bisection method, it turns out that some of the bifurcation points are accurately located, but in most of them the accuracy is poor. We enumerate from 1 to 23 the bifurcation points which were computed by finding a change of sign of χ_k (others were computed as extrema of the parameter μ). In Fig. 3 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 (10). We observe that out from the 23 points, only in 8 of them the error was below the demanded level of 5×10^{-6} (marked with a dotted line in the plot).

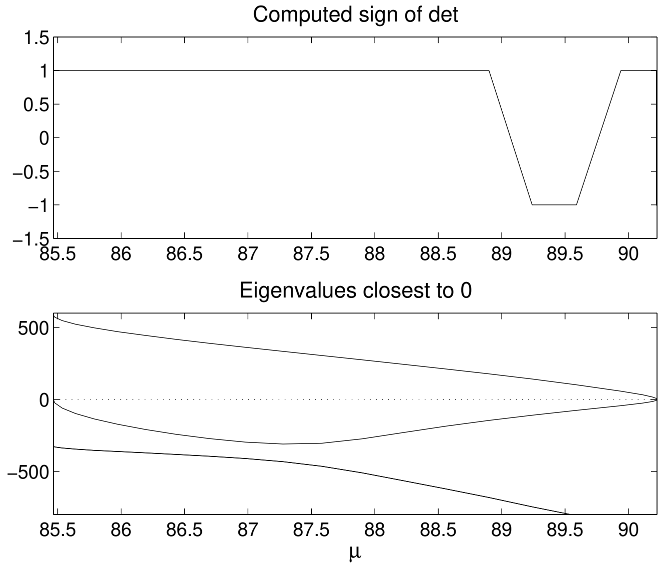

Figure 2: Top: evolution of $\chi_k = \text{sign}(\det(\mathcal{P})\text{sign}(\det(H_k^{(m)}))$ on branch A. Bottom: the three eigenvalues of f_u closest to zero are shown.

The reason of this poor resolution is that χ_k changes sign at wrong places. This can be seen in Fig. 4 where we show the value of χ_k as a function of the parameter μ near the 7th bifurcation point in Fig 3 (also marked 7 in Fig. 1) and represented in the plot with a vertical discontinuous line indicating the parameter value μ_0 at the bifurcation point. It is obvious that χ_k does not changes sign at the correct value of μ. Observe also that we have marked with + signs the points computed in the bisection method. The point where they accumulate show how the program located the bifurcation point in the wrong place.

This poor performance of such a promising and cheap technique, lead us to investigate its causes, which, as explained in the next section, are related to the presence of symmetries.

Remark 1 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 [22] as well as, for double checking, with standard QR iteration from LAPACK routines. Reference solutions were computed with a general tolerance of $TOL = 10^{-8}$.

4.1 The effect of symmetries

Of all the branch-crossings (marked with $\circ$ in Fig. 1), only four correspond to transcritical bifurcations, the rest of them being pitchfork bifurcations. It is well-known (see e.g. [5])

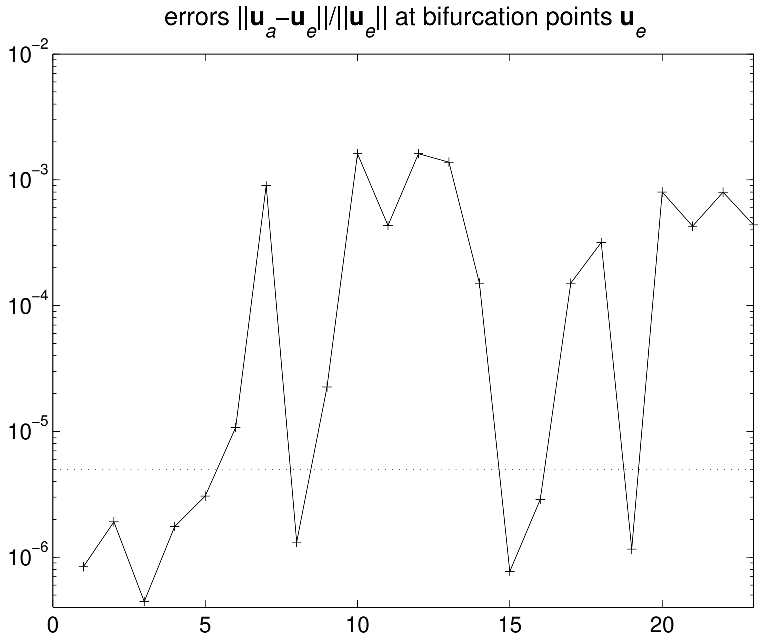

Figure 3: Errors in bifurcation location by zeroing χ_k.

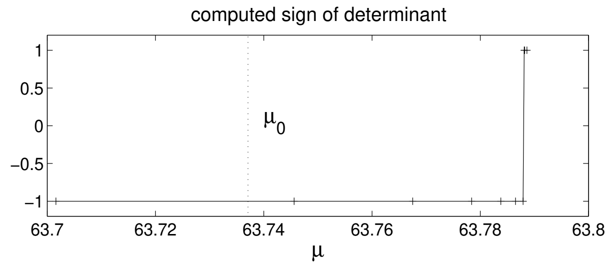

Figure 4: The value of $\mathrm{sign}(\det(H_k^{(m)}))$ near the bifurcation points 7, marked with μ_0

that a system

$$f(u,\mu)=0 \tag{35}$$

such as that in (1) may undergo pitchfork bifurcations if it is invariant under a certain symmetry, that is, a linear mapping such that $S^2=I$ and

$$f(Su,\mu)=Sf(u,\mu), \qquad u\in\mathcal{H}, \qquad \mu\in\mathbb{R}. \tag{36}$$

Observe that given a symmetry S, we can decompose $\mathcal{H}$ (or S_m) as $\mathcal{H}=E_+ + E_-$, where E_+ and E_- are the eigenspaces associated with the eigenvalues 1 and -1 of E respectively.

Assume now that (36) holds and that $\mathbf{u}_0$ is a bifurcation point corresponding to a branch crossing. Let us describe the two crossing branches as

$$\mathbf{u_0}+\mathbf{p}(s)=\left[\begin{array}{c} u_0+p(s) \\ \mu_0+\gamma(s)\end{array}\right], \qquad \mathbf{u_0}+\mathbf{q}(s)=\left[\begin{array}{c} u_0+q(s) \\ \mu_0+\delta(s)\end{array}\right], \qquad \mathbf{p}(0)=\mathbf{q}(0)=\mathbf{0}.$$

The pitchfork bifurcation takes place when one of the branches, say, $\mathbf{u_0}+\mathbf{p}(s)$, satisfies that

$$\mathbf{u}_0+\mathbf{p}(s)\in E_+\times\mathbb{R}, \qquad s\in(-\epsilon,\epsilon), \tag{37}$$

$$\frac{d}{ds}\gamma(0) \neq 0, \tag{38}$$

zero is a simple eigenvalue of $f_u(u_0, \mu_0)$ and its associated eigenvector v satisfies that

$$f_u(\mathbf{u_0})v = 0, \qquad v \notin E_+. \tag{39}$$

In this case, and under further generic assumptions (see e.g. [5], §6.7) it follows that the other branch, $\mathbf{u}_0 + \mathbf{q}(s)$, satisfies

$$\frac{d}{ds}\mathbf{q}(0) = \alpha\mathbf{v}, \qquad \mathbf{v} = \begin{bmatrix} v \\ 0 \end{bmatrix}, \tag{40}$$

where $\alpha \neq 0$ and being v the vector in (39).

Observe also that (36) and (37) imply that

$$f(u_0 + p(s), \mu_0 + \gamma(s)) = f(S(u_0 + p(s)), \mu_0 + \gamma(s)) = Sf(u_0 + p(s), \mu_0 + \gamma(s)),$$

that is

$$f(\mathbf{u}_0 + \mathbf{p}(s)) \in E_+ \qquad s \in (-\epsilon, \epsilon), \tag{41}$$

and, taking derivatives,

$$f_\mu(\mathbf{u}_0 + \mathbf{p}(s)) \in E_+, \quad \text{and} \quad f_u(\mathbf{u}_0 + \mathbf{p}(s))E_+ \subset E_+. \tag{42}$$

In the case of the KS equation, some of the symmetries S for which (36) holds are

$$\begin{aligned} \sum_{j=1}^{\infty} \hat{u}_j \sin(jx) \quad &\mapsto \quad \sum_{j=1}^{\infty} (-1)^j \hat{u}_j \sin(jx), \\ \sum_{j=1}^{\infty} \hat{u}_{2j} \sin(2jx) \quad &\mapsto \quad \sum_{j=1}^{\infty} (-1)^j \hat{u}_{2j} \sin(2jx). \end{aligned}$$

All these symmetries are preserved by the operator $\mathcal{A}$ in (34), and thus, (41–42) also holds with f replaced by $\mathcal{A}^{-1}f$.

Thus, in view of how the mapping F in Keller's pseudo-arclength continuation technique is build in (2–3), when using Newton's method to solve these equations, in the corresponding linear systems (4), the coefficient matrix F_u and the right-hand side $-F$ satisfy that

$$\mathcal{P}^{-1}F \in E_+ \times \mathbb{R}, \qquad \mathcal{P}^{-1}F_{\mathbf{u}}(E_+ \times \mathbb{R}) \subset E_+ \times \mathbb{R}. \tag{43}$$

Consequently, when approximating the solution of system (4) by the preconditioned GMRES method, taking as we did $\mathbf{x}_0 = \mathbf{0}$ as the initial approximation, all the Arnoldi vectors $\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_L$, generated in this process are in $E_+ \times \mathbb{R}$. In other words, the minimal invariant subspace $\mathcal{V}$ containing $\mathbf{v}_1$ satisfies that

$$\mathcal{V} \subset E_+ \times \mathbb{R}.$$

That is, all computations are done not with $F_{\mathbf{u}}$ but with $F_{\mathbf{u}|_{\mathcal{V}}}$.

Now recall how F is build in (3), and observe that, at the bifurcation point, the eigenvectors of $F_{\mathbf{u}}$ associated to the eigenvalue 0 must be in the null space of $f_{\mathbf{u}}$, which is span$(\frac{d}{ds}\mathbf{p}(0), \frac{d}{ds}\mathbf{q}(0))$. But in view of (3), the null space of $F_{\mathbf{u}}$ must be orthogonal to $\frac{d}{ds}\mathbf{p}(0)$ and hence, in view of (39–40), is not contained in E_+. From this, it follows that no eigenvalue of $F_{\mathbf{u}|_{\mathcal{V}}}$ changes sign at the bifurcation point.

The capricious changes of sign in Fig. 2 seem to be due to roundoff error, since changing the general tolerance TOL or the arclength stepsize δ_r in (3) alters the location of these sign changes or removes them completely. This may explain also the wrong location of the change of sign shown in Fig. 4.

4.2 A simple remedy

We have seen that, in a pitchfork bifurcation, the Krylov subspaces generated when solving Newton by preconditioned GMRES are always (in exact arithmetic) in the subspace E_+ where $F_{\mathbf{u}}$ does not possess an eigenspace associated with the eigenvalue that changes sign.

One possible way to avoid this situation is to "pollute" the initial Arnoldi vector $\mathbf{v}_1$, so that $\mathbf{v}_1 \notin E_+ \times \mathbb{R}$. Since $\mathbf{v}_1$ is proportional to the first residual $-\mathcal{P}^{-1}(F + F_{\mathbf{u}}\mathbf{x}_0)$, where $\mathbf{x}_0$ is the initial guess, we tried

$$\mathbf{x}_0 = \alpha \begin{bmatrix} v_a \\ 0 \end{bmatrix}, \qquad \alpha = \|F\|, \qquad v_a \notin E_+. \tag{44}$$

We set v_a to be the eigenvector of f_u associated with the eigenvalue of smallest modulus of f_u in some neighbour point (in our tests, the regular point where a change of sign in one of the eigenvalues of f_u was detected).

The proposed technique greatly improves the location of bifurcation points. In Fig. 5 we show the value of $\chi_k = \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H_k^{(m)}))$ around the same bifurcation point as in Fig. 4. The scales are the same in both plots so that the improvement can be better noticed. The new selection of $\mathbf{x}_0$ succeeds in locating much more accurately the bifurcation points. If we denote by μ_c the computed parameter value of the bifurcation point, the relative error $|\mu - \mu_0| / |\mu_0|$ obtained taking $\mathbf{x}_0$ as indicated in (44) is approximately 500 times smaller than that obtained with $\mathbf{x}_0 = \mathbf{0}$.

In Fig. 6 we show the relative errors as in Fig. 3, with both $\mathbf{x}_0 = \mathbf{0}$ (+ signs joined by continuous lines) and $\mathbf{x}_0$ chosen as in (44) ($\circ$ signs joined by discontinuous lines). The reduction of the error with the new $\mathbf{x}_0$ is dramatic, more than 200 times smaller on average. In fact, all points are computed within the desired accuracy of 5×10^{-6}.

5 Conclusions

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. From the theoretical point of view, a crucial role in the convergence of the Krylov sequence is played by the singular values of T. This has allowed to prove convergence of sign$(\det(H_k))$, not only for the operator $I + T$, but also for its discretization by spectral methods. As a consequence, sign$(\det(H_k)$ is a useful tool for locating bifurcations in branches of

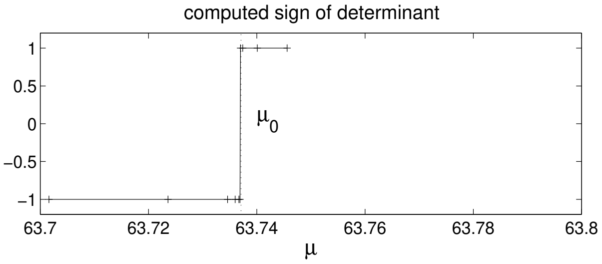

Figure 5: The value of χ_k near the bifurcation points 7, marked with μ_0, with $\mathbf{x_0}$ given by (44).

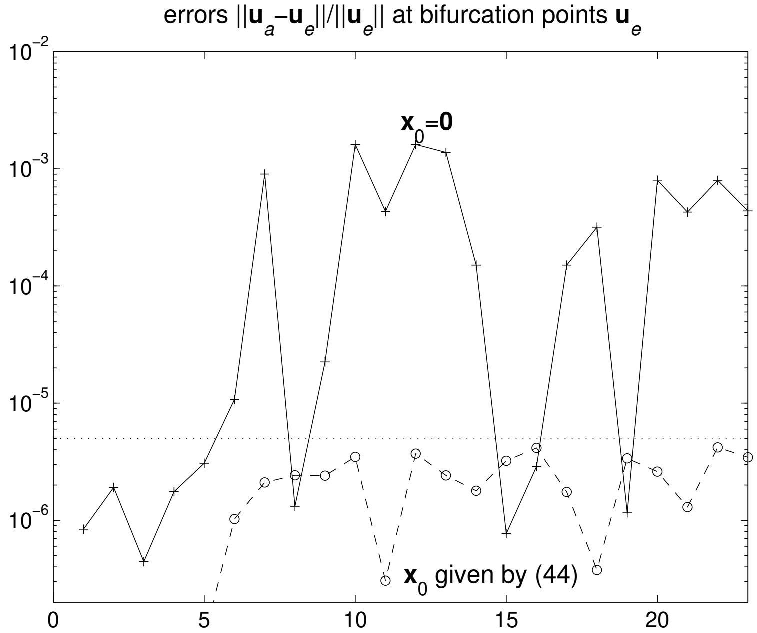

Figure 6: Errors in bifurcation location by zeroing χ_k.

equilibria, and computationally inexpensive 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.

On the practical side, we have seen that some difficulties may arise in the presence of symmetries, which may result in a poor location of bifurcation points. Taking an adequate initial Arnoldi vector overcomes in this shortcoming.

Acknowledgements. The research of B. G.-A. has been partially supported by DGICYT project PB98-0072. 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] 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.

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

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

[6] E. J. Doedel, *Lecture Notes on Numerical Analysis of Bifurcation Problems*, Lecture notes form 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

[7] 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.

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

[9] W. R. Ferng and C. T. Kelly, Mesh independence of matrix-free methods for path following, *SIAM J. Sci. Comput.*, **21** (2000), pp. 18-35–1850.

[10] 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.

[11] W. Govaerts, M. Myers and B. Sturmfelds, Computing Hopf bifurcations, *SIAM J. Numer. Anal.*, **34** (1997), pp. 1–21.

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

[13] 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.

[14] 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.

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

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

[17] 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.

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

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

[20] Y. Kuramoto and T. Tsuzuki, On the formation of dissipative structures in reaction-diffusion systems, *Prog. Theoret. Phys. Suppl.*, **54** (1975), pp. 346–367.

[21] Y. A. Kuznetsov, *Elements of applied bifurcation theory (3rd Ed.)*, Springer, New York, 1998.

[22] 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.

[23] 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.

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

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

[26] 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

[27] R. D. Russell, D. M. Sloan and M. R. Trummer, Some numerical aspects of computing inertial manifolds, *SIAM J. Sci. Comput.*, **14** (1993), pp. 19–43.

[28] 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.

[29] 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.

[30] 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**, 55–102, Springer, 1998.

[31] G. Sivashinsky, On flame propagation under conditions of stoichiometry, *SIAM J. Appl. Math.*, **39** (1980), pp. 67–82.

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

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