Krylov methods and determinants for detecting bifurcations in one parameter dependent partial differential equations

BOSCO GARCÍA-ARCHILLA[1], JUAN SÁNCHEZ[2] and CARLES SIMÓ[3]

[1]*Departamento de Matemática Aplicada II, Universidad de Sevilla*
Escuela Superior de Ingenieros, Camino de los Descubrimientos, s/n,
41092 Sevilla, Spain. email: bosco.garcia@esi.us.es

[2]*Departament de Física Aplicada, Universitat Politècnica de Catalunya,*
Jordi Girona, 1–3, mòdul B4–B5, Campus Nord,
08034 Barcelona, Spain. email: sanchez@fa.upc.es

[2]*Departament de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona*
Gran Via de les Corts Catalanes, 585,
08071 Barcelona, Spain. email: 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). 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 detect and accurately locate bifurcation points. We present an analysis that helps to explain the good performance observed in practice.

AMS subject classification (2000): 65F40, 15A60, 65P30, 35B60, 65N35, 65F10.

Key words: 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., [9], [21], [30], [31]) 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 with spectral methods. 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 [15], [22], and are regularly used in the studies of large dynamical systems (see e.g. [10], [13], [12], [23], [24], [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 [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

$$\text{(1.1)}\quad \text{find} \quad u\in\mathcal{H},\quad \mu\in\mathbb{R},\quad \text{such that}\qquad f(u,\mu)\equiv \mathcal{A}u+R(u,\mu)=0,$$

where $\mathcal{H}$ is a Hilbert space (typically $L^2(\Omega)$ for some domain $\Omega\subset\mathbb{R}^d$) $\mathcal{A}$ is a linear 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, and smooth as function of $u\in D(R(\cdot,\mu))$ and $\mu\in\mathbb{R}$. This is typically the case when R is a smooth function of μ and u and some of its derivatives, and these are of lower order than those in $\mathcal{A}$. The solutions of (1.1) 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 is usually the case with 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 [8], [12], we will analyze first the case where the coefficient matrix is replaced by a compact perturbation of the identity in a Hilbert space, in order to deduce results in $\mathbb{R}^m$ independent of the value of m.

The rest of the paper is as follows. 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. 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 (1.1) by spectral methods, a complete orthonormal set $p_1,p_2,\dots$ 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

$\mathcal{H}_m = \operatorname{span}(p_1, \dots, p_m)$, and let $P_{\mathcal{H}_m} : \mathcal{H} \to \mathcal{H}_m$ be the orthogonal projection onto $\mathcal{H}_m$. Although it is not always the case, quite frequently $P_{\mathcal{H}_m}\mathcal{A} = \mathcal{A}P_{\mathcal{H}_m}$. 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 (1.1) is then replaced by the family of problems

$$\begin{aligned} &\text{find } \mathbf{u}_m = \begin{bmatrix} u_{m-1} \\ \mu \end{bmatrix} \in \mathcal{H}_{m-1} \times \mathbb{R} \\ &\text{such that } \quad f^{(m)}(\mathbf{u}_m) \equiv \mathcal{A}u_{m-1} + P_{\mathcal{H}_{m-1}}R(\mathbf{u}_m) = 0. \end{aligned} \tag{2.1}$$

For simplicity we will drop the subindex m and the superindex (m) when no confusion arises. Also, here and in the sequel, we reserve boldface characters for elements of

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

If we denote by $\langle \cdot, \cdot \rangle$ the inner product in $\mathcal{H}$, we consider in $\hat{\mathcal{H}}$ the inner product

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

so that $\hat{\mathcal{H}}$ is also a Hilbert space. For a closed subspace $\mathcal{V}$ of either $\mathcal{H}$ or $\hat{\mathcal{H}}$, we will denote by $P_{\mathcal{V}}$ the orthogonal projection onto $\mathcal{V}$, although in the case of $\hat{\mathcal{H}}_m = \mathcal{H}_{m-1} \times \mathbb{R}$ we shall write P_m instead of $P_{\hat{\mathcal{H}}_m}$.

Solutions of (2.1), when not isolated, appear in branches $s \mapsto \mathbf{u}(s) \in \hat{\mathcal{H}}$ (we dropped here the subindex m) for a 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 [18], [19], [21], which requires at every step the solution of the system

$$F^{(m)}(\mathbf{u}) = \begin{bmatrix} f^{(m)}(\mathbf{u}) \\ (\mathbf{t}_r, (\mathbf{u} - \mathbf{u}_r)) - \delta_r \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}, \tag{2.2}$$

where $\mathbf{u}_r$ is a reference point, typically the most recently available solution of (2.1), $\mathbf{t}_r$ is an approximation to the tangent $\mathbf{t}(s) = d\mathbf{u}(s)/ds$ of the branch at $\mathbf{u}_r$ and δ_r is the corresponding increment along this tangent. There are different criteria to select δ_r: number of iterations, extreme values, angle between consecutive tangents to the curve, etc.

System (2.2) is typically solved by Newton's method. This requires to solve systems of the form

$$F_{\mathbf{u}}\mathbf{x} = -F. \tag{2.3}$$

Of the various methods to solve (2.3) based on $\mathbf{v} \mapsto F_{\mathbf{u}}\mathbf{v}$ operations (see e.g. [15]) we focus on the GMRES method [28], although our discussion covers other methods based on the Arnoldi decomposition of a matrix such as the Full Orthogonalization Method (FOM) [27], [28].

In practice, it is generally of interest to know how many solutions of (1.1) exist for a given value of μ, or if two branches of solutions intersect in what is known as a branching point, etc. Detection of branching points can be done by monitoring $\operatorname{sign}(\det(F_{\mathbf{u}}))$ [9], and, their location can be done by finding the zero of sign(det($F_{\mathbf{u}}$)) with the help of the bisection method (to avoid possible overflow or underflow, it is advisable to compute sign(det($F_{\mathbf{u}}$)) rather than det($F_{\mathbf{u}}$) [9]) although there are some more sophisticated techniques (see e.g., [1], [2], [12], [29]). Detecting branching points by monitoring sign(det($F_{\mathbf{u}}$)) is particularly appealing if sign(det($F_{\mathbf{u}}$)) can be obtained at little cost as a byproduct of Newton's method when solving (2.2), as it is suggested in [9]. Such is the case if m is small and Gaussian elimination is used to solve the linear systems (2.3). Our interest, though, is to compute sign(det($F_{\mathbf{u}}$)) when (2.3) is solved by means of the GMRES method.

Note however that the systems in (2.3) arise from the discretization of PDEs, and thus, are typically ill-conditioned. This can ruin the computational efficiency of the GMRES method. For this reason, preconditioning is used in practice. Here, we consider left-preconditioning, that is, the systems in (2.3) are replaced by $\mathcal{P}^{-1}F_{\mathbf{u}}\mathbf{x} = -\mathcal{P}^{-1}F$, 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

$$\mathcal{P}\begin{bmatrix} u \\ \mu \end{bmatrix} = \begin{bmatrix} \mathcal{A}u \\ \mu \end{bmatrix}, \tag{2.4}$$

so that, in the sequel, we concentrate on systems of the form

$$(I + T)\mathbf{x} = \mathbf{b}, \tag{2.5}$$

Observe that from (1.1–2.2) it follows that system $\mathcal{P}^{-1}F_{\mathbf{u}}\mathbf{x} = -\mathcal{P}^{-1}F$ naturally translates into (2.5).

Let us mention that, typically, when using spectral methods, fast solvers are available for $\mathcal{P}$. In fact, in Fourier spectral methods, $\mathcal{P}$ is a diagonal matrix in many practical instances.

2.2 Arnoldi decomposition

In an Arnoldi decomposition of an operator $I + T$, starting from an initial vector $\mathbf{v}_1$ (for the linear system (2.5), $\mathbf{v}_1 = (\mathbf{b} - \mathbf{x}_0 - T\mathbf{x}_0)/\left\|\mathbf{b} - \mathbf{x}_0 - T\mathbf{x}_0\right\|$, for an initial approximation $\mathbf{x}_0$ to $\mathbf{x}$), a sequence of orthonormal vectors, the *Arnoldi vectors*, is obtained recursively by

$$\mathbf{v}_{k+1} = \frac{(I - P_{\mathcal{V}_k})T\mathbf{v}_k}{\|(I - P_{\mathcal{V}_k})T\mathbf{v}_k\|}, \qquad k = 1, 2, \ldots, L-1, \tag{2.6}$$

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

$$\mathcal{V}_k = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k), \tag{2.7}$$

and L is such that $(I - P_{\mathcal{V}_L})T\mathbf{v}_L = 0$. Notice that since $\mathbf{v}_k \in \mathcal{V}_k$, we have $(I - P_{\mathcal{V}_k})(I + T)\mathbf{v}_k = (I - P_{\mathcal{V}_k})T\mathbf{v}_k$, and, hence, $\mathcal{V}_L$ is the minimal invariant subspace of $I+T$ containing $\mathbf{v}_1$. However, observe that if $I+T$ is not invertible $\mathbf{v}_1$ may not be in the Krylov subspace $(I + T)\mathcal{V}_L$. We remark that, later in the paper, we will consider operators in infinite-dimensional Hilbert spaces. In that case, L may not exist and the sequence of Arnoldi vectors may be infinite.

Denoting by V_k the row vector $V_k = [\mathbf{v}_1, \dots, \mathbf{v}_k]$ with entries in $\hat{\mathcal{H}}$, (V_k is an $m \times k$ matrix when the Hilbert Space $\hat{\mathcal{H}}$ is $\mathbb{R}^m$) from (2.6) we obtain the Arnoldi decomposition of the operator $I + T$,

$$(I + T)V_k = V_{k+1}\tilde{H}_k = V_k H_k + [\mathbf{0}, \dots, \mathbf{0}, h_{k+1,k}\mathbf{v}_{k+1}], \tag{2.8}$$

where $\tilde{H}_k$ is a $(k + 1) \times k$ unreduced upper Hessenberg 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

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

Observe that H_k is the matrix of the operator $P_{\mathcal{V}_k}(I + T)_{|\mathcal{V}_k}$, given by the restriction of $P_{\mathcal{V}_k}(I+T)$ 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 $(I+T)\mathbf{x} = \mathbf{b}$, the GMRES method approximates $\mathbf{x}$ by $\mathbf{x}_k$, found by solving the least-squares problem $\min_{\mathbf{x}_k - \mathbf{x}_0 \in \mathcal{V}_k} \|\mathbf{b} - (I + T)\mathbf{x}_k\|$, and the FOM [27], [28] approximates $\mathbf{x}$ by $\mathbf{x}_k \in \mathbf{x}_0 + \mathcal{V}_k$ such that the residual $\mathbf{b} - (I + T)\mathbf{x}_k$ 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 approxiamte solution $\mathbf{x}_k$ is considered sufficiently close to the the solution $\mathbf{x}$. In our case, for simplicity, the iteration was stopped when the normalized residual $\|\mathbf{b} - (I + T)\mathbf{x}_k\| / \|\mathbf{b}\|$ satisfied a prescribed tolerance test, although other stopping criteria are possible.

2.3 The computation of sign(det($F_{\mathbf{u}}$))

When an iterative method based on an Arnoldi decomposition of $\mathcal{P}^{-1}F_{\mathbf{u}} = I+T$ is used to solve (2.3), the following argument suggests how to approximate $\text{sign}(\det(F_{\mathbf{u}}))$. Since H_k is a matrix representation of $P_{\mathcal{V}_k}(I + T)_{|\mathcal{V}_k}$, and $\mathcal{V}_1 \subset \mathcal{V}_2 \subset \dots$ with increasing dimensions, we may take

$$\text{sign}(\det(F_{\mathbf{u}})) \approx \chi_k := \text{sign}(\det(\mathcal{P}))\text{sign}(\det(H_k)) \tag{2.9}$$

for the larger k reached in the iterative solution of (2.3). In the next section we will show that under certain hypotheses this approximation is correct. We end this section commenting on how readily available this approximation is 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,

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

where if we express

$$\begin{bmatrix} Q_{k-1} & 0 \\ 0^T & 1 \end{bmatrix} \tilde{H}_k = \begin{bmatrix} R_{k-1} & r_k \\ 0^T & \eta_k \\ 0^T & h_{k+1,k} \end{bmatrix}, \tag{2.10}$$

then

$$G_k = \begin{bmatrix} I_{k-1} & 0 & 0 \\ 0^T & c_k & s_k \\ 0^T & -s_k & c_k \end{bmatrix}, \quad c_k = \frac{\eta_k}{\sqrt{\eta_k^2 + h_{k+1,k}^2}}, \quad s_k = \frac{h_{k+1,k}}{\sqrt{\eta_k^2 + h_{k+1,k}^2}}. \tag{2.11}$$

In view of (2.11), we observe that for $l = 1, \ldots, L-1$, $\det(G_l) = 1$, and, consequently, $\det(Q_l) = 1$. Thus, in view of (2.10), we have $\det(H_k) = \eta_k \det(R_{l-1})$. 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,

$$\operatorname{sign}(\det(H_k)) = \operatorname{sign}(\eta_k) = \operatorname{sign}(c_k), \qquad k = 1, \ldots, L-1. \tag{2.12}$$

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

We finish this section with a simple result that states sufficient conditions for the sequence $\operatorname{sign}(\det(H_k))$, $k = 1, 2, \ldots$, to become stationary. This is a necessary condition for $\operatorname{sign}(\det(H_k)$ to have the correct value for $k \ll L$.

Lemma 2.1. *Assume that the value of s_k in (2.11) satisfies $|s_k| < 1/\sqrt{2}$. If $\|T\mathbf{v}_{k+1}\| < 1/\sqrt{2}$, then*

$$\operatorname{sign}(\det(H_{k+1})) = \operatorname{sign}(\det(H_k)). \tag{2.13}$$

Proof. In view of (2.12) we will prove this result by showing that $\operatorname{sign}(\eta_{k+1}) = \operatorname{sign}(c_k)$. Recall (2.10–2.12), 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

$$(\|f\|^2 + |\gamma|^2)^{1/2} = \left\|P_{\mathcal{V}_k}(I+T)\mathbf{v}_{k+1}\right\| = \left\|P_{\mathcal{V}_k} T\mathbf{v}_{k+1}\right\|, \tag{2.14}$$

so that

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

On the other hand,

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

Thus, η_{k+1} and c_k will have the same sign if $|c_k(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1}) - s_k\gamma|$ is sufficiently small. Applying Hölder's inequality and recalling (2.14) we have

$$\begin{aligned}
|c_k(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1}) - s_k\gamma| &\le (c_k^2 + s_k^2)^{1/2}\Big(|(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|^2 + |\gamma|^2\Big)^{1/2} \\
&\le \Big(|(\mathbf{v}_{k+1}, T\mathbf{v}_{k+1})|^2 + \|P_{\mathcal{V}_k} T\mathbf{v}_{k+1}\|^2\Big)^{1/2} \\
&\le \Big(\|T\mathbf{v}_{k+1}\|^2 - |h_{k+2,k+1}|\Big)^{1/2} < \frac{1}{\sqrt{2}} < |c_k|, \qquad (2.15)
\end{aligned}$$

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

REMARK 2.1. Since $\det(H_k)$ is the product of the Ritz values of $\mathcal{P}^{-1}F_{\mathbf{u}}$ (that is, the egienvalues of H_k), one may wonder if the product of the harmonic Ritz values may have a different sign. Since the harmonic Ritz values are the eigenvalues of the matrix $J_k = H_k + h_{k+1,k}^2 (H_k^T)^{-1} e_k e_k^T$ (see e. g., [25]), then, using (2.10), a simple calculation shows that $\mathrm{sign}(\det(J_k)) = \mathrm{sign}(\eta_k + h_{k+1,k}^2/\eta_k)$. Thus, in view of (2.12), it follows that the product of the Ritz values and that of the harmonic Ritz values have the same sign.

3 A convergence result

Although effective 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 when $m = \infty$ in (2.1), 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 unnecessarily artificial since, as discussed in [17], in finite-dimensional spaces the norm of the residuals in the GMRES method, $\|\mathbf{b} - (I+T)\mathbf{x}_k\| / \|\mathbf{b}\|$, $k = 1, \ldots, L$ may be any nonincreasing finite sequence. The following matrix, used in [17] 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,

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

Taking as v_1 the first coordinate vector in $\mathbb{R}^m$, the resulting matrices H_k are the $k \times k$ principal submatrices of A. Thus, for $k = 1, \ldots, m-1$, $\det(H_k) =$

$0 \neq (-1)^m \mathrm{sign}(\alpha_0) = \mathrm{sign}(\det(A))$. Furthermore, if we replace the diagonal elements $a_{kk} = 0$ by $a_{kk} = (-1)^{l_k}\epsilon$, for l_k integers, $k = 1, \ldots, m-1$, and $\epsilon > 0$ sufficiently small, we do not alter $\mathrm{sign}(\det(A))$, but by adequately choosing the integers l_k, the sequence $\mathrm{sign}(\det(H_k))$ $k = 1, \ldots, m-1$ may be anyone we may a priori set. We will see that this situation cannot happen in the infinite-dimensional case. Then, for finite m, and under some general assumptions, it will also be excluded as $m \to \infty$.

The results we present here are not constrained to the linear operators that arise from the discretization of one parameter dependent systems of PDEs, but they apply, in general, to compact perturbations of the identity. Hence, on the rest of Section 3, we consider operators of the form $I + T$ with T compact in a Hilbert space that, in order not to unnecessarily increase the notation we will also denote by $\hat{\mathcal{H}}$, its inner product by $(\cdot, \cdot)$, and its elements by boldface letters. Also by $\hat{\mathcal{H}}_m$ we denote m-dimensional subspaces of $\hat{\mathcal{H}}$ satisfying $\hat{\mathcal{H}}_m \subset \hat{\mathcal{H}}_{m+1}$, for $m = 1, 2, \ldots$, and whose union is dense in $\hat{\mathcal{H}}$, and by P_m the orthogonal projection onto $\hat{\mathcal{H}}_m$. Observe that the contents of Section 2.2 as well as Lemma 2.1 are valid in this more abstract setting.

3.1 The infinite-dimensional case

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

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

the set of eigenvalues of $I + T$ with negative and non negative real part respectively. Observe that $\sigma_-(I + T)$ is a finite set. For an isolated eigenvalue $\lambda \in \sigma(I + T)$, the invariant subspace associated with λ is the range $\mathcal{R}(E_\lambda) = E_\lambda \hat{\mathcal{H}}$ of E_λ, where

$$E_\lambda \mathbf{x} = \frac{1}{2\pi i} \int_{|z-\lambda|=\delta} (zI - (I + T))^{-1} \mathbf{x} \, \mathrm{d}z, \qquad \mathbf{x} \in \hat{\mathcal{H}},$$

for $\delta > 0$ such that $\sigma(I + T) \cap \{z \in \mathbb{C} \mid |z - \lambda| \leq \delta\} = \{\lambda\}$. Since T is compact, $\mathrm{rank}(E_\lambda) < \infty$, and this number is known as the *algebraic multiplicity* of λ (see e. g. [20]).

For operators in infinite-dimensional spaces, it is not generally possible to extend the idea of determinant [20], § X.1.4. Thus, we consider instead the orientation. If $I + T$ is invertible, we define the *orientation* of $I + T$ as

$$\mathrm{or}(I + T) = (-1)^{d(T)}, \qquad d(T) = \sum_{\lambda \in \sigma_-(I+T)} \mathrm{rank}(E_\lambda). \tag{3.1}$$

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

The following result is a consequence of Theorem 3.11 and Remark 3.2 in [20], § IV.5, and [20], § I.4.6.

THEOREM 3.1. *Let T be a compact operator such that $-1 \notin \sigma(T)$, and Γ a closed curve in $\{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0\}$ such that its interior encloses $\sigma_-(I+T)$. If B is a bounded operator such that*

$$\|B\| \leq \frac{1}{2\max_{\xi\in\Gamma}\|(\xi I - (I+T))^{-1}\|}, \tag{3.2}$$

then

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

Let us now consider the sequence of Arnoldi vectors $(\mathbf{v}_k)_{k=1}^{\infty}$ of $I+T$, that is, the vectors obtained by the process in (2.6–2.8), and recall 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

$$\mathcal{V} = \overline{\cap_{\mathbf{v}_1\in\mathcal{X}}\mathcal{X}}, \qquad \mathcal{X} \text{ subspace of } \hat{\mathcal{H}}, \quad \text{and } T\mathcal{X} \subset \mathcal{X}, \tag{3.3}$$

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

$$T_k = P_{\mathcal{V}_k} T P_{\mathcal{V}_k}, \qquad k = 1, 2, \ldots. \tag{3.4}$$

It is clear then that

$$\mathrm{sign}(\det(H_k)) = \mathrm{or}(I+T_k). \tag{3.5}$$

We will then show that $\mathrm{sign}(\det(H_k))$ converges to $\mathrm{or}\big((I+T)_{|\mathcal{V}}\big)$ as a consequence of Theorem 3.1 after showing that $\left\|T_k - T_{|\mathcal{V}}\right\| \to 0$ as $k \to \infty$. For this purpose, we need the following technical result.

LEMMA 3.2. *Let $\mathcal{X} \subset \mathcal{Y}$ and $\mathcal{U}$ be closed subspaces of the Hilbert space $\hat{\mathcal{H}}$ satisfying that for some positive $\epsilon_s < 1$,*

$$\|P_{\mathcal{X}}\mathbf{u}\| \geq \sqrt{1-\epsilon_s^2}\,\|P_{\mathcal{Y}}\mathbf{u}\|, \qquad \forall \mathbf{u} \in \mathcal{U}. \tag{3.6}$$

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

$$\|P_{\mathcal{U}}\mathbf{y}\| \leq \epsilon_s \|\mathbf{y}\|. \tag{3.7}$$

PROOF. We write

$$\begin{aligned}
\mathbf{y} &= (I-P_{\mathcal{U}})\mathbf{y} + P_{\mathcal{U}}\mathbf{y} && (3.8)\\
&= (I-P_{\mathcal{U}})\mathbf{y} + (I-P_{\mathcal{Y}})P_{\mathcal{U}}\mathbf{y} + P_{\mathcal{Y}}P_{\mathcal{U}}\mathbf{y} \\
&= (I-P_{\mathcal{U}})\mathbf{y} + (I-P_{\mathcal{Y}})P_{\mathcal{U}}\mathbf{y} + (I-P_{\mathcal{X}})P_{\mathcal{Y}}P_{\mathcal{U}}\mathbf{y} + P_{\mathcal{X}}P_{\mathcal{Y}}P_{\mathcal{U}}\mathbf{y}. && (3.9)
\end{aligned}$$

Taking the inner product with $\mathbf{y}$ we have

$$\|\mathbf{y}\|^2 = \|(I-P_{\mathcal{U}})\mathbf{y}\|^2 + (\mathbf{y}, (I-P_{\mathcal{X}})P_{\mathcal{Y}}P_{\mathcal{U}}\mathbf{y}). \tag{3.10}$$

Since $P_{\mathcal{X}} P_{\mathcal{Y}} P_{\mathcal{U}} \mathbf{y} = P_{\mathcal{X}} P_{\mathcal{U}} \mathbf{y}$ and $P_{\mathcal{U}} \mathbf{y} \in \mathcal{U}$, from (3.6) we have

$$\|(I - P_{\mathcal{X}}) P_{\mathcal{Y}} P_{\mathcal{U}} \mathbf{y}\|^2 = \|P_{\mathcal{Y}} P_{\mathcal{U}} \mathbf{y}\|^2 - \|P_{\mathcal{X}} P_{\mathcal{U}} \mathbf{y}\|^2 \leq \epsilon_s^2 \|P_{\mathcal{Y}} P_{\mathcal{U}} \mathbf{y}\|^2 \leq \epsilon_s^2 \|P_{\mathcal{U}} \mathbf{y}\|^2 ,$$

so that (3.10) becomes $\|\mathbf{y}\|^2 \leq \|(I - P_{\mathcal{U}})\mathbf{y}\|^2 + \epsilon_s \|\mathbf{y}\| \|P_{\mathcal{U}} \mathbf{y}\|$, and, hence,

$$\|(I - P_{\mathcal{U}})\mathbf{y}\|^2 \geq \|\mathbf{y}\|^2 - \epsilon_s \|\mathbf{y}\| \|P_{\mathcal{U}} \mathbf{y}\| . \tag{3.11}$$

Applying Pythagoras Theorem in (3.8) we have $\|\mathbf{y}\|^2 = \|(I - P_{\mathcal{U}})\mathbf{y}\|^2 + \|P_{\mathcal{U}} \mathbf{y}\|^2$, and, taking into account (3.11), it follows that

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

from where (3.7) follows. □

REMARK 3.1. Notice that for (3.6) to hold, it must be that $\dim(P_{\mathcal{Y}} \mathcal{U}) \leq \dim(\mathcal{X})$. In fact, if $\dim(\mathcal{U}) < +\infty$ (which will be the case in the analysis that follows), $\epsilon_s = \sin(\theta)$, where θ is the largest principal angle between $\mathcal{U}$ and $\mathcal{X}$ (see e.g. [14], § 12.4.3).

We need the singular value decomposition or canonical expansion of the compact operator T (see e.g., [20, § V.2.3]). This is given by

$$T = \sum_{j=1}^{\infty} \sigma_j (\mathbf{u}_j, \cdot) \mathbf{w}_j . \tag{3.12}$$

where $\sigma_1 \geq \sigma_s \geq \ldots \geq 0$ are the singular values, $\mathbf{u}_1, \mathbf{u}_2, \ldots$ form a complete orthonormal set in $\hat{\mathcal{H}}$, and $\mathbf{w}_j = \sigma_j T \mathbf{u}_j$, for $j = 1, 2, \ldots$. Note that $\|T\| = \sigma_1$. For a given $\epsilon > 0$ let us denote

$$\mathcal{U}_\epsilon = \operatorname{span}\{\mathbf{u}_j \mid \sigma_j \geq \epsilon\}. \tag{3.13}$$

We have the following result.

LEMMA 3.3. *Fix the initial Arnoldi vector $\mathbf{v}_1$. Then for every $\epsilon > 0$ there exist an integer $k_0 > 0$ such that*

$$\|T\mathbf{v}\| \leq \epsilon \|\mathbf{v}\| , \qquad \forall \mathbf{v} \in \mathcal{V}_{k_0}^{\perp} \cap \mathcal{V},$$

with $\mathcal{V}$ as defined in (3.3).

PROOF. If $\mathcal{V}$ is a finite-dimensional space, we take $k_0 = \dim(\mathcal{V})$ and the result is trivial. Let us assume then that $\dim(\mathcal{V}) = \infty$, and let us set $\epsilon_1 = \epsilon/\sqrt{2}$ and $\epsilon_s = \epsilon_1 / \|T\|$. Since the Arnoldi vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots,$ form a complete orthonormal set in $\mathcal{V}$, then, for every $\mathbf{x} \in \hat{\mathcal{H}}$ we have $P_{\mathcal{V}_k} \mathbf{x} \to P_{\mathcal{V}} \mathbf{x}$, as $k \to \infty$. Thus, since $\mathcal{U}_{\epsilon_1}$ is a finite-dimensional space, there exists an integer k_0 such that for $k \geq k_0$ we have

$$\|P_{\mathcal{V}_k} \mathbf{u}\| \geq \sqrt{1 - \epsilon_s^2} \|P_{\mathcal{V}} \mathbf{u}\| , \qquad \forall \mathbf{u} \in \mathcal{U}_{\epsilon_1}.$$

We now apply Lemma 3.2 to $\mathcal{X} = \mathcal{V}_{k_0}$, $\mathcal{Y} = \mathcal{V}$ and $\mathcal{U} = \mathcal{U}_{\epsilon_1}$, so that, for $\mathbf{v} \in \mathcal{V}_{k_0}^{\perp} \cap \mathcal{V}$ we have $\left\|P_{\mathcal{U}_{\epsilon_1}} \mathbf{v}\right\| \leq \epsilon_s \|\mathbf{v}\|$. On the other hand, since $\left\|T(I - P_{\mathcal{U}_{\epsilon_1}})\right\| \leq \epsilon_1$, it follows that

$$\|T\mathbf{v}\|^2 = \left\|T P_{\mathcal{U}_{\epsilon_1}} \mathbf{v}\right\|^2 + \left\|T(I - P_{\mathcal{U}_{\epsilon_1}})\mathbf{v}\right\|^2 \leq \left(\|T\|^2 \epsilon_s^2 + \epsilon_1^2\right) \|\mathbf{v}\|^2 = \epsilon \|\mathbf{v}\|^2 .$$

□

As a direct consequence of the previous Lemma we have the following result.

THEOREM 3.4. *Let the initial Arnoldi vector* $\mathbf{v}_1$ *be such that* $\mathcal{V}$ *is an infinite-dimensional space. Then for every* $\epsilon > 0$ *there exist an integer* $k_0 > 0$ *such that the Arnoldi vectors* $(\mathbf{v}_k)_{k=1}^{\infty}$ *of* $I + T$ *satisfy*

$$\|T\mathbf{v}_k\| < \epsilon, \quad \text{for } k \geq k_0, \tag{3.14}$$

COROLLARY 3.5. *If*

$$\bigcup_{\lambda \in \sigma_-(I+T)\cap\mathbb{R}} E_\lambda \hat{\mathcal{H}} \subset \mathcal{V}, \tag{3.15}$$

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

PROOF. If $\dim(\mathcal{V}) < \infty$, the result is trivial. Otherwise, let us consider the operators T_k in (3.4) and let us recall (3.5). Due to (3.15) we have that $\operatorname{or}(I + T) = \operatorname{or}\big((I + T)_{|\mathcal{V}}\big)$. For the difference $T_{|\mathcal{V}} - T_k$ we write

$$T_{|\mathcal{V}} - T_k = T_{|\mathcal{V}}(P_{\mathcal{V}_k} + (I - P_{\mathcal{V}_k})) - T_k = (I - P_{\mathcal{V}_k})TP_{\mathcal{V}_k} + T_{|\mathcal{V}}(I - P_{\mathcal{V}_k}).$$

As a consequence of Theorem 3.4, we have that $\|T_{|\mathcal{V}}(I - P_{\mathcal{V}_k})\| \to 0$, as $k \to \infty$. On the other hand, a simple calculation shows that

$$\|(I - P_{\mathcal{V}_k})TP_{\mathcal{V}_k}\| = |(\mathbf{v}_{k+1}, T\mathbf{v}_k)| \leq \|T\mathbf{v}_k\|,$$

so that applying again Theorem 3.4 we have $\|T - T_k\| \to 0$ as $k \to \infty$. Then, in view of (3.5) and since in Theorem 3.1 we can replace the Hilbert space H by $\mathcal{V}$, the proof is easily finished. □

REMARK 3.2. Observe that since $\mathcal{V}$ is an invariant subspace of $I+T$, if (3.15) does not hold, then $\lim_{k\to\infty} \operatorname{sign}(\det(H_k)) = \operatorname{or}\big((I + T)_{|\mathcal{V}}\big)$.

3.2 The finite-dimensional case

Besides the notation introduced in Section 3.1, we will use the following one in the present section. For every $m = 1, 2, \ldots$, we denote by $(\mathbf{v}_k^{(m)})_{k=1}^m$ the Arnoldi vectors of $(I+P_mT)_{|\mathcal{H}_m}$ (see Remark 3.5 below), by $\mathcal{V}_k^{(m)}$ their Krylov subspaces, by $\tilde{H}_k^{(m)}$ and $H_k^{(m)}$ the corresponding matrices of the Arnoldi decomposition, and by $L^{(m)}$ the first integer for which $T\mathbf{v}_{L^{(m)}}^{(m)} \in \mathcal{V}_{L^{(m)}}^{(m)}$. For simplicity, we will assume that $\mathbf{v}_1^{(m)} = P_m\mathbf{v}_1$, although any other sequence of initial vectors $\mathbf{v}_1^{(m)} \in \hat{\mathcal{H}}_m$ converging to $\mathbf{v}_1^{(m)}$ as $m \to \infty$ may be chosen.

It must be noticed that the same arguments that lead to Corollary 3.5 also show that, as $m \to \infty$, $\operatorname{sign}(\det((I+P_mT)_{|\mathcal{H}_m})) \to \operatorname{or}(I+T)$. However, this only guarantees that $\operatorname{sign}(\det(H_k^{(m)})) = \operatorname{or}(I + T)$ when $k = m$. This is undesirable when m is large in practical computations, since one usually aims at keeping $k \ll m$ when using the GMRES method. The purpose of this section is to show that the correct value of $\operatorname{sign}(\det((I + P_mT)_{|\mathcal{H}_m}))$ may be obtained for $k \geq k_1$ with k_1 independent of m as $m \to \infty$, and thus, eventually, for $k \ll m$.

The main result of this Section is Theorem 3.6 below. Its proof relies on Lemma 2.1. For the compact operator T, observe that Theorem 3.4 guarantees that from a given k_0 onwards we have $\|T(\mathbf{v}_k)\|$ as small as we wish, in particular smaller that $1/\sqrt{2}$. With respect to the value of s_k, Lemma 6.11 in [11] states that

$$|s_k| \leq \left\|(I+T)^{-1}\right\| |(\mathbf{v}_{k+1}, T\mathbf{v}_k)|, \qquad k = 1, 2, \ldots . \tag{3.16}$$

Since $|(\mathbf{v}_{k+1}, T\mathbf{v}_k)| \leq \|T\mathbf{v}_k\|$, then setting $\epsilon = \epsilon_0$ given by

$$\epsilon_0 = \frac{1}{\sqrt{2}\max(1, \|(I+T)^{-1}\|)}, \tag{3.17}$$

as a consequence of Theorem 3.4 and Lemma 2.1 we have that the sequence $\mathrm{sign}(\det(H_k))$, $k = 1, 2, \ldots$ becomes stationary from a value k_0 onwards, and, thus, according to Theorem 3.6, equal to $\mathrm{or}(I+T)$ if (3.15) holds.

However, for the sequences $\mathrm{sign}(\det(H_k^{(m)}))$ to have the same behaviour, we need to ensure that the value k_0 does not change under small perturbations of the initial Arnoldi vector $\mathbf{v}_1$ (See Remark 3.3 below). This can be achieved for example if $\mathcal{V}$ is such that for the value ϵ_0 in (3.17) the following inequality holds:

$$\left\|T_{|_{\mathcal{V}^\perp}}\right\| = \sup_{\mathbf{v}\in\mathcal{V}^\perp - \{\mathbf{0}\}} \frac{\|T\mathbf{v}\|}{\|\mathbf{v}\|} < \epsilon_0. \tag{3.18}$$

The fact that the invariant subspaces $E_\lambda\hat{\mathcal{H}}$ are finite dimensional allows to easily show that if (3.18) holds, then, for any λ such that $E_\lambda\hat{\mathcal{H}} \not\subset \mathcal{V}$, it is possible to find $\mathbf{w} \in \mathcal{V}^\perp$ such that $T\mathbf{w} = \lambda\mathbf{w} + \mathbf{v}$, where $\mathbf{v} \in \mathcal{V}$, so that $\|T\mathbf{w}\| \geq |\lambda|$. Since for $1 + \lambda \in \sigma_-(I+T)$ we have $|\lambda| > 1 > \epsilon_0$, it follows that (3.18) implies (3.15).

Theorem 3.6. *Let $\mathbf{v}_1$ be such that (3.18) holds. Then, there exist positive integers k_1 and m_1 depending on $\mathbf{v}_1$ and $\left\|T_{|_{\mathcal{V}^\perp}}\right\|$ such that for $m \geq m_1$*

$$\mathrm{sign}(\det(H_k^{(m)})) = \mathrm{sign}(\det(H_{k_1})) = \mathrm{or}(I+T), \qquad k_1 \leq k \leq L^{(m)}. \tag{3.19}$$

Proof. Let us denote $\delta = (\epsilon_0 - \left\|T_{|_{\mathcal{V}^\perp}}\right\|)/3$. Applying Lemma 3.3, there exists an integer k_1 such that

$$\|T\mathbf{v}\| \leq \delta, \qquad \forall \mathbf{v} \in \mathcal{V} \cap \mathcal{V}_{k_1}^\perp. \tag{3.20}$$

If $\dim(\mathcal{V}) < \infty$, taking $k_1 = \dim(\mathcal{V})$, then, since (3.18) implies (3.15), the second equality in (3.19) follows. Otherwise, if $\dim(\mathcal{V}) = \infty$, in view of Lemma 2.1, the fact that $\delta < \epsilon_0$ and (3.16), it follows from (3.20) that, for $k \geq k_1$, the sequence $\mathrm{sign}(\det(H_k))$ is stationary, and, hence, since (3.18) holds, also the second equality in (3.19) follows.

On the other hand, in view of (2.6), every Arnoldi vector is the quotient of two continuous functions of previous Arnoldi vectors. Since $\dim(\mathcal{V}) \geq k_1$, we have $\|(I - P_{\mathcal{V}_k})T\mathbf{v}_k\| > 0$, for $k = 1, \ldots, k_1$, and, thus, the generation of the first k_1 Arnoldi vectors is a continuous mapping in some neighbourhood of $\mathbf{v}_1$ (even if $\mathbf{v}_1$ is deficient in an eigendirection, so that a small perturbation in that direction may result in an abrupt change of some $\mathbf{v}_{k'}$, we have $k' > k_1$ if the perturbation is sufficiently small). Then, since $\mathbf{v}_1^{(m)} \to \mathbf{v}_1$ as $m \to \infty$, there exist an integer m_1 such that

$$\left\|P_{\mathcal{V}_{k_1}^{(m)}} - P_{\mathcal{V}_{k_1}}\right\| \le \delta / \left\|T\right\|, \qquad \text{for all } m \ge m_1 \tag{3.21}$$

and

$$\operatorname{sign}(\det(H_{k_1}^{(m)})) = \operatorname{sign}(\det(H_{k_1})), \qquad m \ge m_1. \tag{3.22}$$

Also, since $P_m\mathbf{x} \to \mathbf{x}$ as $m \to \infty$ for every $\mathbf{x} \in \hat{\mathcal{H}}$, it follows that $\left\|(I + P_mT)^{-1}_{|\mathcal{H}_m}\right\| \to \left\|(I+T)^{-1}\right\|$ as $m \to \infty$, so that we may choose m_1 such that

$$\left\|(I + P_mT)^{-1}_{|\mathcal{H}_m}\right\| \le \frac{\epsilon_0}{\epsilon_0 - \delta}\left\|(I+T)^{-1}\right\|, \qquad m \ge m_1. \tag{3.23}$$

Let us take $\mathbf{v} \in (\mathcal{V}_k^{(m)})^\perp$. We notice that $P_{\mathcal{V}_{k_1}}\mathbf{v} = (P_{\mathcal{V}_{k_1}} - P_{\mathcal{V}_{k_1}^{(m)}})\mathbf{v}$, so that in view of (3.21) we have

$$\left\|P_{\mathcal{V}_{k_1}}\mathbf{v}\right\| \le \delta / \left\|T\right\|, \qquad \forall \mathbf{v} \in (\mathcal{V}_k^{(m)})^\perp. \tag{3.24}$$

Thus, by writing

$$\mathbf{v} = P_{\mathcal{V}_{k_1}}\mathbf{v} + P_{\mathcal{V}\cap\mathcal{V}_{k_1}}\mathbf{v} + P_{\mathcal{V}^\perp}\mathbf{v},$$

then, from (3.24), (3.20) and (3.18) if follows that $\|T\mathbf{v}\| \le \delta + \delta + \beta = \epsilon_0 - \delta$, and, in particular, $\left\|T\mathbf{v}_k^{(m)}\right\| \le \epsilon_0 - \delta$, for $k_1 \le k \le L^{(m)}$. Then, in view of (3.16), (3.17) and (3.23) we have $\operatorname{sign}(\det(H_k^{(m)})) = \operatorname{sign}(\det(H_{k_1}^{(m)}))$, for $k_1 \le k \le L^{(m)}$, so that in view of (3.22) and the second equality in (3.19), the proof is finished. $\square$

Remark 3.3. With small changes in the proof of the previous theorem it can be shown that if (3.18) holds, then the same sequence of $\operatorname{sign}(\det(H_k))$ for $k \ge 1$ is obtained if we replace the initial Arnoldi vector $\mathbf{v}_1$ by any other of unit norm in some neighbourhood of $\mathbf{v}_1$. That is, although small perturbations of $\mathbf{v}_1$ may alter greatly the sequence of Arnoldi vectors $\mathbf{v}_k$ as $k \to \infty$, (3.18) ensures that these changes are not enough to alter the sequence $\operatorname{sign}(\det(H_k))$, for $k = 1, 2, \ldots$.

Remark 3.4. Observe also that, in view of (3.16), in order to guarantee that $|s_k| < 1/\sqrt{2}$ in Lemma 2.1, so that the sequence of values $\operatorname{sign}(\det(H_k))$ may become stationary, 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.

Remark 3.5. Returning to the one parameter systems of PDEs that motivate the present paper, when $I + T = \mathcal{P}^{-1}F_{\mathbf{u}}$, observe that $(I + P_mT)_{|\mathcal{H}_m}$ does not coincide with $\mathcal{P}^{-1}F_{\mathbf{u}_m}^{(m)}$ since in the former R_u is evaluated at $\mathbf{u}$ and, in the latter, at $\mathbf{u}_m$. However, since we are assuming that the discretization is convergent and

R is a smooth function of its arguments, we have $\mathbf{u}_m \to \mathbf{u}$ as $m \to \infty$, and all results about $(I + P_m T)_{|\mathcal{H}_m}$ apply also to $\mathcal{P}^{-1} F^{(m)}_{\mathbf{u}_m}$.

It must be noticed that, in practice, it may be difficult to check whether conditions (3.15) and (3.18) hold, or that the value of k_1 in Theorem 3.6 has been reached. Furthermore, by choosing as initial Arnoldi vector small perturbations of a vector $\mathbf{v}_1$ deficient in an adequate eigendirection, we may have that (3.18) holds but, at the same time, we may make the value k_1 in Theorem 3.6 as large as we wish. Notice, however, that either the violation of (3.18) or $\mathbf{v}_1$ being almost deficient in an eigendirection imply that either $\mathbf{v}_1$ is in a proper closed subspace of $\hat{\mathcal{H}}$ or very close to one such a subspace, so that, in practice, choosing random initial Arnoldi vector makes either of these two scenarios extremely unlikely. The numerical evidence in Section 5 seems to confirm this view.

4 Practical implementation

Let us recall, though, that, in our case, the Arnoldi decomposition appears when solving by the GMRES method the linear systems $\mathcal{P}^{-1} F_{\mathbf{u}} \mathbf{x} = -\mathcal{P}^{-1} F$ when applying Newton's method to solve (2.2). The initial Arnoldi vector is proportional to the first residual $-\mathcal{P}^{-1}(F + F_{\mathbf{u}} \mathbf{x}_0)$ for an initial guess $\mathbf{x}_0$. The arguments in the previous section suggest then that $\mathbf{x}_0$ should be a random vector. On the other hand, taking $\mathbf{x}_0 = \mathbf{0}$ may result in a smaller number of iterations (see Figs. 5.6 and 5.7 in Section 5). However, as we argue in the present section (and confirm in Section 5) $\mathbf{x}_0 = \mathbf{0}$ is a particularly bad choice from the point of view of detecting bifurcation points, in case the problem has some symmetry.

Indeed, if $\mathbf{x}_0$ is set to $\mathbf{0}$, $\chi_k = \operatorname{sign}(\det(\mathcal{P}))\operatorname{sign}(\det(H_k^{(m)}))$ should not be expected 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 under an appropriate group Γ of symmetries (see e.g., [4], [5], [16]), that is

$$f(Su, \mu) = Sf(u, \mu), \qquad S \in \Gamma, \quad \mu \in \mathbb{R}, \tag{4.1}$$

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 $\mathcal{E}$ of Γ,

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

is then an invariant set of both f and its differential f_u.

For a branch

$$s \mapsto \mathbf{u}(s) \in \hat{\mathcal{E}} = \mathcal{E} \times \mathbb{R}, \tag{4.2}$$

branching takes place at $\mathbf{u}_0 = \mathbf{u}(s_0)$ if, for instance, besides some further assumptions [4, § 2.3.2], a single eigenvalue of f_u changes sign at s_0 and

$$\operatorname{Ker}(f_u(u_0, \mu_0)) = \operatorname{span}(v), \qquad v \notin \mathcal{E}. \tag{4.3}$$

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 the subspace $\mathcal{E}$ is an invariant set of f, (4.3) 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 (2.1–2.2), and then it is a simple exercise to check that from (4.1) it follows that, in the case when $m = \infty$, the subspace $\hat{\mathcal{E}}$ is an invariant set of both F and $F_{\mathbf{u}}$. If, in addition, $\hat{\mathcal{E}}$ is also an invariant subspace of the preconditioner $\mathcal{P}$, we also have

$$(4.4) \qquad -\mathcal{P}^{-1}F(\mathbf{u}) \in \hat{\mathcal{E}}, \quad \forall \mathbf{u} \in \hat{\mathcal{E}}, \qquad \mathcal{P}^{-1}F_{\mathbf{u}}(\mathbf{u})\mathbf{x} \in \hat{\mathcal{E}}, \quad \forall \mathbf{u} \in \hat{\mathcal{E}}, \quad \forall \mathbf{x} \in \hat{\mathcal{E}}.$$

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 (4.3). Observe 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 $\alpha \in \mathbb{R}$,

$$(4.5) \quad \mathrm{Ker}(F_{\mathbf{u}}(\mathbf{u}_0)) = \mathrm{Ker}(\mathcal{P}^{-1}F_{\mathbf{u}}(\mathbf{u}_0)) = \mathrm{span}(\mathbf{v} + \alpha\mathbf{t}_0), \qquad \mathbf{v} + \alpha\mathbf{t}_0 \notin \hat{\mathcal{E}}.$$

(In fact, since we are assuming that the elements in Γ are isometries, $\alpha = 0$; however, for $\mathcal{E}$ a general invariant subspace, (4.5) follows from (4.2–4.4) with α not necessarily null). 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 $F_{\mathbf{u}}$ changes sign at the bifurcation point $\mathbf{u}_0$.

Now, when using Newton's method to solve equations (2.1–2.2) (for $m = \infty$) along the branch $s \mapsto \mathbf{u}(s) \in \hat{\mathcal{E}}$, if the initial guess $\mathbf{u}^{[0]}$ for Newton's method satisfies $\mathbf{u}^{[0]} \in \hat{\mathcal{E}}$ (which will be the case if $\mathbf{u}^{[0]}$ is computed by any of the standard extrapolation procedures used in practice) taking $\mathbf{x}_0 = \mathbf{0}$ as the initial guess in the GMRES method for the system (2.3) will make the first residual $\mathbf{r}_0 = -\mathcal{P}^{-1}(F + F_{\mathbf{u}}\mathbf{x}_0)$ to be also in $\hat{\mathcal{E}}$. Hence, 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

$$(4.6) \qquad \mathcal{V} \subset \hat{\mathcal{E}}.$$

Thus, according to Remark 3.2, the value of $\mathrm{sign}(\det(H_k))$ 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 (4.5), no eigenvalue changes sign at $\mathbf{u}_0$. This will also be the case with a symmetry-breaking double turning point, where, instead of (4.3), we have $\mathrm{Ker}(f_u(u_0, \mu_0)) = \mathrm{span}(v, (du/ds(s_0)),\ v \notin \mathcal{E}$ 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, so that (4.4) holds with F, $\mathbf{u}$ and $\hat{\mathcal{E}}$ replaced $F^{(m)}$, $\mathbf{u}_m$ and $\hat{\mathcal{E}} \cap \hat{\mathcal{H}}_m$ respectively. Thus, the arguments above also show that $\chi_k = \mathrm{sign}(\det(\mathcal{P}))\mathrm{sign}(\det(H_k^{(m)}))$ will fail to change sign when crossing a branching point if $\mathbf{x}_0 = \mathbf{0}$. 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.

Thus as argued before, since the probability of any given vector to be in the proper subspace $\hat{\mathcal{E}} \cap \hat{\mathcal{H}}_m$ of $\hat{\mathcal{H}}_m$ is 0, a random initial guess $\mathbf{x}_0$ is advisable.

Even with a random $\mathbf{x}_0$, $\det(H_k)$ will fail to change sign at some double singular points. These, for example, are generic in two parameter dependent systems with 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 $S_1^2 = S_2^2 = I$ and $S_1S_2 = S_2S_1$, or in systems with $\Gamma = \mathcal{O}(2)$, the isotropy group of the unit circle (see e.g. [4]). In the case where there is a $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry, at a double symmetry-breaking point $\mathbf{u}_0 = \mathbf{u}(s_0) \in \hat{\mathcal{E}}$, two eigenvalues of f_u change sign, and their associated eigenvectors v_j, $j = 1, 2$ satisfy that $v_j \in \mathcal{E}_j^a = (I - S_j)\mathcal{H}$. Then, from (4.2) and (4.4) if follows that two eigenvalues of $F_{\mathbf{u}}(\mathbf{u}(s))$ change sign at s_0, and their associated eigenvectors are in each of the invariant subspaces $\hat{\mathcal{E}}_j^a = \mathcal{E}_j^a \times \{0\}$.

However, a double symmetry-breaking point can be detected (and, hence, located) with the following strategy. After computing a point $\mathbf{u}_m$ on the branch, we can compute the sign of the determinant of $\mathcal{P}^{-1}F_{\mathbf{u}_m}^{(m)}(\mathbf{u}_m)$ as a mapping from each of of the subspaces $\hat{\mathcal{H}}_m \cap \hat{\mathcal{E}}_j^a$ onto itself. This can be done by means of GMRES method with zero initial guess, if we solve two systems of the form $\mathcal{P}^{-1}F_{\mathbf{u}_m}^{(m)}\mathbf{x} = \mathbf{b}$, where in each case $\mathbf{b}$ is $\mathbf{b} = \mathcal{P}^{-1}F_{\mathbf{u}_m}^{(m)}\mathbf{x}_j$, where $x_j = (I - S_j)x_0$ for some random element $x_0 \in \mathcal{H}$.

Similarly, for the $\mathcal{O}(2)$ symmetry, following the ideas in [7], an adequate $S \in \Gamma$, with $S^2 = I$, can be selected to detect double singular points by solving an extra linear system whose solution is in $(I - S)\mathcal{H} \times \{0\}$. As shown in [6], [7] and [34], at the price of further programming f as a mapping from the different invariant subspaces of selected symmetries in Γ onto themselves, the computational cost of all the linear system to be solved in every point can be substantially reduced.

5 Numerical experiments

In this section we test numerically how correctly $\mathrm{sign}(\det(F_{\mathbf{u}_m}^{(m)}))$ is approximated by $\chi_k = \mathrm{sign}(\det(\mathcal{P}))\mathrm{sign}(\det(H_k^{(m)}))$, 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 (chosen for simplicity)

$$\|\mathbf{b} - A\mathbf{x}_k\| \,/\, \|\mathbf{b}\| \leq TOL_{MR}. \tag{5.1}$$

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

$$
\begin{array}{rcll}
u_t + \sqrt{\mu} v(\nabla u - e_3) & = & \Delta u, & \\
-\nabla p - v + \sqrt{\mu} u e_3 & = & 0, & \\
\nabla \cdot v & = & 0, & \\
-\nabla u \cdot n & = & 0, & y = \pm\frac{1}{2}, \quad z \in [0,1], \\
u & = & 0, & y \in [-1/2, 1/2], \quad z = 0, 1, \\
v \cdot n & = & 0, & (y,z) \in \partial\Omega,
\end{array}
\tag{5.2}
$$

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. [6], [16], [26], [32], [33] and the references cited therein). The variable u represents the deviation 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 gradient ∇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

$$
f(u,\mu) \equiv -\Delta u + \sqrt{\mu} v(u) \cdot (\nabla u - e_3) = 0. \tag{5.3}
$$

Observe that the eigenfunctions of the Laplacian operator subject to the boundary conditions imposed on u in (5.2) 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 (5.3) is equivariant under the group of symmetries generated by S_y and S_z

$$
S_y u(y,z) = u(-y,z), \qquad S_z u(y,z) = -u(y, 1-z). \tag{5.4}
$$

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

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

are invariant sets of f and have the translational invariance $T_p^s u(y,z) = u(y + \frac{2s}{p}, z)$, for $s = 1, \ldots, p-1$ (see e.g. [26]).

Fig. 5.1 shows a bifurcation diagram of the equation (5.3), that is a representation of the branches of its steady-state solutions parameter μ for $0 \le \mu \le 325$ (See Remark 5.1 below for details on the computation of this diagram). A two-dimensional version of this diagram can be found for example in [26]. The vertical axis represents the L^2 norm of the solution u and the transversal axis the value of u at the centre of the left wall, that is $u(-1/2, 1/2)$. For the values

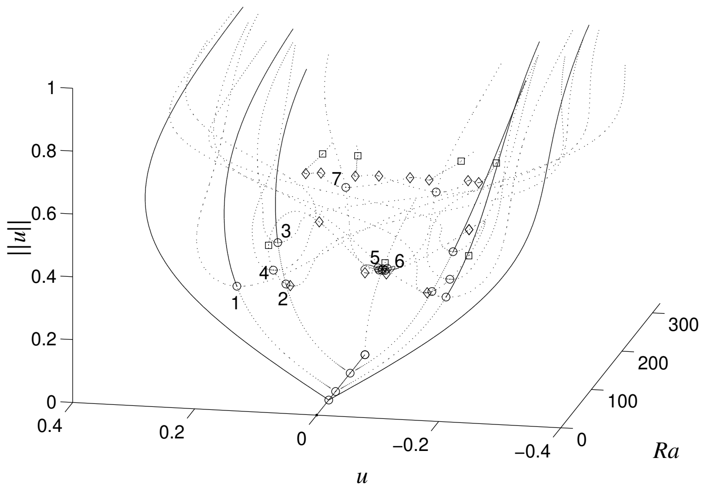

Figure 5.1: Bifurcation diagram of system (5.3), $\mu = Ra$ the Rayleigh number. Bifurcation points: ○, pitchfork or transcritical; ◇, saddle-node; □, Hopf bifurcation (See Remark 5.1)

.

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 solid and dashed lines respectively.

For the discretization of (5.3) we use a standard pseudospectral Fourier method [3]. That is, following the notation of Section 2.1,

$$\mathcal{H}_{m-1} = \operatorname{span}\{p_{j,k}, \quad 0 \le j \le N,\ 1 \le k \le N-1\}, \quad m = N^2.$$

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 (5.3) can be computed with a relative accuracy close to machine precision. This can be seen in Fig. 5.3, where we show the first nonzero Fourier coefficients of bifurcation point labelled 7 in Fig. 5.1, 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 known 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},

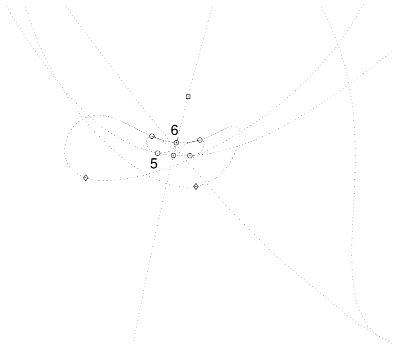

Figure 5.2: Detail of bifurcation points 5 and 6.

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.

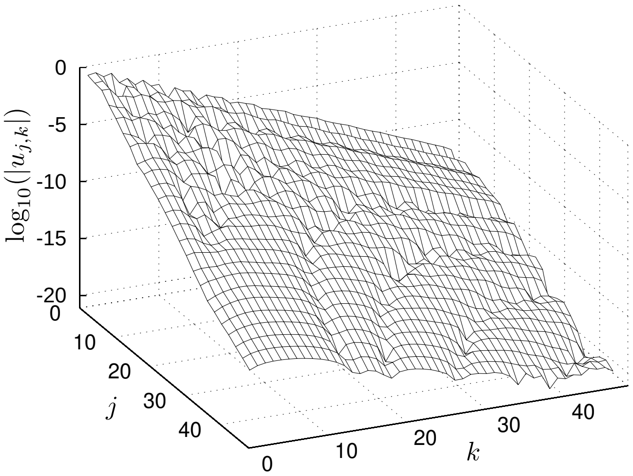

Figure 5.3: Nonzero Fourier coefficients of bifurcation point 7.

As preconditioner $\mathcal{P}$ in (2.4), we took $\mathcal{A} = -\Delta$, subject to the boundary conditions specified in (5.2), which, in terms of the Fourier coefficients, is a diagonal operator.

Despite being a very simple preconditioner, its use resulted in a small number of GMRES iterations (see Figs. 5.7 and 5.6 below) even for Ra values greater than the ones used in this experiment. 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 $\chi_k = \mathrm{sign}(\det(\mathcal{P}))\mathrm{sign}(\det(H_k^{(m)}))$ is given by the sign of $\det(\mathcal{P})$, 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 (2.3) in Newton's method,

$$\mathbf{x}_0 = \alpha_r \|F\| \begin{bmatrix} v_r \\ 0 \end{bmatrix}, \tag{5.5}$$

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. 5.4, where we show χ_k as a function of $\mu = Ra$ in the branch between bifurcation points marked 2 and 3 in Fig. 5.1, taking $\mathbf{x}_0 = \mathbf{0}$ and $\mathbf{x}_0$ given by (5.5). 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 random vector (formula (5.5)), as argued in Section 4, χ_k behaves as the theory predicts.

The spurious changes of sign of χ_k in Fig. 5.4 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 $\mathbf{v} \in (I - P_{\mathcal{E}})\mathcal{H} \times \{\mathbf{0}\}$. Factors such as round-off in FFT routines and, in Newton's 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 GMRES 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 $\mathbf{r}_0 = \mathbf{b} + \delta\mathbf{b}$, and whether computations 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 (5.5), 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.

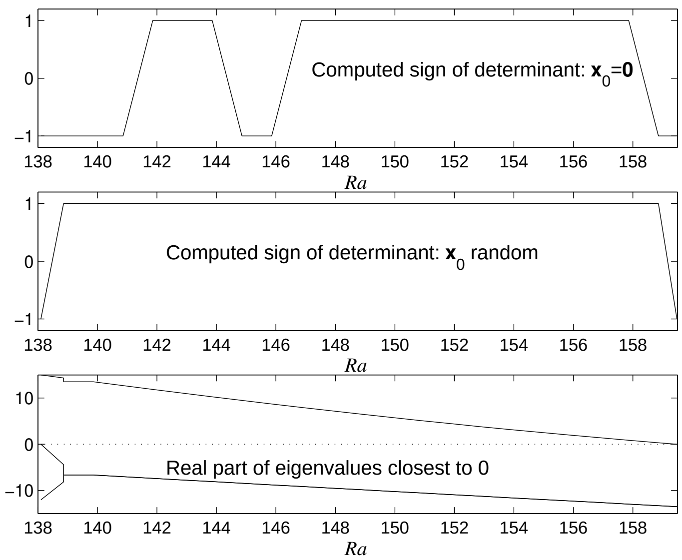

Figure 5.4: Top: evolution of $\chi_k = \text{sign}(\det(\mathcal{P})\text{sign}(\det(H_k^{(m)}))$ on branch between bifurcation points 2 and 3. Bottom: the three eigenvalues of f_u closest to zero are shown.

Of all the branching points shown in Fig. 5.1 (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 (5.5)), and furthermore, accurately located as zeros of χ_k by the bisection method. The (approximate) values of the Rayleigh number for these points are, respectively, 80.95, 138.01, 159.48, 146.74, 214.23, 216.04 and 243.06.

We remark that system (5.3) has a $\mathbb{Z}_2 \times \mathbb{Z}_2$ symmetry, and, as argued in [6] 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 [6], [26], those double symmetry-breaking bifurcation points are not to be found in the range of the Rayleigh number considered here, so that $\det(F_{\mathbf{u}})$ changes sign at all branching points in Fig. 5.1. Had we chosen a domain Ω with a different aspect ratio, double symmetry-breaking points would have been located either as extrema of the transversal branches or in the stability study at a later stage.

We now study how accurate the location of bifurcation points as zeros of χ_k can be. In Fig. 5.5 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 (2.9) by the bisection method taking (5.5) 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 dashed line. The corresponding precisions demanded in the bifurcation points, $10 \times TOL$, are marked in the plot with dotted lines. It can be seen that the errors are below

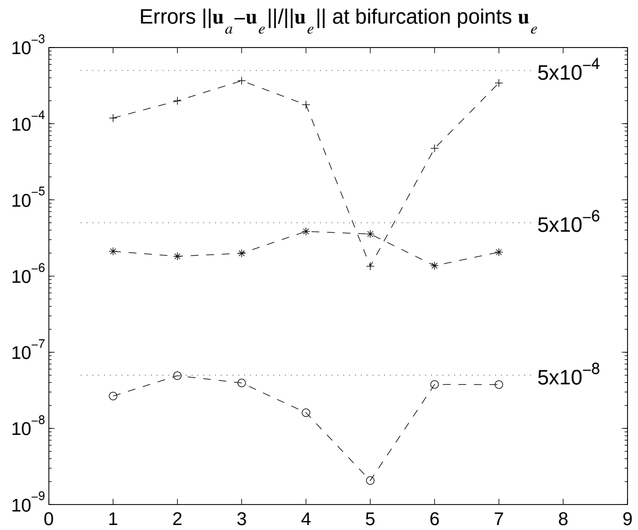

Figure 5.5: Errors in bifurcation location by zeroing χ_k: $+$, $TOL = 5 \times 10^{-5}$; $*$, $TOL = 5 \times 10^{-7}$; $\circ$, $TOL = 5 \times 10^{-7}$. Accuracy demanded is $10 \times TOL$.

the precision demanded.

To achieve these errors, it was also necessary to reduce the tolerance TOL_{MR} of the stopping criterion (5.1) in the GMRES method. Thus, for $TOL = 5 \times 10^{-(2k+1)}$, $k = 2, 3, 4$, we used $TOL_{MR} = 5 \times 10^{-(k+2)}$. 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. 5.5); 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 3.4. 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 in (5.5), 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. 5.6 we show the numbers of GMRES iterations

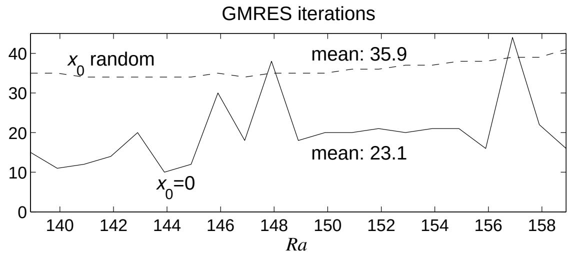

Figure 5.6: Iterations k in GMRES method between bifurcation points 2 and 3.

corresponding to the computation of regular points between bifurcation points 2 and 3, that is, corresponding to results shown in Fig. 5.4. 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, (5.5) results in a 50% increase in cost.

In Fig. 5.7, 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. 5.5. Also, joined by a solid line, we show the results corresponding

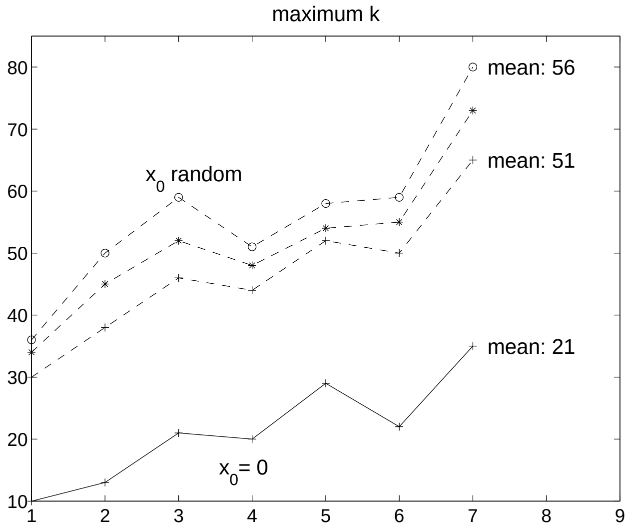

Figure 5.7: Maximum number of iterations in GMRES method on bifurcation points labeled 1 to 7.

to taking $\mathbf{x}_0 = \mathbf{0}$ and $TOL_{MR} = 5{\times}10^{-4}$. Comparing the results marking with + signs, we see that now, when computing bifurcation points, using (5.5) 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 dashed 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. 5.6 and 5.7 we notice that the major increase in cost is due to using (5.5) 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. 5.5, 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 (5.5), 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, and due to shift and invert operations, around 300 matrix-vector products were required to compute the relevant rightmost part of the spectrum of f_u.

Remark 5.1. 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 $TOL = 10^{-9}$. 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 (2.8) 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 $\operatorname{sign}(\det(H_k))$, not only for the operator $I + T$, but also for its discretization by spectral methods.

As a consequence, $\operatorname{sign}(\det(H_k))$ 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.

Although an initial random guess also increases the number of the iteration in the GMRES method, the technique studied here is a computationally efficient alternative to costly eigenvalue computations to detect and locate branching points and, as argued in Section 2.3, to techniques based on augmented systems. These, although efficient and more robust, demand more from the end user, since they require 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 which require matrix-free methods), which, in view of systems like (5.3), may discourage some potential users.

Thus, looking for changes of sign in $\det(H_k)$, as well as for extrema of the parameter along branches of equilibria, is a simple technique which allows the computation of complex bifurcation diagrams, with computations of eigenvalues only at the located bifurcation points. Some multiple singular points (and the corresponding emerging branches) can no doubt be missed in systems equivariant under symmetry groups larger than $\mathbb{Z}_2$, if, as indicated in Section 4, no extra linear system are solved. The missing bifurcations as well as the Hopf points and the stability of the computed branches, however, can be located at a later stage by computing eigenvalues. This should not greatly add to the computational cost of the first stage since it can be focused at specific locations.

Acknowledgement.

The research of B. G.-A. has been partially supported by DGICYT project BFM2003-00336. The research of J. S. has been partially supported by grants DGICYT FIS2004-01066 and CIRIT 2005-SGR-1028. The research of C. S. has been supported by grants DGICYT BFM2003-09504-C02-01 and CIRIT 2005-SGR-1028.

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. P. Chossat, and R. Lauterbach, *Methods in Equivariant Bifurcations and Dynamical Systems*, World Scientific Publishing Co., Inc., River Edge, NJ, 2000.
5. S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.

6. K. A. Cliffe and K. H. Winters, *The use of symmetry in bifurcation calculations and its application to the Bénard Problem*, J. Comput. Phys., 67 (1986), pp. 310–326.
7. K. A. Cliffe, A. Spence and S. J. Tavener, O(2)*-symmetry breaking bifurcation: with application to the flow past a sphere in a pipe*, Internat. J. Numer. Methods Fluids, 32 (2000), pp. 175–200.
8. B. D. Davidson, *Large-scale continuation and numerical bifurcation for partial differential equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2008–2027.
9. 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 pub/doedel/doc/hamburg.ps.Z)
10. 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.
11. M. Eiermann and O. G. Ernst, *Geometric aspects of Krylov subspace methods*, Acta Numerica 2001, Cambridge University Press, Cambridge, 2000, pp. 251–312.
12. 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.
13. 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.
14. G. H. Golub & C. F. Van Loan, *Matrix Computations*, 2nd ed., The John Hopkins University Press, London, 1989.
15. 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.
16. M. Golubitsky, I. Stewart and D. G. Schaeffer, *Singularities and groups in bifurcation theory. Vol. II.*, Springer-Verlag, New York, 1988.
17. 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.
18. 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.
19. 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.
20. T. Kato, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1995.
21. Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd 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. J.M. López, F. Marqués and J. Sánchez, *Oscillatory modes in enclosed swirling flow*, J. Fluid Mech., 439 (2001), pp. 109-129.
24. C. K. Mamun and L. S. Tuckerman, *Asymmetry and Hopf bifurcation in spherical Couette flow*, Phys. of Fluids, 7 (1995), pp. 80-91.

25. C. C. Paige, B. N. Parlett and H. A. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–133.

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

27. Y. Saad, *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2003.

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. A. G. Salinger, N. M. Bou-Rabee, R. P. Pawlowski, E. D. Wilkes, E. A. Burroughs, R. B. Lehoucq and L. A. Romero, *LOCA 1.1. Library of continuation algorithms. Theory and implementation manual*, Sandia Technical Report, SAND2002-0293, 2002.

30. R. Seidel, *Practical Bifurcation and Stability Analysis. From Equilibrium to Chaos*, 2nd ed., Springer, New York, 1994.

31. C. Simó, *Analytical and numerical computation of invariant manifolds*, in Modern Methods in Celestial Mechanics, D. Benest and C. Froeschlé, eds., Editions Frontières, 1990, pp. 285-330. Also at `http://www.maia.ub.es/dsg/2004/`.

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

33. P.H. Steen, *Container geometry and the transition to unsteady Bénard convection in porous media*, Phys. Fluids, 29 (1986), pp. 925–933.

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