

# Rigorous Numerics for Global Dynamics: a study of the Swift-Hohenberg equation

Sarah Day\*    Yasuaki Hiraoka<sup>†</sup>    Konstantin Mischaikow<sup>‡</sup>  
Toshiyuki Ogawa<sup>§</sup>

February 24, 2004

## Abstract

This paper presents a rigorous numerical method for the study and verification of global dynamics. In particular, this method produces a conjugacy or semi-conjugacy between an attractor for the Swift-Hohenberg equation and a model system. The procedure involved relies on first verifying bifurcation diagrams produced via continuation methods, including proving the existence and uniqueness of computed branches as well as showing the nonexistence of additional stationary solutions. Topological information in the form of the Conley index, also computed during this verification procedure, is then used to build a model for the attractor consisting of stationary solutions and connecting orbits.

*Keywords:* rigorous numerics, semi-conjugacy, Conley index

*AMS subject classifications.* 35B45, 35B60, 37L25, 37B30

## 1 Introduction

The Swift-Hohenberg equation,

$$\begin{aligned} u_t = E(u) &= \left\{ \nu - \left( 1 + \frac{\partial^2}{\partial x^2} \right)^2 \right\} u - u^3, & u(\cdot, t) \in L^2 \left( 0, \frac{2\pi}{L} \right), \\ u(x, t) &= u \left( x + \frac{2\pi}{L}, t \right), & u(-x, t) = u(x, t), & \nu > 0, \end{aligned} \quad (1)$$

---

\*Afdeling Wiskunde, Faculteit der Exacte Wetenschappen, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, Nederlands ([sday.math03@gtalumni.org](mailto:sday.math03@gtalumni.org)).

<sup>†</sup>Department of Mathematical Science, Graduate School of Engineering Science, Osaka University, Japan ([hiraoka@sigmath.es.osaka-u.ac.jp](mailto:hiraoka@sigmath.es.osaka-u.ac.jp)).

<sup>‡</sup>Center for Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology, Atlanta, Georgia 30322 USA ([mischaik@math.gatech.edu](mailto:mischaik@math.gatech.edu)).

<sup>§</sup>Department of Mathematical Science, Graduate School of Engineering Science, Osaka University, Japan ([ogawa@sigmath.es.osaka-u.ac.jp](mailto:ogawa@sigmath.es.osaka-u.ac.jp)).

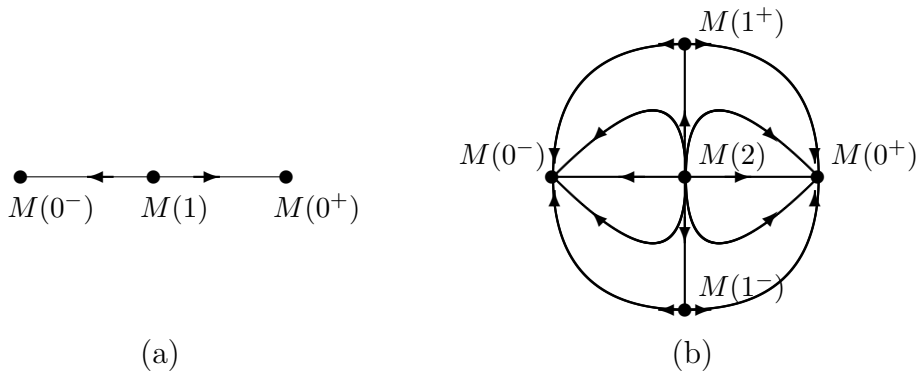


Figure 1: (a) Dynamics on a compact interval. (b) Model gradient dynamics on the unit disk.

was originally introduced to describe the onset of Rayleigh-Bénard heat convection [15], where  $L$  is a fundamental wave number for the system size  $l = 2\pi/L$ . The parameter  $\nu$  corresponds to the Rayleigh number and its increase is associated with the appearance of multiple solutions that exhibit complicated patterns. The focus of this paper is on the development of rigorous numerical techniques that can be used to capture the global dynamics of this equation for various values of  $L$  and  $\nu$ .

Our results are summarized in the following theorems.

**Theorem 1.1** *Let  $(L_1, \nu_1) = (0.62, 0.35)$  and  $(L_2, \nu_2) = (0.65, 0.4)$ . For  $i = 1, 2$  and parameter values sufficiently close to  $(L_i, \nu_i)$ , the set  $J_i \subset L^2(0, 2\pi/L_i)$  (see Tables 1 and 2) is positively invariant and contains exactly three equilibria  $M_i(p)$ ,  $p \in \{0^\pm, 1\}$ . Moreover the dynamics of the maximal invariant set in  $J_i$  is conjugate to the flow on the interval indicated in Figure 1(a).*

**Theorem 1.2** *Let  $(L_3, \nu_3) = (0.62, 0.38)$  and  $(L_4, \nu_4) = (0.65, 0.48)$ . For  $i = 3, 4$  and parameter values sufficiently close to  $(L_i, \nu_i)$ , the set  $J_i \subset L^2(0, 2\pi/L_i)$  (see Tables 3 and 4) is positively invariant and contains exactly five equilibria  $M_i(p)$ ,  $p \in \{0^\pm, 1^\pm, 2\}$ . Moreover the dynamics of the maximal invariant set in  $J_i$  is semi-conjugate to the flow on the unit disk indicated in Figure 1(b).*

Observe that these theorems clarify not only the existence of equilibria but also the global dynamical structures of (1). To be more precise, recall that a flow  $\phi : \mathbf{R} \times X \rightarrow X$  is *semi-conjugate* to a flow  $\psi : \mathbf{R} \times Y \rightarrow Y$  if there exists a continuous surjective mapping  $h : X \rightarrow Y$  such that the following diagram commutes

$$\begin{array}{ccc}
 \mathbf{R} \times X & \xrightarrow{\phi} & X \\
 \downarrow \text{id} \times h & & \downarrow h \\
 \mathbf{R} \times Y & \xrightarrow{\psi} & Y
 \end{array}$$

that is,  $h \circ \phi = \psi \circ (id \times h)$ . If  $h$  is a homeomorphism, then  $\phi$  is *conjugate* to  $\psi$ .

In the case of a conjugacy, the dynamics on  $X$  and  $Y$  are topologically identical. For a semi-conjugacy one can conclude that for every orbit defined by  $\psi$  there exists a corresponding orbit of  $\phi$ . However, this correspondence need not be 1-1. Thus, the dynamics of  $\psi$  provides a lower bound on the complexity of the dynamics of  $\phi$ .

As should be expected, the proofs of Theorems 1.1 and 1.2 involve several distinct steps. The first is to reduce the problem from its infinite dimensional setting to a finite dimensional problem on which numerical calculations can be performed. This is done via a standard Galerkin approximation which is discussed in greater detail in Section 2. For the moment it is sufficient to remark that we use the Fourier basis  $\{\cos(kLx) \mid k = 0, 1, 2, \dots\}$  for  $L^2(0, 2\pi/L)$ . Setting

$$u(x, t) = a_0 + 2 \sum_{k=1}^{\infty} a_k(t) \cos(kLx)$$

leads to the following alternative expression for (1),

$$\dot{a}_k = \left( \nu - (1 - k^2 L^2)^2 \right) a_k - \sum_{\substack{n_1+n_2+n_3=k \\ n_i \in \mathbb{Z}}} a_{n_1} a_{n_2} a_{n_3} \quad (2)$$

where  $a_{-k} = a_k$ . Finally, projecting onto the first  $m$  modes yields the system of ordinary differential equations:

$$\dot{a}_k = \left( \nu - (1 - k^2 L^2)^2 \right) a_k - \sum_{\substack{n_1+n_2+n_3=k \\ |n_i| < m}} a_{n_1} a_{n_2} a_{n_3}, \quad k = 0, 1, \dots, m-1. \quad (3)$$

Having obtained a finite dimensional system the second step is to identify the set of equilibria. This is done as follows. Observe that an equilibrium solution to (3) is given by  $\{a_k = 0 \mid k = 0, 1, \dots, m-1\}$  independent of the parameter values  $L$  and  $\nu$ . This provides a starting point for numerical continuation based on the pseudo-arclength method [8]. Using both  $L$  and  $\nu$  as parameter values we obtain the bifurcation diagrams indicated in Figure 2. In these bifurcation diagrams, we labeled equilibria  $M(0^+)$  and  $M(1^+)$  which correspond to the stable and unstable equilibria in Figure 1. The trivial solution may be labeled  $M(1)$  or  $M(2)$  depending on the existence of the second bifurcation branch.

Since the equilibria indicated in Figure 2 are computed using (3) it is not obvious that they represent equilibria for (1). The following theorem resolves this in the affirmative, except near the bifurcation points. A preliminary observation is that

$$u(x, t) \rightarrow -u(x, t) \quad (4)$$

is an equivariant action for (1).

**Theorem 1.3** *Except, perhaps in a small neighborhood of the bifurcation points, each nontrivial curve in Figure 2 represents exactly two equilibria for (1). These equilibria are related by the symmetry (4)*

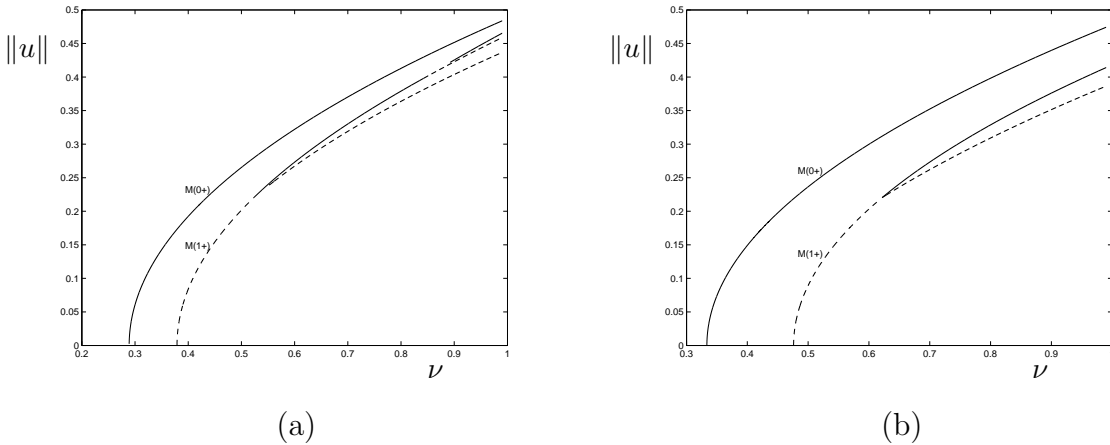


Figure 2: Bifurcation diagrams at (a)  $L = 0.62$  and (b)  $L = 0.65$ . The solid and the dotted lines indicate stable and unstable equilibria, respectively.

The proof of this theorem is presented in Section 4.1. We use a technique due to N. Yamamoto [16] which involves verifying the conditions for a contraction mapping theorem using interval arithmetic. It should be noted that there are alternative approaches to this problem. One such method is described in a forthcoming work by P. Zgliczyński and the third author on the Kuramoto-Sivashinsky equations. In particular, in addition to proving the existence and uniqueness of solutions the structure of the bifurcation points is demonstrated.

Recall that the equilibria indicated in Figure 2 were obtained using a continuation method from the trivial solution. This implies that if (1) possesses an equilibrium which cannot be continued to the origin, then it cannot be identified using these techniques. However, it must be shown that no other equilibria exist within the sets  $J_i$  for  $i = 1, \dots, 4$ , in order to prove Theorems 1.1 and 1.2. This statement about the nonexistence of additional solutions is proven in Section 4.2 using a version of the mean value theorem.

Before proceeding further, it is worth discussing the sets  $J_i$  in greater detail. Using the basis for  $L^2(0, 2\pi/L_i)$  mentioned earlier, each set takes the form of a product of intervals. That is,

$$J = \prod_{k=0}^{\infty} [a_k^-, a_k^+] \quad (5)$$

where  $a_k^\pm = \pm C/k^s$  for all  $k \geq m$ . As can be seen from Tables 1-4, for the results presented in this paper  $s = 4$ ,  $C = 1$ , and  $m = 7$ .

There are several observations that can be made at this point.

1. The set  $J$  is a compact subset of  $L^2$ . Thus, restricting our attention to the dynamics on  $J$  allows us to immediately apply topological tools such as the Conley index which are applicable to locally compact Hausdorff spaces. The

Table 1: The set  $J_1 = \prod_{k \in \mathbb{N}} [a_k^-, a_k^+] \subset L^2(0, 2\pi/0.62)$ .

$k$	$a_k^-$	$a_k^+$
0	$-4.3380010295 \times 10^{-4}$	$4.3380010295 \times 10^{-4}$
1	$-3.4374821943 \times 10^{-3}$	$3.4374821943 \times 10^{-3}$
2	$-1.4440654070 \times 10^{-1}$	$1.4440654070 \times 10^{-1}$
3	$-4.5735140818 \times 10^{-5}$	$4.5735140819 \times 10^{-5}$
4	$-1.0 \times 10^{-4}$	$1.0 \times 10^{-4}$
5	$-1.0 \times 10^{-4}$	$1.0 \times 10^{-4}$
6	$-1.0 \times 10^{-4}$	$1.0 \times 10^{-4}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 2: The set  $J_2 = \prod_{k \in \mathbb{N}} [a_k^-, a_k^+] \subset L^2(0, 2\pi/0.65)$ .

$k$	$a_k^-$	$a_k^+$
0	$-9.2266695451 \times 10^{-4}$	$9.2266695451 \times 10^{-4}$
1	$-1.5081791552 \times 10^{-1}$	$1.5081791552 \times 10^{-1}$
2	$-3.3066549558 \times 10^{-3}$	$3.3066549558 \times 10^{-3}$
3	$-4.9968172502 \times 10^{-4}$	$4.9918701659 \times 10^{-4}$
4	$-1.0 \times 10^{-4}$	$1.0 \times 10^{-4}$
5	$-1.0 \times 10^{-4}$	$1.0 \times 10^{-4}$
6	$-1.0 \times 10^{-4}$	$1.0 \times 10^{-4}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

importance of this will become clear in Section 5 where these tools are used to prove the existence of conjugacies and semi-conjugacies.

2. Because of the polynomial decay in the size of the intervals  $[a_k^-, a_k^+]$  for  $k \geq m$ , useful explicit bounds on the truncation errors that arise from the Galerkin approximation can be obtained. This is explained in Section 3.
3. By similar arguments of [13], (1) possesses a global attractor  $\mathcal{A} = \mathcal{A}(L, \nu)$ . Furthermore, for  $s \geq 1$  and  $C$  sufficiently large,  $\mathcal{A} \subset J$ .

Our choice of  $C = 1$  was made based on numerical experimentation. While it is possible to use energy estimates to analytically derive a value of  $C$  such that  $\mathcal{A} \subset J$ , the resulting number is too large to be of computational value. Thus we are limited to making the following conjecture.

**Conjecture 1.4** *For  $i = 1, 2, 3, 4$ , let  $\mathcal{A}_i$  denote the global attractor of (1) at the parameter values  $(L_i, \nu_i)$ . Then  $\mathcal{A}_i \subset J_i$ .*

Table 3: The set  $J_3 = \prod_{k \in \mathbb{N}} [a_k^-, a_k^+] \subset L^2(0, 2\pi/0.62)$ .

$k$	$a_k^-$	$a_k^+$
0	$-7.5618224050 \times 10^{-4}$	$7.5618224050 \times 10^{-4}$
1	$-2.0750361309 \times 10^{-2}$	$2.0750361309 \times 10^{-2}$
2	$-1.7594153098 \times 10^{-1}$	$1.7594153098 \times 10^{-1}$
3	$-3.4403863059 \times 10^{-4}$	$3.4403863059 \times 10^{-4}$
4	$-1.8 \times 10^{-4}$	$1.8 \times 10^{-4}$
5	$-1.8 \times 10^{-4}$	$1.8 \times 10^{-4}$
6	$-1.8 \times 10^{-4}$	$1.8 \times 10^{-4}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 4: The set  $J_4 = \prod_{k \in \mathbb{N}} [a_k^-, a_k^+] \subset L^2(0, 2\pi/0.65)$ .

$k$	$a_k^-$	$a_k^+$
0	$-1.7963292553 \times 10^{-2}$	$1.7963292554 \times 10^{-2}$
1	$-2.2277203072 \times 10^{-1}$	$2.2277203072 \times 10^{-1}$
2	$-3.6512349385 \times 10^{-2}$	$3.6512349385 \times 10^{-2}$
3	$-2.8824986746 \times 10^{-3}$	$2.8824986746 \times 10^{-3}$
4	$-1.8 \times 10^{-4}$	$1.8 \times 10^{-4}$
5	$-1.8 \times 10^{-4}$	$1.8 \times 10^{-4}$
6	$-1.8 \times 10^{-4}$	$1.8 \times 10^{-4}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Returning to the discussion of the proofs of Theorems 1.1 and 1.2, the above mentioned arguments provide sharp information concerning the existence of equilibria in the regions  $J_i$ . Recall that the the energy functional

$$F(u) = \int_0^{2\pi/L} \left[ \frac{1}{4}u^4 - \frac{\nu}{2}u^2 + \frac{1}{2} \left( (1 + \Delta) u \right)^2 \right] dx \quad (6)$$

acts as a Lyapunov function for Swift-Hohenberg. Given a gradient vector field on a compact manifold, Morse theory relates the global topology of the manifold to the equilibria and their heteroclinic connections. This suggests a strategy for obtaining the global structure of the maximal invariant set in  $J$ . However, some of the essential ingredients of the classical Morse theory are lacking. In particular, we do not know that the maximal invariant set is a compact manifold nor do we know that the heteroclinic orbits arise as transverse intersections of stable and unstable manifolds of hyperbolic equilibria. For this reason we employ the Conley index theory which is a purely topological generalization of Morse theory. Rather than being an index of a hyperbolic fixed point, the Conley index is an index of isolated invariant sets,

which for (1) consists of equilibria and heteroclinic orbits between these equilibria (see Section 4.3 for further details).

For our purposes we need to compute the Conley indices of the equilibria and the maximal invariant sets of the  $J_i$ . This is done in Section 4.3. Furthermore, we prove Theorem 4.10 which guarantees that if the uniqueness of the equilibria can be determined by the method presented in Section 4.1, then the Conley index of the equilibrium can be computed with no extra computational cost. Finally, in Section 5 these indices are combined with existing results from the Conley index theory to complete the proofs of Theorems 1.1 and 1.2.

We conclude in Section 6 with a discussion of the general applicability of the techniques presented here.

## 2 The setting

Consider first an evolution equation:

$$u_t = E(u), \quad E : X \longrightarrow X \quad (7)$$

where  $X$  is a function space on a subset  $\Omega$  in  $\mathbf{R}$ . We now rewrite this system using two types of bases. The first basis ( $a$ -coordinates) is useful for initial simulations and analytic computations while the second basis ( $x$ -coordinates) is more natural for the verification of numerical solutions.

Assume that  $X$  is a Hilbert space and so has an orthonormal basis  $\{\varphi_k\}$ . Using this basis,  $u(x, t)$  can be expressed as follows,

$$u(x, t) = \sum_{k=0}^{\infty} a_k(t) \varphi_k(x).$$

By expanding the original evolution equation in this basis, we obtain the countable system of differential equations:

$$\dot{a} = f(a), \quad (8)$$

where  $a = (a_0, a_1, \dots)$  and for  $k = 0, 1, \dots$ ,

$$\dot{a}_k = f_k(a) = \langle E(a), \varphi_k \rangle.$$

Here,  $\langle \cdot, \cdot \rangle$  denotes the inner product on  $X$  and  $E(a) := E(u)$  for  $u(x, t) = \sum_{k=0}^{\infty} a_k(t) \varphi_k(x)$ . Note that on sets of the form  $W = \prod_k [a_k^-, a_k^+]$  with sufficient (at least quadratic) decay, (8) is equivalent to (7).

For the Swift-Hohenberg equation, consider the following expansion using the Fourier cosine series:

$$u(x, t) = \sum_{k \in \mathbf{Z}} a_k(t) \cos(kLx)$$

with  $a_{-k} = a_k$ . Expanding (1) in this basis, we obtain the following system:

$$\dot{a}_k = f_k(a) = \mu_k a_k - \sum_{\substack{n_1+n_2+n_3=k \\ n_i \in \mathbb{Z}}} a_{n_1} a_{n_2} a_{n_3} \quad (9)$$

where  $\mu_k := \nu - (1 - k^2 L^2)^2$ .

Since we would like to rigorously study equilibria of (9), a natural first step is to numerically compute a solution  $\bar{a}$  with  $f(\bar{a}) \approx 0$ . For this purpose, let us define the orthogonal projection onto the  $m$ -dimensional subspace

$$P_m : X \rightarrow X_m := \text{span}\{\varphi_k | k = 0, 1, \dots, m-1\}$$

and its complementary orthogonal projection  $Q_m := I - P_m$ . We also introduce the notation  $a_F := P_m(a)$  and  $a_I := Q_m(a)$ , which expresses the finite part and infinite part, respectively. Then  $f_k^{(m)}(a_F) := f_k(a_F, 0)$ ,  $k = 0, 1, \dots, m-1$ , is a Galerkin projection onto the first  $m$  modes of the system. For (9), this projection yields the following system of ordinary differential equations:

$$\dot{a}_k = f_k^{(m)}(a_F) = \mu_k a_k - \sum_{\substack{n_1+n_2+n_3=k \\ |n_i| < m}} a_{n_1} a_{n_2} a_{n_3}, \quad k = 0, 1, \dots, m-1. \quad (10)$$

Let  $\bar{a} = (\bar{a}_F, 0)$  where  $\bar{a}_F$  is a numerically computed zero of  $f^{(m)}$ . For our purposes, in what follows we consider only solutions  $\bar{a}_F$  away from bifurcation points for which the eigenvalues of  $Df^{(m)}(\bar{a}_F)$  are both nonzero and nondegenerate.

In order to study  $\bar{a}$  in the infinite-dimensional setting, we construct a second, more natural, coordinate system. The key idea is to perform a transformation to a system where the (hyperbolic) linear behavior may be studied more explicitly. More concretely, we wish to study an equivalent system to (9), given by  $T : x \mapsto a$ , of the form

$$\dot{x}_k = \lambda_k x_k + \epsilon_k(x)$$

for  $k \in \mathbb{N}$ , where  $\lambda_k \neq 0$ , and  $\epsilon_k(x)$  is small near 0 (corresponding to  $T(0) = \bar{a}$  in the original system).

For large  $k$ ,  $|\mu_k| \gg 0$ , and, as will be shown in Section 3, the cubic term  $\sum a_{n_1} a_{n_2} a_{n_3}$  in (9) may be made small near  $\bar{a}$ . Therefore, in the infinite modes, it suffices to consider the linear term  $\lambda_k a_k$  (letting  $\lambda_k := \mu_k$ ) and incorporate the cubic term  $\sum a_{n_1} a_{n_2} a_{n_3}$  into  $\epsilon_k$ .

In the lower order modes, however, the desired linear structure may not be naturally aligned with the directions given by  $a_k$ . In this case, we wish instead to exploit the local eigenstructure given by the Galerkin projection  $f^{(m)}$ . Note that if the projection dimension  $m$  is suitably large, this Galerkin projection is expected to

capture the essential dynamics. That is to say, in the low modes,  $k = 0, 1, \dots, m-1$ , the truncation error term

$$\begin{aligned} r_k(a) &:= f_k(a) - f_k^{(m)}(a_F) \\ &= - \sum_{\substack{n_1+n_2+n_3=k \\ \max\{n_1, n_2, n_3\} \geq m}} a_{n_1} a_{n_2} a_{n_3} \end{aligned}$$

has a small bound (see again Section 3). Next, consider a Taylor expansion of  $f^{(m)}(a_F)$  around  $\bar{a}_F$ . Then

$$f_F(a) = Df^{(m)}(\bar{a}_F)(a_F - \bar{a}_F) + R(a)$$

where

$$\begin{aligned} f_F(a) &= (f_0(a), \dots, f_{m-1}(a))^t, & f^{(m)}(\bar{a}_F) &= (f_0^{(m)}(\bar{a}_F), \dots, f_{m-1}^{(m)}(\bar{a}_F))^t, \\ R(a) &= f^{(m)}(\bar{a}_F) + \frac{1}{2!} D^2 f^{(m)}(\bar{a}_F)(a_F - \bar{a}_F)^2 + \frac{1}{3!} D^3 f^{(m)}(\bar{a}_F)(a_F - \bar{a}_F)^3 + r(a), \\ r(a) &= (r_0(a), \dots, r_{m-1}(a))^t. \end{aligned}$$

Finally, we use the eigenstructure of  $Df^{(m)}(\bar{a}_F)$  to diagonalize this linear map. Suppose  $p_k$  is an eigenvector of  $Df^{(m)}(\bar{a}_F)$  corresponding to the (simple) eigenvalue  $\lambda_k$  and let  $P = [p_0 p_1 \dots p_{m-1}]$ . Then  $P^{-1} Df^{(m)}(\bar{a}_F) P$  is the diagonal matrix with diagonal entries  $\lambda_0, \dots, \lambda_{m-1}$ .

The above description naturally introduces a new variable  $x = (x_F, x_I)$  and the affine transformation  $T : x \mapsto a$  given by

$$Tx = (Px_F + \bar{a}_F, x_I) = (a_F, a_I). \quad (11)$$

In the  $x$ -coordinate system, (8) becomes

$$\dot{x} = g(x)$$

where for  $k = 0, 1, \dots$

$$\dot{x}_k = g_k(x) = \lambda_k x_k + \epsilon_k(x), \quad (12)$$

and  $\lambda_k = \mu_k$  for  $k \geq m$ . Note that  $\epsilon_k(x)$  is given by

$$(\epsilon_0(x), \epsilon_1(x), \dots, \epsilon_{m-1}(x))^t = P^{-1} R(Tx)$$

in the low modes, and

$$\epsilon_k(x) = - \sum_{\substack{n_1+n_2+n_3=k \\ n_i \in \mathbb{Z}}} (Tx)_{n_1} (Tx)_{n_2} (Tx)_{n_3}$$

in the higher order modes ( $k \geq m$ ).

This  $x$ -coordinate system is the most natural setting for our computations which depend on local linear properties. Since (12) was obtained via a change of coordinates, it is equivalent to (9). In addition, note that this transformation leaves the truncated modes unchanged ( $x_k = a_k$  for  $k \geq m$ ). This structure may be exploited when finding bounds for  $\epsilon(x)$  and similar truncation terms in Section 3.

### 3 Error bounds

The success of the numerical techniques presented in this paper depends heavily on finding appropriate interval bounds of truncation terms (for example  $\epsilon(x)$  in (12)). Therefore, we now discuss formulas which may be used to produce the desired bounds. For a more complete discussion, see [3] and [4].

Recall that we restrict our study to subsets of the form  $W = \prod_{k \in \mathbb{N}} [x_k^-, x_k^+]$  with a power decay. As previously mentioned, the original  $a$ -coordinate system is more convenient for the following analytic computations. Therefore, let  $\tilde{W} = \prod_{k=0}^{\infty} [a_k^-, a_k^+]$  be the box in  $a$ -coordinates satisfying

$$[a_k^-, a_k^+] = [\min(TW)|_k, \max(TW)|_k] \quad (13)$$

for all  $k \in \mathbb{Z}$ . Since  $TW \subset \tilde{W}$ , bounds computed for  $\tilde{W}$  in the  $a$ -coordinate system hold for all values in the set  $W$  in the  $x$ -coordinate system. Note also that for  $k \geq m$ ,  $[a_k^-, a_k^+] = [x_k^-, x_k^+]$  and  $\tilde{W}$  exhibits the same decay property as  $W$ . We here present a couple of useful formulas for computing bounds for infinite sums of the form

$$\sum_{\substack{n_1, \dots, n_p \in \mathbb{Z} \\ n_1 + \dots + n_p = k}} (a_{n_1} - a_{n_1}^*) (a_{n_2} - a_{n_2}^*) \cdots (a_{n_p} - a_{n_p}^*)$$

which arise in the following numerical procedures. This is the sum of products of elements in the shifted set  $\tilde{W} - a^*$ , where, for example, the shift  $a^* = \bar{a}$  has the effect of moving the box  $\tilde{W}$  around  $\bar{a}$  to a box  $\tilde{W} - \bar{a}$  around 0. For our purposes, the shift  $a^*$  will be 0 in the higher modes, so that the decay property also holds for  $\tilde{W} - a^*$ .

In order to find error bounds for all  $a \in \tilde{W}$ , we consider as input for  $a_k$  the interval denoted by  $\tilde{a}_k := [a_k^-, a_k^+]$ . As previously discussed, we assume that the intervals  $\tilde{a}_k - a_k^*$  satisfy a power decay law. In other words, there exist constants  $A_s > 0$ ,  $s > 1$ , and  $M > 0$  such that  $\tilde{a}_k - a_k^* \subseteq \frac{A_s}{|k|^s} [-1, 1]$  for all  $k > M$ .

Define  $A$  to be the constant

$$A = \max\{A_s, \max_{a_0 \in \tilde{a}_0} |a_0 - a_0^*|, \max_{0 < k \leq M, a_k \in \tilde{a}_k} |k|^s |a_k - a_k^*|\}.$$

Then  $\tilde{a}_k - a_k^* \subseteq \frac{A}{|k|^s} [-1, 1]$  for all  $k \in \mathbf{Z} \setminus \{0\}$  and  $\tilde{a}_0 \subseteq A[-1, 1]$ .

**Lemma 3.1** ([3, 4]) *Let  $\alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$ . Then*

$$\sum_{n_1 + \dots + n_p = k} (\tilde{a}_{n_1} - a_{n_1}^*) (\tilde{a}_{n_2} - a_{n_2}^*) \cdots (\tilde{a}_{n_p} - a_{n_p}^*) \subseteq I_k$$

where

$$I_k = \begin{cases} \frac{\alpha^{p-1} A^p}{|k|^s} [-1, 1] & k \neq 0 \\ \alpha^{p-1} A^p [-1, 1] & k = 0. \end{cases}$$

We now improve these bounds by taking advantage of the explicit interval  $\tilde{a}_k - a_k^*$  for  $|k| \leq \bar{M}$  for some cut-off value  $\bar{M} > 0$  rather than the extended asymptotic bounds.

**Lemma 3.2** ([3, 4]) For  $0 \leq k < \bar{M}$ ,

$$\sum_{n_1 + \dots + n_p = k} (\tilde{a}_{n_1} - a_{n_1}^*) (\tilde{a}_{n_2} - a_{n_2}^*) \cdots (\tilde{a}_{n_p} - a_{n_p}^*) \subseteq I_k$$

where

$$I_k := \sum_{\substack{n_1 + \dots + n_p = k \\ |n_1|, \dots, |n_p| \leq \bar{M}}} (\tilde{a}_{n_1} - a_{n_1}^*) (\tilde{a}_{n_2} - a_{n_2}^*) \cdots (\tilde{a}_{n_p} - a_{n_p}^*) \\ + \frac{(p-1)\alpha^{p-2}A^{p-1}A_s}{\bar{M}^{s-1}(s-1)} \left[ \frac{1}{(\bar{M}-k)^s} + \frac{1}{(\bar{M}+k)^s} \right] [-1, 1].$$

Using the formula given in Lemma 3.2 requires evaluation of the finite sum. This becomes increasingly expensive for higher cut-off values  $\bar{M}$ , especially if the degree,  $p$ , is large. On the other hand, increasing  $\bar{M}$  decreases the amount of overestimation resulting from using the extended asymptotic bounds with constant  $A$ . Therefore, in practice, computing the improved bounds given in Lemma 3.2 involves balancing computational costs with obtaining tighter bounds.

The following formula represents a first level of modification of Lemma 3.2 and may be used to compute a bound for truncation terms.

**Corollary 3.3** ([3, 4]) For  $0 \leq k < m$ ,

$$\sum_{\substack{n_1 + \dots + n_p = k \\ \max\{|n_i|\} \geq m}} (a_{n_1} - a_{n_1}^*) (a_{n_2} - a_{n_2}^*) \cdots (a_{n_p} - a_{n_p}^*) \subseteq I_k^{(m)}$$

with

$$I_k^{(m)} := \sum_{\substack{n_1 + \dots + n_p = k \\ |n_1|, \dots, |n_p| \leq \bar{M}}} \overline{(a_{n_1} - a_{n_1}^*) (a_{n_2} - a_{n_2}^*) \cdots (a_{n_p} - a_{n_p}^*)} \\ + \frac{(p-1)\alpha^{p-2}A^{p-1}A_s}{\bar{M}^{s-1}(s-1)} \left[ \frac{1}{(\bar{M}-k)^s} + \frac{1}{(\bar{M}+k)^s} \right] [-1, 1]$$

where  $\overline{(a_{n_1} - a_{n_1}^*) (a_{n_2} - a_{n_2}^*) \cdots (a_{n_p} - a_{n_p}^*)}$  is 0 if all of the indices have absolute value less than  $m$  and is the interval bound  $(\tilde{a}_{n_1} - a_{n_1}^*) (\tilde{a}_{n_2} - a_{n_2}^*) \cdots (\tilde{a}_{n_p} - a_{n_p}^*)$  otherwise.

These bounds may be further modified for cubic sums as in Appendix 1 and [6].

## 4 Numerical verification method

This section is devoted to exploring rigorous numerical techniques for the verification of bifurcation diagrams and includes sample results for the Swift-Hohenberg equation. The method described in Section 4.1 is a natural extension of verification techniques presented in [16] and is used to study the existence and uniqueness of bifurcation branches. In Section 4.2, an algorithm for proving the nonexistence of additional stationary solutions is presented. Finally, stability properties of the bifurcation branches are studied via index techniques in Section 4.3.

### 4.1 Existence and uniqueness

In this section we describe a method for the verification of the existence and uniqueness of solutions corresponding to the numerical bifurcation branches. This method is based on techniques used by N. Yamamoto in [16] to rigorously ensure the existence and uniqueness of solutions for nonlinear elliptic problems. The key idea is to apply Banach's fixed point theorem to an appropriate contraction (Newton-like) map whose unique fixed point corresponds to the solution we wish to study. In what follows, we briefly describe this technique in this setting and apply it to the Swift-Hohenberg equation. Many of the results described in this part are presented without proof, since the proofs are quite similar to those in [16].

Our goal in this section is to construct a (verification) set  $W = \prod_k [x_k^-, x_k^+]$  around the origin and show that there exists a unique solution  $x_* \in W$  with  $g(x_*) = 0$ . Via the appropriate change of coordinates, this solution  $x_*$  corresponds to the unique stationary solution  $a_* := Tx_*$  in the neighborhood  $\tilde{W} = T \prod_k [x_k^+, x_k^+]$  of the approximate numerical solution  $\bar{a}$  (see Section 2).

Recall that in the  $x$ -coordinate system,

$$\dot{x}_k = g_k(x) := \lambda_k x + \epsilon_k(x), \quad k \in \mathbb{N}.$$

Assume, as before, that  $\lambda_k \neq 0$  for all  $k \in \mathbb{N}$ . Then  $x$  is a zero of  $g$  if and only if  $x$  is a fixed point of the Newton-like operator  $G$  given by

$$G_k(x) = -\lambda_k^{-1} \epsilon_k(x). \tag{14}$$

Hence, it suffices to study the fixed point equation (14).

We now show that under certain conditions,  $G$  is a contraction mapping on an appropriate subset  $W$  (called a candidate set in [16]), and use Banach's fixed point theorem to conclude that  $G$  has the desired unique fixed point  $x_*$ . Following the work of Yamamoto, this involves checking a series of inequalities. The key ideas follow.

For  $W = \prod_k [x_k^+, x_k^+]$ , let

$$\Omega := \left\{ x = (x_0, x_1, \dots) \mid \sup_{k \in \mathbb{N}} \frac{|x_k|}{r_k} < \infty \right\}, \quad \text{where } r_k = \sup_{x \in W} |x_k|.$$

Then one may check that

$$\|x\|_W := \sup_{k \in \mathbb{N}} \frac{|x_k|}{r_k}$$

is a norm on  $\Omega$ . Furthermore,

- $\Omega$  is a Banach space with the norm  $\|\cdot\|_W$
- $W$  is a closed set under  $\|\cdot\|_W$

Finally, suppose there exist constants  $Y_k \neq 0$  and  $Z_k$  for  $k \in \mathbb{N}$ , such that

$$\begin{aligned} |G_k(0)| &\leq Y_k, \\ |[G'(x)y]_k| &\leq Z_k, \quad \text{for all } x, y \in W \end{aligned}$$

where  $G'(x)$  denotes the Fréchet derivative of  $G$ .

**Theorem 4.1** (Yamamoto, [16]) *If  $(Y_k + Z_k)[-1, 1] \subseteq [x_k^-, x_k^+]$  for all  $k \in \mathbb{N}$ , then  $G$  has a unique fixed point  $x_* \in W = \prod_k [x_k^-, x_k^+]$ .*

*Sketch of Proof.* (For a more complete discussion, see [16].) Yamamoto proves a couple of preliminary inequalities which together may be used to show that if  $(Y_k + Z_k)[-1, 1] \subseteq [x_k^-, x_k^+]$  for all  $k \in \mathbb{N}$  then

1.  $G(W) \subset W$
2. there exists a contraction constant  $0 \leq \kappa < 1$  such that

$$\|G(x) - G(y)\|_W \leq \kappa \|x - y\|_W$$

for all  $x, y \in W$ .

The result then follows from Banach's fixed point theorem. ■

**Definition 4.2** If bounds  $\{Y_k, Z_k\}$  may be found for  $W = \prod_k [x_k^-, x_k^+]$  which satisfy the conditions of Theorem 4.1, then  $W$  is called a *verification set*.

#### 4.1.1 Constructing a verification set

In practice, we prescribe an initial, small set  $W = \prod_k [x_k^-, x_k^+]$ , with  $x_k^- = -x_k^+$ , which exhibits power decay. This decay property is given as for all  $k \geq M$ ,  $\tilde{x}_k := \frac{A_s}{k^s}[-1, 1]$  for some constants  $A_s$  and  $s \geq 2$ . Next, we compute the bounds  $Y_k$  and  $Z_k$ .

As in Section 3, let

$$\tilde{a}_k := [\min(TW)|_k, \max(TW)|_k].$$

In the low modes,

$$\begin{aligned}
G_F(0) &= -\Lambda_F^{-1} P^{-1} R(T(0)) \\
&= -\Lambda_F^{-1} P^{-1} R(\bar{a}_F) \\
&= -\Lambda_F^{-1} P^{-1} f^{(m)}(\bar{a}_F).
\end{aligned} \tag{15}$$

In practice  $f^{(m)}(\bar{a}_F)$  lies in a known small interval centered around zero computed using interval arithmetic. This interval bound combined with (15) yields bounds  $Y_k$ ,  $0 \leq k < m$ .

Also, for all  $x, y \in W$ ,

$$\begin{aligned}
G'_F(x)y &= -\Lambda_F^{-1} P^{-1} (R(Tx))' y \\
&= -\Lambda_F^{-1} P^{-1} R'(Tx)(Ty - \bar{a}) \\
&\in -\Lambda_F^{-1} P^{-1} R'(\tilde{a})(\tilde{a} - \bar{a})
\end{aligned} \tag{16}$$

where

$$\Lambda_F = \begin{pmatrix} \lambda_0 & & 0 \\ & \ddots & \\ 0 & & \lambda_{m-1} \end{pmatrix},$$

and for  $k = 0, \dots, m-1$ ,

$$\begin{aligned}
[R'(\tilde{a})(\tilde{a} - \bar{a})]_k &= -6 \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} \bar{a}_{n_1}(\tilde{a}_{n_2} - \bar{a}_{n_2})(\tilde{a}_{n_3} - \bar{a}_{n_3}) \\
&\quad -3 \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} (\tilde{a}_{n_1} - \bar{a}_{n_1})(\tilde{a}_{n_2} - \bar{a}_{n_2})(\tilde{a}_{n_3} - \bar{a}_{n_3}) \\
&\quad -3 \sum_{\substack{n_1+n_2+n_3=k \\ \max\{n_1, n_2, n_3\} \geq m}} \tilde{a}_{n_1} \tilde{a}_{n_2} (\tilde{a}_{n_3} - \bar{a}_{n_3}).
\end{aligned} \tag{17}$$

The first two sums in (17) require finite interval arithmetic while the third sum may be split into two infinite sums to which the the formulas in Section 3 may be applied to find bounds. Combining these bounds with (16) yields values for  $Z_k$ ,  $0 \leq k < m$ .

For  $k \geq m$ ,

$$\begin{aligned}
G_k(0) &= \frac{1}{\lambda_k} \sum_{n_1+n_2+n_3=k} (T0)_{n_1} (T0)_{n_2} (T0)_{n_3} \\
&= \frac{1}{\lambda_k} \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} \bar{a}_{n_1} \bar{a}_{n_2} \bar{a}_{n_3}
\end{aligned} \tag{18}$$

and

$$G'_k(x)y \in \frac{3}{\lambda_k} \sum_{n_1+n_2+n_3=k} \tilde{a}_{n_1} \tilde{a}_{n_2} (\tilde{a}_{n_3} - \bar{a}_{n_3})$$

$$= \frac{3}{\lambda_k} \left( \sum_{\substack{n_1+n_2+n_3=k \\ |n_3|<m}} \tilde{a}_{n_1} \tilde{a}_{n_2} (\tilde{a}_{n_3} - \bar{a}_{n_3}) + \sum_{\substack{n_1+n_2+n_3=k \\ |n_3|\geq m}} \tilde{a}_{n_1} \tilde{a}_{n_2} \tilde{a}_{n_3} \right). \quad (19)$$

Again, finite interval computations and the formulas in Section 3 give the bounds  $Y_k$  and  $Z_k$ ,  $k \geq m$ . Note that for  $k \geq \bar{M} \geq M$ ,  $Y_k + Z_k = \frac{C}{|\mu_k|k^s} \leq \frac{C}{|\mu_{\bar{M}}|k^s}$  for some constant  $C$ , so checking the inequalities in Theorem 4.1 for the tail modes only requires verifying that  $\frac{C}{|\mu_{\bar{M}}|} \leq A_s$ . This leaves a finite number of inequalities to check explicitly.

If  $W$  is not a verification set, we shrink  $W$  (thereby shrinking the bounds  $Y_k$  and  $Z_k$ ) until we obtain a verification set. Since we are evaluating  $G$  very close to zero, the higher order terms which we are bounding should become very small as we shrink  $W$ . Therefore, obtaining a verification set in this manner relies only on the bounds proving to be sufficient for particular numerical values.

We here comment that by considering a small interval of parameter values  $\nu$ , this technique may be combined with the pseudo-arclength method to verify the existence and uniqueness of bifurcation branches in appropriate neighborhoods.

#### 4.1.2 A sample result

For interval ranges of parameter values, we construct verification sets as described in the previous section. These sets are used to prove the existence and uniqueness (in the given box neighborhoods) of portions of bifurcation branches given in Figure 2. Combining these results, we verify the existence and uniqueness of the full branches in Figure 2, except near bifurcation points, and obtain a proof for Theorem 1.3.

The corresponding code using the interval arithmetic package C-XSC [9] is available [18].

## 4.2 Nonexistence

We begin this section by presenting a condition based on the mean value theorem which, if satisfied, may be used to prove the nonexistence of stationary solutions in a box  $B = B_F \times B_I$ , where  $B_F = \prod_{k=0}^{m-1} [a_k^-, a_k^+]$  and  $B_I = \prod_{k=m}^{\infty} [a_k^-, a_k^+]$ . First, let us define for each point  $A \in B_F$  an interval vector  $v(A)$  with

$$v_k(A) = \begin{cases} \tilde{a}_k - A_k, & \text{for } |k| < m \\ 0, & \text{for } |k| \geq m. \end{cases}$$

**Proposition 4.3** *If there exists a point  $A \in B_F$  and  $k \in \mathbb{N}$  such that*

$$0 \notin f_k(A, B_I) + (Df_k(B) \cdot v(A))$$

*then  $B = B_F \times B_I$  contains no stationary solutions.*

*Proof.* Suppose there exists a point  $b = (b_F, b_I) \in B$  such that  $f(b) = 0$ . Then, in particular,  $f_k(b) = 0$ . By the mean value theorem, there exists a point  $c = (c_F, c_I)$  on the line segment connecting  $a := (A, b_I)$  to  $b = (b_F, b_I)$  in  $B$  such that

$$\begin{aligned} f_k(a) + Df_k(c) \cdot (b - a) &= f_k(b) \\ &= 0. \end{aligned}$$

However,

$$f_k(a) + (Df_k(c) \cdot (b - a)) \in f_k(A, B_I) + Df_k(B) \cdot v(A)$$

contradicting the assumption. Therefore, there can be no such point  $b \in B$ . ■

In practice, we set  $k$  to be one of the low modes ( $k \in \{0, \dots, m-1\}$ ) and  $A$  to be a vertex of the constructed box  $B_F$ .

#### 4.2.1 Nonexistence construction

The goal is to completely determine the entire set of equilibria in a region  $J = J_F \times V$ , where  $J_F = \prod_{k=0}^{m-1} [A_k^-, A_k^+]$  and  $V = \prod_{k=m}^{\infty} [a_k^-, a_k^+]$ . As described in Section 4.1, we have procedures for proving the existence and uniqueness in small neighborhoods of hyperbolic equilibria. The method in this section will be used to prove the nonexistence of additional equilibria in  $J$ . The general strategy here is to decompose  $J$  into smaller boxes on which we check the nonexistence of equilibria as described above.

The procedure given in Section 4.1 results in the verification of a unique equilibrium in a set of the form  $W = \prod_{k=0}^{m-1} [x_k^-, x_k^+] \times V$  in  $x$ -coordinates, with the transformation  $T : x \mapsto a$  relating the  $x$ -coordinates to the original  $a$ -coordinate system. Since we will be applying the nonexistence algorithms in the  $a$ -coordinate system, we first think of this set in  $a$ -coordinates as  $TW$ . In addition, for computational purposes we wish to work with boxes aligned with the  $a$ -coordinate directions. We therefore consider a smaller set  $\Omega = \prod_{k=0}^{m-1} [a_k^-, a_k^+] \times V \subset TW$  (recall that  $Q_m T(\cdot, V) = V$ ). Let  $\{\Omega(p)\}_{p=1}^q$  be the collection of such sets. We now wish to show the nonexistence of equilibria in  $J \setminus \bigcup_p \Omega(p)$ . Note that if this can be done, then we have also shown that each of the proven equilibria is actually contained in the corresponding  $\Omega(p)$ .

In order to apply Proposition 4.3, we first decompose  $J \setminus \bigcup_p \Omega(p)$  into some boxes of the form  $B = \prod_{k=0}^{m-1} [a_k^-, a_k^+] \times V$  (Figure 3 is an example in the case of  $m = 2$  with two verified equilibria). Given a box  $B = B_F \times B_I$  in this decomposition, we choose  $k \in \{0, \dots, m-1\}$  and a vertex  $A$  of  $B_F$  and compute bounds for the two terms  $f_k(A, B_I)$  and  $Df_k(B) \cdot v(A)$ . For the first, consider the splitting  $f_k(A, B_I) = f_k^{(m)}(A) + r_k(A, B_I)$ , containing the first, finite sum and a second sum which may be bound using the formulas in Section 3. For the second term,

$$Df_k(B) \cdot v(A) = \sum_{|n| < m} (\tilde{a}_n - A_n) \left( \mu_k \delta_{kn} - 3 \sum_{\substack{n_1 + n_2 = k - n \\ n_i \in \mathbb{Z}}} \tilde{a}_{n_1} \tilde{a}_{n_2} \right).$$

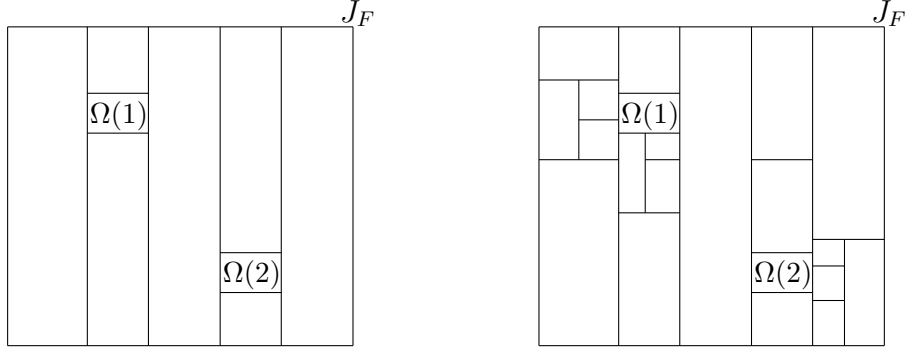


Figure 3: Decomposition of  $J_F$

The infinite sum,  $\sum_{n_1+n_2=k-n} \tilde{a}_{n_1} \tilde{a}_{n_2}$  may be bound by the formulas in Section 3, leaving a finite computation to perform. These bounds may now be used to check whether the condition in Proposition 4.3 is satisfied.

The boxes used in this procedure may be constructed adaptively – larger boxes may be checked for nonexistence of equilibria where  $f_k(A)$  is large relative to the bound for  $Df_k(B) \cdot v(A)$  whereas smaller boxes may be used where these values are closer, i.e. in regions near equilibria. Since we would like to keep the total number of computations low, we begin by dividing the set into a small number of large boxes (always of the form  $B = \prod_{k=0}^{m-1} [a_k^-, a_k^+] \times V$ ) where we check the conditions of Proposition 4.3. If these conditions are not satisfied, we subdivide boxes in the first  $m$  directions into two parts, continuing this procedure until we have proven the nonexistence of equilibria in all of  $J \setminus \bigcup_p \Omega(p)$ .

#### 4.2.2 A sample result

By restricting to the sets  $J_i$ , we use the above method to strengthen Theorem 1.3 to the following two lemmas.

**Lemma 4.4** *For parameter values sufficiently close to  $(L_1, \nu_1) = (0.62, 0.35)$  and  $(L_2, \nu_2) = (0.65, 0.4)$ ,  $J_i \subset L^2(0, 2\pi/L_i)$  contains exactly three equilibria  $M_i(p)$ ,  $p \in \{0, 1^\pm\}$ .*

*Proof.* The above method, combined with the verification sets used in the proof of Theorem 1.3, is applied to the sets  $J_1, J_2$  as listed in Table 1 and Table 2 at the corresponding parameter values. ■

**Lemma 4.5** *For parameter values sufficiently close to  $(L_3, \nu_3) = (0.62, 0.38)$  and  $(L_4, \nu_4) = (0.65, 0.48)$ ,  $J_i \subset L^2(0, 2\pi/L_i)$  contains exactly five equilibria  $M_i(p)$ ,  $p \in \{0, 1^\pm, 2^\pm\}$ .*

*Proof.* The above method combined with the verification sets used in the proof of Theorem 1.3, is applied to the sets  $J_3, J_4$  as listed in Table 3 and Table 4 at the corresponding parameter values. ■

For  $i = 1, 2$ ,  $\Omega_i(p)$ , which include equilibria  $M_i(p)$  in Theorem 1.1, are listed in Appendix 2. Similarly,  $\Omega_i(p)$ ,  $i = 3, 4$ , listed in Appendix 2 include equilibria  $M_i(p)$  in Theorem 1.2. The corresponding code to prove the nonexistence of equilibria is available [18].

### 4.3 Index information

The results of the previous two subsections provide us with complete information concerning the existence of equilibria. However, as was mentioned in the introduction, in order to determine the global dynamics we need to make use of the Conley index. In this subsection we recall the definition of this index (see [2, 12] and references therein for a more detailed introduction) and compute it for the equilibria.

Let us begin by establishing notation that will be used for the remainder of this paper. Let  $X$  be a locally compact metric space on which the local semiflow  $\phi$  is defined. A full solution through  $x \in X$  is a function  $\gamma_x : \mathbb{R} \rightarrow X$  satisfying:  $\gamma_x(0) = x$ , and  $\phi(t, \gamma_x(s)) = \gamma_x(t + s)$  for all  $t \geq 0$  and  $s \in \mathbb{R}$ . Given a set  $N \subset X$ , the *maximal invariant set* in  $N$  is defined by

$$\text{Inv}(N) := \{x \in N \mid \exists \text{ a full solution } \gamma_x : \mathbb{R} \rightarrow N\}.$$

**Definition 4.6** ([14]) A pair of compact sets  $(N, N^-)$  is an *index pair* if

1.  $\text{Inv}(\text{cl}(N \setminus N^-), \phi) \subset \text{int}(N \setminus N^-)$ , and
2. the induced semiflow  $\varphi_{\#} : [0, \infty) \times N/N^- \rightarrow N/N^-$  given by

$$\varphi_{\#}(t, x) := \begin{cases} \varphi(t, x) & \text{if } \varphi([0, t], x) \subset N \setminus N^- \\ [N^-] & \text{otherwise,} \end{cases}$$

is continuous.

**Definition 4.7** Let  $(N, N^-)$  be an index pair. The *cohomological Conley index* of the maximal invariant set  $S := \text{Inv}(\text{cl}(N \setminus N^-))$  is the relative Alexander-Spanier cohomology of  $N$  and its exit set  $N^-$ . For our purposes it is easier to work with the field coefficients  $\mathbb{Z}_2$  and in an abuse of notation we will occasionally write

$$CH^*(S) = CH^*(S; \mathbb{Z}_2) := \bar{H}^*(N, N^-; \mathbb{Z}_2).$$

We remark that the Conley index is well-defined. More precisely, if  $(N, N^-)$  and  $(N', N'^-)$  are index pairs with the property that

$$\text{Inv}(\text{cl}(N \setminus N^-)) = S = \text{Inv}(\text{cl}(N' \setminus N'^-))$$

then

$$\bar{H}^*(N, N^-; \mathbb{Z}_2) = CH^*(S) = \bar{H}^*(N', N'^-; \mathbb{Z}_2).$$

In particular, to compute the Conley index of an equilibrium we can make use of any index pair for which the maximal invariant set is precisely that equilibrium. As will be indicated below, projections of the verification sets of Section 4.1 onto the low modes yield index pairs.

As with the previous numerical procedures, the computation of the index is performed using a finite dimensional truncation. The following definition outlines the properties that allow us to lift index computations from the  $m$ -dimensional system to the original, infinite dimensional system.

**Definition 4.8** Recall that  $\{\varphi_k | k \in \mathbb{N}\}$  is a basis for  $X$  with  $a_k := \langle a, \varphi_k \rangle$ . A compact set  $W = N \times V \subset \prod_k [a_k^-, a_k^+]$ , is called a *lifting set* if, perhaps following a reordering of the basis, there exists a finite  $m \geq 0$  such that the following hold:

1.  $N$  and  $\{(a_k^-, a_k^+)\}$  form self-consistent a priori bounds [17] for (1).
2.  $(N, N^-) \subset \prod_{k=0}^{m-1} [a_k^-, a_k^+]$  is an index pair for all flows generated by the multi-valued ordinary differential equation  $\dot{a}_F = P_m E(a_F, V)$ .
3.  $V = \prod_{k=m}^{\infty} [a_k^-, a_k^+]$  and for all  $k \geq m$  either

(a)  $k$  is a *contracting direction*

$$\begin{aligned} \langle E(a), \varphi_k \rangle &< 0 && \text{for all } a \in W \text{ with } a_k = a_k^+ \\ \langle E(a), \varphi_k \rangle &> 0 && \text{for all } a \in W \text{ with } a_k = a_k^- \end{aligned}$$

or

(b)  $k$  is a *expanding direction*

$$\begin{aligned} \langle E(a), \varphi_k \rangle &> 0 && \text{for all } a \in W \text{ with } a_k = a_k^+ \\ \langle E(a), \varphi_k \rangle &< 0 && \text{for all } a \in W \text{ with } a_k = a_k^- \end{aligned}$$

The second condition requires that the computed index pair is an index pair for the multivalued,  $m$ -dimensional system containing a priori bounded error. The third condition requires that the property of isolation is preserved in lifting to higher dimensions. Namely, in each truncated direction, the projection of the system should be either contracting or expanding on the boundary. Note that these assumptions essentially boil down to checking a series of inequalities which again contain bounds computed as in Section 3. With these inequalities in mind, we once again turn to the  $x$ -coordinate system and, as with verification sets, construct lifting sets as boxes in the  $x$ -coordinate system of the form  $W = \prod_k [x_k^-, x_k^+]$ .

The following theorem relates the finite dimensional index to the index for the full, infinite dimensional system.

**Theorem 4.9** *Let  $W = N \times V$  be a lifting set and let  $CH^*(\text{Inv } K)$  denote the Conley index of  $K := \text{cl}(N \setminus N^-)$  under the flow induced by  $x_F = g_F(x_F)$ . If*

$$l = \text{card} \{k \mid k \text{ is an expanding direction}\}$$

*is finite, then the Conley index of  $\text{Inv}(K \times V)$  under the infinite dimensional flow is*

$$CH^{n+l}(\text{Inv}(K \times V)) \cong CH^n(\text{Inv}(K)), \quad n = 0, 1, 2, \dots$$

*Proof.* Observe that the relative topology of  $W$  inherited from  $L^2$  is equivalent to the product topology. Choose any integer  $M$  such that if  $k$  is an expanding direction, then  $k \leq M$ . Let

$$N' = N \times \prod_{k=m}^M [a_k^-, a_k^+].$$

Since there are  $l$  expanding directions by [2, 6.1.D]

$$CH^{n+l} \left( \text{Inv} \left( K \times \prod_{k=m}^M [a_k^-, a_k^+] \right) \right) \cong CH^n(K), \quad n = 0, 1, 2, \dots$$

Since all the remaining directions are contracting, again, by [2, 6.1.D]

$$CH^n(\text{Inv}(K \times V)) \cong CH^n \left( \text{Inv} \left( K \times \prod_{k=m}^M [a_k^-, a_k^+] \right) \right), \quad n = 0, 1, 2, \dots$$

■

In all the examples considered in this paper,  $l = 0$ .

### 4.3.1 Constructing a lifting set

We now construct a lifting set in the  $x$ -coordinate system, where

$$\dot{x}_k = g_k(x) = \lambda_k x_k + \epsilon_k(x).$$

Again, we begin by constructing a (small) initial set  $W = \prod_k [x_k^-, x_k^+]$  around the origin with the property that for some  $M > 0$ ,  $\tilde{x}_k := \frac{A_s}{k^s} [-1, 1]$  for all  $k \geq M$  where  $A_s$ ,  $s \geq 2$  are constants. We now use the formulas given in Section 3, this time to bound the term  $\epsilon(x)$ .

As before, let  $\tilde{W} := \prod_{k=0}^{\infty} [a_k^-, a_k^+]$  where  $\tilde{a}_k = [a_k^-, a_k^+] = [\min(TW)|_k, \max(TW)|_k]$ . In the high modes,  $k \geq m$ ,

$$\begin{aligned} \epsilon_k(x) &= \sum_{\substack{n_1+n_2+n_3=k \\ n_i \in \mathbb{Z}}} (Tx)_{n_1} (Tx)_{n_2} (Tx)_{n_3} \\ &\subseteq \sum_{\substack{n_1+n_2+n_3=k \\ n_i \in \mathbb{Z}}} \tilde{a}_{n_1} \tilde{a}_{n_2} \tilde{a}_{n_3} \end{aligned}$$

for all  $x \in W$  and has bound  $I_k$  as computed in Section 3.

In the low modes,

$$(\epsilon_0(x), \epsilon_1(x), \dots, \epsilon_{m-1}(x))^t = P^{-1}R(Tx), \quad (20)$$

where

$$R(a) = f^{(m)}(\bar{a}_F) + \frac{1}{2!}D^2 f^{(m)}(\bar{a}_F)(a_F - \bar{a}_F)^2 + \frac{1}{3!}D^3 f^{(m)}(\bar{a}_F)(a_F - \bar{a}_F)^3 + r(a).$$

For all  $x \in W$ ,

$$\begin{aligned} R_k(Tx) \in & f^{(m)}(\bar{a}_F) - 3 \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} \bar{a}_{n_1}(\tilde{a}_{n_2} - \bar{a}_{n_2})(\tilde{a}_{n_3} - \bar{a}_{n_3}) \\ & - \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} (\tilde{a}_{n_1} - \bar{a}_{n_1})(\tilde{a}_{n_2} - \bar{a}_{n_2})(\tilde{a}_{n_3} - \bar{a}_{n_3}) \\ & - \sum_{\substack{n_1+n_2+n_3=k \\ \max\{n_1, n_2, n_3\} \geq m}} \tilde{a}_{n_1} \tilde{a}_{n_2} \tilde{a}_{n_3}. \end{aligned}$$

Each of these terms may be bound through finite interval arithmetic computations and the formulas in Section 3. Combining these bounds with (20) gives bounds for  $\epsilon_k(x)$ ,  $0 \leq k < m$ .

As in the construction of a verification set, this set  $W$  may be updated, or refined, until the conditions in Definition 4.8 are met. The following theorem further illustrates the natural relationship between a verification set and a lifting set.

**Theorem 4.10** *If  $W = \prod_k [x_k^-, x_k^+]$  is a verification set (as constructed in Section 4.1) then  $W$  is a lifting set.*

*Proof.* If  $W = \prod_k [x_k^-, x_k^+]$  is a verification set (constructed as in Section 4.1), then  $\lambda_k^{-1}(\epsilon_k(W)) = G_k(W) \subset [x_k^-, x_k^+]$ . In other words, for all  $x \in W$ ,

$$\lambda_k^{-1} \epsilon_k(x) \in \tilde{x}_k$$

or equivalently if  $x_k^- = -x_k^+$ ,

$$|\epsilon_k(x)| \in |\lambda_k| \tilde{x}_k. \quad (21)$$

Consider the case  $\lambda_k < 0$ . We would like to show that on the portion of the boundary  $\{x \in W | x_k = x_k^+\} \subset \partial W$ , the vector field is pointing inwards. Here, (21) implies that

$$\begin{aligned} g_k(x) &= \lambda_k x_k^+ + \epsilon_k(x) \\ &< 0. \end{aligned}$$

The remaining inequality in this contracting case and the inequalities in the expanding case follow similarly. ■

### 4.3.2 A sample result

Consider once again the equilibria  $M(p)$  described in theorems 1.1 and 1.2. In order to show the existence of the connecting orbits between the equilibria, we need to prepare the following lemmas.

**Lemma 4.11** *The Conley indices of the equilibria  $M_i(p)$ ,  $p \in \{0^\pm, 1\}$  are*

$$CH^n(M_i(0^\pm)) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = 0 \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad CH^n(M_i(1)) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (22)$$

for  $i = 1, 2$ .

*Proof.* For  $i = 1, 2$  we use the verification sets constructed for the proof of Theorem 1.3 to compute the indices of  $M_i(p)$ . By Theorem 4.10, the verification sets are also lifting sets. For  $k = 0, \dots, 6$  the number of  $\lambda_k > 0$  (see (12)) is 0 if  $p = 0^\pm$  and is 1 if  $p = 1$ . The number of expanding directions is  $l = 0$ . Hence the result follows by Theorem 4.9. ■

A similar argument produces the following result.

**Lemma 4.12** *The Conley indices of the equilibria  $M_i(p)$ ,  $p \in \{0^\pm, 1^\pm, 2\}$  are, for  $q = 0, 1$ ,*

$$CH^n(M_i(q^\pm)) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = q \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad CH^n(M_i(2)) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = 2 \\ 0 & \text{otherwise.} \end{cases} \quad (23)$$

for  $i = 3, 4$ .

## 5 Conjugacy and semi-conjugacy

In this section, we complete the proofs of Theorems 1.1 and 1.2 by demonstrating the existence of a conjugacy or semi-conjugacy. As was indicated in the Introduction, it is this step that requires the machinery of the Conley index theory, thus we begin by recalling several essential definitions.

As before  $\phi : \mathbf{R} \times X \rightarrow X$  is a local flow on a locally compact metric space  $X$ . For  $x \in X$ , let  $\alpha(x)$  and  $\omega(x)$  denote the alpha and omega limit sets of  $x$  under  $\phi$ .

**Definition 5.1** A *Morse decomposition* of a compact invariant set  $S$  is a finite collection

$$\mathcal{M}(S) := \{M(p) \mid p \in (\mathcal{P}, >)\}$$

of mutually disjoint compact invariant subsets  $M(p)$ , called *Morse sets*, which admits a partial order  $>$  on the indexing set  $\mathcal{P}$  such that if  $x \in S \setminus \bigcup_{p \in \mathcal{P}} M(p)$ , then there exist  $p, q \in \mathcal{P}$  with  $\alpha(x) \subset M(p)$ ,  $\omega(x) \subset M(q)$ , and  $q > p$ .

In the context of this paper  $S_i := \text{Inv}(J_i)$  and the Morse sets are the equilibria in  $J_i$  with an admissible order given by the relative values of the Lyapunov function (6) on the equilibria. A key ingredient of the construction of the semi-conjugacy is the connection matrix [5] which relates the Conley indices of the equilibria, which were computed in Section 4.3.2, to the Conley index of the global invariant set.

**Definition 5.2** Let  $\mathcal{M}(S) = \{M(p) \mid p \in (\mathcal{P}, >)\}$  be a Morse decomposition with admissible ordering  $>$ . An associated *connection matrix*

$$\Delta : \bigoplus_{p \in \mathcal{P}} CH^*(M(p)) \rightarrow \bigoplus_{p \in \mathcal{P}} CH^*(M(p))$$

where

$$\Delta(p, q) : CH^*(M(q)) \rightarrow CH^*(M(p))$$

satisfies the following conditions.

1. It is *lower triangular*; that is if  $p \not> q$ , then

$$\Delta(p, q)CH^*(M(q)) = 0$$

2. It is a *co-boundary operator*; that is

$$\Delta(p, q)CH^n(M(q)) \subset CH^{n+1}(M(p))$$

and  $\Delta \circ \Delta = 0$ .

3. The relation between the Conley indices of the Morse sets and the Conley index of the total invariant set  $S$  is

$$\frac{\text{kernel } \Delta}{\text{image } \Delta} \cong CH^*(S).$$

The third condition suggests the need to compute the index of  $S_i$ .

**Lemma 5.3** For  $i = 1, 2, 3, 4$ , the set  $J_i = \prod_k [a_{k,i}^-, a_{k,i}^+]$  is a lifting set and for  $a \in J_i$ ,

$$\begin{aligned} a_k = a_{k,i}^+ &\Rightarrow f_k(a) < 0 \\ a_k = a_{k,i}^- &\Rightarrow f_k(a) > 0. \end{aligned} \tag{24}$$

Therefore, the cohomological Conley index of  $S_i := \text{Inv}(J_i)$  is

$$CH^n(S_i) \cong \begin{cases} \mathbb{Z}_2 & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

*Proof.* To prove that  $J_i$  is a lifting set, it is sufficient to check that the inequalities (24) hold. This is done numerically using the error bound formulae discussed in Section 3 and interval arithmetic.

Observe that (24) implies that the vector field points inward on the boundary of  $J$ . In particular, there is no exit set, thus  $CH^*(S_i) \cong \bar{H}^*(J_i; \mathbb{Z}_2)$ . The result now follows from the fact that  $J_i = \prod_k [a_{k,i}^-, a_{k,i}^+]$  is an acyclic set. ■

The final step before completing the proofs of Theorems 1.1 and 1.2 is to compute the appropriate connection matrices.

**Lemma 5.4** *For  $i = 1, 2$ , consider the compact invariant set  $S_i := \text{Inv}(J_i)$  with Morse decomposition*

$$\mathcal{M}(S_i) := \{M_i(p) \mid p \in \{0^\pm, 1\}\}.$$

*An admissible partial order on the indexing set  $\{0^\pm, 1\}$  is  $1 > 0^\pm$  and the associated connection matrix  $\Delta_i$  defined on*

$$CH^*(M_i(0^-)) \oplus CH^*(M_i(0^+)) \oplus CH^*(M_i(1))$$

*has the form*

$$\Delta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

*Proof.* The fact that  $\mathcal{M}(S_i) := \{M_i(p) \mid p \in \{0^\pm, 1\}\}$  is a Morse decomposition follows from Lemma 4.4 and the existence of a Lyapunov function  $F$  (6). By the symmetry of the Swift-Hohenberg equation,  $F(M_i(0^-)) = F(-M_i(0^+))$ , which implies that there cannot be any connecting orbits between  $M_i(0^-)$  and  $M_i(0^+)$ . Therefore we can restrict our attention to those admissible orderings for which  $0^\pm$  are unrelated. This, in turn, implies that  $0^\pm$  and 1 are adjacent.

By Lemma 4.11 and the fact that  $\Delta$  is a co-boundary operator

$$\Delta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \alpha & \beta & 0 \end{bmatrix}.$$

Since  $0^\pm$  and 1 are adjacent  $\Delta(1, 0^\pm)$  are determined by the long exact sequence of an index triple [12, 1]. Thus the symmetry relation (4) implies that  $\alpha = \beta$ .

Finally by Lemma 5.3 and the third condition of Definition 5.2 we can conclude that  $\alpha = 1$  and hence  $1 > 0^\pm$ . ■

Though the proof of the following lemma is similar we provide it here for the sake of completeness .

**Lemma 5.5** For  $i = 3, 4$ , consider the compact invariant set  $S_i := \text{Inv}(J_i)$  with Morse decomposition

$$\mathcal{M}(S_i) := \{M_i(p) \mid p \in \{0^\pm, 1^\pm, 2\}\}.$$

An admissible partial order on the indexing set  $\{0^\pm, 1^\pm, 2\}$  is  $2 > 1^\pm > 0^\pm$  and the associated connection matrix  $\Delta_i$  defined on

$$CH^*(M_i(0^-)) \oplus CH^*(M_i(0^+)) \oplus CH^*(M_i(1^-)) \oplus CH^*(M_i(1^+)) \oplus CH^*(M_i(2))$$

has the form

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

*Proof.* The fact that  $\mathcal{M}(S_i) := \{M_i(p) \mid p \in \{0^\pm, 1^\pm, 2\}\}$  is a Morse decomposition follows from Lemma 4.5 and the existence of the Lyapunov function  $F$  (6). As in the proof of Lemma 5.4 we can choose an admissible order such that  $0^\pm$  and  $1^\pm$  are unrelated, respectively. To determined the ordering we turn to the connection matrix.

Since  $\Delta$  is a co-boundary operator, by Lemma 4.12 it must take the form

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \gamma & \delta & 0 & 0 & 0 \\ \eta & \mu & 0 & 0 & 0 \\ 0 & 0 & \alpha & \beta & 0 \end{bmatrix}.$$

By Lemma 5.3 and the third condition of Definition 5.2 we can conclude that the pair  $(\alpha, \beta) \neq (0, 0)$  and the quadruple  $(\gamma, \delta, \eta, \mu) \neq (0, 0, 0, 0)$ . This implies that we can choose  $2 > 1^\pm > 0^\pm$  as an admissible ordering. In particular, it also implies that all entries are determined by long exact sequences of index triples. Thus applying the symmetry relation (4) again we can conclude that

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \gamma & \delta & 0 & 0 & 0 \\ \delta & \gamma & 0 & 0 & 0 \\ 0 & 0 & \alpha & \alpha & 0 \end{bmatrix}$$

The result now follows from the fact that  $\alpha = 1$ ,  $(\gamma, \delta) \neq (0, 0)$  and  $\Delta \circ \Delta = 0$ .  $\blacksquare$

We now state a series of assumptions which are minor modifications of the assumptions **A1-A4** of [10], but for which the proof of [10, Theorem 1.2] remains the same. The proofs of theorems 1.1 and 1.2 will follow from this special case of [10, Theorem 1.2].

**A1** Let  $\mathcal{A}$  be the maximal invariant set within a contractible set  $X$ .

**A2** Let  $\mathcal{M}(\mathcal{A}) = \{M(p^\pm) \mid p = 0, \dots, P-1\} \cup \{M(P)\}$  with ordering  $P > P^\pm > \dots > 0^\pm$  be a Morse decomposition for  $\mathcal{A}$ .

**A3** The cohomology Conley indices of the Morse sets are

$$CH^n(M(P); \mathbb{Z}_2) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = P \\ 0 & \text{otherwise,} \end{cases}$$

and

$$CH^n(M(p^\pm); \mathbb{Z}_2) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = p \\ 0 & \text{otherwise,} \end{cases}$$

for  $p = 0, \dots, P-1$ .

**A4** The connection matrix for  $\mathcal{M}(\mathcal{A})$  is given by

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ D_0 & 0 & 0 & 0 & 0 \\ 0 & D_1 & 0 & 0 & 0 \\ \vdots & & \ddots & 0 & 0 \\ 0 & 0 & 0 & D_{P-1} & 0 \end{bmatrix}$$

where

$$D_{P-1} : CH^{P-1}(M(P-1^-)) \oplus CH^{P-1}(M(P-1^+)) \rightarrow CH^P(M(P))$$

is given by  $D_{P-1} = [1, 1]$  and, for  $p = 0, \dots, P-2$

$$D_p : CH^p(M(p^-)) \oplus CH^p(M(p^+)) \rightarrow CH^{p+1}(M(p+1^-)) \oplus CH^{p+1}(M(p+1^+))$$

is given by

$$D_p = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

*Proofs of theorems 1.1 and 1.2.* For  $i = 1, 2, 3, 4$  let  $X = J_i$  and let  $S_i := \text{Inv}(J_i)$ . Since  $J_i$  is contractible, **A1** is satisfied. Choosing  $P = 1$  if  $i = 1, 2$  and  $P = 2$  if  $i = 3, 4$ , **A2** and **A4** follow from Lemmas 5.4 and 5.5. Lemma 4.11 and Lemma 4.12 guarantee that **A3** is satisfied.

Therefore, [10, Theorem 1.2] guarantees the existence of semi-conjugacies to the appropriate flows in Figure 1. Observe that in the case  $i = 1, 2$ , the equilibrium  $M_i(1)$  has a one-dimensional unstable manifold and therefore the semi-conjugacy is in fact a conjugacy. ■

## 6 Conclusions

Though the results presented in this paper are directed toward the Swift-Hohenberg equation it should be clear that the methods are, in principle, applicable to a large class of infinite dimensional systems. Nevertheless it is worth commenting on three shortcomings of this work and suggest possible remedies. The first involves the lack of analysis of the bifurcation points. As was mentioned earlier, techniques for the rigorous numerical identification of saddle-node and pitchfork bifurcations will be discussed in a forthcoming paper by P. Zgliczynski and the third author.

The second issue is the obvious discrepancy between the parameter ranges on which Theorem 1.3, which provides the local uniqueness of the bifurcation branches, and Theorems 1.1 and 1.2, which describe the global dynamics, are stated. While it is clear that in general the cost of demonstrating nonexistence (Section 4.2) clearly exceeds that of proving existence and uniqueness (Section 4.1) the fundamental problem is that of determining an appropriate isolating neighborhood  $J$ . In particular,  $J$  must satisfy three conditions: (1) it must isolate the desired invariant set, (2) it must be an isolating block so that the total index can be computed, and (3) it must be as small as possible to minimize the cost of establishing the nonexistence of equilibria. The careful reader may have observed that the parameter values  $(L_i, \nu_i)$  are all located close to bifurcation points. This allowed us to choose regions  $J_i$  which took the form of a product of intervals. Extending these results to a large range of parameter values will require constructing isolating blocks that are defined in terms of polyhedral regions cross intervals. Work in this direction is underway using ideas presented in [11] and [7].

A third reasonable criticism involves the choice of parameter values. As was indicated in the introduction, Swift-Hohenberg was introduced to describe patterns associated with Rayleigh-Bénard heat convection. Thus the primary interest is in small  $L$  and large  $\nu$ . In this case, there are an enormous number of equilibria, they are highly unstable, and the dimension of the global attractor is large. On the other hand, as is suggested by Theorem 4.9 the index of an isolating neighborhood can be computed using a small number of modes if the expansion or contraction rate in the complementary modes is sufficiently large. For a problem such as Swift-Hohenberg this is essentially determined by the ratios of the adjacent eigenvalues of the associated fourth order operator. This leads us to believe that coupling appropriate coordinate transformations with the ideas of [7] will lead to techniques that can be employed to study the dynamics of solutions involving higher modes. The large number of equilibria also implies that the simple, purely topological arguments used to prove lemmas 5.4 and 5.5 are insufficient. However, in many situations individual entries in the connection matrix can be computed directly if appropriate isolating neighborhoods can be extracted. Again, we believe that coupling the previous techniques with those of [11] will lead to progress on this front.

## Acknowledgements

The authors would like to thank N. Yamamoto for useful comments on the material in Section 4.1. The authors also wish to acknowledge Z. Arai and H. Kokubu for useful comments on this paper. This work is partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for J.S.P.S. Japan-US cooperative science program and for J.S.P.S. Scientific Research (B), 12440026. The second author is partially supported by Grant-in-Aid for J.S.P.S. Fellows, 03948. The first and third authors were partially supported by NSF INT-0089631 and NSF DMS-0107396.

## References

- [1] L Arnold, C. Jones, K. Mischaikow, and G. Raugel, *Dynamical Systems*, Lecture Notes in Math. **1609**, (R. Johnson, ed.) 1995, Springer-Verlag.
- [2] C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Lecture Notes **38** A.M.S. Providence, R.I. (1978).
- [3] S. Day, Towards a Rigorous Numerical Study of the Kot-Schaffer Model, *Dynamic Systems and Applications* **12**, 1-2, 87-97 (2003).
- [4] S. Day, O. Junge, K. Mishchaikow, A Rigorous Numerical Method for the Global analysis of Infinite dimensional Discrete Dynamical Systems, in preparation.
- [5] R. Franzosa, The Connection Matrix Theory for Morse decompositions, *Trans. Amer. Math. Soc.* **311**, 561 (1989).
- [6] Y. Hiraoka, T. Ogawa, K. Mischaikow, Conley Index Based Numerical Verification Method for Global Bifurcations of the Stationary Solutions to the Swift-Hohenberg Equation, *Trans. Jpn. Soc. Ind. Appl. Math* **13**, No.2 (2003).
- [7] E. Boczek, W. D. Kalies, and K. Mischaikow, Polygonal Approximations of Flows, in preparation.
- [8] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Problems*, Springer Verlag. Notes by A. K. Nandakumaran and Mythily Ramaswamy, Indian Institute of Science, Bangalore (1987).
- [9] R. Klatte, U. Kulisch, A. Wiethoff, C. Lawo, and M. Rauch, *C-XSC: A C++ Class Library for Extended Scientific Computing*, Springer-Verlag (1993).
- [10] K. Mischaikow, Global asymptotic dynamics of gradient-like bistable equations, *SIAM J. Math. Anal.*, **26** (1995) 1199-1224.
- [11] K. Mischaikow, Topological techniques for efficient rigorous computations in dynamics, *Acta Numerica 2002*, pages 435-478. Cambridge University Press, 2002.
- [12] K. Mischaikow and M. Mrozek, Conley Index, *Handbook of Dynamical Systems, Vol. 2*, 393-460, North-Holland, 2002.
- [13] B. Nicolaenko, B. Scheurer, and R. Temam, Some global dynamical properties of the Kuramoto-Sivashinsky equations: nonlinear stability and attractors, *Physica 16D* , 155 (1985).

- [14] J. Robbin, and D. Salamon, *Dynamical systems, shape theory and the Conley index*, Ergodic Theory Dynam. Systems **8**, Charles Conley Memorial Issue, 375–393, (1988).
- [15] J. B. Swift and P. C. Hohenberg, Hydrodynamic fluctuations at the convective instability, *Phys. Rev. A* **15**, 319 (1977).
- [16] N. Yamamoto, A Numerical Verification Method for Solutions of Boundary Value Problems with Local Uniqueness by Banach’s Fixed-Point Theorem, *SIAM J. Numer. Anal.* **35**, No.5, 2004 (1998).
- [17] P. Zgliczyński and K. Mischaikow, Rigorous Numerics for Partial Differential Equations: The Kuramoto-Sivashinsky Equation, *Found. Comput. Math.* **1**, 255 (2001).
- [18] <http://www.math.gatech.edu/>

## Appendix 1

We here describe better error bounds of cubic nonlinear terms by using explicit interval bounds for  $a_j$ ,  $|j| \leq M$ . Before proceeding to the cubic nonlinearity, let us review the quadratic nonlinearity:

$$\sum_{\substack{m_1+m_2=k \\ m_1, m_2 \in \mathbb{Z}}} a_{m_1} a_{m_2} = \sum_{\substack{m_1+m_2=k \\ |m_1|, |m_2| \leq M}} a_{m_1} a_{m_2} + 2 \sum_{\substack{m_1+m_2=k \\ |m_1| > M, |m_2| \leq M}} a_{m_1} a_{m_2} + \sum_{\substack{m_1+m_2=k \\ |m_1|, |m_2| > M}} a_{m_1} a_{m_2},$$

for  $k \in \mathbb{Z}^+$ . Note that the original infinite sum is decomposed into three parts. We treat error bounds of each divided part separately as follows,

$$\begin{aligned} \sum_{\substack{m_1+m_2=k \\ |m_1|, |m_2| \leq M}} a_{m_1} a_{m_2} &\subset \tilde{\eta}_1^{(2)}(k), & \sum_{\substack{m_1+m_2=k \\ |m_1| > M, |m_2| \leq M}} a_{m_1} a_{m_2} &\subset \eta_2^{(2)}(k)I, \\ \sum_{\substack{m_1+m_2=k \\ |m_1|, |m_2| > M}} a_{m_1} a_{m_2} &\subset \eta_3^{(2)}(k)I, \end{aligned}$$

where  $\tilde{\eta}_1^{(2)}(k)$  is an interval and  $\eta_2^{(2)}(k), \eta_3^{(2)}(k) \in \mathbb{R}$  for each  $k$  and  $I = [-1, 1]$ .

First of all, since  $\tilde{\eta}_1^{(2)}(k)$  consists of the finite sum and only exists for  $0 \leq k \leq 2M$ , we can directly calculate the error bounds for these terms by the interval arithmetic. Next, for  $\eta_2^{(2)}(k)$  and  $\eta_3^{(2)}(k)$ , the following error bounds can be obtained by estimating the infinite sums as integrals (e.g. [17]).

- $\eta_2^{(2)}(k)$

$$\begin{aligned} \eta_2^{(2)}(k) &= \sum_{m_1=M+1}^{M+k} \frac{c}{m_1^s} |\tilde{a}_{k-m_1}|, & \text{for } 0 \leq k \leq 2M, \\ \eta_2^{(2)}(k) &= \frac{\bar{\eta}_2^{(2)}}{k^s}, & \text{for } k > 2M, \end{aligned}$$

where  $|\tilde{a}_k| = \sup_{a_k \in \tilde{a}_k} |a_k|$  and

$$\bar{\eta}_2^{(2)} = c \left[ \sum_{l=1}^M \left\{ \frac{1}{\left(1 - \frac{l}{2M+1}\right)^s} + 1 \right\} |a_l| + |a_0| \right].$$

- $\eta_3^{(2)}(k)$

$$\begin{aligned} \eta_3^{(2)}(k) &= \frac{2c^2}{(s-1)(M+1)^s(M+k)^{s-1}}, & \text{for } 0 \leq k \leq 2M, \\ \eta_3^{(2)}(k) &= \frac{\bar{\eta}_3^{(2)}}{k^s}, & \text{for } k > 2M, \end{aligned}$$

where

$$\bar{\eta}_3^{(2)} = 2^s c^2 \left[ \frac{2}{(s-1)M^{s-1}} + \frac{1}{\left(M + \frac{1}{2}\right)^s} \right] + \frac{2c^2}{(s-1)M^{s-1}}.$$

Let us next show the error bounds for cubic nonlinear term:

$$\begin{aligned} \sum_{\substack{m_1+m_2+m_3=k \\ m_1, m_2, m_3 \in \mathbb{Z}}} a_{m_1} a_{m_2} a_{m_3} &= \sum_{\substack{m_1+m_2+m_3=k \\ |m_1|, |m_2|, |m_3| \leq M}} a_{m_1} a_{m_2} a_{m_3} + 3 \sum_{\substack{m_1+m_2+m_3=k \\ |m_1| > M \\ |m_2|, |m_3| \leq M}} a_{m_1} a_{m_2} a_{m_3} \\ &+ 3 \sum_{\substack{m_1+m_2+m_3=k \\ |m_1|, |m_2| > M \\ |m_3| \leq M}} a_{m_1} a_{m_2} a_{m_3} + \sum_{\substack{m_1+m_2+m_3=k \\ |m_1|, |m_2|, |m_3| > M}} a_{m_1} a_{m_2} a_{m_3}. \end{aligned}$$

We again treat error bounds of each devided part separately as follows,

$$\begin{aligned} \sum_{\substack{m_1+m_2+m_3=k \\ |m_1|, |m_2|, |m_3| \leq M}} a_{m_1} a_{m_2} a_{m_3} &\subset \tilde{\eta}_1^{(3)}(k), & \sum_{\substack{m_1+m_2+m_3=k \\ |m_1| > M \\ |m_2|, |m_3| \leq M}} a_{m_1} a_{m_2} a_{m_3} &\subset \eta_2^{(3)}(k)I, \\ \sum_{\substack{m_1+m_2+m_3=k \\ |m_1|, |m_2| > M \\ |m_3| \leq M}} a_{m_1} a_{m_2} a_{m_3} &\subset \eta_3^{(3)}(k)I, & \sum_{\substack{m_1+m_2+m_3=k \\ |m_1|, |m_2|, |m_3| > M}} a_{m_1} a_{m_2} a_{m_3} &\subset \eta_4^{(3)}(k)I. \end{aligned}$$

By similar procedures to the quadratic case, we have the following error bound formulas.

- $\eta_2^{(3)}(k)$  and  $\eta_3^{(3)}(k)$

$$\eta_\alpha^{(3)}(k) = \sum_{m_3=-M}^M |a_{m_3}| \eta_\alpha^{(2)}(|k - m_3|), \quad \text{for } 0 \leq k \leq M,$$

$$\eta_\alpha^{(3)}(k) = \sum_{m_3=-M}^{k-2M-1} |a_{m_3}| \frac{\bar{\eta}_\alpha^{(2)}}{(k - m_3)^s} + \sum_{m_3=k-2M}^M |a_{m_3}| \eta_\alpha^{(2)}(|k - m_3|), \quad \text{for } M < k \leq 3M,$$

$$\eta_\alpha^{(3)}(k) = \frac{\bar{\eta}_\alpha^{(3)}}{k^s}, \quad \text{for } k > 3M,$$

where  $\alpha = 2, 3$  and

$$\bar{\eta}_\alpha^{(3)} = \bar{\eta}_\alpha^{(2)} \left[ \sum_{m_3=1}^M \left\{ \frac{1}{\left(1 - \frac{m_3}{3M+1}\right)^s} + 1 \right\} |a_{m_3}| + |a_0| \right].$$

- $\eta_4^{(3)}(k)$

$$\eta_4^{(3)}(k) = \sum_{m_3=M+1}^{k+2M} \frac{c}{m_3^s} \eta_3^{(2)}(|k - m_3|) + \frac{c \bar{\eta}_3^{(2)}}{(s-1)(2M+1)^s (k+2M)^{s-1}}$$

$$\begin{aligned}
& + \sum_{m_3=M+1}^{2M-k} \frac{c}{m_3^s} \eta_3^{(2)}(k+m_3) + \frac{c\bar{\eta}_3^{(2)}}{(s-1)(2M+1)^s(2M-k)^{s-1}}, \quad \text{for } 0 \leq k < M, \\
\eta_4^{(3)}(k) & = \sum_{m_3=M+1}^{k+2M} \frac{c}{m_3^s} \eta_3^{(2)}(|k-m_3|) + \frac{c\bar{\eta}_3^{(2)}}{(s-1)(2M+1)^s(k+2M)^{s-1}} \\
& + \frac{c\bar{\eta}_3^{(2)}}{(s-1)(M+1)^s(M+k)^{s-1}}, \quad \text{for } M \leq k \leq 3M, \\
\eta_4^{(3)}(k) & = \frac{\bar{\eta}_4^{(3)}}{k^s}, \quad \text{for } k > 3M,
\end{aligned}$$

where

$$\begin{aligned}
\bar{\eta}_4^{(3)} & = 2^s c \bar{\eta}_3^{(2)} \left[ \frac{2}{(s-1)M^{s-1}} + \left\{ \frac{2}{3M+1} \right\}^s \right] + c \left[ \sum_{l=1}^{2M} \left\{ \frac{1}{\left(1 - \frac{l}{3M+1}\right)^s} + 1 \right\} \eta_3^{(2)}(l) + \eta_3^{(2)}(0) \right] \\
& + \frac{c\bar{\eta}_3^{(2)}}{(s-1)(2M)^{s-1}} + \frac{c\bar{\eta}_3^{(2)}}{(s-1)M^{s-1}}.
\end{aligned}$$

## Appendix 2

We here list the sets  $\Omega_i(p)$  described in Section 4.2. For  $i = 1, 2, 3, 4$ ,  $\Omega_i(p)$  is given in  $a$ -coordinates and contains the unique equilibrium  $M_i(p)$  found in Theorem 1.1 or Theorem 1.2.

Table 5:  $\Omega_1(1)$

$k$	$a_k^-$	$a_k^+$
0	$-1.5317082198 \times 10^{-4}$	$1.5317082198 \times 10^{-4}$
1	$-3.4374821943 \times 10^{-3}$	$3.4374821943 \times 10^{-3}$
2	$-1.6325163560 \times 10^{-3}$	$1.6325163560 \times 10^{-3}$
3	$-1.7467975387 \times 10^{-5}$	$1.7467975387 \times 10^{-5}$
4	$-1.1338774960 \times 10^{-6}$	$1.1338774960 \times 10^{-6}$
5	$-2.9922175570 \times 10^{-7}$	$2.9922175570 \times 10^{-7}$
6	$-9.5783470215 \times 10^{-8}$	$9.5783470215 \times 10^{-8}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 6:  $\Omega_1(0^+)$  ( $\Omega_1(0^-) = -\Omega_1(0^+)$ )

$k$	$a_k^-$	$a_k^+$
0	$-3.6982410561 \times 10^{-4}$	$4.3380010295 \times 10^{-4}$
1	$-2.0837003819 \times 10^{-3}$	$2.0837003819 \times 10^{-3}$
2	$-1.4440654070 \times 10^{-1}$	$-1.3931824846 \times 10^{-1}$
3	$-3.0490093879 \times 10^{-5}$	$3.0490093879 \times 10^{-5}$
4	$-6.3585131416 \times 10^{-6}$	$6.3585131416 \times 10^{-6}$
5	$-7.5694490026 \times 10^{-6}$	$7.5694490026 \times 10^{-6}$
6	$-2.1693705892 \times 10^{-5}$	$2.1693705892 \times 10^{-5}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 7:  $\Omega_2(1)$

$k$	$a_k^-$	$a_k^+$
0	$-4.1939407023 \times 10^{-4}$	$4.1939407023 \times 10^{-4}$
1	$-3.7843623218 \times 10^{-3}$	$3.7843623218 \times 10^{-3}$
2	$-3.3066549558 \times 10^{-3}$	$3.3066549558 \times 10^{-3}$
3	$-3.3758549926 \times 10^{-5}$	$3.3758549926 \times 10^{-5}$
4	$-2.2595380987 \times 10^{-6}$	$2.2595380987 \times 10^{-6}$
5	$-5.9893415999 \times 10^{-7}$	$5.9893415999 \times 10^{-7}$
6	$-1.8991402709 \times 10^{-7}$	$1.8991402709 \times 10^{-7}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 8:  $\Omega_2(0^+)$  ( $\Omega_2(0^-) = -\Omega_2(0^+)$ )

$k$	$a_k^-$	$a_k^+$
0	$-9.4260800515 \times 10^{-5}$	$9.4260800515 \times 10^{-5}$
1	$-1.5081791552 \times 10^{-1}$	$-1.4655005417 \times 10^{-1}$
2	$-1.1781504261 \times 10^{-3}$	$1.1781504261 \times 10^{-3}$
3	$4.1782758443 \times 10^{-4}$	$4.5380637872 \times 10^{-4}$
4	$-1.2145663676 \times 10^{-5}$	$1.2145663676 \times 10^{-5}$
5	$-1.1784002385 \times 10^{-6}$	$1.1784002385 \times 10^{-6}$
6	$-2.3893991465 \times 10^{-7}$	$2.3893991465 \times 10^{-7}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 9:  $\Omega_3(2)$ 

$k$	$a_k^-$	$a_k^+$
0	$-1.1124876219 \times 10^{-6}$	$1.1124876219 \times 10^{-6}$
1	$-6.6536341020 \times 10^{-4}$	$6.6536341020 \times 10^{-4}$
2	$-7.5807322685 \times 10^{-6}$	$7.5807322685 \times 10^{-6}$
3	$-1.2165556884 \times 10^{-7}$	$1.2165556884 \times 10^{-7}$
4	$-9.7660704376 \times 10^{-9}$	$9.7660704376 \times 10^{-9}$
5	$-2.9747177429 \times 10^{-9}$	$2.9747177429 \times 10^{-9}$
6	$-1.1482066812 \times 10^{-9}$	$1.1482066812 \times 10^{-9}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 10:  $\Omega_3(1^+)$  ( $\Omega_3(1^-) = -\Omega_3(1^+)$ )

$k$	$a_k^-$	$a_k^+$
0	$-1.3740415981 \times 10^{-6}$	$1.3740415981 \times 10^{-6}$
1	$-1.8863964826 \times 10^{-2}$	$-1.7976479763 \times 10^{-2}$
2	$-1.0459873959 \times 10^{-5}$	$1.0459873959 \times 10^{-5}$
3	$1.0411971668 \times 10^{-6}$	$1.2034438744 \times 10^{-6}$
4	$-3.0208353199 \times 10^{-8}$	$3.0208353199 \times 10^{-8}$
5	$-1.4058460589 \times 10^{-8}$	$1.4058460589 \times 10^{-8}$
6	$-6.3175140437 \times 10^{-9}$	$6.3175140437 \times 10^{-9}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 11:  $\Omega_3(0^+)$  ( $\Omega_3(0^-) = -\Omega_3(0^+)$ )

$k$	$a_k^-$	$a_k^+$
0	$-4.8199952889 \times 10^{-4}$	$5.8167864654 \times 10^{-4}$
1	$-2.3748793849 \times 10^{-3}$	$2.3748793849 \times 10^{-3}$
2	$-1.7594153098 \times 10^{-1}$	$-1.7125411581 \times 10^{-1}$
3	$-3.4403863059 \times 10^{-5}$	$3.4403863059 \times 10^{-5}$
4	$-1.1850510301 \times 10^{-5}$	$1.1850510301 \times 10^{-5}$
5	$-1.2711347360 \times 10^{-5}$	$1.2711347360 \times 10^{-5}$
6	$-3.7556092806 \times 10^{-5}$	$3.7556092806 \times 10^{-5}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 12:  $\Omega_4(2)$ 

$k$	$a_k^-$	$a_k^+$
0	$-9.6930702748 \times 10^{-6}$	$9.6930702748 \times 10^{-6}$
1	$-3.4406905024 \times 10^{-5}$	$3.4406905024 \times 10^{-5}$
2	$-1.2924093700 \times 10^{-3}$	$1.2924093700 \times 10^{-3}$
3	$-6.8353570257 \times 10^{-7}$	$6.8353570257 \times 10^{-7}$
4	$-5.0899700401 \times 10^{-8}$	$5.0899700401 \times 10^{-8}$
5	$-1.4635716113 \times 10^{-8}$	$1.4635716113 \times 10^{-8}$
6	$-5.2094286906 \times 10^{-9}$	$5.2094286906 \times 10^{-9}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 13:  $\Omega_4(1^+)$  ( $\Omega_4(1^-) = -\Omega_4(1^+)$ )

$k$	$a_k^-$	$a_k^+$
0	$-9.6792794986 \times 10^{-6}$	$9.6887022005 \times 10^{-6}$
1	$-3.6861573395 \times 10^{-5}$	$3.6861573395 \times 10^{-5}$
2	$-3.6512349385 \times 10^{-2}$	$-3.5208637098 \times 10^{-2}$
3	$-6.6849439319 \times 10^{-7}$	$6.6849439319 \times 10^{-7}$
4	$-9.4028043898 \times 10^{-8}$	$9.4028043898 \times 10^{-8}$
5	$-3.4676484566 \times 10^{-8}$	$3.4676484566 \times 10^{-8}$
6	$-3.0964442109 \times 10^{-7}$	$3.0964442109 \times 10^{-7}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$

Table 14:  $\Omega_4(0^-) = -\Omega_4(0^+)$ 

$k$	$a_k^-$	$a_k^+$
0	$-3.5926585107 \times 10^{-4}$	$3.5926585107 \times 10^{-4}$
1	$-2.2277203072 \times 10^{-1}$	$-2.2031324460 \times 10^{-1}$
2	$-1.5857924271 \times 10^{-4}$	$1.5857924271 \times 10^{-4}$
3	$1.3984926023 \times 10^{-3}$	$1.4412493373 \times 10^{-3}$
4	$-2.7590137347 \times 10^{-5}$	$2.7590137347 \times 10^{-5}$
5	$-4.3759890385 \times 10^{-6}$	$4.3759890385 \times 10^{-6}$
6	$-7.0338045128 \times 10^{-7}$	$7.0338045128 \times 10^{-7}$
$k \geq 7$	$-1.0/k^4$	$1.0/k^4$