Rigorous Numerics for the Cahn-Hilliard Equation on the Unit Square

Stanislaus Maier-Paape, Ulrich Miller, Konstantin Mischaikow, and Thomas Wanner

Institut für Mathematik
RWTH Aachen
52062 Aachen — Germany
maier@instmath.rwth-aachen.de

SKS Unternehmensberatung
65239 Hochheim am Main — Germany
ulrich.miller@sks-ub.de

Department of Mathematics
Rutgers University
Piscataway, NJ 08854 — USA
mischaik@math.rutgers.edu

Department of Mathematical Sciences
George Mason University
Fairfax, VA 22030 — USA
wanner@math.gmu.edu

Received: April 9, 2007
Accepted: November 15, 2007

ABSTRACT

While the structure of the set of stationary solutions of the Cahn-Hilliard equation on one-dimensional domains is completely understood, only partial results are available for two-dimensional base domains. In this paper, we demonstrate how rigorous computational techniques can be employed to establish computer-assisted existence proofs for equilibria of the Cahn-Hilliard equation on the unit square. Our method is based on results by Mischaikow and Zgliczyński [22], and combines rigorous computations with Conley index techniques. We are able to establish branches of equilibria and, under more restrictive conditions, even the local uniqueness of specific equilibrium solutions. Sample computations for several branches are presented, which illustrate the resulting patterns.

Key words: Cahn-Hilliard equation, stationary solutions, bifurcation diagram, continuation, computer-assisted proof.

2000 Mathematics Subject Classification: 37L10, 35B05, 35K35, 35K55.

Konstantin Mischaikow was partially supported by NSF grants DMS-0511115 and DMS-0107396, DOE grant DE-FG02-05ER25711, and DARPA. The work of Thomas Wanner was partially supported by NSF grant DMS-0406231 and the U.S. Department of Energy under Contract DE-FG02-05ER25712.

ISSN: 1139-1138
http://dx.doi.org/10.5209/rev_REMA.2008.v21.n2.16380

Contents

Introduction

The Cahn-Hilliard equation

$$\begin{aligned} u_t &= -\Delta(\Delta u + \lambda f(u)), \quad \text{in} \quad \Omega \subset \mathbb{R}^n, \\ \partial_\nu u &= \partial_\nu \Delta u = 0, \qquad\quad \text{on} \quad \partial\Omega, \end{aligned} \tag{1}$$

was introduced in [2] and [3] as a model for the process of phase separation of a binary alloy at a fixed temperature. The function $u(t,x)$ represents the concentration of one of the two components of the binary alloy, and $\partial_\nu u(x)$ denotes the outer normal derivative of u at a point $x \in \partial\Omega$. The physical "interaction length" is given by $\lambda^{-1/2}$, and thus, this parameter effectively measures the size of the material. Equation (1) is mass preserving, that is,

$$\frac{d}{dt}\left(\frac{1}{|\Omega|}\int_\Omega u(t,x)\,dx\right) = 0,$$

so a second natural parameter is the total mass

$$\mu := \frac{1}{|\Omega|}\int_\Omega u(t,x)\,dx. \tag{2}$$

Of fundamental interest is the development and evolution of the concentration patterns of the alloy components as a function of time. Since f is generally chosen as a cubic-like nonlinearity, in this paper we use the standard choice

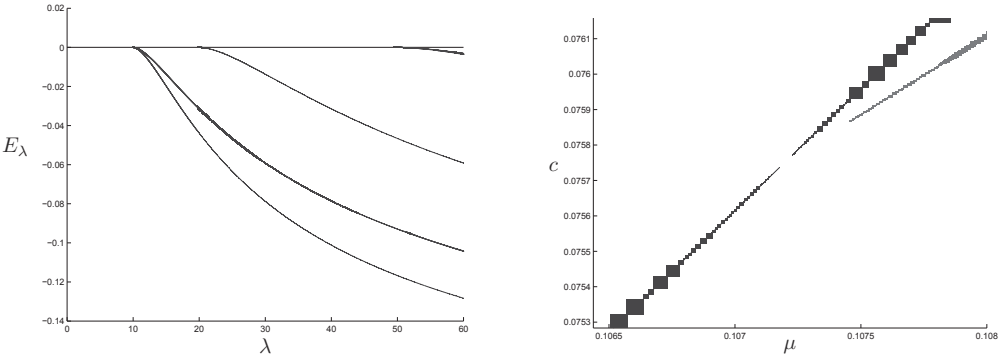
$$f(u) := u - u^3, \tag{3}$$

understanding the global dynamics is extremely difficult. Moreover, (1) is an $H^{-1}(\Omega)$-gradient system with respect to the van der Waals free energy functional

$$E_\lambda(u) := \int_\Omega \left(\frac{1}{2\lambda}\cdot|\nabla u|^2 - \frac{1}{2}u^2 + \frac{1}{4}u^4\right) dx. \tag{4}$$

For more details we refer the reader to Fife [8]. Thus, the first step towards describing the dynamics of (1) is to identify all its equilibrium solutions. Observe that the stationary states are given by the solutions of the nonlinear elliptic boundary value problem

$$\begin{aligned} \Delta u + \lambda f(u) &= \lambda c \quad \text{in} \quad \Omega, \\ \partial_\nu u &= 0 \qquad \text{on} \quad \partial\Omega, \end{aligned} \tag{5}$$



(a) (b)

Figure 1 – Two sample bifurcation diagrams. The left diagram is for fixed mass μ and shows solution branches and their associated energies E_λ as a function of λ. The right diagram contains regions where solution branches for fixed λ are contained, this time as functions of the total mass μ and the integration constant c.

which introduces the additional parameter

$$c := \frac{1}{|\Omega|} \int_\Omega f(u). \tag{6}$$

In the case of the one-dimensional domain $\Omega = (0, 1)$, the set of solutions of (5) is completely understood. In particular, the work of Novick-Cohen and Peletier [20] and of Grinfeld and Novick-Cohen [11] should be mentioned. However, for higher-dimensional domains very little is known, and in fact, given current techniques there appears to be scant hope that the problem can be resolved by purely analytic techniques. Therefore, we have chosen to pursue this problem using newly developed methods that lead to rigorous numerical proofs of the existence of equilibria for partial differential equations [4, 6, 7, 22].

As might be imagined, a proper description of our results is fairly technical and thus is presented in section 1. However, the diagrams in figure 1 serve as representative samples. The (slightly thickened) lines in the left diagram actually represent intervals within which we establish the existence of unique branches of equilibrium solutions as functions of the parameter λ, for fixed mass μ. The vertical axis indicates the corresponding energies E_λ. Similarly, the diagram (b) of figure 1 provides an amplified view of two solution branches in the vicinity of a bifurcation point as functions of the parameters μ and c, this time for fixed λ. Notice that due to inherent limitations of the techniques presented in this paper, these lines do not extend to the bifurcation point. However, this gap can most likely be addressed using ideas developed by P. Zgliczyński and one of the authors of this paper.

The results of section 1 are obtained via a two step procedure. First, the expected equilibria are computed numerically using a path-following algorithm developed for the numerical analysis of the equilibria of the Cahn-Hilliard equation in [14], which is applied to a Galerkin approximation of (5). The algorithm is based on a predictor-corrector procedure with step-length adaption as discussed for example in Allgower and Georg [1, chapter 3]. Details can be found in the appendix of [14] and will therefore not be repeated in the present paper. It suffices to note that the path-following computations merely provide the input information for the subsequent rigorous computations. If this input information turns out to be poor, then the rigorous computations fail to verify the existence of an equilibrium.

The rigorous computations are the subject of the second step. Here, topological techniques in combination with rigorous estimates are used to prove the existence and uniqueness of the equilibria. The majority of this paper is devoted to presenting the details of this approach. However, a brief outline is appropriate at this point. Observe that in case $\Omega = [0,1]^2$ the definitions

$$\phi_{0,0} = 1 \quad \text{and} \quad \phi_{i,j}(x,y) = \cos(i\pi x)\cos(j\pi y), \text{ for all } (i,j) \in \mathbb{N}^2 \setminus \{(0,0)\} \tag{7}$$

furnish a complete orthogonal basis for the Hilbert space $L^2(\Omega)$. Using the associated Fourier expansion, an element $u(t,\cdot,\cdot) \in L^2(\Omega)$ can be written as

$$u(t,x,y) = \sum_{i,j=0}^{\infty} u_{i,j}(t) \cdot \phi_{i,j}(x,y). \tag{8}$$

Substituting this expression into (5) with the cubic nonlinearity f defined in (3), we obtain the infinite system of ordinary differential equations

$$\dot{u}_{i,j} = -(i^2+j^2)^2\pi^4 u_{i,j} + \lambda(i^2+j^2)\pi^2\left(u_{i,j} - \frac{c_{i,j}}{4}\sum_{p,q,r,s\in\mathbb{Z}} \tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}\right) \tag{9}$$

for $i,j \in \mathbb{N}_0$, with suitable fixed constants $c_{i,j}$ (cf. Lemma A.1 and (55)) and

$$\tilde{u}_{i,j} \coloneqq \begin{cases} 4u_{|i|,|j|} & \text{for } (i,j) = (0,0), \\ 2u_{|i|,|j|} & \text{for } i = 0, j \neq 0 \text{ or } i \neq 0, j = 0, \\ u_{|i|,|j|} & \text{otherwise,} \end{cases} \qquad \text{for} \quad (i,j) \in \mathbb{Z}^2.$$

Using the standard idea of a Galerkin projection we restrict our attention to the finite subset of equations of (9) associated with $i,j < M$ and $u_{p,q} = 0$ for all $p,q \geq M$. This yields the finite-dimensional system of ordinary differential equations given by

$$\dot{u}_{i,j} = -(i^2+j^2)^2\pi^4 u_{i,j} + \lambda(i^2+j^2)\pi^2\left(u_{i,j} - \frac{c_{i,j}}{4}\sum_{p,q,r,s} \tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}\right), \tag{10}$$

where $0 \leq i,j < M$.

Clearly, passing from (9) to (10) introduces errors. However, the fact that solutions to the elliptic problem (5) are smooth implies at least a polynomial decay of the Fourier coefficients in (8). As is presented in section A in the appendix, this allows us to obtain explicit estimates for the truncation errors.

To deduce the existence of equilibrium solutions for the Cahn-Hilliard equation from a study of the finite-dimensional system (10) additionally requires existence results that are robust with respect to fixed size perturbations — in particular perturbations of the size of the truncation errors. This suggests the use of topology and for the purposes of this paper we employ two concepts, namely *self-consistent a priori bounds* and the *Conley index*. These concepts are recalled in section 2, which for the most part does not provide new information, but is rather used to introduce the ideas and notation employed in the final two sections. The exception is Theorem 2.6 which provides sufficient conditions for the verification of self-consistent a priori bounds for the Cahn-Hilliard equation on the unit square $(0,1)^2$. The verification of sets and sequences which satisfy *strict topologically self-consistent a priori bounds* and specific conditions on the associated Conley index then guarantee the existence of equilibria satisfying specific bounds for the Cahn-Hilliard equation.

With the existence question satisfied, section 3 contains results that establish the uniqueness of equilibrium solutions. Finally, section 4 concludes the paper with a brief discussion of the interactive use of the path-following algorithms in combination with the rigorous computations. This is necessary to establish existence results over regions of parameter space as opposed to particular points in parameter space.

As already mentioned, for higher dimensions analytical approaches to equilibria of (1) are quite rare. Maybe the most striking result is due to Kielhöfer [13]. He views (5) and (2) on $\Omega = (0,1)^2$ as a bifurcation problem for $\lambda = \lambda_0$ fixed and sufficiently large. It is easy to see that there are two possible bifurcation points from the trivial solution line $c = f(\mu)$ for each mode. The modes are the eigenfunctions of (5) linearized around $u \equiv \mu$. In case of the unit square the modes are given by $w_{ij} = \phi_{i,j}$ given in (7). Kielhöfer shows in [13] that the continua for modes of the form w_{kk} and $w_{k0} + w_{0k}$ connect those two bifurcation points and are separated from each other (see Theorem 1.3). In the same spirit is a result of Maier-Paape and Miller [15]. They show that for modes of the form w_{kl} the two bifurcation points are connected by a continuum of nontrivial solutions, either separated from other branches as in Kielhöfer's result, or they have an extra connection to the trivial solution (see Theorem 1.4). Furthermore they prove that the continua for modes of the form w_{kl} and $w_{k0} + w_{0k}$ and fixed $\mu_0 = 0$ continue as smooth curves in the parameter λ and are separated from each other.

Here, besides that we check the analytical results with our rigorous numerical method, we also find solutions that obviously do not lie on one of the known branches. We refer to them as secondary bifurcations. Additionally, we gathered results for the mode $w_{12} + w_{21}$, for which at present there is no analytical result. We give results for the modes w_{01}, $w_{10} + w_{01}$, w_{11}, w_{12}, and $w_{12} + w_{21}$. Due to limitations in computer

space and time, we only calculated equilibria for $\lambda \leq 60$. Confer also the results of Maier-Paape and Miller in [14] which cover a much larger λ range, but there the used standard numerical method is not rigorous.

We would like to point out that the existence of equilibria could also be obtained by using purely degree-theoretic arguments, i.e., without employing Conley index methods. However, the current paper serves as the foundation for the paper [16], which determines the global dynamics of the Cahn-Hilliard equation on the unit square. For this, one has to combine the existence results for equilibria and knowledge of their Conley indices with the topological machinery of the Conley index theory, most notably connection and transition matrices, to provide possible characterizations of the global attractor and the existence of global bifurcations as a function of the parameters λ and μ. This is the main reason for our use of Conley index methods in the current paper.

1. Equilibrium solutions on the square

In this section we present our main results for the Cahn-Hilliard equation on the unit square in detail. This will be accomplished in sections 1.3 through 1.6, where we describe the bifurcation scenario for $\lambda \leq 60$. As a preparation, sections 1.1 and 1.2 review known analytical results on the structure of the equilibrium set and survey useful symmetry arguments.

1.1. Bifurcation analysis and analytical results

We begin by reviewing some results from bifurcation theory. It can easily be seen that the stationary problem (5) subject to the mass constraint (2) has the trivial solution $u \equiv \mu$ and $c = f(\mu) = \mu - \mu^3$ for arbitrary $\lambda > 0$. (Recall that we assume the specific cubic nonlinearity defined in (3).) In order to study bifurcations from this trivial solution, we consider the kernel of the linearization of (5) at the homogeneous state $u \equiv \mu$, which is given by all solutions of the problem

$$\begin{aligned} \Delta v + \lambda(1 - 3\mu^2)v = 0, \quad \text{in} \quad \Omega, \\ \partial_\nu v = 0, \quad \text{on} \quad \partial\Omega. \end{aligned} \tag{11}$$

It can easily be seen that the solutions of this equation have to be eigenfunctions of $-\Delta$ on Ω subject to homogeneous Neumann boundary conditions. More precisely, if w denotes a non-constant eigenfunction of the negative Laplacian with eigenvalue $\kappa > 0$, then $v = w$ solves (11) for

$$\lambda = \frac{\kappa}{1 - 3\mu^2} \quad \text{for} \quad |\mu| < \frac{1}{\sqrt{3}}.$$

Thus, bifurcations from the homogeneous state μ are only possible if μ is contained in the spinodal region, and in this case

$$\mu = \pm\sqrt{\frac{1}{3} - \frac{\kappa}{3\lambda}} \quad \text{for} \quad \lambda > \kappa.$$

The relevant eigenfunctions or modes w for $-\Delta$ on the square $\Omega = (0,1)^2$ are given by

$$w_{kl}(x_1, x_2) = \cos(\pi k x_1) \cdot \cos(\pi l x_2), \quad \text{for} \quad (x_1, x_2) \in [0,1]^2,$$

as well as $k, l \in \mathbb{N}_0$, with corresponding eigenvalues $\kappa_{kl} := (k^2 + l^2) \cdot \pi^2$. In other words, the set of possible bifurcation points from the homogeneous state is given by

$$\lambda_{ij} = \lambda_{ij}(\mu) = \frac{\kappa_{ij}}{1 - 3\mu^2}, \qquad \text{for} \quad |\mu| < \frac{1}{\sqrt{3}}, \tag{12}$$

or

$$\mu_{ij}^{\pm} = \mu_{ij}^{\pm}(\lambda) = \pm\sqrt{\frac{1}{3} - \frac{\kappa_{ij}}{3\lambda}}, \quad \text{for} \quad \lambda > \kappa_{ij}. \tag{13}$$

In order to state some Rabinowitz-type results for bifurcations occurring at the modes $w_{i0} + w_{0i}$ and w_{ij}, for $i, j \in \mathbb{N}$, we first have to introduce some notation. Consider the closed subspaces

$$X_{ij}^{k+\alpha} = \left\{ v \;\middle|\; \int_\Omega v \, dx = 0 \right\} \cap C^{k,\alpha}(\mathbb{R}^2) \cap S_{ij}$$

of the usual Hölder spaces for $k \in \mathbb{N}_0$ and $\alpha \in [0,1]$, where

$$S_{ij} = \left\{ w \;\middle|\; w(x,y) = w(-x,y) = w(x,-y) = w\left(\frac{2}{i} - x, y\right) \right.$$
$$\left. = w\left(x, \frac{2}{j} - y\right) = w\left(\frac{1}{i} - x, \frac{1}{j} - y\right) \right\},$$

as well as

$$X_i^{k+\alpha} = \left\{ v \;\middle|\; \int_\Omega v \, dx = 0 \right\} \cap C^{k,\alpha}(\mathbb{R}^2) \cap S_i,$$

where

$$S_i = \left\{ w \;\middle|\; w(x,y) = w(-x,y) = w(x,-y) = w\left(\frac{2}{i} - x, y\right) \right.$$
$$\left. = w\left(x, \frac{2}{i} - y\right) = w(y,x) \right\}.$$

Then the mapping G defined by

$$G : \mathbb{R} \times \mathbb{R} \times X_{ij}^{2+\alpha} \longrightarrow X_{ij}^{\alpha} \quad \text{or} \quad G : \mathbb{R} \times \mathbb{R} \times X_i^{2+\alpha} \longrightarrow X_i^{\alpha},$$

with

$$G(\mu,\lambda,v) = \Delta v + \lambda f(v+\mu) - \lambda \int_\Omega f(v+\mu)dx$$

is smooth. Moreover, zeros of G correspond to solutions $u = v + \mu$ of (5).

Now fix $\lambda_0 > \kappa_{ij}$ or $\lambda_0 > \kappa_{i0}$. It can easily be shown that 0 is a simple eigenvalue of $D_v G(\mu_{ij}^\pm, \lambda_0, 0)$ or $D_v G(\mu_{i0}^\pm, \lambda_0, 0)$ with eigenfunction w_{ij} or $w_{i0}+w_{0i}$, respectively. Thus, the following Rabinowitz-type results are evident; see [15] for more details.

Remark 1.1. Fix $\lambda_0 > \kappa_{ij}$ and $\mu_{ij}^\pm = \mu_{ij}^\pm(\lambda_0)$. Then for all $i, j \in \mathbb{N}$ the points $(\mu_{ij}^+, 0)$ and $(\mu_{ij}^-, 0)$ are bifurcation points of global nontrivial continua

$$\mathcal{C}_{ij}^+(\lambda_0) = \mathrm{cl}\{\, (\mu,v) \in \mathbb{R} \times X_{ij}^{2+\alpha} \mid G(\mu,\lambda_0,v) = 0,\ \ v \not\equiv 0 \,\} \ni (\mu_{ij}^+, 0)$$

and

$$\mathcal{C}_{ij}^-(\lambda_0) = \mathrm{cl}\{\, (\mu,v) \in \mathbb{R} \times X_{ij}^{2+\alpha} \mid G(\mu,\lambda_0,v) = 0,\ \ v \not\equiv 0 \,\} \ni (\mu_{ij}^-, 0)$$

subject to the Rabinowitz alternative, i.e., the branches are either unbounded in $\mathbb{R} \times X_{ij}^{2+\alpha}$ or meet the trivial solution line at a different bifurcation point of the form $(\tilde{\mu}, 0)$.

Remark 1.2. Analogously, fix $\lambda_0 > \kappa_{i0}$ and $\mu_{i0}^\pm = \mu_{i0}^\pm(\lambda_0)$. Then for all $i \in \mathbb{N}$ the points $(\mu_{i0}^+, 0)$ and $(\mu_{i0}^-, 0)$ are bifurcation points of global nontrivial continua

$$\mathcal{C}_{i}^+(\lambda_0) = \mathrm{cl}\{\, (\mu,v) \in \mathbb{R} \times X_{i}^{2+\alpha} \mid G(\mu,\lambda_0,v) = 0,\ \ v \not\equiv 0 \,\} \ni (\mu_{i0}^+, 0)$$

and

$$\mathcal{C}_{i}^-(\lambda_0) = \mathrm{cl}\{\, (\mu,v) \in \mathbb{R} \times X_{i}^{2+\alpha} \mid G(\mu,\lambda_0,v) = 0,\ \ v \not\equiv 0 \,\} \ni (\mu_{i0}^-, 0)$$

subject to a Rabinowitz alternative.

In [13], Kielhöfer showed that for fixed $\lambda_0 > \kappa_{i0}$ the two bifurcation points $\mu_{i0}^+ = \mu_{i0}^+(\lambda_0)$ and $\mu_{i0}^- = \mu_{i0}^-(\lambda_0)$ are connected through the continua $\mathcal{C}_i^+(\lambda_0)$ and $\mathcal{C}_i^-(\lambda_0)$, i.e., $\mathcal{C}_i^+(\lambda_0)$ and $\mathcal{C}_i^-(\lambda_0)$ coincide. For different modes, however, these global continua are separated, i.e., $\mathcal{C}_i^\pm(\lambda_0) \cap \mathcal{C}_j^\pm(\lambda_0) = \emptyset$ for $i \neq j$. He also obtained a similar result for the bifurcation points $\mu_{ii}^+ = \mu_{ii}^+(\lambda_0)$ and $\mu_{ii}^- = \mu_{ii}^-(\lambda_0)$, with $\lambda_0 > \kappa_{ii}$. For this, Kielhöfer considers the spaces $X_{ii}^{k+\alpha}$ together with an additional symmetry, namely $\widehat{S} \coloneqq \{\, u \mid u(x,y) = u(y,x) \,\}$. This gives rise to nontrivial continua $\widehat{\mathcal{C}}_{ii}^\pm(\lambda_0)$ which connect the two bifurcation points and are therefore equal. Again, they are separated for different modes, i.e., $\widehat{\mathcal{C}}_{ii}^\pm(\lambda_0) \cap \widehat{\mathcal{C}}_{jj}^\pm(\lambda_0) = \emptyset$ for $i \neq j$. Using our notation, Kielhöfer obtained the following result in [13].

Theorem 1.3. *For fixed $i \in \mathbb{N}$ and $\lambda_0 > \kappa_{i0}$, we have $(\mu_{i0}^-, 0) \in \mathcal{C}_i^+(\lambda_0)$ and $(\mu_{i0}^+, 0) \in \mathcal{C}_i^-(\lambda_0)$, and therefore $\mathcal{C}_i^+(\lambda_0) = \mathcal{C}_i^-(\lambda_0)$. For $\lambda_0 > \kappa_{ii}$, one obtains both $(\mu_{ii}^-, 0) \in \widehat{\mathcal{C}}_{ii}^+(\lambda_0)$ and $(\mu_{ii}^+, 0) \in \widehat{\mathcal{C}}_{ii}^-(\lambda_0)$, i.e., $\widehat{\mathcal{C}}_{ii}^+(\lambda_0) = \widehat{\mathcal{C}}_{ii}^-(\lambda_0)$. Furthermore, these global continua are separated from each other.*

Another result along these lines can be found in [15].

Theorem 1.4. *Choose $i,j \in \mathbb{N}$ and $\lambda_0 > \kappa_{ij}$ fixed. Then the continuum $\mathcal{C}^+_{ij}(\lambda_0)$ of nontrivial solutions of (5) corresponding to w_{ij}, which bifurcates from the trivial solution at the point $(\mu^+_{ij}(\lambda_0),0)$ is equal to the continuum $\mathcal{C}^-_{ij}(\lambda_0)$, which bifurcates from the point $(\mu^-_{ij}(\lambda_0),0)$. It either is separated from the continua $\mathcal{C}^\pm_{\tilde{i}\tilde{j}}(\lambda_0)$ for all $(\tilde{i},\tilde{j}) \in \mathbb{N}^2 \setminus \{(i,j)\}$ with $i|\tilde{i}$ and $j|\tilde{j}$, or, there is some other trivial solution $(\tilde{m},0) \in \mathcal{C}^+_{ij}(\lambda_0) = \mathcal{C}^-_{ij}(\lambda_0)$ with $\tilde{m} \neq m^\pm_{ij}(\lambda_0)$. In the latter case, the continuum $\mathcal{C}^+_{ij}(\lambda_0)$ contains a loop, i.e., the two parts of $\mathcal{C}^+_{ij}(\lambda_0)$ bifurcating at $(\mu^+_{ij}(\lambda_0),0)$ in different directions are connected through a path in $\mathcal{C}^+_{ij}(\lambda_0)$ that does not meet the trivial solution line.*

In addition, [15] contains a result for fixed $\mu_0 = 0$. For this, one views equation (5) as a bifurcation problem in λ for fixed mass $\mu = \mu_0$, where $|\mu_0| < 1/\sqrt{3}$. Then 0 is a simple eigenvalue of $D_vG(\mu_0,\lambda_{ij},0)$ or $D_vG(\mu_0,\lambda_{i0},0)$, and a similar argument furnishes the following result.

Remark 1.5. Fix $\mu_0 \in (-1/\sqrt{3},1/\sqrt{3})$, $\lambda_{ij} = \lambda_{ij}(\mu_0)$, and $\lambda_{i0} = \lambda_{i0}(\mu_0)$. Then for all $i,j \in \mathbb{N}$ the points $(\lambda_{ij},0)$ or $(\lambda_{i0},0)$ are bifurcation points of global nontrivial continua

$$\mathcal{C}_{ij}(\mu_0) = \mathrm{cl}\{\,(\lambda,v) \in \mathbb{R} \times X^{2+\alpha}_{ij} \mid G(\mu_0,\lambda,v) = 0,\ \ v \not\equiv 0\,\} \ni (\lambda_{ij},0)$$

and

$$\mathcal{C}_{i}(\mu_0) = \mathrm{cl}\{\,(\lambda,v) \in \mathbb{R} \times X^{2+\alpha}_{i} \mid G(\mu_0,\lambda,v) = 0,\ \ v \not\equiv 0,\} \ni (\lambda_{i0},0)$$

subject to a Rabinowitz alternative.

In particular for $\mu_0 = 0$, we add an additional symmetry to our spaces X_i and X_{ij}, namely

$$s_i \coloneqq \left\{ w \,\middle|\, w(x,y) = -w\left(\frac{1}{i} - y, \frac{1}{i} - x\right) \right\}$$

and

$$s_{ij} \coloneqq \left\{ w \,\middle|\, w(x,y) = -w\left(\frac{1}{i} - x, y\right) \right\},$$

respectively, to define the spaces

$$\widetilde{X}^{k+\alpha}_i \coloneqq X^{k+\alpha}_i \cap s_i \quad \text{and} \quad \widetilde{X}^{k+\alpha}_{ij} \coloneqq X^{k+\alpha}_{ij} \cap s_{ij}.$$

Then the smooth mapping $\widetilde{G}$ defined by

$$\widetilde{G} : \mathbb{R} \times \widetilde{X}^{2+\alpha}_{ij} \longrightarrow \widetilde{X}^{\alpha}_{ij} \qquad \text{or} \qquad \widetilde{G} : \mathbb{R} \times \widetilde{X}^{2+\alpha}_{i} \longrightarrow \widetilde{X}^{\alpha}_{i},$$
$$\widetilde{G}(\lambda,v) = \Delta v + \lambda f(v),$$

is well defined. For $\mu_0 = 0$, one obtains the possible bifurcation points $\lambda_{ij}(0) = \kappa_{ij}$ and $\lambda_{i0}(0) = \kappa_{i0}$, and arguing as above one can establish the following result.

Remark 1.6. For all $i, j \in \mathbb{N}$ the points $(\kappa_{ij}, 0)$ and $(\kappa_{i0}, 0)$ are bifurcation points of global nontrivial continua

$$\widetilde{\mathcal{C}}_{ij} = \mathrm{cl}\{\, (\lambda, v) \in \mathbb{R} \times \widetilde{X}_{ij}^{2+\alpha} \mid \widetilde{G}(\lambda, v) = 0,\ \ v \not\equiv 0 \,\} \ni (\kappa_{ij}, 0)$$

and

$$\widetilde{\mathcal{C}}_{i} = \mathrm{cl}\{\, (\lambda, v) \in \mathbb{R} \times \widetilde{X}_{i}^{2+\alpha} \mid \widetilde{G}(\lambda, v) = 0, v \not\equiv 0 \,\} \ni (\kappa_{i0}, 0)$$

subject to a Rabinowitz alternative.

Finally, [15] contains the following result.

Theorem 1.7. *Choose $i, j \in \mathbb{N}$ and $\mu_0 = 0$. The continuum $\widetilde{\mathcal{C}}_{ij}$ (respectively $\widetilde{\mathcal{C}}_{i}$) of nontrivial solutions of (5) corresponding to w_{ij} (respectively $w_{i0} + w_{0i}$), which bifurcates from the trivial solution at the point $\lambda_{ij}(0)$ (respectively $\lambda_{i0}(0)$), consists of two differentiable curves which are parameterized with respect to λ. Furthermore the continuum does not return to the trivial solution line.*

1.2. Transformations and symmetry of equilibria

Due to the symmetries of our base domain, many of the equilibrium solutions of (5) can be transformed into each other by suitable symmetry operations. In this brief section we introduce the terminology which will be used in this context.

Let u_0 be a solution of (5) subject to (2) on the unit square $\Omega = (0,1)^2$, and with parameter values (μ_0, λ_0, c_0). Then we can extend u_0 by even reflections to the whole of $\mathbb{R}^2$, and the resulting function is smooth. If we now define the function

$$v^k(x, y) \coloneqq u_0(kx, ky), \quad \text{for} \quad x, y \in (0,1)^2,$$

and $k \in \mathbb{N}$, then v^k is also a solution of (5) and (2), yet this time with parameter values $(\mu_0, k^2\lambda_0, c_0)$. In addition, if we let $\mathbf{m} : u \mapsto -u$ denote the multiplication of a function by -1, then $\mathbf{m}u_0$ is also a solution of (5) and (2), but this time for the parameter values $\mu = -\mu_0$, $\lambda = \lambda_0$, and $c = -c_0$. Finally, let $\mathcal{R}$ denote the counter-clockwise rotation about 90 degrees around the center $(1/2, 1/2)$ of Ω, and let $\mathcal{T}$ be the reflection at the line $x = 1/2$. More precisely, for $u : [0,1]^2 \to \mathbb{R}$, we have

$$(\mathcal{R}u)(x, y) = u(y, 1 - x) \quad \text{and} \quad (\mathcal{T}u)(x, y) = u(1 - x, y).$$

The symmetry group of the unit square, the dihedral group D_4, consists of four rotations around $(1/2, 1/2)$ by multiples of 90 degrees, as well as four reflections, namely the reflections at the line $x = 1/2$, the line $y = 1/2$, as well as the diagonal and the anti-diagonal of the square. Note that the actions $\mathcal{R}$ and $\mathcal{T}$ generate all symmetries of the square. For example, the reflection at the diagonal of the square is given by $\mathcal{T}\mathcal{R}$, since

$$(\mathcal{T}\mathcal{R}u)(x, y) = (\mathcal{R}u)(1 - x, y) = u(y, x).$$

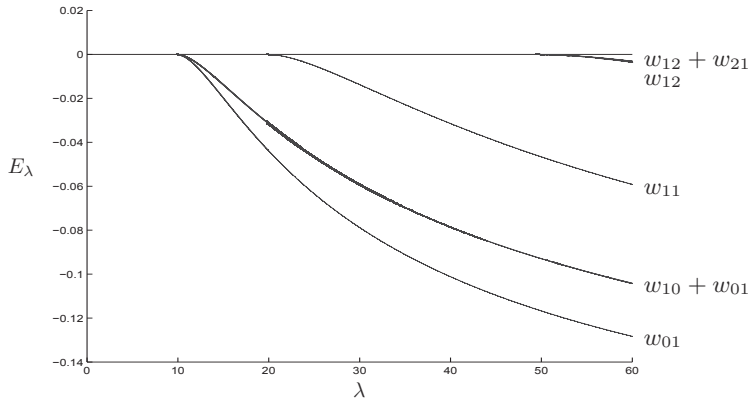


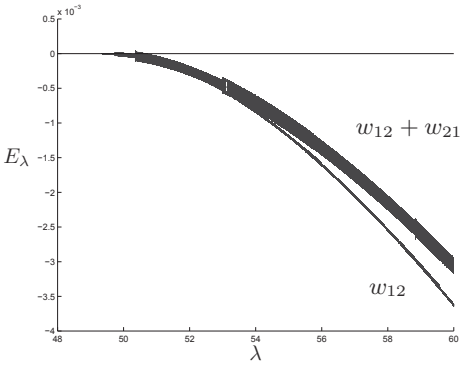
Figure 2 – Bifurcation diagram for solutions of (5) subject to (2) for $\mu_0 = 0$

Similarly, the reflection at the antidiagonal is given by $\mathcal{RT}$, which shows that $\mathcal{R}$ and $\mathcal{T}$ do not commute. Now choose $\gamma \in D_4$ arbitrary. If u_0 is a solution of (5) subject to the mass constraint (2) on $\Omega = (0,1)^2$, and with parameters (μ_0, λ_0, c_0), then γu_0 is also a solution with the same parameter values.

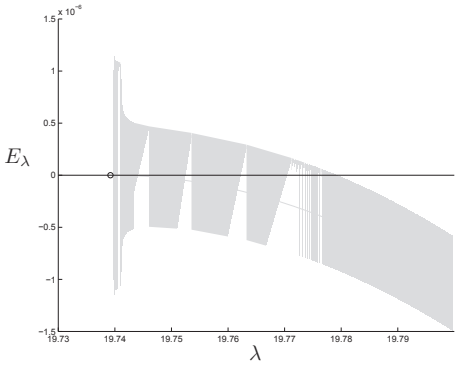
1.3. Bifurcation diagram for $\mu_0 = 0$

After these preliminary discussions, we begin with presenting the main results of this paper, and the current section focuses on an overview in terms of the bifurcation diagram for vanishing total mass.

Due to (12) there are four possible bifurcation points for $\mu_0 = 0$ and $\lambda \leq 60$, namely $\lambda_{01} = \lambda_{10} = \kappa_{10} = \pi^2$, $\lambda_{11} = \kappa_{11} = 2\pi^2$, $\lambda_{02} = \lambda_{20} = \kappa_{20} = 4\pi^2$, and $\lambda_{12} = \lambda_{21} = \kappa_{12} = 5\pi^2$. Figure 2 contains the bifurcation diagram for $\mu_0 = 0$. In this figure, the straight line at the top represents the trivial solution line $u \equiv 0$. The solution branches are shown as functions of the parameter λ, the vertical axis indicates the energy E_λ of the equilibria defined in (4). Each curve corresponds to a particular eigenfunction of the linearization of the Cahn-Hilliard equation at the respective bifurcation point. For example, for the first bifurcation point λ_{10} two branches are shown, one for the mode w_{01}, and one for the superposition $w_{10} + w_{01}$. Yet, these branches are only the representatives within a symmetry class. Thus, while figure 2 shows only two bifurcating branches at λ_{10}, there are in fact four such branches, two each for the two depicted ones. The remaining branches correspond to the modes w_{10} and $w_{10} - w_{01}$, and the collection of all four branches together gives rise to eight equilibria for each $\lambda > \lambda_{10}$. Also the remaining branches in figure 2 show only the relevant modes in each symmetry class. Related branches are suppressed, i.e., we omit modes w_{20}, w_{02}, $w_{20} \pm w_{02}$, w_{21}, and $w_{12} - w_{21}$.



(a) Close-up of figure 2, showing the modes w_{12} and $w_{12} + w_{21}$



(b) Close-up of the bifurcation point $\lambda_{11} = 2\pi^2$

Figure 3

A closer look at the branches in figure 2 reveals that the shown curves exhibit a certain nonuniform thickness. This is due to the fact that for each value of λ the branch contains an energy interval, and we can actually prove the existence of equilibria within these intervals. More details can be found in the following sections.

In figure 2 it is hard to distinguish between the branches that correspond to the modes w_{12} and $w_{12} + w_{21}$. A more detailed representation is contained in figure 3a, which shows for example that the mode w_{12} solutions have lower energy. By using a sufficient condition for uniqueness which will be presented in Theorem 3.7 below, we can establish the existence of a unique solution in each of the computed regions for every fixed λ. See also Theorem 1.15. Hence, the two branches shown in figure 3a are in fact separated, and only due to the projection onto the (λ, E_λ)-plane do they seem to overlap. One deficiency of our method is that it does not work close to the bifurcation points. Figure 3b shows a close-up of a neighborhood of the bifurcation point $\lambda_{11} = 2\pi^2$ in figure 2. The black line corresponds to the trivial solution, the point $(\lambda_{11}, 0)$ is marked with a circle. Notice that the area which contains the branch of nontrivial solutions starts only at $\lambda = 2\pi^2 + 0.0005$. Similarly, for the remaining branches in figure 2, the actually computed nontrivial branches start at $\lambda = \pi^2 + 0.0025$ and $\lambda = 5\pi^2 + 0.0025$, respectively. We will see below that our uniqueness result applies to the complete w_{11}-branch in figure 2, and therefore the overlap of the trivial solution and the area for the nontrivial solution branch in figure 3b is again a consequence of the projection on the (λ, E_λ)-plane.

Finally, the $w_{10} + w_{01}$-branch in figure 2 undergoes a secondary bifurcation of pitchfork-type at $\lambda \approx 51.8485$. Unfortunately, the energy values of the two branches are extremely close, which makes it impossible to resolve them in figure 2. We will address this issue in more detail later. See section 1.5, in particular figures 13a

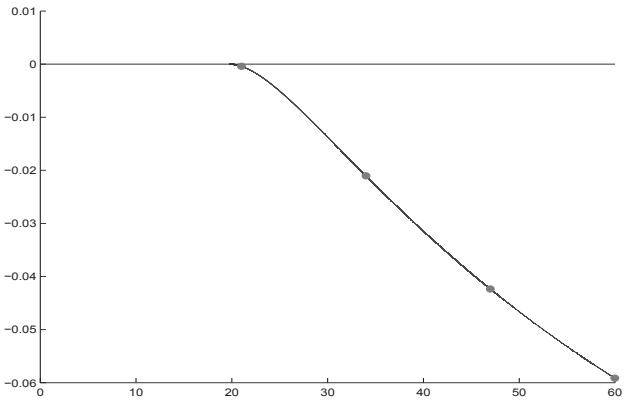


Figure 4 – Energy for mode w_{11} solutions of (5) subject to (2) for $\mu_0 = 0$

and 13b.

1.4. The mode w_{11}

In this and the following two sections we focus on each of the three bifurcation points in figure 2 and discuss the bifurcating branches. Unlike in the previous section, we will also consider the effects of mass variation. We begin in this section with the bifurcation point λ_{11}. For total mass $\mu_0 = 0$, the w_{11}-branch emanating at the bifurcation is shown in figure 4. In fact, the methods of this paper allow us to derive the following result.

Theorem 1.8 (Mode w_{11} for $\mu_0 = 0$)**.** *There exists a branch of solutions of* (5) *subject to the constraint* (2) *for* $\mu_0 = 0$ *and* $\lambda \in [2\pi^2 + 0.0005, 60]$ *as shown in Figure 4. For fixed* λ*, these solutions are unique in a small neighborhood. The errors in the maximum norm along the path, denoted by* δ_{C^0}*, are less than* 0.00314*. The errors in the energy, denoted by* δ_{E_λ}*, are less than* 0.00016*. More precisely, this error is the length of the boxes in the energy direction of the branch shown in figure 4.*

Theorem 1.8 was obtained using the rigorous computational techniques which were outlined in the introduction, and which will be described in more detail in the sections to come. In order to show the geometry of the solutions along the branch guaranteed by the theorem, figure 5 contains contour plots of sample numerically determined solutions along the branch. The position of each of these solutions is marked by a dot in figure 4. Above each contour plot, the corresponding λ-value is shown, together with the maximal distance δ_{C^0}, within which there actually exists a true equilibrium solution. The contour plot is given with a grey scale intensity, varying from white, which corresponds to the value $+1$, to black, which represents the value -1. This

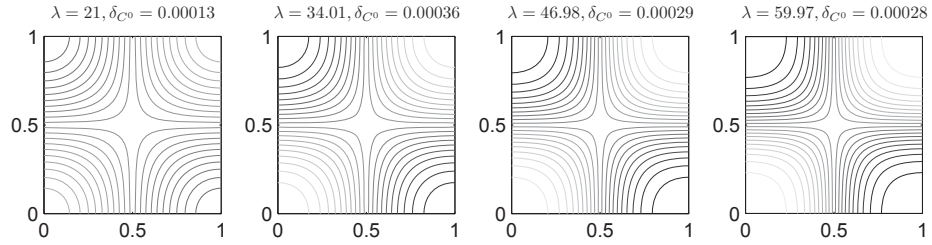
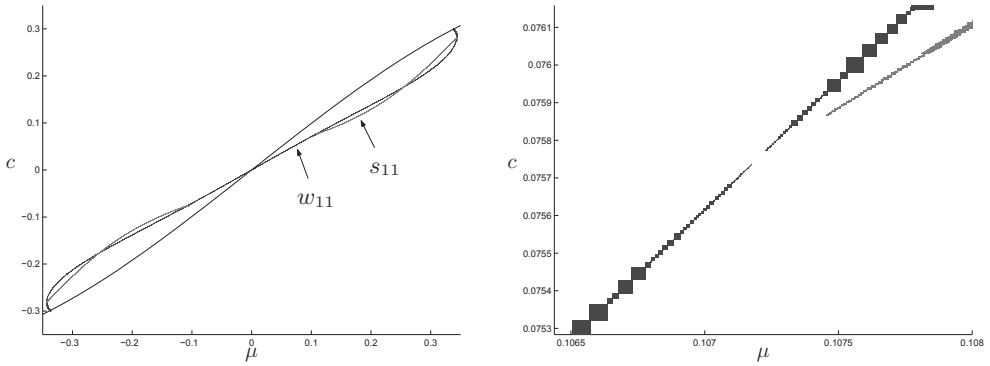


Figure 5 – Snapshots for mode w_{11} solutions of (5) for $\mu_0 = 0$



(a) Bifurcation diagram for mode w_{11} solutions of (5) for fixed $\lambda_0 = 30$

(b) Close-up of figure 6a

Figure 6

convention will be used for all the contour plots in this paper. The respective colored graphics might be downloaded from `http://chomp.rutgers.edu`.

As we mentioned in the previous section, our method cannot resolve the structure at the bifurcation point. Hence, we are not able to prove that the branch of Theorem 1.8 is a part of the branch $\widetilde{\mathcal{C}}_{11}$ introduced in section 1.1. However, the fact that this branch starts near the bifurcation point and that its solutions exhibit the correct geometry supports this conjecture. Of course, there are two curves bifurcating from the trivial solution $u \equiv 0$ at $\lambda = 2\pi^2$. In addition to the one shown in figures 4 and 5, there is also the solution curve which can be obtained by an application of the action $\mathbf{m}$, i.e., by multiplication with -1.

We now turn our attention to the effects of mass variation. Figure 6a shows the solution curve of the w_{11} mode for fixed $\lambda_0 = 30$ in the (μ, c)-plane, see (2) and (6). The black line corresponds to the trivial solution $u \equiv 0$, $c = \mu - \mu^3$. The blue and the

red curves are in fact areas which contain a solution branch. The intersection point of the blue area and the trivial solution line in the center of the figure is artificial and only due to the projection of the branch onto the (μ, c)-plane. This is also true for the intersections of the red and blue curves. By using our method, we can establish the following result.

Theorem 1.9 (Mode w_{11} at $\lambda_0 = 30$)**.**

(i) *There exists a solution branch for* $\lambda_0 = 30$ *from* $(\mu, c) = (-0.107176, -0.075736)$ *to* $(\mu, c) = (0.107176, 0.075736)$. *These solutions are unique in a small neighborhood (for* μ *or* c *fixed) between* $(\mu, c) = (-0.104175, -0.073673)$ *and* $(\mu, c) = (0.104175, 0.073673)$. *The errors* δ_{C^0} *in the maximum norm along the branch are less than* 0.00608. *The error in* μ, *denoted by* δ_μ, *is less than* 0.00025, *the error in* c, *denoted by* δ_c, *is less than* 0.00038. *These errors are the length of the boxes in the* μ *and* c *direction. Furthermore, from* $(\mu, c) = (0.107228, 0.075772)$ *to* $(\mu, c) = (0.343395, 0.280698)$, *and similarly from* $(\mu, c) = (-0.107228, -0.075772)$ *to* $(\mu, c) = (-0.343395, -0.280698)$, *there is a branch for* $\lambda_0 = 30$ *with* $\delta_{C^0} < 0.00052$, $\delta_\mu < 0.00113$, *and* $\delta_c < 0.00046$. *The solutions are unique from* $(\mu, c) = (0.114287, 0.080610)$ *to* $(\mu, c) = (0.342174, 0.272221)$, *as well as from* $(\mu, c) = (-0.114287, -0.080610)$ *to* $(\mu, c) = (-0.342174, -0.272221)$. *Finally, there exists a branch of solutions from* $(\mu, c) = (0.343396, 0.280699)$ *to* $(\mu, c) = (0.337654, 0.299154)$, *and similarly from* $(\mu, c) = (-0.343396, -0.280699)$ *to* $(\mu, c) = (-0.337654, -0.299154)$, *with* $\delta_{C^0} < 0.00129$, $\delta_\mu < 0.00005$, *and* $\delta_c < 0.00002$. *This branch consists of unique solutions between* $(\mu, c) = (0.343397, 0.280721)$ *and* $(\mu, c) = (0.338226, 0.298463)$, *and similarly between* $(\mu, c) = (-0.343397, -0.280721)$ *and* $(\mu, c) = (-0.338226, -0.298463)$. *These five branches are denoted by* w_{11} *(see figure 6a).*

(ii) *There are solution branches for fixed* $\lambda_0 = 30$ *from* $(\mu, c) = (0.107459, 0.075868)$ *to* $(\mu, c) = (0.343392, 0.280694)$, *and similarly from* $(\mu, c) = (-0.107459, -0.075868)$ *to* $(\mu, c) = (-0.343392, -0.280694)$. *They are denoted by* s_{11} *(see figure 6a). Solutions are unique from* $(\mu, c) = (0.110591, 0.077269)$ *to* $(\mu, c) = (0.341452, 0.278271)$, *and we have* $\delta_{C^0} < 0.0043$, $\delta_\mu < 0.00042$ *and* $\delta_c < 0.00042$.

We summarize the essential information pertaining to the branches of Theorem 1.9 for positive μ in a table. Note that in table 1 the data is given with only four decimal places. The errors δ_{C^0}, δ_μ, and δ_c are rounded up, the starting and end points of the curves are rounded towards the directions of the paths, i.e., if a path starts at the point $(\mu, c) = (0.107228, 0.075772)$ and increases in both μ and c directions, then both values are rounded up to $(\mu, c) = (0.1073, 0.0758)$. We will follow this convention in all future tables as well.

Table 1 – Table of branches from (μ_1, c_1) to (μ_2, c_2) at $\lambda_0 = 30$, for $\mu \geq 0$

mode	μ_1	c_1	μ_2	c_2	δ_{C^0}	δ_μ	δ_c	uniqueness
w_{11}	0	0	0.1071	0.0757	0.0061	0.0003	0.0004	partly
w_{11}	0.1073	0.0758	0.3433	0.2807	0.0006	0.0012	0.0005	partly
w_{11}	0.3434	0.2808	0.3377	0.2991	0.0013	0.0001	0.0001	partly
s_{11}	0.1075	0.0759	0.3433	0.2806	0.0043	0.0005	0.0005	partly

Table 2 – Uniqueness of the solutions on the branches in table 1 can be established between the parameter values $(\tilde{\mu}_1, \tilde{c}_1)$ and $(\tilde{\mu}_2, \tilde{c}_2)$

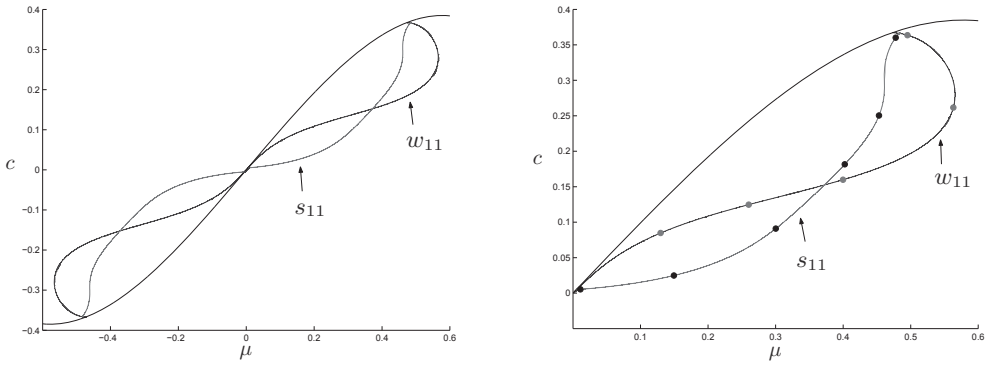
mode	$\tilde{\mu}_1$	$\tilde{c}_1$	$\tilde{\mu}_2$	$\tilde{c}_2$
w_{11}	0	0	0.1041	0.0736
w_{11}	0.1143	0.0807	0.3421	0.2722
w_{11}	0.3434	0.2808	0.3383	0.2964
s_{11}	0.1106	0.0773	0.3414	0.2782

As expected, our method does not apply near the bifurcation points

$$\mu_{11}^{\pm} = \pm\sqrt{\frac{1}{3} - \frac{2\pi^2}{3\lambda_0}}, \qquad \text{with} \quad \lambda_0 = 30.$$

However, the formulation of Theorem 1.9 shows that the method also fails near the four points $(\mu, c) = \pm(0.3433955, 0.2806985)$ and $(\mu, c) = \pm(0.1072, 0.07575)$. Figure 6b contains a close-up of one of these locations. It can clearly be seen that there is a gap in the thick path w_{11} — and that the end of the thin branch s_{11} is very close to this gap. As we will see in section 3, the starting point for our method has to be a hyperbolic equilibrium of the finite-dimensional system (10). Thus, the linearization of (10) at this equilibrium cannot have a vanishing eigenvalue, and even eigenvalues close to 0 will lead to the failure of our approach. This situation certainly arises if the linearization of the full infinite-dimensional system (55) at an equilibrium has a nontrivial kernel — which is the case at any secondary bifurcation point. We believe that this is exactly what happens in figure 6b, see also figure 6a. It seems plausible that the thin branch s_{11} in these figures bifurcates from the thick path w_{11}. Scenarios such as this will be encountered frequently in the following, and we will henceforth simply refer to this situation as secondary bifurcation. In fact there are two branches bifurcating from the w_{11} curve. They are linked by symmetry and therefore have the same (μ, c)-values, see also figure 10 below.

Now consider the fixed parameter value $\lambda_0 = 60$. Figures 7a and 7b contain areas that enclose actual solutions on the primary bifurcating branch w_{11} and on the secondary s_{11}, respectively. As before, the curve $c = \mu - \mu^3$ represents the trivial solution, and the intersection between this and the w_{11} curve in figure 7a at $(\mu, c) = (0, 0)$ is only due to the projection onto the (μ, c)-plane. Figure 7b contains a



(a) Bifurcation diagram for mode w_{11} solutions of (5) for fixed $\lambda_0 = 60$

(b) Close-up of figure 7a

Figure 7

more detailed view of the upper half of the branches in figure 7a, i.e., of the branches with nonnegative μ and c. Finally, in figure 8 we present snapshots of functions which are contained in the regions describing the branches, and are therefore prototypical of the observed geometries. The locations of these functions are indicated by dots in figure 7b. Notice that all of these snapshots are located on the upper branch. The corresponding functions on the lower branch can be obtained by applying the action $\mathbf{m}\mathcal{R}$.

Recall that for the mode w_{11} we can apply Kielhöfer's result for continua with fixed parameter $\lambda = \lambda_0$. Therefore, the continuum $\widehat{\mathcal{C}}^+_{11}(\lambda_0) = \widehat{\mathcal{C}}^-_{11}(\lambda_0)$ connects the two bifurcation points μ^+_{11} and μ^-_{11}, see also Theorem 1.3. A local analysis near the bifurcation point μ^+_{11} shows that the continuum consist of two paths which can be parameterized as

$$v(s) = sw_{11} + o(s) \quad \text{and} \quad \mu(s) = \mu^+_{11} + o(s) \quad \text{for} \quad s \in (-\delta, \delta), \tag{14}$$

and $\delta > 0$ sufficiently small. Now let $\mathcal{K}_{w_{11}} = \mathcal{K}_{w_{11}}(\lambda_0)$ denote the branch which contains $v(s)$ for $s \in (0, \delta)$, and let $\mathcal{K}_{-w_{11}} = \mathcal{K}_{-w_{11}}(\lambda_0)$ be the one with $v(s)$ for $s \in (-\delta, 0)$. Based on the geometries shown in figure 8, we suggest that the branch described in table 3 is part of $\mathcal{K}_{w_{11}} \subset \widehat{\mathcal{C}}^+_{11}(\lambda_0)$ for $\lambda_0 = 60$. Remember that $\widehat{\mathcal{C}}^+_{11}(\lambda_0) \subset X^{2+\alpha}_{11} \cap \widehat{S}$, where $u \in X^{2+\alpha}_{11}$ implies $u = \mathcal{R}^2 u$, and from $u \in \widehat{S}$ we obtain $u = \mathcal{T}\mathcal{R}u$. Together, we have $u = \mathcal{R}^2 u = \mathcal{R}^2\mathcal{T}\mathcal{R}u = \mathcal{R}\mathcal{T}u$, where we used the identity $\mathcal{R}\mathcal{T}\mathcal{R} = \mathcal{T}$. In other words, u is invariant under the reflection at the diagonals. By applying the action $\mathcal{R}$, one obtains another branch of solutions with identical values of μ and c. This is the other branch $\mathcal{K}_{-w_{11}}$ of $\widehat{\mathcal{C}}^+_{11}(\lambda_0)$ which bifurcates at μ^+_{11}. We summarize the results of our computations relating to figures 7a and 7b in the following theorem.

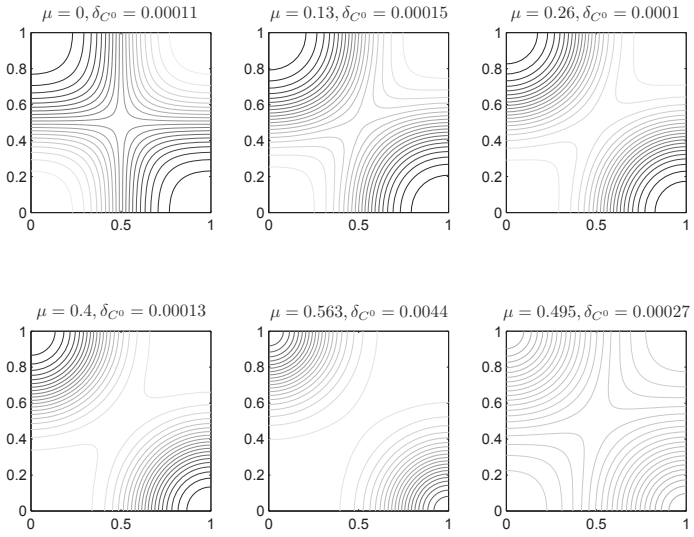


Figure 8 – Snapshots for mode w_{11} solutions of (5) for $\lambda_0 = 60$

Table 3 – Table of branches from (μ_1, c_1) to (μ_2, c_2) at $\lambda_0 = 60$ for $\mu > 0$

mode	μ_1	c_1	μ_2	c_2	δ_{C^0}	δ_μ	δ_c	uniqueness
w_{11}	0	0	0.0052	0.0048	0.0003	0.0001	0.0001	no
w_{11}	0.0054	0.0049	0.4835	0.3661	0.0045	0.0007	0.0008	partly
w_{11}	0.4834	0.3662	0.4736	0.3671	0.0095	0.0012	0.0001	no
s_{11}	0.0054	0.0049	0.4834	0.3661	0.0091	0.0008	0.0006	partly

Table 4 – Uniqueness of the branches in table 3 holds from $(\tilde{\mu}_1, \tilde{c}_1)$ to $(\tilde{\mu}_2, \tilde{c}_2)$

mode	$\tilde{\mu}_1$	$\tilde{c}_1$	$\tilde{\mu}_2$	$\tilde{c}_2$
w_{11}	0.0292	0.0261	0.491	0.3647
s_{11}	0.4652	0.3355	0.4795	0.3622

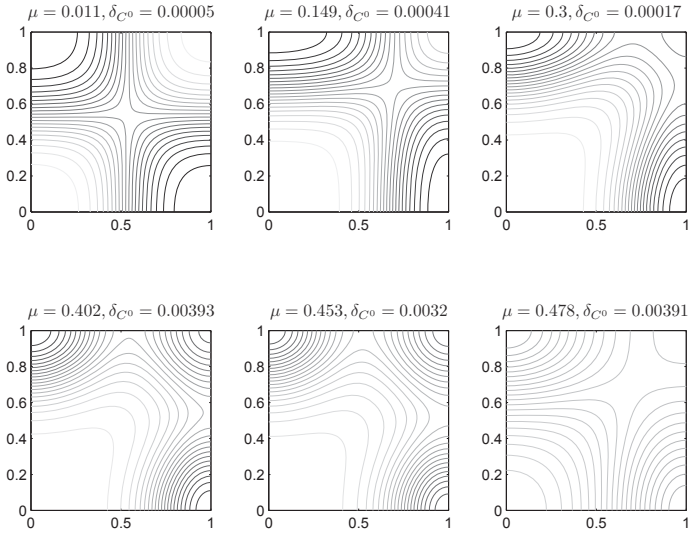


Figure 9 – Snapshots for solutions on the branch s_{11}

Theorem 1.10 (Mode w_{11} for $\lambda_0 = 60$)**.** *There exist branches of solutions of* (5) *subject to the constraint* (2) *for* $\lambda_0 = 60$ *as shown in figure 7b. Detailed information on the endpoints of these branches in the* (μ, c)*-plane, as well as uniqueness assertions, can be found in tables 3 and 4.*

The branch s_{11} described in table 3 is a secondary bifurcation of the part $\mathcal{K}_{w_{11}}$ of the continuum $\widehat{\mathcal{C}}^+_{11}(\lambda_0)$ in the sense mentioned earlier. Figure 9 contains snapshots of sample solutions along the path. Their locations are indicated by black dots in figure 7b. Elements on the red branch in the lower left part of Figure 7a can be obtained by the action $\mathbf{m}\mathcal{R}$. Notice that according to the geometry of the functions in figure 9, s_{11} returns to the branch from which it bifurcates, thereby breaking the symmetry $\mathcal{RT}$. Hence, by applying $\mathcal{RT}$ we obtain another branch with identical μ and c values which bifurcates from and returns to $\mathcal{K}_{w_{11}} \subset \widehat{\mathcal{C}}^+_{11}(\lambda_0)$ at the same points. In figure 10 we give a sketch of this secondary bifurcation. It indicates that the branch $\mathcal{K}_{w_{11}}$ for fixed $\lambda_0 = 60$ is connected with itself through the branches s_{11} and $\mathcal{RT}(s_{11})$. Applying the action $\mathcal{R}$ furnishes a connection of $\mathcal{K}_{-w_{11}}$ with itself through the branches $\mathcal{R}(s_{11})$ and $\mathcal{R}^2\mathcal{T}(s_{11})$.

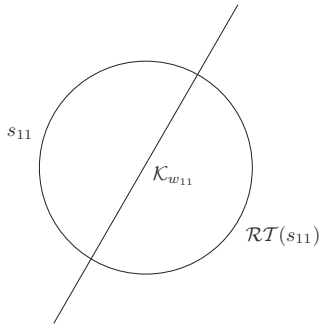


Figure 10 – Sketch of the secondary bifurcation related to s_{11}

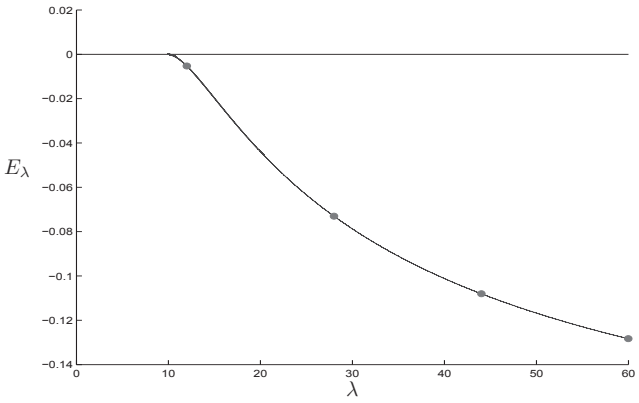


Figure 11 – Energy for mode w_{01} solutions with $\mu_0 = 0$

1.5. The modes w_{01} and $w_{10} + w_{01}$

In this section we consider the solution branches associated with modes w_{01} and $w_{10} + w_{01}$. We begin with the mode w_{01} for $\mu_0 = 0$.

Theorem 1.11 (Mode w_{01} for $\mu_0 = 0$)**.** *For total mass $\mu_0 = 0$ there exists a branch of unique solutions of* (5) *subject to* (2) *for $\lambda \in [\pi^2 + 0.0025, 60]$ as shown in figure 11, with $\delta_{C^0} < 0.00032$ and $\delta_{E_\lambda} < 0.00029$.*

The geometry of sample solutions on this branch is shown in figure 12, the locations of these solutions are indicated by dots in figure 11. Notice that due to the shape of the solutions, this branch is presumably the ODE-branch. Through the action $\mathbf{m}$ one obtains the other branch of w_{01} solutions of (5) subject to (2) for $\mu_0 = 0$. The w_{10} solutions can be generated by applying $\mathcal{TR}$. Next, we turn our attention to mode $w_{10} + w_{01}$ solutions for vanishing total mass $\mu_0 = 0$.

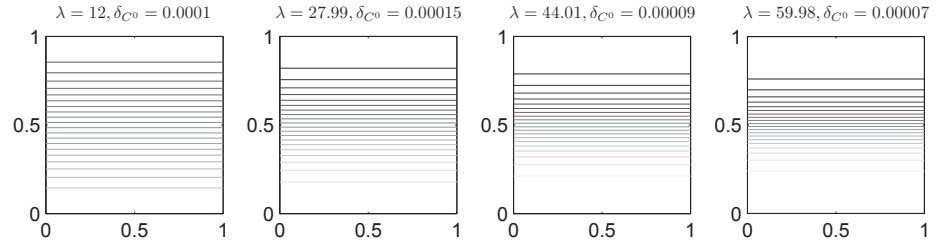
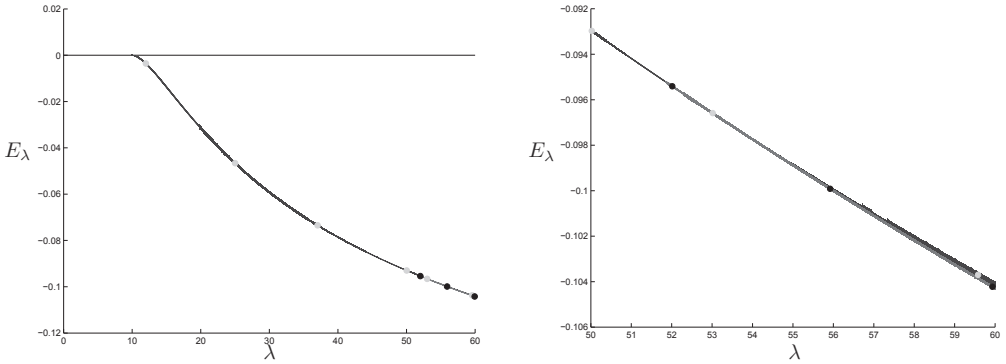


Figure 12 – Snapshots for mode w_{01} solutions with $\mu_0 = 0$



(a) Energy for mode $w_{10} + w_{01}$ solutions with $\mu_0 = 0$

(b) Close-up of figure 13a

Figure 13

Theorem 1.12 (Mode $w_{10} + w_{01}$ for $\mu_0 = 0$)**.** *We consider problem* (5) *subject to the mass constraint* (2) *and with nonlinearity* (3)*.*

(i) *There exists a branch of solutions for* $\lambda \in [\pi^2 + 0.0025, 51.84797]$ *as shown in figure 13a, with* $\delta_{C^0} < 0.00285$ *and* $\delta_{E_\lambda} < 0.0011$*. It is unique for* $\lambda \in [\pi^2 + 0.0025, 51.26511]$*. Furthermore there exists a branch of solutions for* $\lambda \in [51.84914, 60]$ *as shown in figure 13a, with* $\delta_{C^0} < 0.00055$*,* $\delta_{E_\lambda} < 0.00017$ *and uniqueness for* $\lambda \in [52.84354, 60]$*.*

(ii) *There exists a branch of solutions for* $\lambda \in [51.84828, 60]$ *as shown in figure 13a, with error bounds* $\delta_{C^0} < 0.00497$ *and* $\delta_{E_\lambda} < 0.00022$*. It is unique for* $\lambda \in [54.19267, 60]$*.*

The branches in (i) *and* (ii) *to the right of the bifurcation point at* $\lambda \approx 51.85$ *are distinct.*

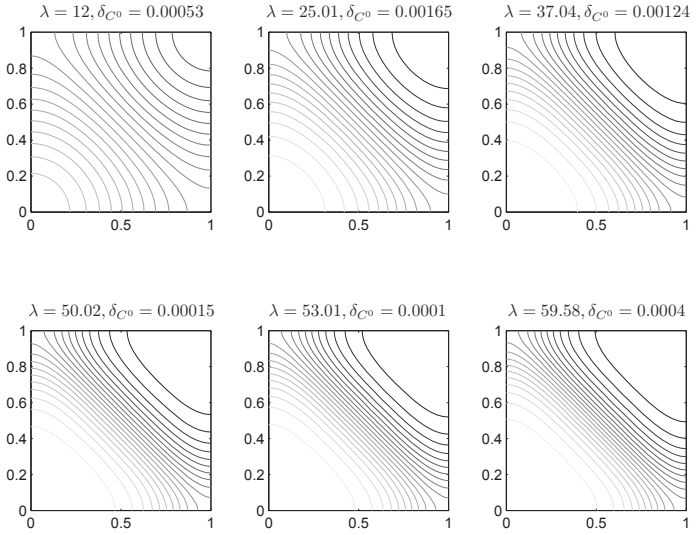


Figure 14 – Snapshots for mode $w_{10} + w_{01}$ solutions with $\mu_0 = 0$ on the primary branch

It was already mentioned in section 1.3 that the primary branch undergoes a pitchfork bifurcation at $\lambda \approx 51.8485$. The energy values of the solutions on the secondary branches are very close to the ones on the primary branch, and therefore the branches could not be easily distinguished in Figure 2. In Figure 13a besides the primary branch also the bifurcating path described in Theorem 1.12 (ii) is given. A more detailed view of the branches for $\lambda \in [50, 60]$ can be found in figure 13b. We would like to point out that at least for $\lambda = 60$, the energy of the secondary branch is lower than the energy of the primary branch. As before, we cannot prove the existence of the bifurcation rigorously. We observe that our method fails for $\lambda \approx 51.8485$, since the linearization has an eigenvalue near zero. For $\lambda \in [\pi^2 + 0.0025, 51.84797]$ we find one branch, for $\lambda \in [51.8492, 60]$ there are two branches — which is indicative of the bifurcation.

Figure 14 shows level sets of sample functions which are close to actual equilibria on the branch in Theorem 1.12 (i), i.e., this branch appears to be part of $\widetilde{\mathcal{C}}_1$ which was introduced in section 1.1; see Remark 1.6. According to Theorem 1.7, $\widetilde{\mathcal{C}}_1$ consists of two curves bifurcating from the trivial solution line $u \equiv 0$ at $\lambda = \pi^2$, and one can generate the second branch by applying $\mathbf{m}$ to the computed one. There is no secondary bifurcation from $\widetilde{\mathcal{C}}_1$ in the corresponding fixed-point space; see Remark 1.6. Hence, the new branch must break one of the symmetries, in fact, it breaks the symmetry $w = \mathbf{m}\mathcal{R}\mathcal{T}w$. This can clearly be seen in figure 15, where snapshots of functions in the regions of figures 13a and 13b are depicted. The locations of these functions are

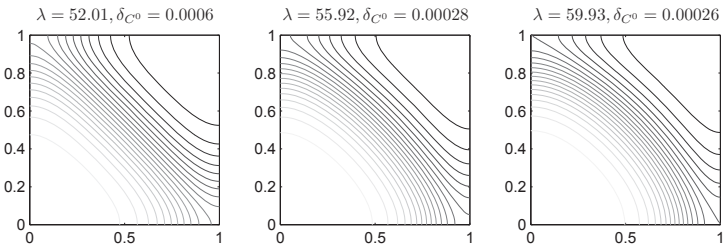


Figure 15 – Snapshots for mode $w_{10} + w_{01}$ solutions with $\mu_0 = 0$ on the secondary branch

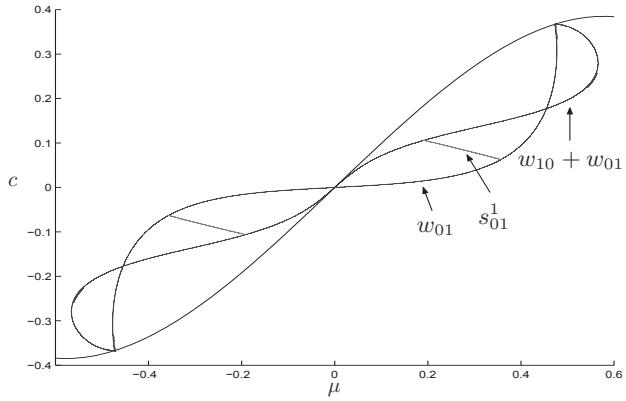


Figure 16 – Bifurcation diagram for mode w_{01} and $w_{10} + w_{01}$ solutions for $\lambda_0 = 30$

indicated by dots. By applying $\mathbf{m}\mathcal{RT}$ to this branch, we get a branch of solutions with the same values of λ and E_λ, but with negated c values. Moreover, if this secondary branch actually bifurcates from $\widetilde{\mathcal{C}_1}$, then the transformed branch has to bifurcate from $\widetilde{\mathcal{C}_1}$ at the same bifurcation point, since $\widetilde{\mathcal{C}_1}$ is invariant under the action $\mathbf{m}\mathcal{RT}$.

We now turn our attention to fixing the parameter $\lambda = \lambda_0$ and studying variations in the total mass μ. Figure 16 shows the situation for $\lambda_0 = 30$. In addition to continuations of the w_{01} and $w_{10} + w_{01}$ branches, we found a secondary bifurcation that creates a connecting branch between them. More precisely, we have the following result.

Theorem 1.13 (Modes w_{01} and $w_{10} + w_{01}$ for $\lambda_0 = 30$)**.** *There exist branches of solutions of* (5) *subject to the constraint* (2) *for* $\lambda_0 = 30$ *as shown in figure 16. Detailed information on the endpoints of these branches in the* (μ, c)*-plane, as well*

Table 5 – Table of branches from (μ_1, c_1) to (μ_2, c_2) at $\lambda_0 = 30$ for $\mu > 0$

mode	μ_1	c_1	μ_2	c_2	δ_{C^0}	δ_μ	δ_c	uniqueness
w_{01}	0	0	0.3568	0.0628	0.0003	0.0003	0.0006	partly
w_{01}	0.3569	0.0629	0.473	0.3671	0.0072	0.0012	0.0006	partly
$w_{10} + w_{01}$	0	0	0.1914	0.106	0.0003	0.0003	0.0006	partly
$w_{10} + w_{01}$	0.1915	0.1061	0.4732	0.3671	0.0234	0.0008	0.0006	partly
s^1_{01}	0.1915	0.106	0.3568	0.0629	0.0003	0.0007	0.0007	partly

Table 6 – Uniqueness of the branches in table 5 holds from $(\tilde\mu_1, \tilde c_1)$ to $(\tilde\mu_2, \tilde c_2)$

mode	$\tilde\mu_1$	$\tilde c_1$	$\tilde\mu_2$	$\tilde c_2$
w_{01}	0	0	0.355	0.0618
w_{01}	0.3587	0.0639	0.473	0.3667
$w_{10} + w_{01}$	0	0	0.1841	0.1038
$w_{10} + w_{01}$	0.1988	0.1083	0.4808	0.3665
s^1_{01}	0.1958	0.1049	0.3562	0.063

as uniqueness assertions, can be found in tables 5 and 6.

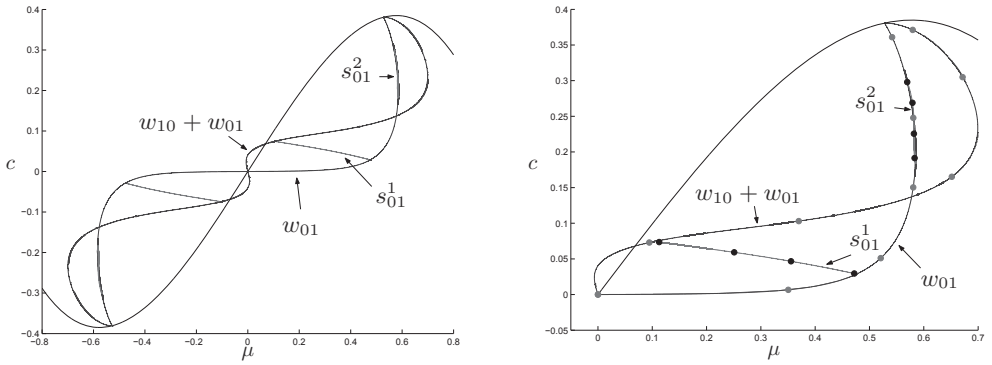
The locations of the branches guaranteed by this theorem are indicated in figure 16. The branches corresponding to the w_{01} and the $w_{10} + w_{01}$ modes are labeled, and the secondary connection between them is shown as well. The situation is similar for $\lambda_0 = 60$. One can establish the existence of a connecting branch s^1_{01} between the w_{01} and $w_{10} + w_{01}$ solution branches. Yet in addition, we also obtain another small branch which bifurcates from the w_{01}-path and returns to it. This new branch is denoted by s^2_{01}. Figure 17a shows the bifurcation diagram for $\lambda_0 = 60$, and a close-up can be found in Figure 17b. The branches in this latter figure are guaranteed by the following result.

Theorem 1.14 (Modes w_{01} and $w_{10} + w_{01}$ for $\lambda_0 = 60$)**.** *There exist branches of solutions of* (5) *subject to the constraint* (2) *for* $\lambda_0 = 60$ *as shown in figure 17a. Detailed information on the endpoints of these branches in the* (μ, c)*-plane, as well as uniqueness assertions, can be found in Tables 7 and 8.*

The geometry of solutions on the w_{01} branch guaranteed by the above theorem is shown in figure 18, the corresponding locations are indicated by red dots in figure 17b. Thus, this branch appears to be the ODE-branch. Moreover, this branch is part of the continuum that connects the bifurcation points

$$\mu^+_{01} = \sqrt{\frac{1}{3} - \frac{\kappa_{01}}{180}} \qquad \text{and} \qquad \mu^-_{01} = -\sqrt{\frac{1}{3} - \frac{\kappa_{01}}{180}}.$$

See also (13) and [20]. The branches for positive and negative total mass μ are related by the action $\mathbf{m}\mathcal{R}^2\mathcal{T}$, and these solutions have the symmetry $\mathcal{T}$. There are



(a) Bifurcation diagram for w_{01} and $w_{10}+w_{01}$ solutions at $\lambda_0 = 60$

(b) Close-up of figure 17a

Figure 17

Table 7 – Table of branches from (μ_1, c_1) to (μ_2, c_2) at $\lambda_0 = 60$ for $\mu > 0$

mode	μ_1	c_1	μ_2	c_2	δ_{C^0}	δ_μ	δ_c	uniqueness
w_{01}	0	0	0.4762	0.0287	0.005	0.0003	0.0008	partly
w_{01}	0.4763	0.0288	0.5843	0.1845	0.0013	0.0012	0.006	partly
w_{01}	0.5844	0.1845	0.5687	0.2997	0.0007	0.0001	0.0001	partly
w_{01}	0.5686	0.2998	0.5278	0.3807	0.0082	0.001	0.0002	partly
$w_{10}+w_{01}$	0	0	0.1031	0.0743	0.0156	0.0007	0.0002	no
$w_{10}+w_{01}$	0.1045	0.0746	0.5299	0.3807	0.0095	0.0017	0.0019	partly
s^1_{01}	0.1043	0.0744	0.4757	0.0289	0.0106	0.0009	0.0008	partly
s^2_{01}	0.5843	0.1846	0.5687	0.2996	0.0081	0.0004	0.0002	partly

Table 8 – Uniqueness of the branches in table 7 holds from $(\tilde{\mu}_1, \tilde{c}_1)$ to $(\tilde{\mu}_2, \tilde{c}_2)$

mode	$\tilde{\mu}_1$	$\tilde{c}_1$	$\tilde{\mu}_2$	$\tilde{c}_2$
w_{01}	0	0	0.4712	0.027
w_{01}	0.4807	0.0031	0.5842	0.1823
w_{01}	0.5845	0.1868	0.5691	0.2982
w_{01}	0.5684	0.3005	0.528	0.3805
$w_{10}+w_{01}$	0.6874	0.1958	0.5543	0.378
s^1_{01}	0.1641	0.0682	0.4286	0.0364
s^2_{01}	0.5818	0.2042	0.577	0.2782

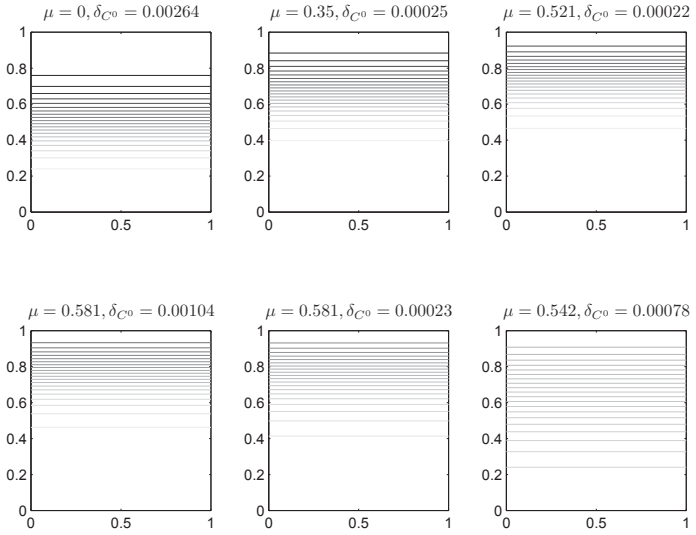


Figure 18 – Snapshots for solutions on the branch w_{01} for $\lambda_0 = 60$

two branches which correspond to w_{01} and bifurcate from μ_{01}^+. These branches are called $\mathcal{K}_{w_{01}}$ and $\mathcal{K}_{-w_{01}}$, and their existence follows readily from a local analysis at the bifurcation point μ_{01}^+, see also the similar situation (14). The branch $\mathcal{K}_{-w_{01}}$ can be obtained from $\mathcal{K}_{w_{01}}$ by applying the action $\mathcal{R}^2\mathcal{T}$. In addition, an application of the actions $\mathcal{TR}$ and $\mathcal{RT}$ to $\mathcal{K}_{w_{01}}$ generates new branches $\mathcal{K}_{w_{10}}$ and $\mathcal{K}_{-w_{10}}$, respectively. In other words, we have

$$\mathcal{K}_{w_{01}} = \mathcal{T}(\mathcal{K}_{w_{01}}),\ \ \mathcal{K}_{-w_{01}} = \mathcal{R}^2\mathcal{T}(\mathcal{K}_{w_{01}}),\ \ \mathcal{K}_{w_{10}} = \mathcal{TR}(\mathcal{K}_{w_{01}}),\ \ \mathcal{K}_{-w_{10}} = \mathcal{RT}(\mathcal{K}_{w_{01}}).$$

Thus, every solution on $\mathcal{K}_{w_{01}}$ gives rise to four different solutions with unchanged values of μ and c, through the application of suitable symmetry actions.

The snapshots in figure 19 contain contour plots of sample solutions on the $w_{10}+w_{01}$ branch described in table 7, the locations of these solutions are indicated by dots in figure 17b. The parameter values for the first functions in figures 18 and 19 are the same, in both cases we have $(\mu, c) = (0, 0)$, hence only one dot is visible in figure 17b. It is possible to apply Kielhöfer's connection result, Theorem 1.3, to mode $w_{10} + w_{01}$ solutions. It states that for $\lambda_0 > \kappa_{01} = \pi^2$ the continuum $\mathcal{C}_1^+(\lambda_0) = \mathcal{C}_1^-(\lambda_0)$ connects the two bifurcation points μ_{10}^+ and μ_{10}^-. This continuum splits into two parts, denoted by $\mathcal{K}_{w_{10}+w_{01}}$ and $\mathcal{K}_{-w_{10}-w_{01}}$, see also (14). Assuming that the path described in table 7 is a part of $\mathcal{C}_1^+(\lambda_0)$ for $\lambda_0 = 60$, and therefore of $\mathcal{K}_{w_{10}+w_{01}}$, the solutions on this path have the symmetry $\mathcal{TR}$. The elements of the branch with negative μ are obtained by applying the action $\mathbf{m}\mathcal{RT}$ to the path with positive mass. Through the application of $\mathcal{RT}$ one obtains $\mathcal{K}_{-w_{10}-w_{01}}$. Finally, the paths $\mathcal{K}_{w_{10}-w_{01}}$ and $\mathcal{K}_{-w_{10}+w_{01}}$ are

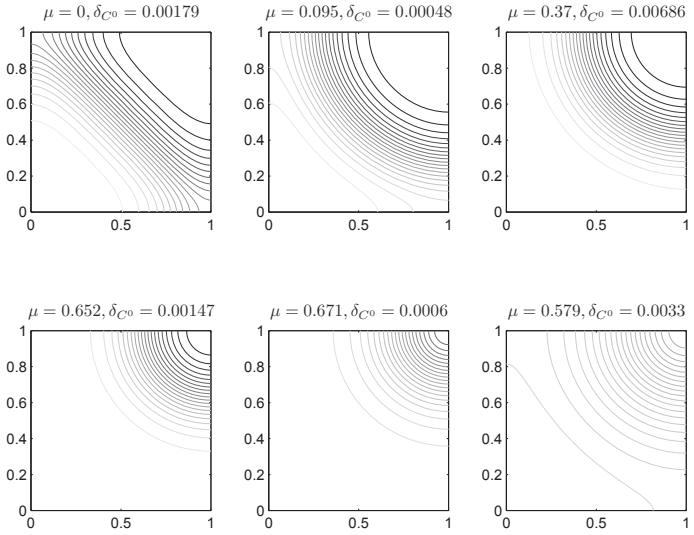


Figure 19 – Snapshots for solutions on the branch $w_{10} + w_{01}$ for $\lambda_0 = 60$

generated from $\mathcal{K}_{w_{10}+w_{01}}$ through the actions $\mathcal{R}^3$ and $\mathcal{T}$. Altogether, we have

$$\mathcal{K}_{w_{10}+w_{01}} = \mathcal{T}\mathcal{R}(\mathcal{K}_{w_{10}+w_{01}}), \qquad \mathcal{K}_{-w_{10}-w_{01}} = \mathcal{R}\mathcal{T}(\mathcal{K}_{w_{10}+w_{01}}),$$
$$\mathcal{K}_{w_{10}-w_{01}} = \mathcal{R}^3(\mathcal{K}_{w_{10}+w_{01}}), \qquad \mathcal{K}_{-w_{10}+w_{01}} = \mathcal{T}(\mathcal{K}_{w_{10}+w_{01}}).$$

Besides the above primary branches, Theorem 1.14 also guaranteed secondary branches as shown in figures 17a and 17b. One of these connects the continuum $\mathcal{C}_1^+(\lambda_0)$ of solutions corresponding to $w_{10} + w_{01}$ with the continuum of the ODE solutions. Figure 20 shows snapshots of functions on this connecting branch, the locations of which are indicated by black dots in figure 17b. Notice that the branch s_{01}^1 breaks the symmetry $\mathcal{T}\mathcal{R}$ of $\mathcal{C}_1^+(\lambda_0)$, as well as the symmetry $\mathcal{T}$ of the ODE-branch. Its elements

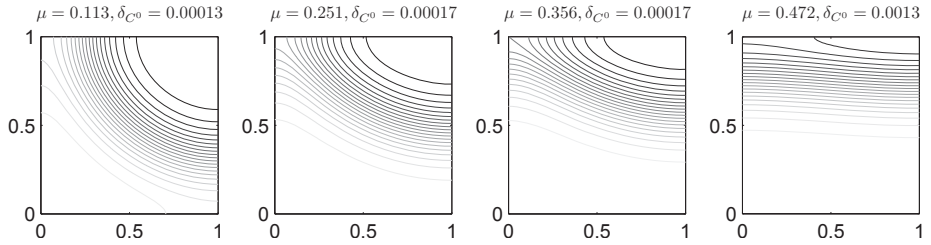
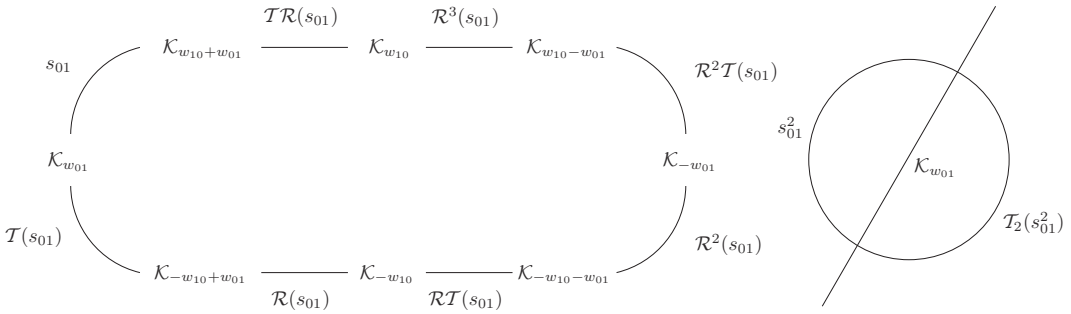


Figure 20 – Snapshots for solutions on the branch s_{01}^1 for $\lambda_0 = 60$



(a) Sketch of the secondary bifurcations related to s_{01}^1

(b) Sketch of the secondary bifurcation related to s_{01}^2

Figure 21

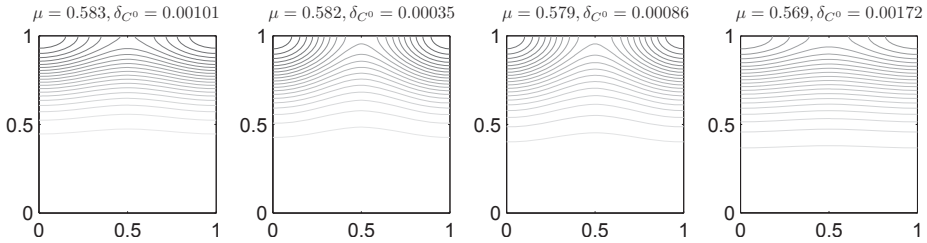
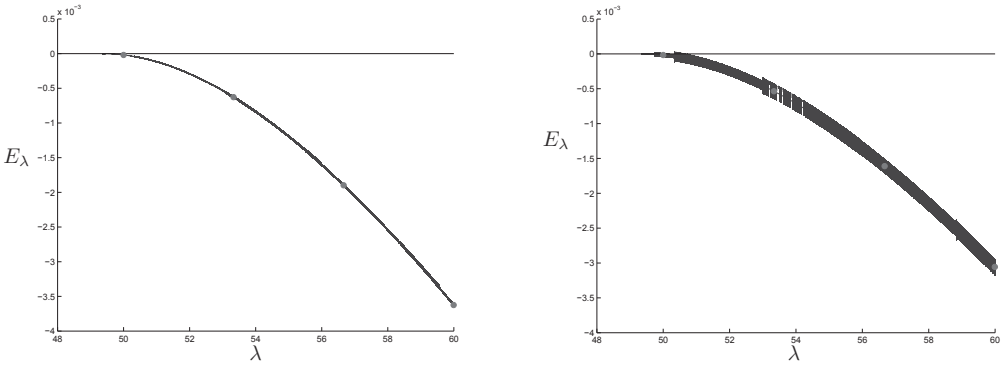


Figure 22 – Snapshots for solutions on the branch s_{01}^2 for $\lambda_0 = 60$

are no longer constant in any of the coordinate directions, and therefore do not correspond to solutions of the one-dimensional Cahn-Hilliard equation. Applying the action $\mathcal{T}$ to the branch s_{01}^1 furnishes a path between $\mathcal{K}_{w_{01}}$ and $\mathcal{K}_{w_{10}-w_{01}}$, since the solutions in $\mathcal{K}_{w_{01}}$ are invariant under $\mathcal{T}$. A similar argument shows that $\mathcal{TR}(s_{01}^1)$ connects $\mathcal{K}_{w_{10}+w_{01}}$ and $\mathcal{K}_{w_{10}}$. The full secondary bifurcation scheme is sketched in figure 21a.

Finally, figure 22 indicates the geometries of functions on the s_{01}^2-branch described in table 7, see also figures 17a and 17b. Notice that this secondary branch returns to the branch it bifurcated from. Moreover, since it does not break the symmetry $\mathcal{T}$, it is impossible to generate the other bifurcating branch by applying this action. To address this issue, let $\mathcal{T}_2$ denote the reflection at the line $x = 1/4$. As before we assume implicitly that all solutions on the unit cube have been extended to all of $\mathbb{R}^2$ by even reflections. If we now apply $\mathcal{T}_2$, one obtains a solution branch which bifurcates from the ODE-branch and returns to $\mathcal{K}_{w_{01}}$ at the same points as s_{01}^2, since the elements of the ODE-branch are constant in the x-direction, and hence invariant under $\mathcal{T}_2$. These



(a) Energy for mode w_{12} solutions with $\mu_0 = 0$

(b) Energy for mode $w_{12} + w_{21}$ solutions with $\mu_0 = 0$

Figure 23

secondary bifurcations are sketched in figure 21b. By applying appropriate actions to s^2_{01}, similar statements can be made about $\mathcal{K}_{-w_{01}}$, $\mathcal{K}_{w_{10}}$, and $\mathcal{K}_{-w_{10}}$.

1.6. The modes w_{12} and $w_{12} + w_{21}$

To conclude this section, we finally address the bifurcation structure associated with the mode w_{12}. For vanishing total mass $\mu_0 = 0$ our rigorous computations furnish the following result.

Theorem 1.15 (Modes w_{12} and $w_{12} + w_{21}$ at $\mu_0 = 0$)**.**

(i) *There exists a branch of unique solutions of* (5) *subject to the constraint* (2) *for* $\mu_0 = 0$ *and* $\lambda \in [5\pi^2 + 0.0025, 60]$ *as shown in figure 23a, with* $\delta_{C^0} < 0.00013$ *and* $\delta_{E_\lambda} < 0.00004$.

(ii) *There exists a branch of unique solutions of* (5) *subject to the constraint* (2) *for* $\mu_0 = 0$ *and* $\lambda \in [5\pi^2 + 0.0025, 60]$ *as shown in figure 23b, with* $\delta_{C^0} < 0.00142$ *and* $\delta_{E_\lambda} < 0.00016$.

The branch guaranteed by Theorem 1.15 (i) is contained in figure 23a. For the sample solutions indicated by dots, the corresponding geometries are shown in figure 24. The shape of these solutions suggests that solutions on this path have the symmetries $\mathbf{m}\mathcal{T}$ and $\mathcal{R}^2\mathcal{T}$, and that the path is contained in $\widetilde{\mathcal{C}}_{12}$. See also Remark 1.6. The other part of $\widetilde{\mathcal{C}}_{12}$ can be generated through the action $\mathbf{m}$.

The branch described in Theorem 1.15 (ii) is contained in figure 23b. The geometries of sample solutions on this branch are indicated in figure 25. These geometries suggest that the branch is part of the solution continuum that corresponds to the

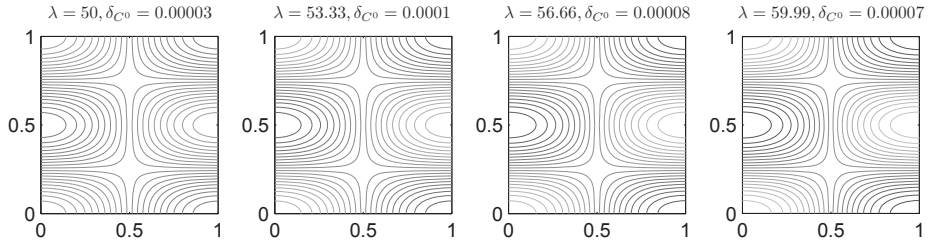


Figure 24 – Snapshots for mode w_{12} solutions with $\mu_0 = 0$

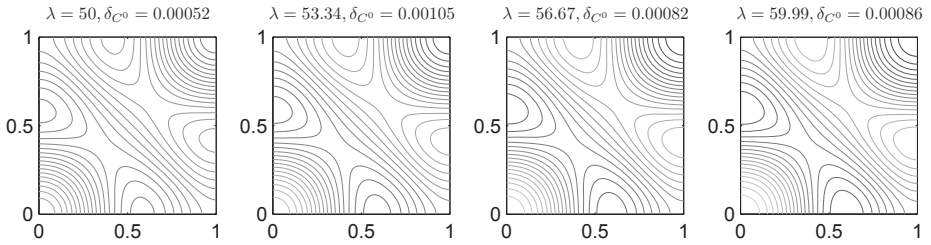


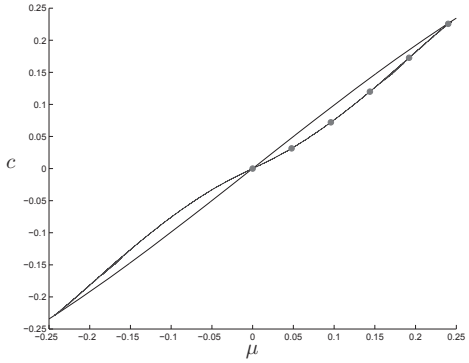
Figure 25 – Snapshots for mode $w_{12} + w_{21}$ solutions with $\mu_0 = 0$

mode $w_{12} + w_{21}$. Notice that none of the results in section 1.1 applies to this mode. If we assume that the symmetry of $w_{12} + w_{21}$ is preserved along the path, i.e., elements of the path are symmetric with respect to $\mathcal{TR}$ and $\mathbf{m}\mathcal{RT}$, then the action $\mathbf{m}\mathcal{TR}$ furnishes the other path of $w_{12} + w_{21}$ solutions bifurcating at κ_{12}.

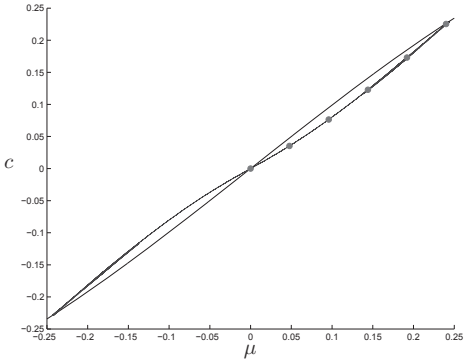
For the remainder of this section we consider the effects of mass variation, i.e., we fix the parameter $\lambda = \lambda_0$. In particular, for $\lambda_0 = 60$ our results are summarized in figures 26a and 26b. More precisely, we have the following result.

Theorem 1.16 (Modes w_{12} and $w_{12} + w_{21}$ for $\lambda_0 = 60$)**.** *There exist branches of solutions of* (5) *subject to the constraint* (2) *for* $\lambda_0 = 60$ *as shown in figures 26a, 26b, and 27a. The* w_{12}*-branches are shown in figure 26a together with the trivial solution* $c = \mu - \mu^3$*. The* $w_{12} + w_{21}$*-branch is shown in figure 26b, while the* s_{12}*-branch is shown in figure 27a. The latter branch is the consequence of a secondary bifurcation. Detailed information on the endpoints of these branches in the* (μ, c)*-plane, as well as uniqueness assertions, can be found in tables 9 and 10.*

The dots in figure 26a correspond to the solution snapshots depicted in figure 28. These solutions are symmetric with respect to the action $\mathcal{R}^2\mathcal{T}$, and the branch for negative μ can be obtained from the one for positive μ by applying $\mathbf{m}\mathcal{T}$. Now assume that the branch w_{12} described in table 9 is contained in $\mathcal{C}_{12}^+(\lambda_0)$ ($\mathcal{C}_{12}^-(\lambda_0)$) for $\lambda_0 = 60$, see also section 1.1. Notice that elements of $\mathcal{C}_{12}^+(\lambda_0)$ ($\mathcal{C}_{12}^-(\lambda_0)$) are point symmetric

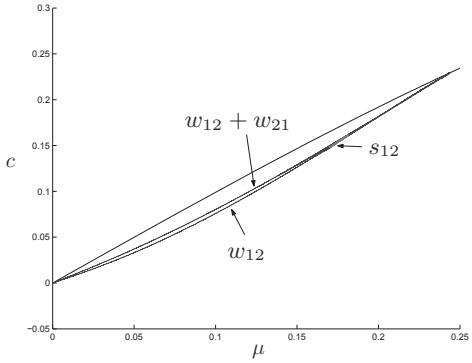


(a) Bifurcation diagram for w_{12} solutions with $\lambda_0 = 60$

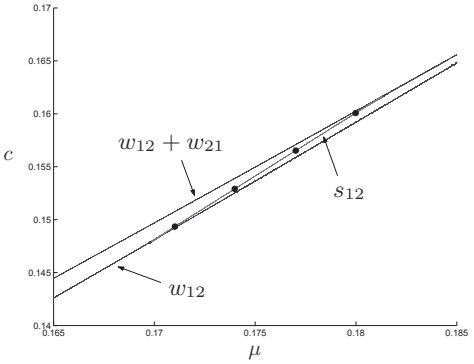


(b) Bifurcation diagram for mode $w_{12} + w_{21}$ solutions with $\lambda_0 = 60$

Figure 26



(a) Bifurcation diagram for mode w_{12} and $w_{12} + w_{21}$ solutions with $\lambda_0 = 60$



(b) Close-up of figure 27a

Figure 27

Table 9 – Branches from (μ_1, c_1) to (μ_2, c_2) for $\lambda_0 = 60$ and $\mu \geq 0$

mode	μ_1	c_1	μ_2	c_2	δ_{C^0}	δ_μ	δ_c	uniqueness
w_{12}	0	0	0.1697	0.1478	0.0007	0.0003	0.0004	partly
w_{12}	0.1698	0.1479	0.2432	0.2288	0.0022	0.0001	0.0001	partly
$w_{12} + w_{21}$	0	0	0.1813	0.1616	0.0004	0.0003	0.0004	partly
$w_{12} + w_{21}$	0.1814	0.1617	0.2432	0.2288	0.0015	0.0001	0.0001	partly
s_{12}	0.1699	0.148	0.1812	0.1614	0.0013	0.0001	0.0001	no

Table 10 – Uniqueness of the branches in table 9 holds from $(\tilde{\mu}_1, \tilde{c}_1)$ to $(\tilde{\mu}_2, \tilde{c}_2)$

mode	$\tilde{\mu}_1$	$\tilde{c}_1$	$\tilde{\mu}_2$	$\tilde{c}_2$
w_{12}	0	0	0.1683	0.1463
w_{12}	0.1718	0.1501	0.2407	0.2262
$w_{12} + w_{21}$	0	0	0.1761	0.1561
$w_{12} + w_{21}$	0.1831	0.1636	0.2394	0.2246

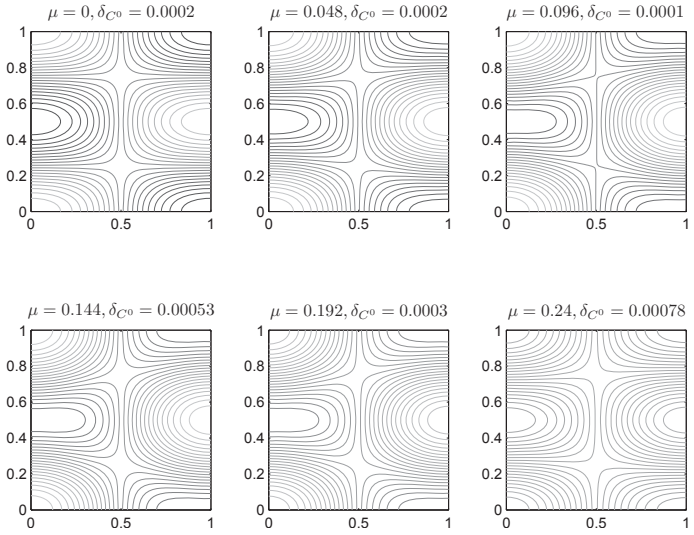


Figure 28 – Snapshots for mode w_{12} solutions with $\lambda_0 = 60$

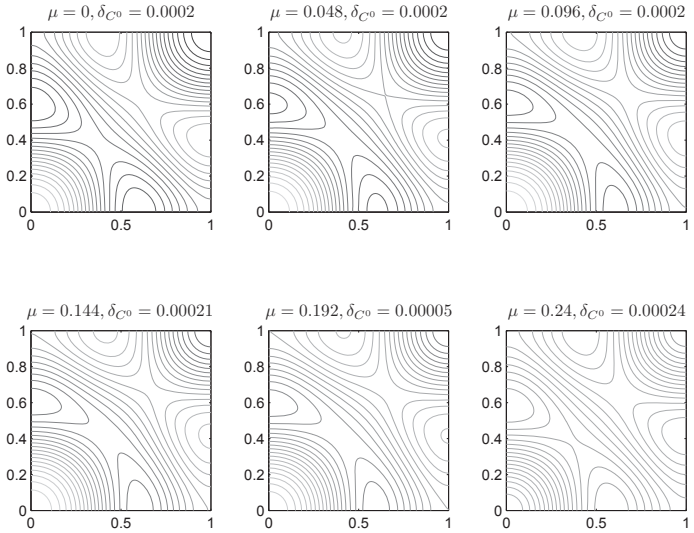


Figure 29 – Snapshots for mode $w_{12} + w_{21}$ solutions with $\lambda_0 = 60$

around the points $(1/2, 1/4)$ and $(1/2, 3/4)$, which is supported by the structure of the level sets in figure 28. According to Theorem 1.4, the branch $\mathcal{C}^+_{12}(\lambda_0)$ ($\mathcal{C}^-_{12}(\lambda_0)$) connects the two bifurcation points μ^+_{12} and μ^-_{12}, i.e., $\mathcal{C}^+_{12}(\lambda_0) = \mathcal{C}^-_{12}(\lambda_0)$. Due to Theorem 1.4, there is a possibility that $\mathcal{C}^+_{12}(\lambda_0)$ meets the trivial solution in some point $(\tilde{m}, 0)$ with $\tilde{m} \neq \mu^\pm_{12}$, but since we could not find such connections, there seems to be in fact exactly two nontrivial paths of solutions between $(\mu^+_{12}, 0)$ and $(\mu^-_{12}, 0)$, i.e., $\mathcal{C}^+_{12}(\lambda_0)$ splits into two parts $\mathcal{K}_{w_{12}}$ and $\mathcal{K}_{-w_{12}}$. By symmetry, these are related by the action $\mathcal{T}$. Nevertheless, due to the gaps at the bifurcation point and the gaps along the path, a slight uncertainty remains. Using the actions $\mathcal{R}$ and $\mathcal{RT}$, one can generate the branches $\mathcal{K}_{w_{21}}$ and $\mathcal{K}_{-w_{21}}$ from $\mathcal{K}_{w_{12}}$. Altogether, we have

$$\mathcal{K}_{w_{12}} = \mathcal{R}^2\mathcal{T}(\mathcal{K}_{w_{12}}),\ \ \mathcal{K}_{-w_{12}} = \mathcal{T}(\mathcal{K}_{w_{12}}),\ \ \mathcal{K}_{w_{21}} = \mathcal{R}(\mathcal{K}_{w_{12}}),\ \ \mathcal{K}_{-w_{21}} = \mathcal{RT}(\mathcal{K}_{w_{12}}).$$

Now consider the $w_{12} + w_{21}$ path in figure 26b, as described in table 9. The solution geometry on this branch is indicated in figure 29. Thus, it seems reasonable to assume that all the solutions on the $w_{12} + w_{21}$ path have the symmetry $\mathcal{TR}$, and that the part with negative mass can be obtained from the one with positive mass by applying the action $\mathbf{m}\mathcal{RT}$. Note that it also seems plausible that the branch connects the bifurcation points $\mu^\pm_{12}$. Therefore, we assume that there exists a continuum which can be split into two pieces, denoted by $\mathcal{K}_{w_{12}+w_{21}}$ and $\mathcal{K}_{-w_{12}-w_{21}}$, which are related through $\mathcal{R}^2$. Furthermore, by applying $\mathcal{R}^3$ and $\mathcal{R}$ to the elements of $\mathcal{K}_{w_{12}+w_{21}}$, one can

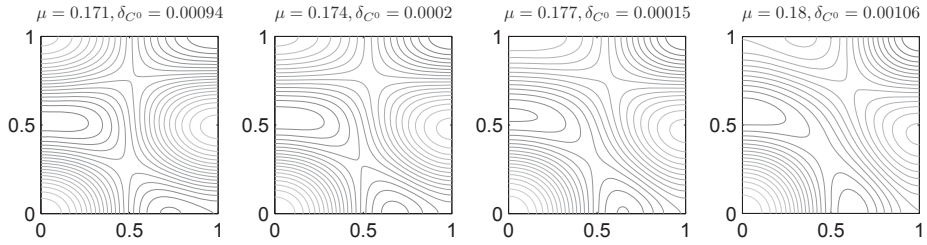


Figure 30 – Snapshots for solutions on the branch s_{12}

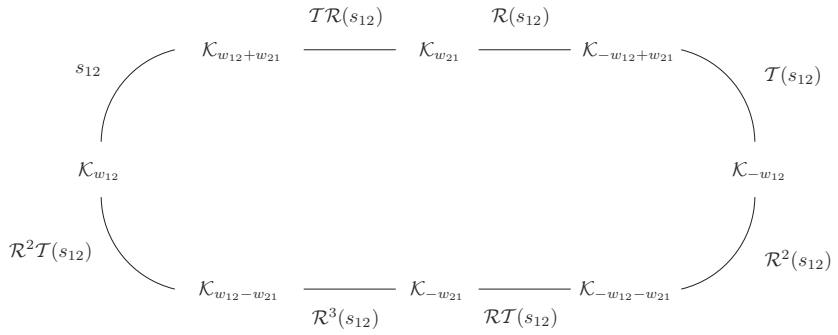


Figure 31 – Sketch of the secondary bifurcations related to s_{12}

generate corresponding paths $\mathcal{K}_{w_{12}-w_{21}}$ and $\mathcal{K}_{-w_{12}+w_{21}}$. Altogether, we now obtain

$$\mathcal{K}_{w_{12}+w_{21}} = \mathcal{TR}(\mathcal{K}_{w_{12}+w_{21}}), \qquad \mathcal{K}_{-w_{12}-w_{21}} = \mathcal{R}^2(\mathcal{K}_{w_{12}+w_{21}}),$$
$$\mathcal{K}_{w_{12}-w_{21}} = \mathcal{R}^3(\mathcal{K}_{w_{12}+w_{21}}), \qquad \mathcal{K}_{-w_{12}+w_{21}} = \mathcal{R}(\mathcal{K}_{w_{12}+w_{21}}).$$

Figure 27a provides a more detailed view of the branches in figures 26a and 26b, yet only for positive mass μ. In addition, the secondary branch s_{12} is shown. This branch is s_{12} from table 9. It connects the $\mathcal{K}_{w_{12}}$ branch of $\mathcal{C}^+_{12}(\lambda_0)$ with the $w_{12}+w_{21}$ solution branch, more precisely, with $\mathcal{K}_{w_{12}+w_{21}}$. Figure 27b shows a close-up of figure 27a in order to resolve the situation better. The secondary branch breaks the symmetry $\mathcal{R}^2\mathcal{T}$ of $\mathcal{K}_{w_{12}}$, and the symmetry $\mathcal{TR}$ of $\mathcal{K}_{w_{12}+w_{21}}$. This can be seen in figure 30, where snapshots of functions on the secondary branch are depicted, corresponding to the dots in figure 27b. Applying the D_4 actions to the branch s_{12} gives new branches. A schematic description of these connections is given in figure 31.

2. Tools from Conley index theory

2.1. The general framework

In [22], Mischaikow and Zgliczyński developed a method for rigorous numerics for partial differential equations using Conley index theory. Their method allowed them to give computer-assisted existence proofs of equilibria for the one-dimensional Kuramoto-Shivashinsky equation for fixed parameter values. Our method is based on their approach, and we therefore briefly recall their setting — yet using notation which is more appropriate for our two-dimensional setting. Consider the abstract evolution equation

$$u_t = F(u) \tag{15}$$

in a Hilbert space H, assume that $F : D(F) \to H$, and that the domain $D(F)$ of F is dense in H. As was mentioned in the introduction, the approach in [22] is based on a Fourier-type representation of the solution u. Thus, we choose a complete orthogonal basis $\{\phi_{i,j}\}_{(i,j)\in\mathbb{N}_0^2\setminus\{(0,0)\}}$ in H, and assume that $\phi_{i,j} \in D(F)$ for all $(i,j) \in \mathbb{N}_0^2 \setminus \{(0,0)\}$.

Specifically for the case of the Cahn-Hilliard equation (1), we consider $\Omega = (0,1)^2$ and define the Hilbert space H as

$$H \coloneqq \left\{ u \in L^2(\Omega) \;\middle|\; \int_\Omega u(x)\,dx = 0 \right\} \subset L^2(\Omega). \tag{16}$$

For given total mass μ as in (2), we define the nonlinearity F as

$$F(u) = -\Delta(\Delta u + \lambda f(\mu + u)), \quad \text{with} \quad f(u) = u - u^3. \tag{17}$$

In view of the Neumann boundary conditions in (1) and our two-dimensional domain, it seems natural to consider the basis given by

$$\phi_{i,j}(x,y) = \cos(i\pi x)\cos(j\pi y) \quad \text{for all} \quad (i,j) \in \mathbb{N}_0^2 \setminus \{(0,0)\}. \tag{18}$$

Notice that these functions are not normalized in $L^2(\Omega)$.

For any choice of the indices $(k,\ell) \in \mathbb{N}_0^2 \setminus \{(0,0)\}$ we define the finite-dimensional subspace

$$X_{k,\ell} \coloneqq \operatorname{span}\{\, \phi_{i,j} \mid (i,j) \neq (0,0) \text{ and } 0 \le i \le k \text{ and } 0 \le j \le \ell \,\}, \tag{19}$$

and let

$$P^{(k,\ell)} : H \longrightarrow X_{k,\ell}$$

denote the orthogonal projection of H onto $X_{k,\ell}$. Moreover, the orthogonal complement of $X_{k,\ell}$ is denoted by $Y_{k,\ell}$, and the corresponding complementary projection by

$$Q^{(k,\ell)} = I - P^{(k,\ell)} : H \longrightarrow Y_{k,\ell}.$$

Finally, we define the operator

$$\mathcal{P}_{k,\ell} : H \longrightarrow \mathbb{R} \qquad \text{by} \qquad \mathcal{P}_{k,\ell}(u) \coloneqq \frac{(u, \phi_{k,\ell})}{(\phi_{k,\ell}, \phi_{k,\ell})}, \tag{20}$$

where $(\cdot, \cdot)$ denotes the scalar product in H. In other words, $\mathcal{P}_{k,\ell}(u)$ denotes the coefficient of $\phi_{k,\ell}$ in the Fourier series representation of $u \in H$ with respect to the basis functions $\{\phi_{i,j}\}_{(i,j)\in\mathbb{N}_0^2\setminus\{(0,0)\}}$, i.e., we have

$$u = \sum_{(i,j)\in\mathbb{N}_0^2\setminus\{(0,0)\}} u_{i,j}\phi_{i,j} \qquad \text{where} \qquad u_{i,j} = \mathcal{P}_{i,j}(u).$$

Now fix an integer $m \in \mathbb{N}$, and set $p = P^{(m,m)}u$ and $q = Q^{(m,m)}u$ for any $u \in H$. Then (15) can be rewritten in the new variables as

$$p_t = P^{(m,m)}F(p+q), \tag{21}$$

$$q_t = Q^{(m,m)}F(p+q). \tag{22}$$

The basic strategy for establishing the existence of equilibrium solutions of this system stems from the following idea. For suitable choices of the parameter m, it should be possible to obtain information on the dynamics of the evolution equation (15) from knowledge of the finite-dimensional system (21). More precisely, assume we know the location of an equilibrium $p = v$ of the finite-dimensional system (21) for the choice $q = 0$, either analytically or numerically. Our intention is then to compute a neighborhood $U = U_p \times U_q$, with $v \in U_p$, such that the existence of an equilibrium of (15) in U can be guaranteed. Since the finite-dimensional system (21) depends on the infinite-dimensional parameter q, this can only be accomplished if we can control the infinite-dimensional complementary equation (22). It will be shown later in this section, using the estimates of Section A, that this can actually be achieved. The computation of U is done numerically. For this, we derive specific conditions which guarantee the existence of an equilibrium of (15). In view of the applications in [16], these derivations rely on Conley index theory, and the resulting conditions are formulated in such a way that they can be checked with the aid of a computer. Since these checks can be performed in rigorous interval arithmetic, we have thus obtained an analytical proof for the existence of an equilibrium of (15). The details of this approach will be presented in the remainder of this section.

We begin our adaptation of the results in [22] by defining the notion of self-consistent a priori bounds. In order to simplify our presentation, we introduce some notation.

Definition 2.1. Consider the abstract situation described above and let

$$I_* \coloneqq \mathbb{N}_0^2 \setminus \{(0,0)\}.$$

Let H denote a Hilbert space with complete orthogonal set $\{\phi_{i,j}\}_{(i,j)\in I_*}$, and let $a^{\pm}_{i,j} \in \mathbb{R}$ denote a family of real numbers with $a^-_{i,j} < a^+_{i,j}$ for all $(i,j) \in I_*$. For any nonempty subset $I \subset I_*$ we then define

$$\prod_{(i,j)\in I} [a^-_{i,j}, a^+_{i,j}] \coloneqq \left\{ u = \sum_{(i,j)\in I} u_{i,j}\phi_{i,j} \;\middle|\; a^-_{i,j} \le u_{i,j} \le a^+_{i,j} \quad \text{for all} \quad (i,j) \in I \right\}.$$

Furthermore, for any integer $m \in \mathbb{N}$ we define the special index set

$$I_m \coloneqq \{\, (i,j) \in \mathbb{N}_0^2 \mid i > m \quad \text{or} \quad j > m \,\}.$$

This implies that we have both

$$\prod_{(i,j)\in I_*\setminus I_m} [a^-_{i,j}, a^+_{i,j}] \subset X_{m,m} \qquad \text{and} \qquad \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}] \subset Y_{m,m},$$

as long as the $a^{\pm}_{i,j}$ decay sufficiently fast as $i,j \to \infty$. The set $X_{m,m}$ was defined in (19), and $Y_{m,m}$ is its orthogonal complement in H.

The above notation is a convenient way to describe subsets of the Hilbert space H, whose images under the mappings $\mathcal{P}_{i,j}$ defined in (20) are compact intervals. Such sets lie at the heart of the following definition.

Definition 2.2. For fixed integers $0 < m < M$ consider a compact set $W \subset X_{m,m}$ and a collection of real numbers $a^{\pm}_{i,j}$ which satisfy $a^-_{i,j} < a^+_{i,j}$ for all $(i,j) \in I_m$. Then W and $\{a^{\pm}_{i,j}\}_{(i,j)\in I_m}$ are called *self-consistent a priori bounds* for the abstract evolution equation (15), if the following conditions hold:

(i) For all $(i,j) \in I_{M-1}$ we have $a^-_{i,j} < 0 < a^+_{i,j}$.

(ii) Every formal series in the definition of $\prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$ is in fact convergent in H, i.e., we have $W \times \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}] \subset H$.

(iii) For all $r, \ell > m$ the composition $P^{(r,\ell)} \circ F : X_{r,\ell} \to X_{r,\ell}$ is Lipschitz continuous on the intersection of $X_{r,\ell}$ with the set $W \times \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$. Here $P^{(r,\ell)}$ denotes the orthogonal projection onto $X_{r,\ell}$.

(iv) If $(z^{(r)})_{r>M}$ is an arbitrary sequence of functions such that for every $r > M$ the function $z^{(r)}$ is contained in the intersection of the set $W \times \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$ and the finite-dimensional space $X_{r,r}$, if in addition we have $P^{(r,r)} \circ F(z^{(r)}) = 0$ for every $r > M$, and if $(z^{(r)})_{r>M}$ has an accumulation point $z^{(\infty)}$ with convergence in H, then we have both

$$z^{(\infty)} \in D(F) \qquad \text{and} \qquad F(z^{(\infty)}) = 0.$$

This definition is in some sense the first step towards establishing the existence of an equilibrium solution of (15) from knowledge of the finite-dimensional system (21). According to (iv), if one identifies a sequence of equilibrium solutions of (21) for $q = 0$, then any accumulation point of this sequence will in fact provide an equilibrium for the original infinite-dimensional system (15). Moreover, the above definition singles out sets of the form given in Definition 2.1 as basis for the method.

The above notion leaves one aspect unanswered: How can we establish the existence of a stationary solution of the finite-dimensional system (21) for all sufficiently large values of m, within sets of the form given in Definition 2.2? To answer this question we need to employ tools from Conley index theory. Due to space limitations, we will not be able to present all definitions here, but rather refer the reader to [18, 22]. The central definition is as follows.

Definition 2.3. Let W and $\{a^{\pm}_{i,j}\}_{i,j\in I_m}$ denote self-consistent a priori bounds for (15) as in Definition 2.2. In addition, let $N \subset W$ be a compact set. Then N, W, and $\{a^{\pm}_{i,j}\}_{i,j\in I_m}$ are called *strict topologically self-consistent a priori bounds* for (15), if the following holds:

(i) For every $u \in W \times \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$ and all $(r,\ell) \in I_m$ we have

$$\begin{aligned} &\text{if} \quad \mathcal{P}_{r,\ell}u = a^+_{r,\ell}, \quad \text{then} \quad \mathcal{P}_{r,\ell}F(u) < 0, \\ &\qquad\qquad \text{as well as} \\ &\text{if} \quad \mathcal{P}_{r,\ell}u = a^-_{r,\ell}, \quad \text{then} \quad \mathcal{P}_{r,\ell}F(u) > 0. \end{aligned} \tag{23}$$

(ii) There exists a closed subset $N^- \subset N$ such that for every $q \in \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$ the set N is an isolating block for (21) with exit set N^-.

Notice that in (ii) we consider (21) as a finite-dimensional ordinary differential equation for $p \in X_{m,m}$, which depends on the parameter q. Thus, the resulting flow $\varphi = \varphi_q$ depends on the choice of q as well, and (ii) states that the fixed set $N \subset X_{m,m}$ is an isolating block with exit set N^- for all flows φ_q, where $q \in \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$ is arbitrary. We would also like to point out that (ii) in fact implies

$$h(\mathrm{Inv}(N), \varphi_q) = \text{constant} \qquad \text{for all} \quad q \in \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}],$$

were h denotes the Conley index of the largest invariant set $\mathrm{Inv}(N)$ in N. As before, we refer the reader to [18, 22] for more details.

We are now finally in a position to present the main tool for establishing the existence of stationary solutions for the infinite-dimensional system (15).

Theorem 2.4. *In the situation of Definition 2.3, assume that N, W, and $\{a^{\pm}_{i,j}\}_{i,j\in I_m}$ are strict topologically self-consistent a priori bounds for* (15)*. Furthermore, suppose that*

$$h(\mathrm{Inv}(N), \varphi_{q_0}) = [\Sigma^{\ell_0}]$$

for some $\ell_0 \in \mathbb{N}_0$ and some $q_0 \in \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$, where $[\Sigma^{\ell_0}]$ denotes the homotopy type of a pointed ℓ_0-sphere. Then there exists an equilibrium v^* of (15) *in $N \times \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$.*

The above result is due to Mischaikow and Zglicszynski [22], who in contrast to our formulation use the homological Conley index. Since the proof of the above result is more or less analogous to their result, we refrain from presenting it in detail and refer the reader to [18]. The basic proof idea is the fact that property (23) of Definition 2.3 allows one to lift the isolating block N to higher dimensions without changing its Conley index. A result due to McCord [17] then furnishes the existence of a fixed point in such an isolating block, provided the Conley index has the form $[\Sigma^{\ell_0}]$ for some $\ell_0 \in \mathbb{N}_0$. In this way, it is possible to construct a sequence of fixed points, and due to Definition 2.2, any accumulation point of this sequence is an equilibrium of (15).

From a computational point of view, it is of course crucial to verify that specific sets N and W, together with collections $\{a^{\pm}_{i,j}\}_{(i,j)\in I_m}$ indeed give rise to (strict topologically) self-consistent a priori bounds, based on a finite amount of information. This will be accomplished in the remainder of this section. In section 2.2 we present a sufficient condition for self-consistent a priori bounds, Section 2.3 addresses the estimates in Definition 2.3 (i). Finally, section 2.4 demonstrates how Definition 2.3 (ii) can be verified and how the Conley index of N can be computed.

2.2. Self-consistent a priori bounds

As we have seen in section 2.1, self-consistent a priori bounds are used to show that stationary solutions of the finite-dimensional system (21) can be used to approximate equilibria of (15). In order to make this construction amenable to a computational treatment, one needs to be able to verify Definition 2.2 in finitely many steps. For this, we make the following assumption.

Assumption 2.5. *For fixed integers $0 < m < M$, assume that there exist positive constants s and C, as well as positive constants $C_1(i)$ and $C_2(i)$, where $i \in \{0, \dots, M-1\}$, such that the following holds. Let $\{a^{\pm}_{i,j}\}_{(i,j)\in I_*}$ denote collections of real numbers which satisfy $a^-_{i,j} < a^+_{i,j}$ for all $(i,j) \in I_*$, as well as $a^-_{i,j} < 0 < a^+_{i,j}$ for all $(i,j) \in I_{M-1}$. In addition, suppose that*

$$|a^{\pm}_{i,j}| \le \begin{cases} \frac{C_1(i)}{j^s} & \text{for} \quad j \ge M \quad \text{and} \quad 0 \le i < M, \\ \frac{C_2(j)}{i^s} & \text{for} \quad i \ge M \quad \text{and} \quad 0 \le j < M, \\ \frac{C}{i^s j^s} & \text{for} \quad i, j \ge M. \end{cases} \tag{24}$$

One can easily see that collections of numbers $a^{\pm}_{i,j}$ as in Assumption 2.5 satisfy both (i) and (ii) in Definition 2.2 as long as $s > 1/2$, provided the norms of the basis functions $\phi_{i,j}$ are uniformly bounded. However, the remaining parts of Definition 2.2 do depend on the function F in (15). For the case of the Cahn-Hilliard equation (1) on the square, the following result establishes the validity of the situation of Definition 2.2 for all $s \geq 2$ and collections $\{a^{\pm}_{i,j}\}$ as in (24).

Theorem 2.6. *Consider the Cahn-Hilliard equation* (1) *on the unit square* $\Omega = (0,1)^2$. *Define the Hilbert space* H *as in* (16), *equipped with the complete orthogonal set in* (18), *and let* F *be as in* (17). *Moreover, assume the situation of Assumption 2.5 with* $s \geq 2$. *Then the set*

$$W := \prod_{(i,j)\in I_* \setminus I_m} [a^-_{i,j}, a^+_{i,j}] \subset X_{m,m}$$

together with $\{a^{\pm}_{i,j}\}_{(i,j)\in I_m}$ *are self-consistent a priori bounds for* (15).

Proof. Property (i) in Definition 2.2 is clear. As for (ii), let $u \in W \times \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$ be arbitrary, and define $\hat{a}_{i,j} = \max\{\, |a^{\pm}_{i,j}| \,\}$. Then the orthogonality of the basis functions $\phi_{i,j}$ and the fact that their norms are bounded by 1, together with Assumption 2.5, furnishes

$$\begin{aligned}
\|u\|^2_{L^2(\Omega)} &\leq \sum_{(i,j)\in I_*} \hat{a}^2_{i,j} \|\phi_{i,j}\|^2_{L^2(\Omega)} \\
&\leq \sum_{(i,j)\in I_* \setminus I_{M-1}} \hat{a}^2_{i,j} + \sum_{i=0}^{M-1} \sum_{j\geq M} \frac{C_1(i)^2}{j^{2s}} + \sum_{i\geq M} \sum_{j=0}^{M-1} \frac{C_2(j)^2}{i^{2s}} + \sum_{i\geq M} \sum_{j\geq M} \frac{C^2}{i^{2s} j^{2s}} \\
&\leq \sum_{(i,j)\in I_* \setminus I_{M-1}} \hat{a}^2_{i,j} + \sum_{i=0}^{M-1} \frac{C_1(i)^2 + C_2(i)^2}{(2s-1)(M-1)^{2s-1}} + \frac{C^2}{(2s-1)^2 (M-1)^{2(2s-1)}} \\
&< \infty,
\end{aligned}$$

which implies Definition 2.2 (ii).

Next we turn our attention to Definition 2.2 (iii). According to our definitions, the mapping $P^{(\ell,r)}$ denotes the orthogonal projection onto the finite-dimensional space $X_{\ell,r}$ defined in (19), and $F(u) = -\Delta(\Delta u + \lambda f(\mu + u))$ was introduced in (17). Due to (9) (see also section A) the composition $P^{(\ell,r)} \circ F$ is a polynomial in the first $\ell \cdot r - 1$ Fourier coefficients, and therefore Lipschitz continuous on $X_{\ell,r}$. This implies (iii).

Finally, we have to establish Definition 2.2 (iv). We begin by showing that

$$\prod_{(i,j)\in I_*} [a^-_{i,j}, a^+_{i,j}] \subset W^{1,2}(\Omega).$$

For this, choose $v = \sum_{(i,j)\in I_*} v_{i,j}\phi_{i,j} \in \prod_{(i,j)\in I_*}[a^-_{i,j}, a^+_{i,j}]$ arbitrary. Then our assumptions imply

$$\begin{aligned}
\|\partial_x v\|^2_{L^2(\Omega)} &= \sum_{(i,j)\in I_*} i^2\pi^2 v^2_{i,j}\|\sin(i\pi x)\cos(j\pi y)\|^2_{L^2(\Omega)} \\
&\le \sum_{i=1}^{M-1}\sum_{j=0}^{M-1} i^2\pi^2 v^2_{i,j} + \sum_{i=1}^{M-1}\sum_{j\ge M} i^2\pi^2\frac{C_1(i)^2}{j^{2s}} + \sum_{i\ge M}\sum_{j=0}^{M-1} i^2\pi^2\frac{C_2(j)^2}{i^{2s}} \\
&\quad + \sum_{i\ge M}\sum_{j\ge M} i^2\pi^2\frac{C^2}{i^{2s}j^{2s}} \\
&\le \sum_{i=1}^{M-1}\sum_{j=0}^{M-1} i^2\pi^2 v^2_{i,j} + \sum_{i=1}^{M-1}\frac{i^2\pi^2 C_1(i)^2}{(2s-1)(M-1)^{2s-1}} \\
&\quad + \sum_{j=0}^{M-1}\frac{\pi^2 C_2(j)^2}{(2s-3)(M-1)^{2s-3}} + \frac{\pi^2 C^2}{(2s-3)(2s-1)(M-1)^{4s-4}} < \infty.
\end{aligned}$$

Similarly, one can show that both $\partial_y v$ and v are contained in $L^2(\Omega)$, which immediately furnishes $v \in W^{1,2}(\Omega)$. In addition, we have the pointwise estimate

$$\begin{aligned}
|v(x)| &= \Big|\sum_{(i,j)\in I_*} v_{i,j}\phi_{i,j}(x)\Big| \le \sum_{(i,j)\in I_*} |v_{i,j}|\cdot|\phi_{i,j}(x)| \\
&\le \sum_{(i,j)\in I_*\setminus I_{M-1}} \hat{a}_{i,j} + \sum_{i=0}^{M-1}\frac{C_1(i)+C_2(i)}{(s-1)(M-1)^{s-1}} + \frac{C}{(s-1)^2(M-1)^{2(s-1)}} \\
&=: D < \infty, \qquad \text{for almost all } x\in\Omega,
\end{aligned} \tag{25}$$

where $\hat{a}_{i,j} = \max\{|a^\pm_{i,j}|\}$. For an arbitrary sequence $z^{(r)} \in \prod_{(i,j)\in I_*}[a^-_{i,j}, a^+_{i,j}]$ which converges to a limit function $z^{(\infty)}$ in $L^2(\Omega)$, there exists a subsequence $\{z^{(r_n)}\}_{n\in\mathbb{N}}$ such that for almost all $x\in\Omega$ we have $\lim_{n\to\infty} z^{(r_n)}(x) = z^{(\infty)}(x)$. This readily furnishes for almost all $x\in\Omega$ the identity $\lim_{n\to\infty}(\mu + z^{(r_n)}(x))^3 = (\mu + z^{(\infty)}(x))^3$, and due to (25) we have

$$\big|\big(\mu + z^{(r_n)}(x)\big)^3\big|^2 \le \big|\mu + z^{(r_n)}(x)\big|^6 \le D^6 \quad \text{almost everywhere,} \qquad \text{for all} \quad n\in\mathbb{N}.$$

The dominated convergence theorem now furnishes $\lim_{n\to\infty}(\mu + z^{(r_n)})^3 = (\mu + z^{(\infty)})^3$ in $L^2(\Omega)$, as well as

$$f(\mu + z^{(r_n)}) \xrightarrow{n\to\infty} f(\mu + z^{(\infty)}) \qquad \text{in} \qquad L^2(\Omega). \tag{26}$$

Now assume that the functions $z^{(r)} \in X_{r,r} \cap (W \times \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}])$, for $r > M$, satisfy the identities $P^{(r,r)}F(z^{(r)}) = 0$. Furthermore, assume that the sequence $(z^{(r)})_{r>M}$

has an accumulation point $z^{(\infty)}$ in $L^2(\Omega)$. If we now fix $(k,\ell) \in I_*$, then for sufficiently large n we have

$$0 = \mathcal{P}_{k,\ell} F\big(z^{(r_n)}\big) = (k^2+\ell^2)\pi^2\Big(-(k^2+\ell^2)\pi^2 \mathcal{P}_{k,\ell} z^{(r_n)} + \lambda \mathcal{P}_{k,\ell} f\big(\mu + z^{(r_n)}\big)\Big).$$

Due to the continuity of the projections $\mathcal{P}_{k,\ell}$ and (26), and after passing to a subsequence if necessary, one can pass to the limit $n \to \infty$, and this furnishes

$$(k^2+\ell^2)\pi^2\Big(-(k^2+\ell^2)\pi^2 \mathcal{P}_{k,\ell} z^{(\infty)} + \lambda \mathcal{P}_{k,\ell} f\big(\mu + z^{(\infty)}\big)\Big) = 0,$$

and therefore

$$\lambda \mathcal{P}_{k,\ell} f(\mu + z^{(\infty)}) = (k^2+\ell^2)\pi^2 \mathcal{P}_{k,\ell} z^{(\infty)}. \tag{27}$$

According to Theorem A.14, there exist collections of real numbers $\{b^{\pm}_{i,j}\}_{(i,j)\in\mathbb{N}_0^2}$, as well as positive constants B, $B_1(i)$, and $B_2(j)$, where $i,j \in \{0,\dots,M-1\}$, such that

$$\big|b^{\pm}_{i,j}\big| \le \begin{cases} \frac{B_1(i)}{j^s} & \text{for} \quad j \ge M \quad \text{and} \quad 0 \le i < M, \\ \frac{B_2(j)}{i^s} & \text{for} \quad i \ge M \quad \text{and} \quad 0 \le j < M, \\ \frac{B}{i^s j^s} & \text{for} \quad i,j \ge M, \end{cases}$$

as well as

$$f(\mu+u) \in \prod_{(i,j)\in\mathbb{N}_0^2} [b^-_{i,j}, b^+_{i,j}] \qquad \text{for all} \qquad u \in \prod_{(i,j)\in I_*} [a^-_{i,j}, a^+_{i,j}].$$

As before, one can show that in fact $\prod_{(i,j)\in\mathbb{N}_0^2}[b^-_{i,j}, b^+_{i,j}] \subset W^{1,2}(\Omega)$. This in turn implies $f(\mu + z^{(\infty)}) \in W^{1,2}(\Omega)$, and together with (27) we now obtain

$$\begin{aligned} -\Delta z^{(\infty)} &= \sum_{k,\ell=0}^{\infty} (k^2+\ell^2)\pi^2 \mathcal{P}_{k,\ell} z^{(\infty)} \\ &= \lambda \sum_{k,\ell=0}^{\infty} \mathcal{P}_{k,\ell} f\big(\mu + z^{(\infty)}\big) \\ &= \lambda f\big(\mu + z^{(\infty)}\big) \in W^{1,2}(\Omega). \end{aligned} \tag{28}$$

Now extend $z^{(\infty)}$ by even reflections at the boundary of Ω to an element of $W^{1,2}(\Lambda)$, where $\Lambda := (-1,2)^2$, and denote the resulting function again by $z^{(\infty)}$. Then we have $f(\mu + z^{(\infty)}) \in W^{1,2}(\Lambda)$, and in view of (28), the function $z^{(\infty)}$ satisfies

$$\Delta z^{(\infty)} = -\lambda f\big(\mu + z^{(\infty)}\big) \qquad \text{in } \Lambda$$

in the weak sense, i.e., we have

$$\int_\Lambda \big(\partial_x z^{(\infty)} \partial_x \varphi + \partial_y z^{(\infty)} \partial_y \varphi\big)\, dx\, dy = \int_\Lambda \lambda f\big(\mu + z^{(\infty)}\big)\varphi \, dx\, dy \quad \text{for all} \quad \varphi \in C_0^1(\Lambda).$$

For more details see [10, chapter 8, p. 177]. An application of [10, Theorem 8.8] immediately implies $z^{(\infty)} \in W^{2,2}(\Lambda')$ for any subdomain $\Lambda' \subset\subset \Lambda$, in particular $z^{(\infty)} \in W^{2,2}(\Omega)$. Next, define the functions

$$z^K := \sum_{(i,j)\in I_* \setminus I_K} \mathcal{P}_{i,j} z^{(\infty)} \phi_{i,j}.$$

Since $z^{(\infty)} \in W^{2,2}(\Omega)$, we obtain

$$z^K \to z^{(\infty)} \quad \text{in} \quad W^{2,2}(\Omega) \quad \text{for} \quad K \to \infty.$$

The functions $z^K \in C^\infty(\bar{\Omega})$ satisfy Neumann boundary conditions, and therefore are elements of $H^2_N(\Omega)$, and thus $z^{(\infty)} \in H^2_N(\Omega)$. (We refer the reader to [9] for the definition of the space $H^2_N(\Omega)$.) Sobolev's embedding theorem yields $z^{(\infty)} \in C^{0,\alpha}(\bar{\Omega})$, for $\alpha \in [0,1)$, and together we obtain $z^{(\infty)} \in C^{0,\alpha}(\bar{\Omega}) \cap H^2_N(\Omega)$. The latter implies $f(\mu + z^{(\infty)}) \in C^{0,\alpha}(\bar{\Omega})$, and therefore we obtain $z^{(\infty)} \in C^{2,\alpha}(\bar{\Omega}) \cap H^2_N(\Omega)$ (see the appendix in [9]), which in turn implies $f(\mu + z^{(\infty)}) \in C^{2,\alpha}(\bar{\Omega}) \cap H^2_N(\Omega)$. Now (28) furnishes $\Delta z^{(\infty)} \in C^{2,\alpha}(\bar{\Omega}) \cap H^2_N(\Omega)$, and by repeating the argument in [9] one finally obtains $z^{(\infty)} \in C^{4,\alpha}(\bar{\Omega}) \cap H^2_N(\Omega)$. From this we can deduce Definition 2.2 (iv), since $z^{(\infty)} \in D(F) = \{\, v \in C^4(\bar{\Omega}) \mid \partial_\nu v = \partial_\nu \Delta v = 0 \text{ on } \partial\Omega \,\}$, and due to (28), we have $F(z^{(\infty)}) = -\Delta(\Delta z^{(\infty)} + \lambda f(\mu + z^{(\infty)})) = 0$. □

2.3. Strict topologically self-consistent a priori bounds

In view of Theorem 2.4, we have thus far achieved only the first step towards establishing the existence of equilibria of (15). Theorem 2.6 furnishes in the situation of the Cahn-Hilliard equation on the square a sufficient condition for self-consistent a priori bounds, as defined in Definition 2.2. Of course, in order to apply Theorem 2.4, we need to verify strict topologically self-consistent a priori bounds as in Definition 2.3, in particular, we have to derive (23). This is the subject of the present subsection.

We begin by rewriting the original evolution equation (15). For this, let $L := DF(0)$ denote the linearization of F at 0. Then we can write (15) equivalently as

$$u_t = Lu + R(u), \tag{29}$$

where R denotes the nonlinear part of F. We make the following assumption.

Assumption 2.7. *Let $L := DF(0)$ denote the linearization of the function F in* (15). *Furthermore, assume that L has real eigenvalues $\{\kappa_{i,j}\}_{(i,j)\in I_*}$ which satisfy*

$$\kappa_{i,j} \geq \kappa_{k,l} \quad \text{for} \quad i \leq k \quad \text{and} \quad j \leq l, \qquad \text{as well as} \qquad \kappa_{i,i} \to -\infty \quad \text{for} \quad i \to \infty.$$

Denote the eigenfunction of L corresponding to $\kappa_{i,j}$ by $\phi_{i,j}$. Then we assume further that the collection $\{\phi_{i,j}\}_{(i,j)\in I_*}$ forms a complete orthogonal set in H. Finally, suppose that the eigenfunctions are chosen in such a way that

$$\|\phi_{i,j}\|_H \leq B \qquad \text{and} \qquad \|\phi_{i,j}\|_H \geq b > 0 \qquad \text{for all} \quad (i,j) \in I_*, \tag{30}$$

where B and b are positive constants.

One can easily see that for any elliptic and symmetric operator L the above assumption is satisfied. Using Assumption 2.7 we can expand any $u \in H$ and $R(u) \in H$ as

$$u = \sum_{(i,j)\in I_*} u_{i,j}\phi_{i,j} \qquad \text{and} \qquad R(u) = \sum_{(i,j)\in I_*} g_{i,j}\phi_{i,j}. \tag{31}$$

Due to (30), the ℓ^2-norm of the Fourier-coefficients of u and the norm of u on H, defined by the underlying scalar product, are equivalent.

Remark 2.8. In the specific situation of the Cahn-Hilliard equation, with $F = F_\mu$ given in (17), we obtain on the unit square $Lv = L_\mu v = (-\Delta)(\Delta v + \lambda f'(\mu)\cdot v)$ and $\kappa_{i,j} = (i^2+j^2)\pi^2(-(i^2+j^2)\pi^2 + \lambda f'(\mu))$. We want to point out that this used setup with functions in H with no mean (see (16)) is essential to have hyperbolic equilibria as needed in Theorem 2.14. Nevertheless, for computational reasons, it is more convenient to work in $L^2(\Omega)$, i.e., the considered functions do have a constant $\phi_{0,0}$ component given by the a priori fixed mean μ (cf. (55)).

In order to demonstrate how the estimates in (23) can be verified, we fix two positive integers m and M which satisfy

$$\kappa_{0,m} < 0 \quad \text{and} \quad \kappa_{m,0} < 0, \qquad \text{as well as} \qquad m < M. \tag{32}$$

In addition, choose positive constants C, $C_1(i)$, and $C_2(i)$, where $0 \le i < M$, let $s > 0$, let $\{a^{\pm}_{i,j}\}_{(i,j)\in I_*\setminus I_{M-1}}$ denote a collection of real numbers with $a^-_{i,j} < a^+_{i,j}$, and define

$$a^{\pm}_{i,j} := \begin{cases} \pm\frac{C_1(i)}{j^s} & \text{for} \quad j \ge M \quad \text{and} \quad 0 \le i < M, \\ \pm\frac{C_2(j)}{i^s} & \text{for} \quad i \ge M \quad \text{and} \quad 0 \le j < M, \\ \pm\frac{C}{i^s j^s} & \text{for} \quad i,j \ge M. \end{cases}$$

Finally, define the compact set $W = \prod_{(i,j)\in I_*\setminus I_m}[a^-_{i,j}, a^+_{i,j}]$. Then according to Theorem 2.6, for the specific situation of the Cahn-Hilliard equation on the unit square and for $s \ge 2$, the set W and the collection $\{a^{\pm}_{i,j}\}_{(i,j)\in I_m}$ are self-consistent a priori bounds for (15). Our goal is to refine these bounds in such a way that property (23) is satisfied. This will be accomplished by only adjusting the definition of the pairs $a^{\pm}_{i,j}$ for $(i,j) \in I_m$, i.e., for all pairs with $i > m$ or $j > m$. For this, we need the following assumption.

Assumption 2.9. *Consider the nonlinearity R in* (29) *and its expansion in* (31)*. We assume that it is possible to compute bounds for the coefficients $g_{i,j}$ of $R(u)$ for all $u \in W \times \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$ in the following sense. There exist constants $g^-_{i,j}$ and $g^+_{i,j}$, as well as positive constants G, $G_1(i)$, and $G_2(j)$, for $i,j \in \{0,\dots,M-1\}$, such that*

$$g^-_{i,j} < g_{i,j} < g^+_{i,j} \qquad \textit{for all} \qquad (i,j) \in I_* \setminus I_{M-1}$$

and

$$|g_{i,j}| < \begin{cases} \frac{G_1(i)}{j^s} & \text{for} \quad j \geq M \quad \text{and} 0 \leq i < M, \\ \frac{G_2(j)}{i^s} & \text{for} \quad i \geq M \quad \text{and} \quad 0 \leq j < M, \\ \frac{G}{i^s j^s} & \text{for} \quad i,j \geq M, \end{cases}$$

where $s > 0$ is chosen as above.

Notice that the constants $g^{\pm}_{i,j}$ do not necessarily have to have opposite signs. For the specific situation of the Cahn-Hilliard equation (1) on the unit square, Assumption 2.9 is satisfied for $s \geq 2$, and we refer the reader to Section A for details on the computation of these bounds. We would like to point out, however, that all these computations can be done explicitly.

By combining (29) and (31), we now see that (29) is equivalent to the infinite system of coupled ordinary differential equations given by

$$\dot{u}_{i,j} = \kappa_{i,j} u_{i,j} + g_{i,j}, \qquad \text{for all} \qquad (i,j) \in I_*. \tag{33}$$

We now turn our attention to computing sufficient conditions for the validity of (23). According to Definition 2.3, we need to satisfy the inequalities

$$\kappa_{i,j} a^+_{i,j} + g_{i,j} < 0 \quad \text{and} \quad \kappa_{i,j} a^-_{i,j} + g_{i,j} > 0, \qquad \text{whenever} \qquad i > m \quad \text{or} \quad j > m.$$

According to Assumption 2.7 and (32), these estimates are equivalent to the inequalities

$$a^+_{i,j} > -\frac{g_{i,j}}{\kappa_{i,j}} \qquad \text{and} \qquad a^-_{i,j} < -\frac{g_{i,j}}{\kappa_{i,j}}.$$

Using the definitions

$$\tilde{a}^+_{i,j} := -\frac{g^-_{i,j}}{\kappa_{i,j}} \quad \text{and} \quad \tilde{a}^-_{i,j} := -\frac{g^+_{i,j}}{\kappa_{i,j}} \quad \text{for} \quad (i,j) \in I_m \quad \text{with} \quad i < M, \ j < M, \tag{34}$$

as well as

$$\tilde{a}^{\pm}_{i,j} := \begin{cases} a^{\pm}_{i,j} & \text{for} \quad (i,j) \in I_* \setminus I_m \\ \pm\frac{\widetilde{C}_1(i)}{j^s} & \text{for} \quad j \geq M \quad \text{and} \quad 0 \leq i < M, \\ \pm\frac{\widetilde{C}_2(j)}{i^s} & \text{for} \quad i \geq M \quad \text{and} \quad 0 \leq j < M, \\ \pm\frac{\widetilde{C}}{i^s j^s} & \text{for} \quad i,j \geq M, \end{cases} \tag{35}$$

where $\widetilde{C}_1(i) := -C_1(i)/\kappa_{i,M}$, $\widetilde{C}_2(j) := -C_2(j)/\kappa_{M,j}$, and $\widetilde{C} := -C/\kappa_{M,M}$, we have just established the following result.

Lemma 2.10. *Assume that Assumptions 2.7 and 2.9 are satisfied, as well as* (32)*. Furthermore, suppose that the collection $\{\tilde{a}^{\pm}_{i,j}\}_{(i,j)\in I_*}$ is defined as in* (34) *and* (35)*,*

starting from a collection $\{a^{\pm}_{i,j}\}_{(i,j)\in I_*}$ *as described above. Then the original collection* $\{a^{\pm}_{i,j}\}_{(i,j)\in I_*}$ *satisfies property* (23) *of Definition 2.3 if*

$$a^+_{i,j} \geq \tilde{a}^+_{i,j} > \tilde{a}^-_{i,j} \geq a^-_{i,j} \qquad \text{for all} \qquad (i,j) \in I_*. \tag{36}$$

It is immediate that the refined bounds $\tilde{a}^{\pm}_{i,j}$ satisfy Assumption 2.5, and therefore they are self-consistent a priori bounds if the original collection $\{a^{\pm}_{i,j}\}_{(i,j)\in I_*}$ had that property. Thus, the above procedure can be repeated (possibly several times) with the adjusted constants, and at each step one can test for (23) with (36). This furnishes an iterative method for deciding the validity of (23) for given self-consistent a priori bounds.

Remark 2.11. Due to Theorem 2.6 we can use the procedure given in Lemma 2.10 particularly for the Cahn-Hilliard equation on the unit square with $s \geq 2$.

2.4. Isolating blocks and Conley index computation

In this final subsection we demonstrate how Definition 2.3 (ii) can be established, i.e., we construct an isolating block N in $W \subset X_{m,m}$. In addition, it is shown how the Conley index of the largest invariant set in N can be computed.

In order to avoid double indices in the following presentation, we introduce the bijective transformation $\hat{\sigma} : \mathbb{N}_0 \times \mathbb{N}_0 \to \mathbb{N}_0$ defined as

$$\hat{\sigma}(i,j) = \begin{cases} i + j(m+1) & \text{for} \quad 0 \leq i,j \leq m, \\ \sigma(i,j) & \text{otherwise,} \end{cases}$$

where $\sigma(i,j) = \max\{i,j\}^2 + j + (j-i)\mathbf{1}_{\{i<j\}}$. The bijection $\hat{\sigma}$ is introduced so that one can work with coordinate vectors in the following, rather than with "coordinate matrices" which arise when using the basis $\phi_{i,j}$ directly.

Assume that W and $\{a^{\pm}_{i,j}\}_{(i,j)\in I_m}$ are self-consistent a priori bounds for (15) which satisfy (23) of Definition 2.3. Let $n = (m+1)^2 - 1$ and let $v = (v_1, \ldots, v_n)$ be a hyperbolic equilibrium of (21) with $q = 0$ and $v \in W$. More precisely, suppose that the function

$$v = \sum_{l=1}^{n} v_l \phi_{\hat{\sigma}^{-1}(l)} \in X_{m,m}$$

solves (21) with $q = 0$. Our goal is to apply Theorem 2.4. For this, we have to construct strict topologically self-consistent a priori bounds for (15). In view of the previous subsections, this amounts to finding an isolating block $N \subset W$ with closed exit set N^- for (21), for all $q \in \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$. For this, we rewrite (21), or more precisely its interpretation in $\mathbb{R}^n \simeq X_{m,m}$, by using both (29) and (33) as

$$\dot{p}_k = \gamma_k(p_1, \ldots, p_n) + \epsilon_k, \qquad \text{for} \qquad k = 1, \ldots, n,$$

where $\gamma := (\gamma_1, \dots, \gamma_n) : \mathbb{R}^n \to \mathbb{R}^n$ denotes the right-hand side of (21) with $q = 0$, i.e., we have

$$\gamma_k(p) = \mathcal{P}_{\hat{\sigma}^{-1}(k)}(L + R(p, 0)),$$

as well as

$$\epsilon_k(p, q) = \mathcal{P}_{\hat{\sigma}^{-1}(k)}(R(p, q) - R(p, 0)).$$

We make the following assumption.

Assumption 2.12. *Assume that the function $\gamma : \mathbb{R}^n \to \mathbb{R}^n$ is differentiable at v and that the Jacobian $A := D\gamma(v) \in \mathbb{R}^{n\times n}$ is diagonalizable in $\mathbb{R}$, i.e., we can find an invertible matrix $B \in \mathbb{R}^{n\times n}$ with $A = BDB^{-1}$, where $D = (d_{ij})_{i,j=1,\dots,n} \in \mathbb{R}^{n\times n}$ is a diagonal matrix containing the eigenvalues of A, all of which are assumed to be real.*

Remark 2.13. It is possible, though more complicated, to compute an isolating block even if the Jacobian A has complex eigenvalues. However, for our application to the Cahn-Hilliard equation it suffices to only consider the case of real eigenvalues.

If we now define $x := p - v \in \mathbb{R}^n$ and use a Taylor expansion of γ around v, then we obtain

$$\dot{x} = \gamma(v) + Ax + \tilde{\epsilon}(x, q),$$

where the new "error" term $\tilde{\epsilon}$ contains, in addition to the old "error" term ϵ, also the higher-order terms in the expansion of γ, i.e., we have $\tilde{\epsilon}(x, q) = \epsilon(x, q) + o(x^2)$. Recall that γ is the right-hand side of (21) with $q = 0$, which implies $\gamma(v) = 0$, as well as

$$\dot{x} = BDB^{-1}x + \tilde{\epsilon}(x, q), \tag{37}$$

and with $y = B^{-1}x$ this furnishes

$$\dot{y} = Dy + B^{-1}\tilde{\epsilon}(By, q). \tag{38}$$

According to Assumption 2.9, we can compute estimates for the nonlinear term R. Therefore, we may assume that

$$(B^{-1}\tilde{\epsilon}(By, q))_k \subset [s_k, S_k],$$

for all $k \in \{1, \dots, n\}$, as well as all $x = By \in \widetilde{W} := W - v$ and $q \in \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$. Notice that if the equilibrium v is determined numerically, one additionally has to incorporate the numerical error into the estimate for $\tilde{\epsilon}$. This implies

$$d_{kk} \cdot \left(y_k + \frac{s_k}{d_{kk}} \right) < \dot{y}_k < d_{kk} \cdot \left(y_k + \frac{S_k}{d_{kk}} \right). \tag{39}$$

For all k with $d_{kk} < 0$ we now define $r_k := -s_k/d_{kk}$ and $R_k := -S_k/d_{kk}$, for the remaining values of k set $r_k := -S_k/d_{kk}$ and $R_k := -s_k/d_{kk}$. Recall that since v is a hyperbolic equilibrium, all eigenvalues d_{kk} are non-zero. If we now define

$$\widetilde{N} := \bigtimes_{k=1}^{n} [r_k, R_k] \subset B^{-1}(\widetilde{W}) \tag{40}$$

then $\widetilde{N}$ is an isolating block for (38). (Due to (39) the flow is transverse to the boundary of $\widetilde{N}$ at each point on the boundary.) Thus, the set $B(\widetilde{N}) \subset \widetilde{W}$ is an isolating block for (37) and $N := B(\widetilde{N}) + v$ is an isolating block for (21), for all $q \in \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$, which clearly satisfies $N \subset W$.

It remains to show that exit set N^- of N is closed. For this, we consider the exit set $\widetilde{N}^-$ of $\widetilde{N}$ and define

$$E := \{\, k \in \{\, 1, \dots, n \,\} \mid d_{kk} > 0 \,\}, \tag{41}$$

as well as

$$\widetilde{N}^- = \bigcup_{k\in E}\Big(\partial[r_k, R_k] \times \bigtimes_{\substack{1\le i\le n\\ i\neq k}} [r_i, R_i]\Big).$$

One can readily see that the set $\widetilde{N}^-$ is closed, and consequently also the set $B(\widetilde{N}^-)+v$ is closed. Yet, the latter is exactly the exit set N^-, since B is a change of variables that preserves eigenvalues and therefore the directions of the flows of the corresponding differential equations. Altogether, we have shown the following result.

Theorem 2.14. *Let W and $\{a^\pm_{i,j}\}_{(i,j)\in I_m}$ be self-consistent a priori bounds for* (21) *which satisfy* (23) *of Definition 2.3. Assume that W contains a hyperbolic equilibrium v of* (21) *for $q = 0$. Assume further that the set $N = B(\widetilde{N} + v)$ is derived as above from the set $\widetilde{N}$ in* (40)*. If $N \subset W$, then N, W, and $\{a^\pm_{i,j}\}_{(i,j)\in I_m}$ are strict topologically self-consistent a priori bounds for* (15)*.*

Combined with the previous sections, the above result furnishes a complete technique to find strict topologically self-consistent a priori bounds. In order to apply Theorem 2.4, one now only has to determine the Conley index $h(\mathrm{Inv}(N), \varphi_{q_0})$ with the help of the index pair (N, N^-). This will be accomplished in the remainder of this section.

Consider strict topologically self-consistent a priori bounds N, W, and $\{a^\pm_{i,j}\}_{(i,j)\in I_m}$, and assume that the isolating block N was obtained by the method described above. Hence, we have $N = B(\widetilde{N}) + v$, and the exit set N^- is closed and given by $N^- = B(\widetilde{N}^-) + v$. Since $B, B^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ are linear and therefore continuous functions, it is fairly easy to verify that

$$\big[(N/N^-, [N^-])\big] = \Big[\Big((B(\widetilde{N}) + v)/(B(\widetilde{N}^-) + v), [B(\widetilde{N}^-) + v]\Big)\Big] = \big[(\widetilde{N}/\widetilde{N}^-, [\widetilde{N}^-])\big].$$

Recall that according to Definition 2.3, the pair (N, N^-) is an index pair for equation (21) for all $q \in \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$. Now fix $q_0 \in \prod_{(i,j)\in I_m}[a^-_{i,j}, a^+_{i,j}]$ and let φ_{q_0} denote the corresponding flow of (21). Then we have

$$h\left(\mathrm{Inv}\left(N, \varphi_{q_0}\right)\right) = \big[(N/N^-, [N^-])\big] = \big[\big(\widetilde{N}/\widetilde{N}^-, [\widetilde{N}^-]\big)\big]. \tag{42}$$

Due to (40) and (41) we have both

$$\widetilde{N} = \bigtimes_{k=1}^{n} [r_k, R_k] \qquad \text{and} \qquad \widetilde{N}^- = \bigcup_{k\in E} \left(\partial[r_k, R_k] \times \bigtimes_{\substack{1\le i\le n \\ i\neq k}} [r_i, R_i] \right).$$

Now choose $z_0 \in \operatorname{int}(\widetilde{N})$ and define the diagonal matrix $D_E \in \mathbb{R}^{n\times n}$ by

$$(D_E)_{k,k} \coloneqq \begin{cases} 1 & \text{for} \quad k \in E, \\ -1 & \text{otherwise,} \end{cases}$$

where E is defined as in (41). Finally, consider the equation

$$\dot{z} = D_E(z - z_0),$$

with induced flow ψ. This system has a hyperbolic equilibrium z_0 with $|E|$ positive eigenvalues. Furthermore, the set $(\widetilde{N}, \widetilde{N}^-)$ is an index pair for $\operatorname{Inv}(\widetilde{N}, \psi) = \{z_0\}$. But this immediately implies

$$[\Sigma^{|E|}] = h(\{z_0\}) = h(\operatorname{Inv}(\widetilde{N}, \psi)) = \big[(\widetilde{N}/\widetilde{N}^-, [\widetilde{N}^-])\big],$$

and together with (42) one finally obtains for all $q \in \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$ the identity

$$h(\operatorname{Inv}(N, \varphi_q)) = [\Sigma^{|E|}].$$

In other words, the Conley index takes the form required in Corollary 2.4, and we are guaranteed that the set $N \times \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$ contains an equilibrium v^* of (15).

Remark 2.15. It should be pointed out that it is not necessary that $N \times \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$ contains the hyperbolic equilibrium $(v, 0)$ of (21). The latter is only used as a starting point for the iteration mentioned earlier.

3. Uniqueness of equilibrium solutions

In this section we demonstrate how one can establish the uniqueness of the equilibrium guaranteed by Theorem 2.4 in the set $N \times \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$. In order to make our presentation as simple as possible, we will avoid the double index notation and consider a linearly ordered complete orthogonal set $\{\psi_k\}_{k\in\mathbb{N}_0}$ in the underlying Hilbert space H, similar to our proceeding in section 2.4.

To fix our notation, let $X \subseteq D(F) \subseteq H$ denote a suitable Banach space containing the orthogonal basis $\{\psi_k\}_{k\in\mathbb{N}_0}$, and as in (30) we assume that

$$\|\psi_k\|_H \le B \qquad \text{and} \qquad \|\psi_k\|_H \ge b > 0 \qquad \text{for all} \qquad k \in \mathbb{N}_0,$$

where b and B are positive constants. Moreover, similar to (20) we define the operator

$$\mathcal{P}_k : H \longrightarrow \mathbb{R} \qquad \text{by} \qquad \mathcal{P}_k(u) := \frac{(u,\psi_k)}{(\psi_k,\psi_k)},$$

where $(\cdot,\cdot)$ denotes the scalar product in H. Then for all $x \in X$ the sequence $\{x_k\}_{k\in\mathbb{N}_0}$ with $x_k = \mathcal{P}_k x$ converges to zero as $k \to \infty$. For $T \in \mathcal{L}(X,H)$ we now set

$$t_{i,j} := \mathcal{P}_i(T\psi_j),$$

which readily implies

$$Tx = \sum_{j=0}^{\infty} x_j T\psi_j = \sum_{i=0}^{\infty} \mathcal{P}_i\Big(\sum_{j=0}^{\infty} x_j T\psi_j\Big)\psi_i = \sum_{i,j=0}^{\infty} t_{i,j}x_j\psi_i,$$

since T and the projections $\mathcal{P}_k$ are continuous. Here we used again the definition $x_k = \mathcal{P}_k x$ yielding $x = \sum_{j=0}^{\infty} x_j\psi_j$. We begin by proving the following auxiliary result.

Lemma 3.1. *Consider the notation introduced above, and assume that*

$$|t_{k,k}| > \sum_{i\in\mathbb{N}_0\setminus\{k\}} |t_{k,i}| \qquad \textit{for all} \qquad k \in \mathbb{N}_0.$$

Then we have $\|Tx\|_H > 0$ *for all* $x \in X \setminus \{0\}$.

Proof. Let $x = \sum_{i\in\mathbb{N}_0} x_i\psi_i \in X \setminus \{0\}$ be arbitrary and consider an index $i_0 \in \mathbb{N}_0$ such that $|\mathcal{P}_{i_0}x| = |x_{i_0}| > 0$. Then there exists a $p \in \mathbb{N}_0$ such that $|x_q| < |x_{i_0}|$ for all $q > p$. Using the abbreviation $\mu_* := \max_{0\le k\le p}|x_k|$ one obtains $\mu_* \ge |x_{i_0}| > |x_q|$ for all $q > p$. This in turn implies that $\mu_* \ge |x_k|$ for all $k \in \mathbb{N}_0$, and according to the definition of μ_* there exists an index $k_0 \le p$ with $|x_{k_0}| > 0$ and $|x_{k_0}| \ge |x_k|$ for all $k \in \mathbb{N}_0$. This furnishes

$$\begin{aligned}
|\mathcal{P}_{k_0}(Tx) - t_{k_0,k_0}x_{k_0}| &= \Big|\sum_{i\in\mathbb{N}_0} t_{k_0,i}x_i - t_{k_0,k_0}x_{k_0}\Big| = \Big|\sum_{i\in\mathbb{N}_0\setminus\{k_0\}} t_{k_0,i}x_i\Big| \\
&\le \sum_{i\in\mathbb{N}_0\setminus\{k_0\}} |t_{k_0,i}|\cdot|x_i| \le |x_{k_0}|\cdot\sum_{i\in\mathbb{N}_0\setminus\{k_0\}} |t_{k_0,i}| \\
&< |x_{k_0}|\cdot|t_{k_0,k_0}| = |t_{k_0,k_0}x_{k_0}|.
\end{aligned}$$

Therefore we have both $0 < |t_{k_0,k_0}x_{k_0}| - |\mathcal{P}_{k_0}(Tx) - t_{k_0,k_0}x_{k_0}| \le |\mathcal{P}_{k_0}(Tx)|$ and

$$\|Tx\|_H^2 = \sum_{i\in\mathbb{N}_0} |\mathcal{P}_i(Tx)|^2\|\psi_i\|_H^2 \ge |\mathcal{P}_{k_0}(Tx)|^2\cdot\|\psi_{k_0}\|_H^2 > 0,$$

which completes the proof of the lemma. □

The above Lemma 3.1 can be viewed as an infinite-dimensional extension of a theorem of Gershgorin [21] about the location of eigenvalues of matrices. For our applications, we also need the following extension of the lemma.

Lemma 3.2. *Consider the notation introduced above, and let $U \subset X$ denote an open set. Furthermore, let $T : U \to \mathcal{L}(X, H)$ be continuous and set $t(u)_{i,j} \coloneqq \mathcal{P}_i(T(u)\psi_j)$. Finally, suppose that*

$$|t(u)_{k,k}| > \sum_{i \in \mathbb{N}_0 \setminus \{k\}} |t(u)_{k,i}| \qquad \text{for all} \quad k \in \mathbb{N}_0 \quad \text{and} \quad u \in U.$$

Then for every continuous curve $\gamma : [0,1] \to U$ and for all $x \in X \setminus \{0\}$ we have

$$\left\| \int_0^1 T(\gamma(t))x \, dt \right\|_H > 0.$$

Proof. Let $x = \sum_{i \in \mathbb{N}_0} x_i \psi_i \in X \setminus \{0\}$ be arbitrary, and let k_0 be as in the proof of Lemma 3.1. Then we obtain $\mathcal{P}_{k_0}(T(u)x) \neq 0$ for all $u \in U$ as in Lemma 3.1, and this immediately shows that $\mathcal{P}_{k_0}(T(u)x)$ does not change sign on any connected component of U. Due to

$$\int_0^1 T(\gamma(t))x \, dt = \sum_{k=0}^{\infty} \left(\int_0^1 \mathcal{P}_k(T(\gamma(t))x) \, dt \right) \psi_k$$

this furnishes

$$\begin{aligned} \left\| \int_0^1 T(\gamma(t))x \, dt \right\|_H^2 &= \sum_{k=0}^{\infty} \left(\int_0^1 \mathcal{P}_k(T(\gamma(t))x) \, dt \right)^2 \cdot \|\psi_k\|_H^2 \\ &\geq \left(\int_0^1 \mathcal{P}_{k_0}(T(\gamma(t))x) \, dt \right)^2 \cdot \|\psi_{k_0}\|_H^2 > 0, \end{aligned}$$

and the proof of the lemma is complete. □

Now assume that $U \subseteq X$ is convex, suppose that F is Fréchet differentiable in U with derivative $DF(u) \in \mathcal{L}(X, H)$, and set

$$z(u)_{i,j} \coloneqq \mathcal{P}_i(DF(u)\psi_j).$$

Then we have the following result.

Theorem 3.3. *Let $v \in U$ be an equilibrium of (29), i.e., assume that $u = v$ solves*

$$F(u) = Lu + R(u) = 0. \tag{43}$$

If in addition we have

$$|z(u)_{k,k}| > \sum_{i \in \mathbb{N}_0 \setminus \{k\}} |z(u)_{k,i}| \qquad \text{for all} \quad k \in \mathbb{N}_0 \quad \text{and} \quad u \in U, \tag{44}$$

then v is the unique solution of (43) in U.

Proof. Assume there exist two equilibria $v = \sum_{i\in\mathbb{N}_0} v_i\psi_i \in U$ and $w = \sum_{i\in\mathbb{N}_0} w_i\psi_i \in U$ with $w \neq v$. Due to $F(w) = F(v) = 0$, the integral version of the mean value theorem implies

$$\int_0^1 DF(w + t(v - w))\, dt \cdot (v - w) = 0,$$

which contradicts Lemma 3.2 if we set $\gamma(t) = w + t(v - w)$, since $v - w \neq 0$. □

It is well-known that Gershgorin's theorem is a very rough tool for establishing the invertibility of a matrix. Since its infinite-dimensional extension forms the basis for our uniqueness test, it is therefore not surprising that condition (44) often fails — even if our solution seems to be unique. On the other hand, we usually have good information on the finite-dimensional part $A(u) \in \mathbb{R}^{M\times M}$ of the operator $DF(u)$, where

$$(A(u))_{ij} \coloneqq z(u)_{i,j}, \qquad \text{for all} \qquad i,j \in \{0,\dots,M-1\}.$$

By making use of this information, we can improve our approach. Notice that Theorem 3.3 can only be applied if the diagonal of the infinite-dimensional matrix representing $DF(u)$ is extremely dominant. This may not always be the case, especially for the diagonal entries of the above-mentioned finite-dimensional part. In the following we therefore introduce some sort of preconditioning to diagonalize the finite-dimensional part of that matrix.

Assumption 3.4. *Assume that there exists a function $u_0 \in X$ such that $A(u_0)$ is diagonalizable, i.e., suppose there exists an invertible matrix $B = (b_{ij})_{i,j=0,\dots,M-1} \in \mathbb{R}^{M\times M}$ such that $A(u_0) = BDB^{-1}$, where $D = (d_{ij})_{i,j=0,\dots,M-1} \in \mathbb{R}^{M\times M}$ is a diagonal matrix which contains the real eigenvalues of $A(u_0)$.*

In most cases u_0 is the numerically computed equilibrium of the finite-dimensional system (21) with $q = 0$. Note that it is not necessary for the following that u_0 lies in the set U. In fact, it suffices to assume that u_0 lies in X and that F is Fréchet differentiable at u_0. See also Remark 2.15. Let $(\tilde{b}_{ij})_{i,j=0,\dots,M-1}$ denote the entries of B^{-1}. We define

$$\Xi(u) \coloneqq B^{-1}(A(u) - A(u_0))B \in \mathbb{R}^{M\times M},$$

and denote its coefficients by $(\xi(u)_{ij})_{i,j=0,\dots,M-1}$. In addition, define the bijective continuous linear operator $Q : H \to H$ by

$$\mathcal{P}_i(Q\psi_j) = \begin{cases} b_{ij} & \text{for} \quad 0 \le i,j < M, \\ \delta_{ij} & \text{otherwise.} \end{cases}$$

Since we assume that X contains the orthogonal basis $\{\psi_k\}_{k\in\mathbb{N}_0}$, we have $Q(X) = X$. The inverse of Q is given by

$$\mathcal{P}_i\left(Q^{-1}\psi_j\right) = \begin{cases} \tilde{b}_{ij} & \text{for} \quad 0 \le i,j < M, \\ \delta_{ij} & \text{otherwise.} \end{cases}$$

Using this notation, we can now present a refined tool for checking the uniqueness of an equilibrium.

Theorem 3.5. *Using the notation introduced above, suppose that Assumption 3.4 holds and assume that $v \in U$ is an equilibrium of (29), i.e., we have $Lv + R(v) = 0$. In addition, assume that for all $u \in U$ we have*

$$|d_{kk}| > \sum_{i=M}^{\infty} \left| \sum_{l=0}^{M-1} \tilde{b}_{kl} z(u)_{l,i} \right| + \sum_{i=0}^{M-1} |\xi(u)_{ki}| \tag{45}$$

for $k \in \{0, \dots, M-1\}$, and that for $k \geq M$ we have

$$|z(u)_{k,k}| > \sum_{i=0}^{M-1} \left| \sum_{l=0}^{M-1} z(u)_{k,l} b_{li} \right| + \sum_{\substack{i \geq M \\ i \neq k}} |z(u)_{k,i}|. \tag{46}$$

Then v is the unique solution of (43) in U.

Proof. Let $u \in U$ be arbitrary and consider the operator $Q^{-1}DF(u)Q \in \mathcal{L}(X, H)$. Then for all $0 \leq i, j < M$ one obtains

$$\begin{aligned}
\mathcal{P}_i(Q^{-1}DF(u)Q\psi_j) &= \mathcal{P}_i Q^{-1} DF(u) \sum_{k \in \mathbb{N}_0} \mathcal{P}_k(Q\psi_j)\psi_k = \mathcal{P}_i Q^{-1} DF(u) \sum_{k=0}^{M-1} b_{kj}\psi_k \\
&= \mathcal{P}_i Q^{-1} \sum_{k=0}^{M-1} b_{kj} DF(u)\psi_k = \mathcal{P}_i Q^{-1} \sum_{k=0}^{M-1} b_{kj} \sum_{l \in \mathbb{N}_0} z(u)_{l,k}\psi_l \\
&= \mathcal{P}_i \sum_{k,l=0}^{M-1} z(u)_{l,k} b_{kj} Q^{-1}\psi_l + \mathcal{P}_i \sum_{l=M}^{\infty} \sum_{k=0}^{M-1} z(u)_{l,k} b_{kj} Q^{-1}\psi_l \\
&= \mathcal{P}_i \sum_{k,l=0}^{M-1} z(u)_{l,k} b_{kj} \sum_{n=0}^{M-1} \tilde{b}_{nl}\psi_n = \sum_{k,l=0}^{M-1} \tilde{b}_{il} z(u)_{l,k} b_{kj} \\
&= \sum_{k,l=0}^{M-1} \tilde{b}_{il} \big(z(u_0)_{l,k} + z(u)_{l,k} - z(u_0)_{l,k} \big) b_{kj} = d_{ij} + \xi(u)_{ij}.
\end{aligned}$$

Similarly, for $0 \leq i < M$ and $j \geq M$ one obtains

$$\begin{aligned}
\mathcal{P}_i\big(Q^{-1}DF(u)Q\psi_j\big) &= \mathcal{P}_i Q^{-1} DF(u)\psi_j = \mathcal{P}_i Q^{-1} \sum_{l \in \mathbb{N}_0} z(u)_{l,j}\psi_l \\
&= \mathcal{P}_i Q^{-1} \sum_{l=0}^{M-1} z(u)_{l,j}\psi_l + \mathcal{P}_i Q^{-1} \sum_{l=M}^{\infty} z(u)_{l,j}\psi_l \\
&= \mathcal{P}_i \sum_{l=0}^{M-1} z(u)_{l,j} \sum_{n=0}^{M-1} \tilde{b}_{nl}\psi_n = \sum_{l=0}^{M-1} \tilde{b}_{il} z(u)_{l,j},
\end{aligned}$$

for $i \geq M$ and $0 \leq j < M$ one has

$$\begin{aligned}\mathcal{P}_i\big(Q^{-1}DF(u)Q\psi_j\big) &= \mathcal{P}_i \sum_{k,l=0}^{M-1} z(u)_{k,l} b_{lj} Q^{-1}\psi_k + \mathcal{P}_i \sum_{k=M}^{\infty}\sum_{l=0}^{M-1} z(u)_{k,l} b_{lj} Q^{-1}\psi_k \\ &= \sum_{l=0}^{M-1} z(u)_{i,l} b_{lj},\end{aligned}$$

and finally for $i \geq M$ and $j \geq M$ one has

$$\mathcal{P}_i\big(Q^{-1}DF(u)Q\psi_j\big) = \mathcal{P}_i Q^{-1} \sum_{l\in\mathbb{N}_0} z(u)_{l,j}\psi_l = z(u)_{i,j}.$$

An application of (45) now furnishes for all $0 \leq k < M$ the estimate

$$\begin{aligned}\sum_{i\in\mathbb{N}_0\setminus\{k\}} \big|\mathcal{P}_k\big(Q^{-1}DF(u)Q\psi_i\big)\big| &= \sum_{i=M}^{\infty}\left|\sum_{l=0}^{M-1}\tilde{b}_{kl} z(u)_{l,i}\right| + \sum_{\substack{0\leq i<M\\ i\neq k}} |\xi(u)_{ki}| \\ &< |d_{kk}| - |\xi(u)_{kk}| \leq |d_{kk} + \xi(u)_{kk}| \\ &= \big|\mathcal{P}_k\big(Q^{-1}DF(u)Q\psi_k\big)\big|,\end{aligned}$$

and (46) implies for all $k \geq M$ the estimate

$$\begin{aligned}\sum_{i\in\mathbb{N}_0\setminus\{k\}} \big|\mathcal{P}_k\big(Q^{-1}DF(u)Q\psi_i\big)\big| &= \sum_{i=0}^{M-1}\left|\sum_{l=0}^{M-1} z(u)_{k,l} b_{li}\right| + \sum_{\substack{i\geq M\\ i\neq k}} |z(u)_{k,i}| \\ &< |z(u)_{k,k}| = \big|\mathcal{P}_k\big(Q^{-1}DF(u)Q\psi_k\big)\big|.\end{aligned}$$

Using Lemma 3.2 we now obtain $\int_0^1 Q^{-1}DF(\gamma(t))Qx\,dt \neq 0$ for all $x \in X\setminus\{0\}$, which immediately implies $\int_0^1 DF(\gamma(t))x\,dt \neq 0$ for all $x \in X\setminus\{0\}$. Now the result follows as in the proof of Theorem 3.3. □

We now reformulate the results of this section for the case of double index basis functions. For this, define

$$z(u)^{p,q}_{r,s} := \mathcal{P}_{p,q}(DF(u)\phi_{r,s})$$

and consider the finite part $A(u) \in \mathbb{R}^{(M^2-1)\times(M^2-1)}$ of $DF(u)$, with coefficients

$$(A(u))_{i,j=1,\ldots,M^2-1} = z(u)^{\hat{\sigma}^{-1}(i)}_{\hat{\sigma}^{-1}(j)},$$

where $\hat{\sigma} : \mathbb{N}_0\times\mathbb{N}_0 \to \mathbb{N}_0$ is a bijective transformation similar to the one introduced in section 2.4. This time we consider

$$\hat{\sigma}(k,l) = \begin{cases} k + lM & \text{for} \quad 0 \leq k,l < M, \\ \sigma(k,l) & \text{otherwise.}\end{cases}$$

The analogue of Assumption 3.4 now takes the following form.

Assumption 3.6. *Assume that there exists a function $u_0 \in X$, as well as an invertible matrix $B = (b_{ij})_{i,j=1,\dots,M^2-1} \in \mathbb{R}^{(M^2-1)\times(M^2-1)}$ such that the identity $A(u_0) = BDB^{-1}$ holds, where $D = (d_{ij})_{i,j=1,\dots,M^2-1} \in \mathbb{R}^{(M^2-1)\times(M^2-1)}$ denotes a diagonal matrix which contains the real eigenvalues of $A(u_0)$.*

In most cases, u_0 is the numerically computed solution of the finite-dimensional system (21). The entries of the inverse matrix B^{-1} are denoted by $(\tilde{b}_{ij})_{i,j=1,\dots,M^2-1}$, and we define

$$\Xi(u) \coloneqq B^{-1}\big(A(u) - A(u_0)\big)B \in \mathbb{R}^{(M^2-1)\times(M^2-1)},$$

with coefficients $(\xi(u)_{ij})_{i,j=1,\dots,M^2-1}$. Finally, define the bijective continuous linear operator $Q : H \to H$ as

$$\mathcal{P}_{p,q}(Q\phi_{r,s}) = \begin{cases} b_{\hat{\sigma}(p,q)\hat{\sigma}(r,s)} & \text{for} \quad 0 \le p,q,r,s < M, \\ \delta_{pr}\delta_{qs} & \text{otherwise.} \end{cases}$$

As before, the Banach space $X \subseteq D(F) \subseteq H$ is chosen in such a way that it contains the orthogonal basis $\{\phi_{i,j}\}_{(i,j)\in I_*}$. The inverse of Q is given by

$$\mathcal{P}_{p,q}\left(Q^{-1}\phi_{r,s}\right) = \begin{cases} \tilde{b}_{\hat{\sigma}(p,q)\hat{\sigma}(r,s)} & \text{for} \quad 0 \le p,q,r,s < M, \\ \delta_{pr}\delta_{qs} & \text{otherwise.} \end{cases}$$

In the new setting, Theorem 3.5 can now be restated as follows.

Theorem 3.7. *Using the notation introduced above, suppose that Assumption 3.6 holds and assume that $v \in U$ is an equilibrium of (29) on a two-dimensional domain. In addition, assume that for all $u \in U$ we have*

$$|d_{\hat{\sigma}(k,l)\hat{\sigma}(k,l)}| > \sum_{(r,s)\in I_{M-1}} \left| \sum_{(p,q)\in I_*\setminus I_{M-1}} \tilde{b}_{\hat{\sigma}(k,l)\hat{\sigma}(p,q)} z(u)^{p,q}_{r,s} \right| + \sum_{j=1}^{M^2-1} |\xi(u)_{\hat{\sigma}(k,l)j}|$$

for $(k,l) \in I_* \setminus I_{M-1}$, and that otherwise we have

$$|z(u)^{k,l}_{k,l}| > \sum_{(p,q)\in I_*\setminus I_{M-1}} \left| \sum_{(r,s)\in I_*\setminus I_{M-1}} z(u)^{k,l}_{r,s} b_{\hat{\sigma}(r,s)\hat{\sigma}(p,q)} \right| + \sum_{(r,s)\in I_{M-1}\setminus\{(k,l)\}} |z(u)^{k,l}_{r,s}|.$$

Then v is the unique solution of the two-dimensional form of (43) in U.

4. Rigorous path-following

In this section we combine the method described in the preceding sections with a path-following algorithm — such as for example [14] — in order to rigorously compute branches of solutions. We consider the equation

$$u_t = F(u, \lambda), \tag{47}$$

in the Hilbert space H, where F depends continuously on the real parameter λ. By orthogonal projection onto the spaces $X_{m,m}$ and $Y_{m,m}$ one obtains as in (21) and (22) the coupled system

$$\begin{aligned} p_t &= P^{(m,m)} F((p+q), \lambda), \\ q_t &= Q^{(m,m)} F((p+q), \lambda). \end{aligned} \tag{48}$$

Central to our proceeding is the following result.

Theorem 4.1. *Assume that N, W, and $\{a^{\pm}_{i,j}\}_{(i,j)\in I_m}$ are strict topologically self-consistent a priori bounds for (47) as introduced in Definition 2.3, for all $\lambda \in [\lambda^-, \lambda^+] \subset \mathbb{R}$. If we have*

$$h(\mathrm{Inv}(N, \varphi_{q_0,\lambda_0})) = [\Sigma^{l_0}]$$

for some $l_0 \in \mathbb{N}_0$, $q_0 \in \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$ and $\lambda_0 \in [\lambda^-, \lambda^+]$, then for all $\lambda \in [\lambda^-, \lambda^+]$ there exist equilibria $v^*_\lambda \in N \times \prod^{\infty}_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$ of (47).

Proof. According to our assumption, the set N is an isolating block with closed exit set for all $\lambda \in [\lambda^-, \lambda^+]$ and all $q \in \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$. Due to the continuation property of the Conley index [19], we then obtain

$$h(\mathrm{Inv}(N, \varphi_{q_0,\lambda_0})) = h(\mathrm{Inv}(N, \varphi_{q_0,\lambda})),$$

for all $\lambda \in [\lambda^-, \lambda^+]$. Hence,

$$h(\mathrm{Inv}(N, \varphi_{q_0,\lambda})) = [\Sigma^{l_0}]$$

for all $\lambda \in [\lambda^-, \lambda^+]$, and Theorem 2.4 furnishes an equilibrium $v^*_\lambda \in N \times \prod_{(i,j)\in I_m} [a^-_{i,j}, a^+_{i,j}]$ for each λ. □

Our strategy for rigorously computing branches is as follows, see also figure 32. First, we compute an equilibrium of the finite-dimensional equation (48) with $q = 0$ and $\lambda = \lambda_0$. Then we try to find bounds nearby, which are strict topologically self-consistent a priori bounds for all $\lambda \in [\lambda^-, \lambda^+]$. By Theorem 4.1 we then obtain a solution, possibly unique, for each $\lambda \in [\lambda^-, \lambda^+]$. Using a path-following algorithm we now compute another solution of (48) with $q = 0$ and search again for strict

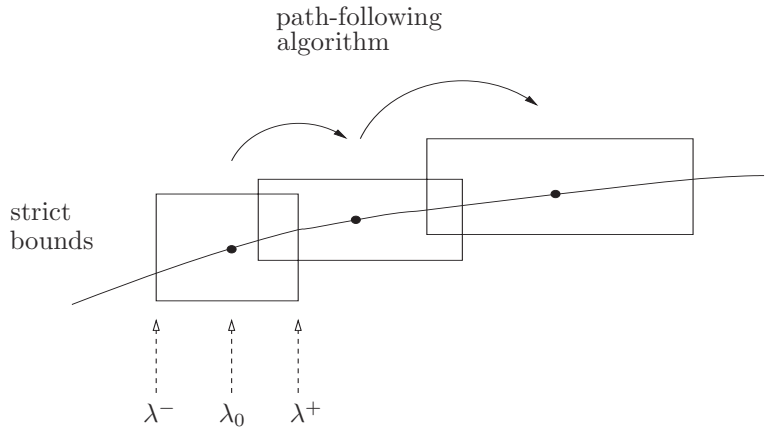


Figure 32 – Strategy for rigorous path-following

topologically self-consistent a priori bounds in a certain λ-interval, in order to apply Theorem 4.1 again. The goal is to cover the whole branch with such λ-intervals.

We now turn our attention to the Cahn-Hilliard equation (1). The equilibria of (1) are given by the three-parameter problem (5) and (2), and we consider the two-dimensional case $\Omega = (0,1)^2$. Note that $\mu = \frac{1}{|\Omega|}\int_\Omega u(x)\,dx = \mathcal{P}_{0,0}u$. The general strategy for path-following a three-parameter problem is to fix one parameter, advance the second parameter, and adjust the third one, as well as all the variables related to the solution. Our goal is to compute paths of solutions of (5), or equilibria of (1), using the strategy described above. Assume we know an equilibrium of the Cahn-Hilliard version of (48) with $q = 0$ and parameters (μ_0, λ_0, c_0). If, for example, we want to determine a piece of the path in λ-direction with fixed mass $\mu = \mu_0$, then we have to take into account an additional error term in (9) (see also (55) and Lemma A.1), namely

$$\begin{aligned}\dot{u}_{i,j} &= -(i^2+j^2)^2\pi^4 u_{i,j} + \lambda(i^2+j^2)\pi^2\Bigg(u_{i,j} - \frac{1}{4}c_{i,j}\sum_{p,q,r,s\in\mathbb{Z}}\tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}\Bigg)\\ &= -(i^2+j^2)^2\pi^4 u_{i,j} + \lambda_0(i^2+j^2)\pi^2\Bigg(u_{i,j} - \frac{1}{4}c_{i,j}\sum_{p,q,r,s\in\mathbb{Z}}\tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}\Bigg)\\ &\quad + \epsilon_{i,j}(\lambda,u),\end{aligned} \tag{49}$$

where

$$\epsilon_{i,j}(\lambda,u) \coloneqq (\lambda-\lambda_0)(i^2+j^2)\pi^2\Bigg(u_{i,j} - \frac{1}{4}c_{i,j}\sum_{p,q,r,s\in\mathbb{Z}}\tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}\Bigg),$$

with $i,j \in \mathbb{N}_0$ and λ in a certain interval $[\lambda^-,\lambda^+]$ around λ_0. Of course, one can

compute bounds for $\epsilon_{i,j}(\lambda, u)$ for all $\lambda \in [\lambda^-, \lambda^+]$ by employing the estimates from section A for the infinite sum. Hence, in order to determine a path of solutions v^*_λ of (5) in a predetermined interval $[\lambda^-, \lambda^+]$ around λ_0, we have to apply our method to (49) with the additional error term $\epsilon_{i,j}(\lambda, u)$. Upon success, one then computes another approximate solution by a path-following algorithm and proceeds as before. In this way one can try to cover the whole branch with λ-intervals. See also figure 32.

For paths in μ-direction with fixed λ, no new error term is necessary. One only has to realize that $\mu = \mathcal{P}_{0,0}u$ and that the estimates of section A are not restricted to fixed $\mathcal{P}_{0,0}u$. Therefore, one can simply use the bounds that are computed there.

Unfortunately, it is not possible to use the above strategy if one wants to consider the parameter c, neither for path-following in c-direction nor for following μ with fixed c. This parameter does not appear in (1) explicitly, so one cannot incorporate a fixed c_0 or a predetermined interval $[c^-, c^+]$ into the equation. Yet, in order to avoid this problem, one can consider the Allen-Cahn equation

$$\begin{aligned} u_t &= \Delta u + \lambda f(u) - \lambda c \qquad \text{in} \quad \Omega, \\ \partial_\nu u &= 0 \qquad\qquad\qquad\quad \text{on} \quad \partial\Omega. \end{aligned} \tag{50}$$

The equilibria of the Allen-Cahn equation and of the Cahn-Hilliard equation coincide. Also, Theorem 2.6 applies to (50) as well, and due to Lemma A.1 we have (see (54) and (53))

$$\dot{u}_{i,j} = -(i^2+j^2)\pi^2 u_{i,j} + \lambda\left(u_{i,j} - \frac{1}{4}c_{i,j} \sum_{p,q,r,s\in\mathbb{Z}} \tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}\right), \tag{51}$$

for $(i,j) \in I_*$. In addition, since $c_{0,0} = \frac{1}{16}$ by Lemma A.1, we have

$$\dot{u}_{0,0} = \lambda\left(u_{0,0} - \frac{1}{64} \sum_{p,q,r,s\in\mathbb{Z}} \tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{p+r,q+s}\right) - \lambda c, \tag{52}$$

because the mass of Allen-Cahn solutions may change with time. Now, if we search for a path in c-direction around c_0 with fixed $\lambda = \lambda_0$, then we replace (52) by

$$\dot{u}_{0,0} = \lambda\left(u_{0,0} - \frac{1}{64} \sum_{p,q,r,s\in\mathbb{Z}} \tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{p+r,q+s}\right) - \lambda c_0 + \epsilon_{0,0}(c,\lambda),$$

where

$$\epsilon_{0,0}(c,\lambda) \coloneqq \lambda(c_0 - c).$$

Note that rigorous path-following in c with fixed μ is not possible with this technique. If we restrict ourselves to a fixed μ, we lose the information gathered by considering (50), namely the dynamics in u_0. Hence, we cannot take equation (52) into account and are again left with the problem that c does not occur explicitly in equations (51). Rigorous path-following for all the other combinations of the parameters is possible, see table 11.

Table 11 – Parameter combinations for rigorous path-following

fixed	advanced	adjusted	technique
λ	μ	c	method with Cahn-Hilliard or Allen-Cahn equation
	c	μ	method with Allen-Cahn equation
μ	λ	c	method with Cahn-Hilliard or Allen-Cahn equation
	c	λ	not possible
c	λ	μ	method with Allen-Cahn equation
	μ	λ	method with Allen-Cahn equation

Appendix

A. Estimates for the truncation error

As we have seen in section 2, our method for establishing computer-assisted existence proofs for equilibria of the Cahn-Hilliard equation relies crucially on the ability to control the truncation error which arises in passing from the infinite system (9) to the truncated finite system (10). This is the subject of the current, fairly technical, section, which is therefore part of the appendix. The form of the estimates are motivated in part by the work of Day [4, 5]

Consider the Cahn-Hilliard equation (1) on the square domain $\Omega = (0,1)^2$, with the specific cubic nonlinearity $f(u) = u - u^3$. Furthermore, we choose the functions

$$\phi_{i,j}(x,y) = \cos(i\pi x)\cos(j\pi y) \qquad \text{for all} \quad i,j \in \mathbb{N}_0$$

defined in (18) as the complete orthogonal system of the underlying Hilbert space $L^2(\Omega)$. Any function in $L^2(\Omega)$ can be written as Fourier series with respect to the above basis, and we use the following notation which was introduced in (20). Let $u \in L^2(\Omega)$ be arbitrary. Then its (i,j)-th Fourier coefficient is denoted by $\mathcal{P}_{i,j}(u)$, i.e., we have

$$u = \sum_{i,j=0}^{\infty} u_{i,j}\phi_{i,j} \qquad \text{where} \qquad u_{i,j} = \mathcal{P}_{i,j}(u). \tag{53}$$

Since we are considering a polynomial nonlinearity, one of the first steps towards bounding the truncation error is to rewrite the product of two functions in terms of their Fourier coefficients. This is accomplished in the following lemma.

Lemma A.1. *Let $v, w \in L^2(\Omega)$ be arbitrary functions such that their pointwise product satisfies $vw \in L^2(\Omega)$. Let $v = \sum_{i,j\geq 0} v_{i,j}\phi_{i,j}$ and $w = \sum_{i,j\geq 0} w_{i,j}\phi_{i,j}$. In addition, define for all $(i,j) \in \mathbb{Z}^2$ the coefficients*

$$\tilde{v}_{i,j} \coloneqq \begin{cases} 4v_{|i|,|j|} & \text{for } (i,j) = (0,0), \\ 2v_{|i|,|j|} & \text{for } i = 0, j \neq 0 \text{ or } i \neq 0, j = 0, \\ v_{|i|,|j|} & \text{otherwise,} \end{cases}$$

and similarly for $\tilde{w}_{i,j}$. *Then the* (i,j)*-th Fourier coefficient* $\mathcal{P}_{i,j}(vw)$ *of the product* vw *as in* (20) *is given by*

$$\mathcal{P}_{i,j}(vw) = c_{i,j} \cdot \sum_{p,q\in\mathbb{Z}} \tilde{v}_{p,q}\tilde{w}_{i-p,j-q} \quad \text{for all} \quad i,j \geq 0,$$

where $c_{0,0} = 1/16$, $c_{i,0} = c_{0,j} = 1/8$, *and* $c_{i,j} = 1/4$, *for* $i,j > 0$.

The proof of the above lemma is rather long, but straightforward, and is therefore omitted. It is based on the formula

$$\phi_{p,q}\phi_{r,s} = \frac{1}{4} \cdot \big(\phi_{p+r,q+s} + \phi_{p+r,|q-s|} + \phi_{|p-r|,q+s} + \phi_{|p-r|,|q-s|}\big).$$

Details can be found in [18, Lemma 5.0.1]. In order to describe the Fourier coefficients of the cubical term in the nonlinearity f one only has to apply Lemma A.1 twice. In this way, one can show that the (i,j)-th Fourier coefficient $g_{i,j}$ of u^3 is given explicitly by

$$g_{i,j} = \mathcal{P}_{i,j}(u^3) = \frac{c_{i,j}}{4} \cdot \sum_{p,q,r,s\in\mathbb{Z}} \tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}. \tag{54}$$

Inserting (53) and (54) into (1), we finally obtain an infinite system of coupled differential equations which is equivalent to the Cahn-Hilliard equation. This system is given by

$$\dot{u}_{i,j} = -(i^2+j^2)^2\pi^4 u_{i,j} + \lambda(i^2+j^2)\pi^2\bigg(u_{i,j} - \frac{c_{i,j}}{4}\sum_{p,q,r,s\in\mathbb{Z}} \tilde{u}_{p,q}\tilde{u}_{r,s}\tilde{u}_{i-p-r,j-q-s}\bigg), \tag{55}$$

where $i,j \in \mathbb{N}_0$. Note that there is no dynamics in $u_{0,0}$, which is a reflection of the mass conservation of the Cahn-Hilliard equation. Nevertheless, particularly for the computations to come, it is convenient to add $\dot{u}_{0,0} = 0$ and $u_{0,0} = \mu$.

After these preliminary comments, we now turn our attention to the main topic of this section, i.e., estimating the truncation error. As a first step, we derive estimates for expressions of the form

$$\mathrm{IS}(a,b,i,j) = \sum_{p,q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q}, \tag{56}$$

for $i,j \in \mathbb{Z}$. In view of the later applications of these estimates, one needs to obtain tight bounds on the possible values of $\mathrm{IS}(a,b,i,j)$, if the numbers $a_{p,q}$ and $b_{p,q}$ are chosen from suitable intervals. It is therefore natural to pursue an interval arithmetic approach, i.e., we consider the factors after the summation sign in $\mathrm{IS}(a,b,i,j)$ as compact intervals of the form $a_{m,n} = [a^-_{m,n}, a^+_{m,n}]$ and $b_{m,n} = [b^-_{m,n}, b^+_{m,n}]$. In addition, we define

$$\|a_{m,n}\|_I \coloneqq \max\{\,|a^-_{m,n}|,\ |a^+_{m,n}|\,\}. \tag{57}$$

In order to make the estimation problem for the sums $\mathrm{IS}(a,b,i,j)$ feasible, additional assumptions are necessary. These are collected in the following.

Assumption A.2. *Consider sums* $\mathrm{IS}(a,b,i,j)$ *of the form defined in* (56)*, where* $i,j \in \mathbb{Z}$*, and let* $a = (a_{p,q})_{p,q\in\mathbb{Z}}$ *and* $b = (b_{p,q})_{p,q\in\mathbb{Z}}$ *denote collections of intervals. Suppose that for* $p,q \in \mathbb{Z}$ *we have* $a_{p,q} = a_{|p|,|q|}$ *and* $b_{p,q} = b_{|p|,|q|}$*, and thus* $\mathrm{IS}(a,b,i,j) = \mathrm{IS}(a,b,|i|,|j|)$.

In addition, we require that there are constants $s \geq 2$*,* $A > 0$*,* $B > 0$*, and an integer* $M \geq 2$*, as well as positive constants* $A_1(p)$ *and* $B_1(p)$ *for* $p \in \{0,\dots,M-1\}$*, and positive constants* $A_2(q)$ *and* $B_2(q)$ *for* $q \in \{0,\dots,M-1\}$ *such that*

$$\|a_{p,q}\|_I \leq \begin{cases} \dfrac{A_1(|p|)}{|q|^s} & \text{for} \quad |q| \geq M \quad \text{and} \quad |p| < M, \\ \dfrac{A_2(|q|)}{|p|^s} & \text{for} \quad |p| \geq M \quad \text{and} \quad |q| < M, \\ \dfrac{A}{|p|^s|q|^s} & \text{for} \quad |p|,|q| \geq M, \end{cases} \tag{58}$$

$$\|b_{p,q}\|_I \leq \begin{cases} \dfrac{B_1(|p|)}{|q|^s} & \text{for} \quad |q| \geq M, \quad \text{and} \quad |p| < M \\ \dfrac{B_2(|q|)}{|p|^s} & \text{for} \quad |p| \geq M \quad \text{and} \quad |q| < M, \\ \dfrac{B}{|p|^s|q|^s} & \text{for} \quad |p|,|q| \geq M, \end{cases} \tag{59}$$

where $\|a_{p,q}\|_I$ *was defined in* (57).

In order to shorten terms in the following derivations, we define

$$A_1(p) = A_2(p) \coloneqq \frac{A}{p^s} \qquad \text{and} \qquad B_1(p) = B_2(p) \coloneqq \frac{B}{p^s},$$

for all $p \geq M$. We start with estimates for $i,j \in \{0,\dots,M-1\}$.

Lemma A.3. *Under Assumption A.2 we have for* $i,j \in \{0,\dots,M-1\}$*, that*

$$\begin{aligned} \mathrm{IS}(a,b,i,j) \subset \mathrm{FS}(a,b,i,j) &+ \frac{4AB\tau^2}{(i+M+1)^s(j+M+1)^s}\,[-1,1] \\ &+ 2\tau\left(\frac{S_2(j)}{(i+M+1)^s} + \frac{S_1(i)}{(j+M+1)^s}\right)[-1,1] \end{aligned}$$

where we use the abbreviations $\tau = 1/((s-1)M^{s-1})$,

$$\begin{aligned} \mathrm{FS}(a,b,i,j) &= \sum_{p=-M}^{M+i} \sum_{q=-M}^{M+j} a_{p,q} b_{i-p,j-q}, \\ S_1(i) = S_{1,M}(i) &= \sum_{p=-M}^{M+i} A_1(|p|) B_1(|i-p|), \\ S_2(j) = S_{2,M}(j) &= \sum_{q=-M}^{M+j} A_2(|q|) B_2(|j-q|). \end{aligned} \tag{60}$$

Proof. In order to estimate $\mathrm{IS}(a,b,i,j)$ we rewrite the sum in the form

$$\begin{aligned} \mathrm{IS}(a,b,i,j) = &\sum_{p=-M}^{i+M} \sum_{q=-M}^{j+M} a_{p,q} b_{i-p,j-q} + \sum_{p<-M} \sum_{q<-M} a_{p,q} b_{i-p,j-q} \\ &+ \sum_{p<-M} \sum_{q=-M}^{M+j} a_{p,q} b_{i-p,j-q} + \sum_{p<-M} \sum_{q>M+j} a_{p,q} b_{i-p,j-q} \\ &+ \sum_{p>M+i} \sum_{q<-M} a_{p,q} b_{i-p,j-q} + \sum_{p>M+i} \sum_{q=-M}^{M+j} a_{p,q} b_{i-p,j-q} \\ &+ \sum_{p>M+i} \sum_{q>M+j} a_{p,q} b_{i-p,j-q} + \sum_{p=-M}^{M+i} \sum_{q<-M} a_{p,q} b_{i-p,j-q} \\ &+ \sum_{p=-M}^{M+i} \sum_{q>M+j} a_{p,q} b_{i-p,j-q}, \end{aligned} \tag{61}$$

and Lemma A.3 is established by deriving estimates for each of the terms in (61) by using (58) and (59). First of all, one obtains

$$\begin{aligned} \sum_{p<-M} \sum_{q<-M} a_{p,q} b_{i-p,j-q} &\subseteq \sum_{p<-M} \sum_{q<-M} \frac{A}{|p|^s |q|^s} \frac{B}{|i-p|^s |j-q|^s} [-1,1] \\ &\subseteq \frac{AB}{(i+M+1)^s (j+M+1)^s} \sum_{p<-M} \sum_{q<-M} \frac{1}{|p|^s} \frac{1}{|q|^s} [-1,1] \\ &\subset \frac{AB}{(i+M+1)^s (j+M+1)^s} \int_M^\infty \int_M^\infty \frac{1}{x^s y^s} \, dx \, dy \, [-1,1] \\ &\subseteq \frac{AB}{(i+M+1)^s (j+M+1)^s (s-1)^2 M^{2(s-1)}} [-1,1] \\ &= \frac{AB\tau^2}{(i+M+1)^s (j+M+1)^s} [-1,1]. \end{aligned}$$

Similarly one obtains

$$\sum_{p<-M}\sum_{q>M+j} a_{p,q}b_{i-p,j-q} \subset \frac{AB\tau^2}{(i+M+1)^s(j+M+1)^s}\,[-1,1],$$

$$\sum_{p>M+i}\sum_{q<-M} a_{p,q}b_{i-p,j-q} \subset \frac{AB\tau^2}{(i+M+1)^s(j+M+1)^s}\,[-1,1],$$

as well as

$$\sum_{p>M+i}\sum_{q>M+j} a_{p,q}b_{i-p,j-q} \subset \frac{AB\tau^2}{(i+M+1)^s(j+M+1)^s}\,[-1,1].$$

In addition, we have

$$\begin{aligned}\sum_{p<-M}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} &\subseteq \sum_{p<-M}\sum_{q=-M}^{M+j} \frac{A_2(|q|)}{|p|^s}\frac{B_2(|j-q|)}{|i-p|^s}\,[-1,1]\\ &\subset \frac{1}{(i+M+1)^s}\cdot\sum_{q=-M}^{M+j} A_2(|q|)B_2(|j-q|)\cdot\int_M^\infty \frac{1}{x^s}\,dx\,[-1,1]\\ &\subseteq \frac{\tau S_2(j)}{(i+M+1)^s}\,[-1,1],\end{aligned}$$

and analogously

$$\sum_{p>M+i}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} \subset \frac{\tau S_2(j)}{(i+M+1)^s}\,[-1,1],$$

$$\sum_{p=-M}^{M+i}\sum_{q<-M} a_{p,q}b_{i-p,j-q} \subset \frac{\tau S_1(i)}{(j+M+1)^s}\,[-1,1],$$

and

$$\sum_{p=-M}^{M+i}\sum_{q>M+j} a_{p,q}b_{i-p,j-q} \subset \frac{\tau S_1(i)}{(j+M+1)^s}\,[-1,1].$$

The result now follows by combining all these estimates. □

Before proceeding, we need the following two auxiliary results.

Lemma A.4. *The sum*

$$S(M) := \sum_{p=1}^{M-1}\frac{M^2}{p^2(M-p)^2}$$

is decreasing in M for all $M \geq 4$.

Proof. Using the identity

$$\frac{M^2}{p^2(M-p)^2} = \frac{1}{p^2} + \frac{2}{M}\left(\frac{1}{p} + \frac{1}{M-p}\right) + \frac{1}{(M-p)^2}$$

one obtains

$$S(M) = 2\sum_{p=1}^{M-1}\frac{1}{p^2} + \frac{4}{M}\sum_{p=1}^{M-1}\frac{1}{p},$$

which furnishes for all $M \geq 4$ the estimate

$$\begin{aligned} S(M) - S(M+1) &= 2\left(\sum_{p=1}^{M-1}\frac{1}{p^2} - \sum_{p=1}^{M}\frac{1}{p^2}\right) + \frac{4}{M}\sum_{p=1}^{M-1}\frac{1}{p} - \frac{4}{M+1}\sum_{p=1}^{M}\frac{1}{p} \\ &= -\frac{2}{M^2} + \frac{4}{M(M+1)}\left((M+1)\sum_{p=1}^{M-1}\frac{1}{p} - M\sum_{p=1}^{M}\frac{1}{p}\right) \\ &= -\frac{2}{M^2} + \frac{4}{M(M+1)}\left(\sum_{p=1}^{M-1}\frac{1}{p} - 1\right) > -\frac{2}{M^2} + \frac{4}{M(M+1)}\cdot\frac{5}{6} \\ &> -\frac{2(M+1)}{M^2(M+1)} + \frac{3M}{M^2(M+1)} = \frac{M-2}{M^2(M+1)} > 0. \end{aligned}$$

This completes the proof of the lemma. □

Lemma A.5. *For $i \geq M$ and $s \geq 2$, we have*

$$\sum_{p=1}^{i-1}\frac{1}{p^s(i-p)^s} \leq \frac{\gamma_{s,M}}{i^s},$$

where

$$\gamma_s = \gamma_{s,M} := \begin{cases} \dfrac{41}{9}\left(\dfrac{M}{M-1}\right)^{s-2} & \text{if} \quad M \in \{2,3\}, \\ S(M)\left(\dfrac{M}{M-1}\right)^{s-2} & \text{if} \quad M \geq 4. \end{cases}$$

Proof. For $s = 2$ and $M \geq 4$ the proof follows from Lemma A.4. Furthermore, we have the identities $\sup_{M>1} S(M) = \max_{1<M\leq 4} S(M) = S(4) = 41/9$, which completes the proof for the case $s = 2$. Now assume that $s > 2$ and $i \geq M$. In this case one obtains

$$\begin{aligned} \sum_{p=1}^{i-1}\frac{1}{p^s(i-p)^s} &= \sum_{p=1}^{i-1}\frac{1}{p^{s-2}(i-p)^{s-2}}\cdot\frac{1}{p^2(i-p)^2} \\ &\leq \frac{1}{(i-1)^{s-2}}\cdot\frac{\gamma_{2,M}}{i^2} = \frac{i^{s-2}}{(i-1)^{s-2}}\cdot\frac{\gamma_{2,M}}{i^s} \leq \frac{\gamma_{s,M}}{i^s}, \end{aligned}$$

and the proof is complete. □

In order to formulate the following estimates more concisely, we have to introduce the following abbreviations.

Definition A.6. We define

$$\begin{aligned}
V_2(j) = V_{2,M,s} &:= \sum_{q=-M}^{M+j} \max_{1\le r\le M-1} \max_{1\le t\le r} \Big(\max(\|a_{r,q}\|_I \cdot r^s, A_2(|q|)) \\
&\qquad \cdot \max(\|b_{M-t,j-q}\|_I \cdot (M-t)^s, B_2(|j-q|)) \Big), \\
S_2^\star(j) = S_{2,M}^\star(j) &:= \sum_{p=-M}^{0} \sum_{q=-M}^{M+j} \big(\|a_{p,q}\|_I \cdot B_2(|j-q|) + A_2(|j-q|) \|b_{p,q}\|_I \big), \\
W_1 = W_{1,M,s} &:= \max_{1\le r\le M-1} \max_{1\le t\le r} \big(\max(A_1(r) r^s, A) \cdot \max(B_1(M-t)(M-t)^s, B) \big), \\
A_1^\star = A_{1,M}^\star &:= \sum_{p=-M}^{0} A_1(|p|), \qquad \text{and} \qquad B_1^\star = B_{1,M}^\star := \sum_{p=-M}^{0} B_1(|p|).
\end{aligned}$$

With these abbreviations one obtains the following estimate.

Lemma A.7. *Under Assumption A.2 and for $i \ge M$ and $j \in \{0, \dots, M-1\}$ we have*

$$\begin{aligned}
\mathrm{IS}(a,b,i,j) \subset \frac{1}{i^s} \cdot \Big(& S_2^\star(j) + \gamma_s V_2(j) + 2\tau S_2(j) \\
& + \frac{2\tau}{(j+M+1)^s} (2AB\tau + BA_1^\star + \gamma_s W_1 + AB_1^\star) \Big) [-1,1]
\end{aligned}$$

Proof. We use the same strategy as in the proof of Lemma A.3 and write

$$\begin{aligned}
\mathrm{IS}(a,b,i,j) = & \sum_{p=-M}^{M+i} \sum_{q=-M}^{M+j} a_{p,q} b_{i-p,j-q} \\
& + \sum_{p<-M} \sum_{q\in\mathbb{Z}} a_{p,q} b_{i-p,j-q} + \sum_{p>M+i} \sum_{q\in\mathbb{Z}} a_{p,q} b_{i-p,j-q} \\
& + \sum_{p=-M}^{M+i} \sum_{q<-M} a_{p,q} b_{i-p,j-q} + \sum_{p=-M}^{M+i} \sum_{q>M+j} a_{p,q} b_{i-p,j-q}.
\end{aligned}$$

Again we consider each of the sums in the above identity separately. Starting with

the first sum, one obtains

$$\begin{aligned}
\sum_{p=-M}^{M+i}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} &= \sum_{p=-M}^{0}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} + \sum_{p=1}^{i-1}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} \\
&\quad + \sum_{p=i}^{M+i}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} \\
&\subseteq \sum_{p=-M}^{0}\sum_{q=-M}^{M+j} \|a_{p,q}\|_I \frac{B_2(|j-q|)}{|i-p|^s}\,[-1,1] \\
&\quad + \sum_{q=-M}^{M+j}\sum_{p=1}^{i-1} \frac{p^s a_{p,q}}{p^s}\cdot\frac{|i-p|^s b_{i-p,j-q}}{|i-p|^s} \\
&\quad + \sum_{p=-M}^{0}\sum_{q=-M}^{M+j} \frac{A_2(|j-q|)}{|i-p|^s}\|b_{p,q}\|_I\,[-1,1] \\
&\subseteq \frac{1}{i^s}\big(S_2^\star(j)+\gamma_s V_2(j)\big)\,[-1,1],
\end{aligned}$$

and

$$\begin{aligned}
\sum_{p<-M}\sum_{q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q} &\subseteq \sum_{p<-M}\sum_{q<-M} \frac{A}{|p|^s|q|^s}\cdot\frac{B}{|i-p|^s|j-q|^s}\,[-1,1] \\
&\quad + \sum_{p<-M}\sum_{q=-M}^{M+j} \frac{A_2(|q|)}{|p|^s}\cdot\frac{B_2(|j-q|)}{|i-p|^s}\,[-1,1] \\
&\quad + \sum_{p<-M}\sum_{q>M+j} \frac{A}{|p|^s|q|^s}\cdot\frac{B}{|i-p|^s|j-q|^s}\,[-1,1] \\
&\subset \frac{1}{(i+M+1)^s}\left(\frac{2AB\tau^2}{(j+M+1)^s}+\tau S_2(j)\right)[-1,1] \\
&\subseteq \frac{1}{i^s}\left(\frac{2AB\tau^2}{(j+M+1)^s}+\tau S_2(j)\right)[-1,1].
\end{aligned}$$

Similarly one can show that

$$\begin{aligned}
\sum_{p>M+i}\sum_{q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q} &\subset \frac{1}{i^s}\left(\frac{2AB\tau^2}{(j+M+1)^s}+\tau S_2(j)\right)[-1,1], \\
\sum_{p=-M}^{M+i}\sum_{q<-M} a_{p,q}b_{i-p,j-q} &\subset \frac{\tau}{i^s(j+M+1)^s}(BA_1^\star+\gamma_s W_1+AB_1^\star)\,[-1,1],
\end{aligned}$$

as well as

$$\sum_{p=-M}^{M+i} \sum_{q>M+j} a_{p,q} b_{i-p,j-q} \subset \frac{\tau}{i^s (j+M+1)^s} (BA_1^\star + \gamma_s W_1 + AB_1^\star)\,[-1,1],$$

which completes the proof. □

In some sense dual to the abbreviations in Definition A.6 are the following constant definitions, and the subsequent lemma.

Definition A.8. We define

$$\begin{aligned}
V_1(i) = V_{1,M,s} &:= \sum_{p=-M}^{M+i} \max_{1\le r\le M-1} \max_{1\le t\le r} \Big(\max\big(\|a_{p,r}\|_I r^s, A_1(|p|) \big) \\
&\qquad \cdot \max\big(\|b_{i-p,M-t}\|_I (M-t)^s, B_1(|i-p|) \big) \Big), \\
S_1^\star(i) = S_{1,M}^\star(i) &:= \sum_{p=-M}^{M+i} \sum_{q=-M}^{0} \big(\|a_{p,q}\|_I B_1(|i-p|) + A_1(|i-p|) \|b_{p,q}\|_I \big), \\
W_2 = W_{2,M,s} &:= \max_{1\le r\le M-1} \max_{1\le t\le r} \big(\max(A_2(r) r^s, A) \cdot \max(B_2(M-t)(M-t)^s, B) \big), \\
A_2^\star = A_{2,M}^\star &:= \sum_{q=-M}^{0} A_2(|q|), \qquad \text{and} \qquad B_2^\star = B_{2,M}^\star := \sum_{q=-M}^{0} B_2(|q|).
\end{aligned}$$

Lemma A.9. *Let Assumption A.2 be satisfied, and assume the notation of Lemma A.5, Lemma A.3, and Definition A.8. Then we have for all $i \in \{0,\dots,M-1\}$ and $j \ge M$*

$$\begin{aligned}
\mathrm{IS}(a,b,i,j) \subset \frac{1}{j^s} \Bigg(& S_1^\star(i) + \gamma_s V_1(i) + 2\tau S_1(i) \\
& + \frac{2\tau}{(i+M+1)^s} (2AB\tau + BA_2^\star + \gamma_s W_2 + AB_2^\star \Bigg)\,[-1,1].
\end{aligned}$$

Proof. The proof is similar to the one of Lemma A.7 and is therefore omitted. □

Definition A.10. For the case $i, j \geq M$ we define

$$R_1 := \sum_{p=-M}^{0} \max_{1\leq r\leq M-1} \max_{1\leq t\leq r} \big(\max(r^s\|a_{p,r}\|_I, A_1(|p|)) \cdot \max(B_2(M-t)(M-t)^s, B)\big),$$

$$R_2 := \sum_{p=-M}^{0} \max_{1\leq r\leq M-1} \max_{1\leq t\leq r} \big(\max(r^s\|b_{p,r}\|_I, B_1(|p|)) \cdot \max(A_2(M-t)(M-t)^s, A)\big),$$

$$T_1 := \sum_{q=-M}^{0} \max_{1\leq r\leq M-1} \max_{1\leq t\leq r} \big(\max(r^s\|a_{r,q}\|_I, A_2(|q|)) \cdot \max(B_1(M-t)(M-t)^s, B)\big),$$

$$T_2 := \sum_{q=-M}^{0} \max_{1\leq r\leq M-1} \max_{1\leq t\leq r} \big(\max(r^s\|b_{r,q}\|_I, B_2(|q|)) \cdot \max(A_1(M-t)(M-t)^s, A)\big),$$

$$a^\star := \sum_{p=-M}^{0}\sum_{q=-M}^{0} \|a_{p,q}\|_I, \qquad b^\star := \sum_{p=-M}^{0}\sum_{q=-M}^{0} \|b_{p,q}\|_I,$$

$$Z := \max_{1\leq k,r\leq M-1} \max_{1\leq l\leq k} \max_{1\leq t\leq r} \Big(\max(k^s r^s\|a_{k,r}\|_I, A_1(k)k^s, A_2(r)r^s) \cdot \max\big((M-l)^s(M-t)^s\|b_{M-l,M-t}\|_I, B_1(M-l)(M-l)^s, B_2(M-t)(M-t)^s\big)\Big)$$

Notice that in fact $R_1 = R_{1,M,s}$, $R_2 = R_{2,M,s}$, $T_1 = T_{1,M,s}$, $T_2 = T_{2,M,s}$, $a^\star = a^\star_M$, $b^\star = b^\star_M$, and $Z = Z_{M,s}$.

Lemma A.11. *Consider the notation of Lemma A.5 and Definitions A.6, A.8, and A.10, and suppose that Assumption A.2 holds. Then for all $i \geq M$ and $j \geq M$ we have*

$$\mathrm{IS}(a,b,i,j) \subset \frac{1}{i^s j^s}\Big(2\tau\big(B\,(A_1^\star + A_2^\star) + A(B_1^\star + B_2^\star) + \gamma_s(W_1+W_2) + 2\tau AB\big) + Ba^\star A_1^\star B_2^\star + A_2^\star B_1^\star + Ab^\star + \gamma_s\big(R_1 + R_2 + \gamma_s Z + T_1 + T_2\big)\Big)\,[-1,1].$$

Proof. Again we rewrite (56), this time in the form

$$\begin{aligned}\mathrm{IS}(a,b,i,j) = &\sum_{p=-M}^{M+i}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} \\ &+ \sum_{p=-M}^{M+i}\sum_{q<-M} a_{p,q}b_{i-p,j-q} + \sum_{p=-M}^{M+i}\sum_{q>M+j} a_{p,q}b_{i-p,j-q} \\ &+ \sum_{p<-M}\sum_{q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q} + \sum_{p>M+i}\sum_{q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q}.\end{aligned}$$

The first sum can be rewritten as

$$\begin{aligned}\sum_{p=-M}^{M+i}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} = &\sum_{p=-M}^{0}\sum_{q=-M}^{0} a_{p,q}b_{i-p,j-q} + \sum_{p=-M}^{0}\sum_{q=1}^{j-1} a_{p,q}b_{i-p,j-q}\\ &+ \sum_{p=-M}^{0}\sum_{q=j}^{M+j} a_{p,q}b_{i-p,j-q} + \sum_{p=1}^{i-1}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q}\\ &+ \sum_{p=i}^{M+i}\sum_{q=-M}^{0} a_{p,q}b_{i-p,j-q} + \sum_{p=i}^{M+i}\sum_{q=1}^{j-1} a_{p,q}b_{i-p,j-q}\\ &+ \sum_{p=i}^{M+i}\sum_{q=j}^{M+j} a_{p,q}b_{i-p,j-q}.\end{aligned}$$

By considering each of the terms separately one obtains

$$\begin{aligned}\sum_{p=-M}^{0}\sum_{q=-M}^{0} a_{p,q}b_{i-p,j-q} &\subseteq \sum_{p=-M}^{0}\sum_{q=-M}^{0} \|a_{p,q}\|_I \frac{B}{|i-p|^s|j-q|^s}\,[-1,1]\\ &\subseteq \frac{Ba^\star}{i^s j^s}\,[-1,1],\end{aligned}$$

$$\begin{aligned}\sum_{p=-M}^{0}\sum_{q=1}^{j-1} a_{p,q}b_{i-p,j-q} &\subseteq \sum_{p=-M}^{0}\sum_{q=1}^{j-1} \|a_{p,q}\|_I \frac{B_2(|j-q|)}{|i-p|^s}\,[-1,1]\\ &\subseteq \frac{1}{i^s}\sum_{p=-M}^{0}\sum_{q=1}^{j-1} \frac{q^s\|a_{p,q}\|_I}{q^s}\cdot\frac{B_2(|j-q|)|j-q|^s}{|j-q|^s}\,[-1,1]\\ &\subseteq \frac{\gamma_s R_1}{i^s j^s}\,[-1,1],\end{aligned}$$

$$\begin{aligned}\sum_{p=-M}^{0}\sum_{q=j}^{M+j} a_{p,q}b_{i-p,j-q} &\subseteq \sum_{p=-M}^{0}\sum_{q=j}^{M+j} \frac{A_1(|p|)B_2(|j-q|)}{q^s|i-p|^s}\,[-1,1]\\ &\subseteq \frac{A_1^\star B_2^\star}{i^s j^s}\,[-1,1],\end{aligned}$$

$$\begin{aligned}\sum_{p=1}^{i-1}\sum_{q=-M}^{M+j} a_{p,q}b_{i-p,j-q} &\subseteq \sum_{q=-M}^{0}\sum_{p=1}^{i-1}\frac{\|a_{p,q}\|_I p^s B_1(|i-p|)|i-p|^s}{p^s|i-p|^s|j-q|^s}\,[-1,1]\\ &\quad+\sum_{p=1}^{i-1}\sum_{q=1}^{j-1}\frac{\|a_{p,q}\|_I p^s q^s}{p^s q^s}\cdot\frac{\|b_{i-p,j-q}\|_I|i-p|^s|j-q|^s}{|i-p|^s|j-q|^s}\,[-1,1]\\ &\quad+\sum_{q=j}^{M+j}\sum_{p=1}^{i-1}\frac{p^s A_1(p)|i-p|^s\|b_{i-p,j-q}\|_I}{p^s|i-p|^s q^s}\,[-1,1]\\ &\subseteq \frac{\gamma_s}{i^s j^s}(T_1+\gamma_s Z+T_2)\,[-1,1],\end{aligned}$$

$$\begin{aligned}&\sum_{p=i}^{M+i}\sum_{q=-M}^{0} a_{p,q}b_{i-p,j-q}+\sum_{p=i}^{M+i}\sum_{q=1}^{j-1} a_{p,q}b_{i-p,j-q}\\ &\qquad\subseteq \frac{A_2^\star B_1^\star}{i^s j^s}\,[-1,1]+\frac{\gamma_s R_2}{i^s j^s}\,[-1,1],\end{aligned}$$

and

$$\sum_{p=i}^{M+i}\sum_{q=j}^{M+j} a_{p,q}b_{i-p,j-q} \subseteq \frac{A}{i^s j^s}\sum_{p=i}^{M+i}\sum_{q=j}^{M+j}\|b_{i-p,j-q}\|_I\,[-1,1] \subseteq \frac{Ab^\star}{i^s j^s}\,[-1,1].$$

The next two terms can be bounded using estimates of the proof of Lemma A.7, namely

$$\sum_{p=-M}^{M+i}\sum_{q<-M} a_{p,q}b_{i-p,j-q} \subset \frac{\tau}{i^s j^s}(BA_1^\star+\gamma_s W_1+AB_1^\star)\,[-1,1],$$

and

$$\sum_{p=-M}^{M+i}\sum_{q>M+j} a_{p,q}b_{i-p,j-q} \subset \frac{\tau}{i^s j^s}(BA_1^\star+\gamma_s W_1+AB_1^\star)\,[-1,1].$$

The remaining estimates follow similarly to the proof of Lemma A.7, see also Lemmas A.3 and A.9):

$$\sum_{p<-M}\sum_{q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q} \subset \frac{\tau}{i^s j^s}(2AB\tau+BA_2^\star+\gamma_s W_2+AB_2^\star)\,[-1,1],$$

as well as

$$\sum_{p>M+i}\sum_{q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q} \subset \frac{\tau}{i^s j^s}(2AB\tau+BA_2^\star+\gamma_s W_2+AB_2^\star)\,[-1,1],$$

which completes the proof of the lemma. □

So far we have established bounds on the auxiliary expression $\mathrm{IS}(a,b,i,j)$ defined in (56). Yet, our main interest lies in bounding the truncation error when passing from the infinite system (55) to the truncated finite system (10). For this, we need to establish estimates for expressions of the form

$$\sum_{p,q,r,s\in\mathbb{Z}} a_{p,q}a_{r,s}a_{i-p-r,j-q-s} = \sum_{p,q\in\mathbb{Z}} a_{p,q}\,\mathrm{IS}(a,a,i-p,j-q). \tag{62}$$

These estimates will be derived under the following assumptions.

Assumption A.12. *Let $a = (a_{m,n})_{n,m\in\mathbb{Z}}$ be a collection of intervals $a_{m,n} = [a^-_{m,n}, a^+_{m,n}]$ with $a_{p,q} = a_{|p|,|q|}$ for all $p,q\in\mathbb{Z}$. Furthermore, assume that there are constants $A>0$, $s\geq 2$ and an integer $M\geq 2$, as well as positive constants $A_1(p)$ for $p\in\{0,\dots,M-1\}$ and positive constants $A_2(q)$ for $q\in\{0,\dots,M-1\}$, such that*

$$\|a_{p,q}\|_I \leq \begin{cases} \dfrac{A_1(|p|)}{|q|^s} & \text{for } |q|\geq M \text{ and } |p|<M,\\[2ex] \dfrac{A_2(|q|)}{|p|^s} & \text{for } |p|\geq M \text{ and } |q|<M,\\[2ex] \dfrac{A}{|p|^s|q|^s} & \text{for } |p|,|q|\geq M, \end{cases}$$

where we use the definition in (57).

Definition A.13. For $|k|,|l|<M$, we set

$$b_{k,l} \coloneqq \mathrm{FS}(a,a,|k|,|l|) + \frac{4A^2\tau^2}{(|k|+M+1)^s(|l|+M+1)^s}\,[-1,1] + 2\tau\left(\frac{\widehat{S}_2(|l|)}{(|k|+M+1)^s} + \frac{\widehat{S}_1(|k|)}{(|l|+M+1)^s}\right)[-1,1],$$

where we use

$$\widehat{S}_1(i) = \sum_{p=-M}^{M+i} A_1(|p|)A_1(|i-p|) \qquad \text{and} \qquad \widehat{S}_2(j) = \sum_{q=-M}^{M+j} A_2(|q|)A_2(|j-q|).$$

Similarly, we use the notation of Definitions A.6, A.8, and A.10 with a hat, when all entries related to $b_{p,q}$ (B, $B_1(p)$, $B_2(q)$) are replaced by $a_{p,q}$ (A, $A_1(p)$, $A_2(q)$).

Finally, we define

$$B_1(|k|) \coloneqq \widehat{S_1^\star}(|k|) + \gamma_s \widehat{V_1}(|k|) + 2\tau \widehat{S_1}(|k|) + \frac{2\tau(2A^2\tau + 2AA_2^\star + \gamma_s \widehat{W_2})}{(|k|+M+1)^s},$$
$$B_2(|l|) \coloneqq \widehat{S_2^\star}(|l|) + \gamma_s \widehat{V_2}(|l|) + 2\tau \widehat{S_2}(|l|) + \frac{2\tau(2A^2\tau + 2AA_1^\star + \gamma_s \widehat{W_1})}{(|l|+M+1)^s},$$
$$\begin{aligned} B \coloneqq{}& 2\tau\big(2A(A_1^\star + A_2^\star) + \gamma_s(\widehat{W_1} + \widehat{W_2}) + 2\tau A^2\big) + 2Aa^\star + 2A_1^\star A_2^\star \\ &+ \gamma_s(\widehat{R_1} + \widehat{R_2} + \gamma_s \widehat{Z} + \widehat{T_1} + \widehat{T_2}), \end{aligned}$$

as well as

$$b_{k,l} \coloneqq \begin{cases} \dfrac{B_1(|k|)}{|l|^s}\,[-1,1] & \text{for } |l| \geq M \text{ and } |k| < M, \\ \dfrac{B_2(|q|)}{|p|^s}\,[-1,1] & \text{for } |k| \geq M \text{ and } |l| < M, \\ \dfrac{B}{|p|^s|q|^s}\,[-1,1] & \text{for } |k|, |l| \geq M. \end{cases}$$

Using Lemmas A.3, A.7, A.9, and A.11 one can readily see that

$$\mathrm{IS}(a,a,k,l) \subset b_{k,l} \qquad \text{for all} \qquad k,l \in \mathbb{Z}, \tag{63}$$

which in combination with (62) furnishes

$$\sum_{p,q,r,s\in\mathbb{Z}} a_{p,q}a_{r,s}a_{i-p-r,j-q-s} \subset \sum_{p,q\in\mathbb{Z}} a_{p,q}b_{i-p,j-q},$$

and by employing our estimates again, we finally obtain the following theorem.

Theorem A.14. *Let $a = (a_{m,n})_{n,m\in\mathbb{Z}}$ denote a collection of intervals $a_{m,n} = [a^-_{m,n}, a^+_{m,n}]$ as in Assumption A.12, and let $b = (b_{m,n})_{n,m\in\mathbb{Z}}$ be defined as in Definition A.13 such that* (63) *holds. Using the notation of Lemmas A.3 and A.5, as well as Definitions A.6, A.8, and A.10, we have the following inclusions.*

(i) *If $i,j \in \{0,\dots,M-1\}$, then*

$$\begin{aligned} \sum_{p,q,r,s\in\mathbb{Z}} a_{p,q}a_{r,s}a_{i-p-r,j-q-s} \subset{}& \mathrm{FS}(a,b,i,j) + \bigg(\frac{4AB\tau^2}{(i+M+1)^s(j+M+1)^s} \\ &+ 2\tau\bigg(\frac{S_2(j)}{(i+M+1)^s} + \frac{S_1(i)}{(j+M+1)^s}\bigg)\bigg)\,[-1,1], \end{aligned}$$

where $\mathrm{FS}(a,b,i,j)$ *was defined in* (60).

(ii) *If* $i \geq M$ *and* $j \in \{0, \ldots, M-1\}$, *then*

$$\sum_{p,q,r,s\in\mathbb{Z}} a_{p,q}a_{r,s}a_{i-p-r,j-q-s} \subset \frac{1}{i^s}\Bigg(S_2^\star(j) + \gamma_s V_2(j) + 2\tau S_2(j) + \frac{2\tau(2AB\tau + BA_1^\star + \gamma_s W_1 + AB_1^\star)}{(j+M+1)^s}\Bigg)[-1,1].$$

(iii) *If* $i \in \{0, \ldots, M-1\}$ *and* $j \geq M$, *then*

$$\sum_{p,q,r,s\in\mathbb{Z}} a_{p,q}a_{r,s}a_{i-p-r,j-q-s} \subset \frac{1}{j^s}\Bigg(S_1^\star(i) + \gamma_s V_1(i) + 2\tau S_1(i) + \frac{2\tau(2AB\tau + BA_2^\star + \gamma_s W_2 + AB_2^\star)}{(i+M+1)^s}\Bigg)[-1,1].$$

(iv) *If* $i, j \geq M$, *then*

$$\sum_{p,q,r,s\in\mathbb{Z}} a_{p,q}a_{r,s}a_{i-p-r,j-q-s} \subset \frac{1}{i^s j^s}\Big(2\tau\big(B(A_1^\star + A_2^\star) + A(B_1^\star + B_2^\star) + 2\tau AB + \gamma_s(W_1 + W_2)\big) + Ba^\star + A_1^\star B_2^\star + A_2^\star B_1^\star + Ab^\star + \gamma_s(R_1 + R_2 + \gamma_s Z + T_1 + T_2)\Big)[-1,1].$$

The above theorem is the main tool for establishing strict topologically self-consistent a priori bounds in section 2. Similar results for one-dimensional base domains can be found in [4, 6, 7].

B. Numerical description

We close this article with a brief description of the numerical and implementational aspects of our methods. The rigorous results described in section 1 using the methods outlined in the remaining sections of this paper were all obtained using MATLAB. The actual programs can be divided into two parts: One set is implementing a path-following algorithm, and is used to calculate approximate solutions along solution paths. The second set implements the rigorous method described in section 2, by employing the explicit estimates of section A. We describe each set separately in the following.

The path-following algorithm was originally developed for the numerical analysis of equilibria of the two-dimensional Cahn-Hilliard equation in [14]. It is based on a predictor-corrector algorithm with step-length adaption. See for example Allgower and Georg [1, chapter 3]. In the appendix of [14], our implementation is described in

more detail. In contrast to [14], for the present paper we approximate the Laplace operator by a discrete Laplacian on a fairly small equidistant grid, usually consisting of 50x50 grid points. The solutions gathered in this way are then used as initial points for the rigorous numerics. Therefore, it is sufficient to consider small grid sizes as above, since the higher accuracy will be achieved in the second part of the implementation.

The second part of the implementation is aimed at computing a box-neighborhood U around an approximative solution. The projection of U onto one Fourier coefficient is usually a closed interval, as described in section 2. Hence, our implementation makes heavy use of interval arithmetic. Note that

$$[a,b] \diamond [c,d] = \big[\min\{x \diamond y | x \in [a,b], y \in [c,d]\},\ \max\{x \diamond y | x \in [a,b], y \in [c,d]\}\big],$$

where '$\diamond$' can be replaced by the operators for addition, subtraction, multiplication, or division. (In the case of division one of course has to assume that $0 \notin [c,d]$.) Of course, the use of interval arithmetic alone does not suffice to obtain rigorous computer-assisted proofs. In addition, one has to be able to keep track of rounding errors. This is accomplished by the MATLAB toolbox INTLAB by Siegfried Rump. This toolbox comprises interval arithmetic for real and complex data including vectors and matrices, including MATLAB's sparse matrix commands. Moreover, it contains rigorous interval versions of standard MATLAB functions such as sin or exp, which do keep track of rounding errors. More precisely, applying such a MATLAB function to an interval returns an interval which contains all possible function values. See [12] or INTLAB's homepage at `www.ti3.tu-harburg.de/~rump/intlab/` for more detail. For detailed information on the implementation of the rigorous numerical method we refer the reader to the appendix of [18].

Acknowledgement. The authors would like to thank the anonymous referee for his useful comments.

References

[1] E. L. Allgower and K. Georg, *Numerical continuation methods*: *An introduction*, Springer Series in Computational Mathematics, vol. 13, Springer-Verlag, Berlin, 1990.

[2] J. W. Cahn, *On spinodal decomposition*, Acta Metallurgica **9** (1961), 795–801.

[3] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system*. I: *Interfacial free energy*, Journal of Chemical Physics **28** (1958), 258–267.

[4] S. Day, *A rigorous numerical method in infinite dimensions*, PhD Thesis, Georgia Institute of Technology, 2003.

[5] ———, *Towards a rigorous numerical study of the Kot-Schaffer model*: *Dynamic equations on time scales*, Dynam. Systems Appl. **12** (2003), no. 1-2, 87–97.

[6] S. Day, Y. Hiraoka, K. Mischaikow, and T. Ogawa, *Rigorous numerics for global dynamics*: *A study of the Swift-Hohenberg equation*, SIAM J. Appl. Dyn. Syst. **4** (2005), no. 1, 1–31.

[7] S. Day, O. Junge, and K. Mischaikow, *A rigorous numerical method for the global analysis of infinite-dimensional discrete dynamical systems*, SIAM J. Appl. Dyn. Syst. **3** (2004), no. 2, 117–160.

[8] P. C. Fife, *Models for phase separation and their mathematics*, Electron. J. Differential Equations **48** (2000), 1–26.

[9] P. C. Fife, H. Kielhöfer, S. Maier-Paape, and T. Wanner, *Perturbation of doubly periodic solution branches with applications to the Cahn-Hilliard equation*, Phys. D **100** (1997), no. 3-4, 257–278.

[10] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, reprint of the 1998 edition, Classics in Mathematics, Springer-Verlag, Berlin, 2001.

[11] M. Grinfeld and A. Novick-Cohen, *Counting stationary solutions of the Cahn-Hilliard equation by transversality arguments*, Proc. Roy. Soc. Edinburgh Sect. A **125** (1995), no. 2, 351–370.

[12] G. Hargreaves, *Interval analysis in Matlab*, Numerical Analysis Reports, vol. 416, University of Manchester, Manchester, 2002.

[13] H. Kielhöfer, *Pattern formation of the stationary Cahn-Hilliard model*, Proc. Roy. Soc. Edinburgh Sect. A **127** (1997), no. 6, 1219–1243.

[14] S. Maier-Paape and U. Miller, *Path-following the equilibria of the Cahn-Hilliard equation on the square*, Comput. Vis. Sci. **5** (2002), no. 3, 115–138.

[15] ———, *Connecting continua and curves of equilibria of the Cahn-Hilliard equation on the square*, Discrete Contin. Dyn. Syst. **15** (2006), no. 4, 1137–1153.

[16] S. Maier-Paape, K. Mischaikow, and T. Wanner, *Structure of the attractor of the Cahn-Hilliard equation on a square*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **17** (2007), no. 4, 1221–1263.

[17] C. McCord, *Mappings and homological properties in the Conley index theory*, Ergodic Theory Dynam. Systems **8*** (1988), 175–198. Charles Conley Memorial Issue.

[18] U. Miller, *Rigorous numerics using Conley index theory*, with a preface by S. Maier-Paape, Augsburger Schriften zur Mathematik, Physik und Informatik, vol. 9, Logos Verlag Berlin, Berlin, 2005. PhD Thesis, Universität Augsburg, 2004.

[19] K. Mischaikow and M. Mrozek, *Conley index theory*, Handbook of dynamical systems, Vol. 2, North-Holland, Amsterdam, 2002, pp. 393–460.

[20] A. Novick-Cohen and L. A. Peletier, *Steady states of the one-dimensional Cahn-Hilliard equation*, Proc. Roy. Soc. Edinburgh Sect. A **123** (1993), no. 6, 1071–1098.

[21] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, Texts in Applied Mathematics, vol. 37, Springer-Verlag, New York, 2000.

[22] P. Zgliczyński and K. Mischaikow, *Rigorous numerics for partial differential equations: The Kuramoto-Sivashinsky equation*, Found. Comput. Math. **1** (2001), no. 3, 255–288.