SIAM J. NUMER. ANAL.
Vol. 0, No. 0, pp. 000–000

VALIDATED CONTINUATION FOR EQUILIBRIA OF PDES*

SARAH DAY[†], JEAN-PHILIPPE LESSARD[‡], AND KONSTANTIN MISCHAIKOW[§]

Abstract. One of the most efficient methods for determining the equilibria of a continuous parameterized family of differential equations is to use predictor-corrector continuation techniques. In the case of partial differential equations this procedure must be applied to some finite dimensional approximation which of course raises the question of the validity of the output. We introduce a new technique that combines the information obtained from the predictor-corrector steps with ideas from rigorous computations and verifies that the numerically produced equilibrium for the finite dimensional system can be used to explicitly define a set which contains a unique equilibrium for the infinite dimensional partial differential equation. Using the Cahn-Hilliard and Swift-Hohenberg equations as models we demonstrate that the cost of this new validated continuation is less than twice the cost of the standard continuation method alone.

Key words. continuation, PDE, Swift-Hohenberg, validation

AMS subject classifications. AUTHOR: PLEASE PROVIDE

DOI. 10.1137/050645968

1. Introduction.

The first step in understanding the dynamics of a nonlinear system of differential equations

$$u_t = f(u, \nu) \tag{1.1}$$

on a Hilbert space is to identify the set of equilibria $\mathcal{E} := \{(u, \nu) \mid f(u, \nu) = 0\}$. For many applications this can only be done using numerical methods. In particular, continuation provides an efficient technique for determining elements on branches of $\mathcal{E}$. Recall, that this method involves a predictor and corrector step: given, within a prescribed tolerance, an equilibrium u_0 at parameter value ν_0, the predictor step produces an approximate equilibrium $\tilde{u}_1$ at nearby parameter value ν_1, and the corrector step, often based on a Newton-like operator, takes $\tilde{u}_1$ as its input and produces, once again within the prescribed tolerance, an equilibrium u_1 at ν_1.

With any numerical method there is the question of validity of the output as compared with the cost of computation. The goal of this paper is to argue that for a large and important class of partial differential equations the cost of validating the existence and uniqueness of equilibria is small when compared to the cost of identifying potential equilibria by means of a continuation method. Our interest in this question was motivated by the increasing development of computer-assisted proofs in the dynamics of infinite dimensional systems (see [3], [10] and references therein). As mathematicians we are willing to argue forcefully for the importance

[*] Received by the editors November 23, 2005; accepted for publication (in revised form) January 16, 2007; published electronically DATE.
http://www.siam.org/journals/sinum/x-x/64596.html

[†] The College of William and Mary, Department of Mathematics, P.O.Box 8795, Williamsburg, VA 23187-8795 (sday@math.wm.edu). This author was partially supported by NSF DMS 9983660.

[‡] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160 (lessard@math.gatech.edu). This author was partially supported by FQRNT, NSF DMS 0511115, and grants from D.O.E. and DARPA.

[§] Department of Mathematics, Rutgers, The State University of New Jersey, Hill Center-Busch Campus, 110 Frelinghusen Rd, Piscataway, NJ 08854-8019, and School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160 (mischaik@math.rutgers.edu). This author was partially supported by NSF DMS 0511115, and grants from D.O.E. and DARPA.

of rigorous verification and thus marginalize the cost. However, in reality for many applications, researchers are often interested in investigating a variety of model partial differential equations at a multitude of parameter values to gain scientific insight rather than an answer to a particular question. This places a premium on minimizing computational cost, often leading to acceptance of the validity of numerical results simply based upon the reproducibility of the result at different levels of refinement. As we shall argue, the results of this paper suggest that this dichotomy need not exist and we provide examples wherein it is demonstrated that by judicious use of the computations involved in the continuation method it is cheaper to validate the results than to re-perform the continuation computation. We refer to the method we propose as *validated continuation*. As is made clear towards the end of the introduction, validated continuation is slightly weaker and computationally cheaper than rigorous continuation.

To the best of our knowledge this is the first attempt to integrate the techniques of rigorous computations with a continuation method, thus we focus on a clear presentation of the ideas as opposed to presenting the results in the most general possible setting. We make use of spectral methods as they provide us with considerable control on truncation errors. To be more precise, assume that (1.1) takes the form

$$u_t = L(u,\nu) + \sum_{p=0}^{d} c_p(\nu)u^p \tag{1.2}$$

where $L(\cdot,\nu)$ is a linear operator at parameter value ν and d is the degree of the polynomial nonlinearity. Typically, $c_1(\nu) = 0$ since linear terms are grouped under $L(\cdot,\nu)$. Expanding (1.2) using an orthogonal basis chosen appropriately in terms of the eigenfunctions of the linear operator $L(\cdot,\nu)$, the particular domain and the boundary conditions, results in a countable system of differential equations on the coefficients of the expanded solution.

To simplify the exposition, let us assume the expansion takes the form

$$\dot{u}_k = f_k(u,\nu) := \mu_k u_k + \sum_{p=0}^{d} \sum_{\sum n_i = k} (c_p)_{n_0} u_{n_1} \cdots u_{n_p} \quad k = 0,1,2,\dots \tag{1.3}$$

where $\mu_k = \mu_k(\nu)$ are the parameter dependent eigenvalues of $L(\cdot,\nu)$ and $\{u_n\}$ and $\{(c_p)_n\}$ are the coefficients of the corresponding expansions of the functions u and $c_p(\nu)$ respectively with $u_n = u_{-n}$ and $(c_p)_n = (c_p)_{-n}$ for all n. In order to simplify the notation, for a fixed parameter ν, we use $f(u)$ to denote $f(u,\nu)$. The continuation method is applied to the m-dimensional system of ODEs of the form

$$\dot{u}_k = \mu_k u_k + \sum_{p=0}^{d} \sum_{\substack{\sum n_i = k \\ |n_i| < m}} (c_p)_{n_0} u_{n_1} \cdots u_{n_p} \quad k = 0,1,\dots,m-1. \tag{1.4}$$

obtained by performing a Galerkin projection on (1.3). It is this truncation that introduces the most substantial concern for the validity of the results of the continuation method. In Section 3 we present estimates that provide us with bounds on the errors. We obtain these bounds under the assumption of power decay rates in the coefficients $\{u_n\}$. Of course, such decay rates are directly related to the spatial smoothness of the equilibria which in turn is governed, at least in part, by the linear operator $L(\cdot,\nu)$.

The theoretical justification for our proof of existence and uniqueness of equilibria is based on a component-wise version of the Banach fixed point theorem (see Theorem 2.1) which itself represents a minor modification of a result of Yamamoto [9, Theorem 2.1]. A similar formulation can also be found in [4]. Recall that to apply the Banach fixed point theorem one must have a contraction mapping $T : X \to X$. With this in mind, we can state that it is appropriate to view our approach as a method by which the Newton-like iteration of the corrector step in the continuation process is used to construct a set X and the above estimates are used to verify that an appropriate generalization of the Newton-like operator is in fact a contraction. More precisely, let $\bar{u}$ be a numerical zero obtained from (1.4). In the orthogonal basis used to obtain (1.3) consider the set $X = \bar{u} + W(r)$ of $\bar{u}$ where $W(r)$ is of the form

$$W(r) = \prod_{k=0}^{m-1} [-r, r] \times \prod_{k=m}^{\infty} \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s} \right].$$

Observe that s indicates the decay rate of the coefficients and r is referred to as the *validation radius.* Our strategy which is described in detail in Section 3 is to produce a set of *radii polynomials*, $\{P_k(r)\}_{k=0,1,\ldots}$, whose coefficients are given explicitly in terms of the constants A_s, s, and (1.3). Theorem 3.4 guarantees that if there exists a validation radius $r > 0$ such that $P_k(r) < 0$ for all k, then there exists a unique equilibrium solution to (1.2) in the set $X = \bar{u} + W(r)$ built around the numerical equilibrium $\bar{u}$ produced by the continuation procedure. It is important to remark that the conditions of Theorem 3.4 can be checked with a finite number of calculations.

As is indicated above the focus of this paper is on the computational efficacy of validated continuation, and hence, the following organization of the material. Section 2 contains the statement and proof of the aforementioned component-wise version of the Banach fixed point theorem, Theorem 2.1, without any indication of how this result can be used in practice. Section 3 provides the opposite extreme, an explicit set of formulas and steps and the assertion that their successful implementation leads, via Theorem 3.4, to the existence of a unique equilibrium in a specified set. The justification of this assertion and the relationship between Theorems 2.1 and 3.4 is presented in Section 6. However, presenting the formulae in this fashion has two advantages. First, they contain all the necessary information should the reader wish to independently code and test the techniques suggested in this paper. Second, it allows for the presentation in Section 4 of the comparison of the computational costs between traditional and validated continuation.

It should be emphasized that how one should best compare the costs between the two methods of continuation is not completely clear. In the standard approach m, the dimension of the system on which continuation is performed, is fixed. Thus traditionally, a particular Galerkin projection dimension is chosen and continuation is performed. The results are checked by choosing a higher dimensional projection, re-performing the continuation and then deciding if the two calculations agree within a certain level of numerical tolerance. In validated continuation, m becomes a variable. In particular, if validation fails then one has the option of choosing a higher dimensional Galerkin projection. Equally important, failure of validation may be an indication that a higher dimensional projection is necessary. In summary, validated continuation provides an internal check of consistency on the dimension of truncation from the infinite to finite dimensional problem a feature which is not present in the traditional application of continuation methods.

With this in mind we have chosen to compare the computational costs as follows.

First we restrict our attention to cubic nonlinearities. As is made clear by the formulae of Section 3 in this case the cost of evaluating the nonlinearities and performing Newton's method are both of order m^3. Thus, we can obtain a rough bound on the ratio of the cost of traditional versus validated continuation by counting the number of m^3 operations which need to be performed. These calculations suggest that for fixed m the cost of validated continuation is less than twice the cost of traditional continuation, that is *it appears that it is cheaper to perform validated continuation, than to perform traditional continuation and then check it against continuation performed on a higher dimensional projection.* In Section 5 this estimate is tested against actual computations for the Swift-Hohenberg equation and the Cahn-Hilliard equation. To ensure that these comparisons are fair, we employ standard floating point computations in both cases.

This last point raises an important distinction: validated continuation versus rigorous continuation. Using floating point calculations at all steps of the validated continuation, does not allow one to control for roundoff errors and hence one cannot rigorously concluded the existence of an equilibrium. Because the current computer technology treats floating point and interval arithmetic differently we chose not to make and present timed comparisons between the two for this paper. However, if specific steps in the validation argument are performed using interval arithmetic, then one obtains rigorous results on the existence of equilibria. Results of this type are presented in Section 5 for a branch of equilibria of the Swift-Hohenberg equation.

We see the results of this paper as a first step in the direction of combining continuation methods with rigorous computations. With this in mind we conclude the paper in Section 7 with a discussion of open questions and on going work. In particular, we return to the issue of the necessity of interval arithmetic computations.

2. Computational proofs for equilibria. Assume that following the expansion of a PDE into an appropriate orthogonal basis, we have a system of the form (1.3). Our goal is to prove that there is a unique equilibrium for (1.3) which lies in a small set containing a computed numerical equilibrium. Suppose $\bar{u}_F$ is a numerical equilibrium computed using an m-dimensional continuation procedure (as described in Section 3) and $\bar{u} := (\bar{u}_F, 0, \dots)$ is the corresponding point in the infinite dimensional space. We will consider a set of the form $\bar{u} + W$ where $W = \Pi_k \tilde{w}_k$,

$$\tilde{w}_k = \begin{cases} [-r, r] & 0 \le k < m \\ \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s}\right] & k \ge m \end{cases} \tag{2.1}$$

for some constants $r, A_s > 0$ and $s \ge 2$.

A particularly nice norm to use for this set (similar to the one used by Yamamoto in [9]) is the normalized sup norm

$$\|u\|_W := \sup_k \left\{ \frac{|u_k|}{|\tilde{w}_k|} \right\}$$

where $|\tilde{w}_k| := \max\{|x| \mid x \in \tilde{w}_k\}$. In this norm, $W = B(0,1)$ is the unit ball around 0, and $\bar{u} + W = B(\bar{u}, 1)$ is the unit ball around $\bar{u}$.

We will now reformulate our problem of studying equilibria for (1.3) by establishing an equivalent fixed point problem on $\bar{u} + W$. Suppose J is an invertible operator. Then u is an equilibrium solution of (1.3) if and only if u is a fixed point of

$$T(u) = u - Jf(u) \tag{2.2}$$

where f is given by (1.3). In practice, T is constructed to be a contraction (Newton-like) operator with $J \approx (Df(\bar{u}))^{-1}$ so that we may use Banach's fixed point theorem. We now frame this fixed point theorem in a more computational setting.

In the process of showing that T is a contraction, we first consider the following Lipschitz condition on $\bar{u} + W$:

$$\|T(x) - T(y)\|_W \leq K\|x - y\|_W \quad \text{for } x, y \in \bar{u} + W. \tag{2.3}$$

The question now becomes whether we can compute a contraction constant $K < 1$ satisfying (2.3). We begin by computing Lipschitz constants, K_n, for the component functions T_n on $\bar{u} + W$ satisfying the following

$$|T_n(x) - T_n(y)| \leq K_n\|x - y\|_W \quad \text{for } x, y \in \bar{u} + W. \tag{2.4}$$

If T is C^1, we may take K_n to be a bound on the derivative of T_n over $\bar{u} + W$. More explicitly,

$$\begin{aligned} K_n &\geq \sup |DT_n(\bar{u} + W) \cdot W| \\ &:= \sup_{b,c \in W} |DT_n(\bar{u} + b) \cdot c| . \end{aligned}$$

A constant K_n computed in this manner satisfies (2.4) by the following argument. For $x, y \in \bar{u} + W$, let $g_n(s) := T_n[sx + (1 - s)y]$. Applying the mean value theorem to g_n, we get the existence of $s_n \in [0, 1]$ such that $g_n(1) - g_n(0) = g'(s_n)$. Since the set $\bar{u} + W$ is convex, we get the existence of $z_n := s_n x + (1 - s_n)y \in \bar{u} + W$ such that

$$\begin{aligned} |T_n(x) - T_n(y)| &= |DT_n(z_n)(x - y)| \\ &= \left| DT_n(z_n) \frac{x - y}{\|x - y\|_W} \right| \|x - y\|_W. \end{aligned}$$

By construction of $\| \cdot \|_W$, $\frac{x-y}{\|x-y\|_W} \in W$. Now if $K := \sup_n \frac{K_n}{|\tilde{w}_n|} < \infty$, then, as the following argument shows, it satisfies (2.3)

$$\begin{aligned} \|T(x) - T(y)\|_W &= \sup_n \frac{|T_n(x) - T_n(y)|}{|\tilde{w}_n|} \\ &= \sup_n \frac{\left| DT_n(z_n) \frac{x-y}{\|x-y\|_W} \right| \|x - y\|_W}{|\tilde{w}_n|} \\ &\leq \sup_n \frac{K_n}{|\tilde{w}_n|} \|x - y\|_W \\ &= K\|x - y\|_W. \end{aligned}$$

Theorem 2.1 (existence and uniqueness). *If for all n there exist bounds $Y_n \geq |T_n(\bar{u}) - \bar{u}_n|$ and K_n satisfying (2.4) such that*

$$Y_n + K_n - |\tilde{w}_n| < 0 \tag{2.5}$$

and

$$K := \sup_n \frac{K_n}{|\tilde{w}_n|} < 1 \tag{2.6}$$

then there exists a unique fixed point of T in $\bar{u} + W$.

Proof. The first inequality ensures that $T(\bar{u}+W) \subset \bar{u}+W$. This is true if and only if for every $u \in \bar{u}+W$, $\|T(u)-\bar{u}\|_W \leq 1$, or equivalently, $\frac{|T_n(u)-\bar{u}_n|}{|\tilde{w}_n|} < 1$ for all n.

Let $u \in \bar{u}+W$. Then $\|u-\bar{u}\|_W \leq 1$ and for each n,

$$\begin{aligned} |T_n(u)-\bar{u}_n| &= |T_n(u)-T_n(\bar{u})+T_n(\bar{u})-\bar{u}_n| \\ &\leq |T_n(u)-T_n(\bar{u})|+|T_n(\bar{u})-\bar{u}_n| \\ &\leq K_n\|u-\bar{u}\|_W+Y_n \\ &\leq Y_n+K_n \\ &< |\tilde{w}_n| \end{aligned}$$

by assumption (2.5). Therefore, $T(\bar{u}+W) \subset \bar{u}+W$. The second inequality guarantees that T is also a contraction. Thus, the result follows from Banach's fixed point theorem. □

Let us make the comment here that sufficient regularity of the equilibrium solutions will effectively reduce the infinite set of conditions listed in Theorem 2.1 to a finite list. In essence, the strong decay in the higher modes may be used to verify (2.5) simultaneously for all $n > N$ for some N. (In our case N is determined by the dimension used for continuation and the degree of the nonlinearity.) Furthermore, regularity of the equilibria may also be used to show that $K_n|\tilde{w}_n|^{-1}$ becomes a decreasing sequence. Therefore, (2.6) follows automatically from (2.5).

Perhaps an even more important point to make for our intended algorithmic approach in this paper is that $Y_n+K_n-|\tilde{w}_n|$ will be given as a polynomial in the validation radius r, the width of the set W in the low modes. Therefore, validating the existence of a unique equilibrium near $\bar{u}$ will amount to showing that it is possible to simultaneously solve a (finite) list of polynomial inequalities in r.

3. Validated continuation.

The ideas outlined in Section 2 for proving the existence of unique equilibria fit naturally with traditional continuation techniques for following branches of numerical equilibria. In particular, an approximation of a projection of the Newton operator given in (2.2) onto the appropriate m-dimensional subspace is an intrinsic element of the continuation algorithm. In this Section, we discuss exploiting this relationship to automatically produce a validation of the existence of unique equilibria at each step of the continuation procedure.

Recall that following the expansion of the system in the appropriate basis, we have

$$\dot{u} = f(u,\nu) \tag{3.1}$$

where for $k=0,1,2,\ldots$, $\mu_k=\mu_k(\nu)$, $(c_p)_n=(c_p(\nu))_n$ and

$$\dot{u}_k = f_k(u) = \mu_k u_k + \sum_{p=0}^{d} \sum_{\sum n_i = k} (c_p)_{n_0} u_{n_1} \cdots u_{n_p} \tag{3.2}$$

A first step for implementing a continuation algorithm for studying a PDE is to perform a Galerkin projection. Let m be a fixed projection dimension and consider the following truncated version of our original expansion of the PDE given in (3.2). For $u_F := (u_0,\ldots,u_{m-1}) \in \mathbb{R}^m$, define $f^{(m)} : \mathbb{R}^m \to \mathbb{R}^m$ by $f^{(m)}(u_F) =$

$(f_0^{(m)}(u_F), \dots, f_{m-1}^{(m)}(u_F))$ where for $k = 0, \dots, m-1$,

$$f_k^{(m)}(u_F) = \mu_k u_k + \sum_{p=0}^{d} \sum_{\substack{\sum n_i = k \\ |n_i| < m}} (c_p)_{n_0} u_{n_1} \cdots u_{n_p}$$

The corresponding Galerkin projection of the original system (3.1) is then

$$\dot{u}_F = f^{(m)}(u_F, \nu) \tag{3.3}$$

This is the m-dimensional system to be studied numerically. Intuitively, we expect that if m is sufficiently large, (3.3) will capture the essential dynamics for the original system (3.1). In particular, given an equilibrium $\bar{u}_F$ for (3.3) we expect that there is a small set around $\bar{u} := (\bar{u}_F, 0, \dots)$ which contains a unique equilibrium solution for (3.1). Our approach is to study this relationship via the tools outlined in Section 2.

3.1. Continuation for ODEs and Newton-like operator.

A traditional continuation procedure involves iteration of predictor and corrector steps to trace out branches of equilibria. Under the assumption that at some parameter $\nu = \nu_0$ we have an equilibrium solution for (3.3), we want to continue the equilibrium as we vary ν.

1) Euler predictor: Given an approximate equilibrium x_0 at ν_0, the *predictor* at $\nu_1 = \nu_0 + \Delta\nu$ is $x_1^{(0)} = x_0 + \dot{x}_0 \Delta\nu$, where

$$\dot{x}_0 = -f_x^{(m)}(x_0, \nu_0)^{-1} f_\nu^{(m)}(x_0, \nu_0). \tag{3.4}$$

2) Quasi-Newton corrector: We now use the following quasi-Newton iterative scheme to improve our approximation at ν_1

$$x_1^{(n+1)} = x_1^{(n)} - f_x^{(m)}(x_1{}^{(0)}, \nu_1)^{-1} f^{(m)}(x_1^{(n)}, \nu_1) \tag{3.5}$$

If k is the total number of iterations of (3.5), then $\bar{u}_F := x_1^{(k)}$ and $f^{(m)}(\bar{u}_F, \nu_1) \approx 0$.

As before, define the corresponding point $\bar{u} = (\bar{u}_F, 0, \dots)$ in the infinite dimensional space. We now use the information required for the next predictor step, the numerical inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$, to construct a Newton-like operator near $\bar{u}$ at the parameter value ν_1. Let $J_{F\times F}$ be the numerical inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$ and define the Newton-like operator T by

$$T(u) = u - Jf(u) \tag{3.6}$$

where

$$J := \begin{bmatrix} J_{F\times F} & & 0 & \\ & \mu_m^{-1} & & \\ 0 & & \mu_{m+1}^{-1} & \\ & & & \ddots \end{bmatrix}$$

is the block diagonal matrix which we expect to be close to $(Df(\bar{u}, \nu_1))^{-1}$. Note that T, J, and f all depend on the parameter ν. As in Section 2, we will attempt to show that T is a contraction on a set of the form $\bar{u} + W$ where W has the form (2.1). We now emphasize the dependence of this set $W = W(r)$ on the validation radius r since this approach relies on finding an appropriate $r > 0$ to satisfy a set of conditions. The constants A_s and s may be determined by regularity arguments or otherwise set prior to the computations. As seen in the definition of $W(r)$, these constants determine the size of the region in which we are attempting to show the unique existence of an equilibrium solution.

3.2. Radii polynomials. We now present the formulae for *radii polynomials*. In order to focus on the applicability of validated continuation the justification that these polynomials do, in fact, encode the required bounds Y_n and K_n in (2.5) for the Newton-like operator constructed in (3.6) is delayed to Section 6.

Since the formulae for the polynomials are rather ungainly, let us begin by explicitly stating the information that is used to construct the coefficients.

- d is the degree of the nonlinearity of (1.2).
- m is the number of modes used in the Galerkin projection.
- $M \geq m$ is a computational parameter that allows for the use of explicit values for coefficients of $M - m$ additional modes to decrease truncation error bounds.
- $m_+ \geq m$ is a computational parameter that allows for the use of additional structure in the model to get tighter truncation error bounds.
- $\bar{u}_F \in \mathbb{R}^m$ is the numerical zero produced by the predictor-corrector step.
- $J_{F\times F}$ is the numerical inverse obtained from the predictor-corrector step.
- $(c_p)_n$, $|n| < m$ are the coefficients from the expansion (1.3).
- μ_k, $k \geq 0$ are the eigenvalues for the linear operator L as expressed in (1.3) and
$$\bar{\mu} := \liminf_{n \geq m_+} |\mu_n|.$$
Note that if $|\mu_n|$ is monotonically increasing for $n \geq m_+$, then $\bar{\mu} = |\mu_{m_+}|$.
- s and A_s are positive constants that are related to the regularity of the equilibria.

Observe that given this information we can evaluate the vector

$$f_F(\bar{u}) := \begin{bmatrix} f_0(\bar{u}) \\ \vdots \\ f_{m-1}(\bar{u}) \end{bmatrix}$$

where

$$f_n(\bar{u}) = \mu_n \bar{u}_n + \sum_{p=0}^{d} \sum_{\substack{n_0+\cdots+n_p=n \\ |n_1|,\ldots,|n_p|<m}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_p} \ .$$

We can also set

$$Y_k \geq \begin{cases} \left|J_{F\times F} f_F(\bar{u})\right|_k & \text{if } 0 \leq k < m \\ \frac{|\sum_{p=2}^{d} (c_p \bar{u}^p)_k|}{|\mu_k|} & \text{if } k \geq m \end{cases} \tag{3.7}$$

where

$$(c_p \bar{u}^p)_k = \sum_{\sum n_i = k} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_p}.$$

The following constants are all related to asymptotic bounds on the expansions of the numerical equilibrium $\bar{u}$, and the set $\bar{u} + W$. As such they are related to the regularity of the equilibrium and the coefficients of (1.2). Define

$$\begin{aligned}
\alpha &:= \frac{2}{s-1} + 2 + 3.5 \cdot 2^s \\
C_p &:= \max_k \{|(c_p)_0|, |(c_p)_k||k|^s\} \\
\bar{A} &:= \max_{1 \le k < m} \{|\bar{u}_0|, |\bar{u}_k||k|^s\} \\
A = A(r) &:= \max\{A_s, r(m-1)^s\} \\
C(\bar{A}, A) &:= \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l \binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A(r)^l \\
C_+(\bar{A}, A) &:= \begin{cases} \sum_{l=1}^{d} \sum_{p=2}^{d} l \binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A^l & \text{if } Y_k, R_k = 0 \text{ for all } k \ge m_+ \\ \sum_{p=0}^{d} \alpha^p C_p \bar{A}^p + \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l \binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A^l & \text{otherwise} \end{cases},
\end{aligned}$$

$$V_F^{(0)} := |J_{F \times F}| \, R_F \quad , \quad V_F^{(1)} := \left| I_{F \times F} - J_{F \times F} \cdot Df^{(m)}(\bar{u}_F) \right| \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

where $| \cdot |$ denotes entry-wise absolute value and for $k \in \{0, \cdots, m-1\}$,

$$R_k := \sum_{\substack{\bar{n}=-\infty \\ |k-\bar{n}| \ge m}}^{\infty} \left| \sum_{p=1}^{d} p \sum_{\sum n_i = \bar{n}} (c_p)_{n_0} \bar{u}_{n_1} \dots \bar{u}_{n_{p-1}} \right| \frac{A_s}{|k-\bar{n}|^s} \; .$$

Note that if all c_p have finite expansions, then $V_F^{(0)}$ requires only a finite computation. Observe also that the above implies that $\bar{u}_k \in \frac{\bar{A}}{k^s}[-1, 1]$ and $\tilde{w}_k \subset \frac{A}{k^s}[-1, 1]$ for all k.

The validation procedure also requires bounds on the errors due to truncating modes $k \ge m$. These bounds come in the following form:

$$\epsilon_n := \sum_{l=1}^{d} \sum_{p=l}^{d} l \binom{p}{l} \epsilon_n(p, l, M) \tag{3.8}$$

where

$$\begin{aligned} &\epsilon_n(p, l, M) := \\ &\min \left\{ \frac{p \alpha^{p-1} C_p \bar{A}^{p-l} A^l}{(M-1)^{s-1}(s-1)} \left[\frac{1}{(M-n)^s} + \frac{1}{(M+n)^s} \right], \frac{\alpha^p C_p \bar{A}^{p-l} A^l}{n^s} \right\}, \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} &C_n(p, j, l, M) := \\ &\sum_{|\bar{n}| < (p-l)(m-1)+M} \left| \sum_{\substack{\sum n_i = \bar{n} \\ |n_0| < M \\ |n_1|, \dots, |n_{p-l}| < m}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right| \left(\sum_{\substack{\sum n_i + \bar{n} = n \\ m \le |n_1|, \dots, |n_j| < M}} \frac{A_s^j}{|n_1|^s \cdots |n_j|^s} \right) . \end{aligned} \tag{3.10}$$

For notational purposes, we also define m-vectors containing these bounds for modes $n = 0, \dots, m-1$ as follows.

$$\epsilon_F := \begin{bmatrix} \epsilon_0 \\ \vdots \\ \epsilon_{m-1} \end{bmatrix}, \quad \text{and} \quad C_F(p,j,l,M) := \begin{bmatrix} C_0(p,j,l,M) \\ \vdots \\ C_{m-1}(p,j,l,M) \end{bmatrix}.$$

Note that these bounds are computable in that they require only a finite number of computations. In addition, increasing the computational parameter M has the effect of increasing the computational work in order to decrease the bounds.

We now use bounds (3.9) and (3.10) to define *radii polynomials*, $P_n(r)$. These polynomials are designed to encode the bounds required by Theorem 2.1. More specifically, as is demonstrated in Section 6, the polynomials are constructed so that $P_n(r) < 0$ implies that $Y_n + K_n - |\tilde{w}_n| < 0$ on the set

$$\bar{u} + W(r) = \bar{u} + \left(\prod_{k=0}^{m-1} [-r, r] \times \prod_{k=m}^{\infty} \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s} \right] \right). \tag{3.11}$$

Definition 3.1. *To simplify notation, the* finite radii polynomials, *$P_0, \dots, P_{m-1}$, are given as an m-vector $P_F(r) = (P_0(r), \dots, P_{m-1}(r))^t$. Define*

$$P_F(r) := \sum_{n=0}^{d} C_F(n) r^n \tag{3.12}$$

where the coefficients are

$$C_F(n) := \begin{cases} C_F^Y + C_F^K(0) & n = 0 \\ C_F^K(1) - 1 & n = 1 \\ C_F^K(n) & n = 2, \dots, d. \end{cases}$$

The right hand terms are defined as follows. The individual terms of the vectors $C_F^K(i)$ are chosen to satisfy

$$C_k^K(i) \geq \left(\sum_{l=\max\{2,i\}}^{d} \sum_{p=l}^{d} l \binom{p}{l} \binom{l}{i} |J_{F\times F}| C_F(p, l-i, l, M) \right. \\ \left. + \begin{cases} |J_{F\times F}|\epsilon_F + V_F^{(0)} & i = 0 \\ V_F^{(1)} & i = 1 \\ 0 & \textit{otherwise} \end{cases} \right)_k . \tag{3.13}$$

and similarly

$$C_F^Y = Y_F \ . \tag{3.14}$$

where $|\cdot|$ and the bounds are computed component-wise.

Observe, again, that determining these bounds require only a finite number of computations.

Definition 3.2. *For $k \geq m$, the* tail radii polynomial *is*

$$P_k(r) = \begin{cases} \frac{|\sum_{p=0}^{d} (c_p \bar{u}^p)_k|}{|\mu_k|} + \frac{C(\bar{A}, A(r))}{|\mu_k| k^s} - \frac{A_s}{k^s} & m \leq k < m_+ \\ \frac{C_+(\bar{A}, A)}{|\mu_k| k^s} - \frac{A_s}{k^s} & k \geq m_+ \end{cases}$$

where, again,

$$(c_p\bar{u}^p)_k = \sum_{\sum n_i = k} (c_p)_{n_0}\bar{u}_{n_1}\cdots\bar{u}_{n_p}.$$

Definition 3.3. *Consider the radii polynomials consisting of the finite radii polynomials P_k, $k = 0,\ldots,m-1$, and the tail radii polynomials, P_k, $k \geq m$. A positive real number r is a* validation radius *if $P_k(r) < 0$ for all $k \geq 0$.*

The proof of the following theorem is presented in Section 6

Theorem 3.4. *If there exists a validation radius $r > 0$ and the eigenvalues μ_k satisfy $|\mu_k| \to \infty$, then there exists a unique equilibrium solution of* (3.1) *in $\bar{u}+W(r)$.*

We now present a procedure for computing a validation radius that satisfies the hypotheses of Theorem 3.4. In particular, this procedure describes a natural order for defining the decay constants A_s, s, and A. The constants A_s and s reflect regularity properties of the equation and should be chosen either from numerical simulations or analysis. In this approach, we choose to treat $A = A(r)$ as a constant. The rationale for this choice is that from a computational perspective, we would like to find $r > 0$ solving simple constructions of the finite radii inequalities $P_0(r) < 0, \cdots, P_{m-1}(r) < 0$ without having to simultaneously control the more complicated effects from A on the coefficients of these polynomials as well as on the tail polynomials P_k, $k \geq m$. A practical way to achieve this goal is to set $A = A_s$ at the beginning of the procedure and then check in the end that a solution $r > 0$ to $P_0(r) < 0, \cdots, P_{m-1}(r) < 0$ also satisfies $r(m-1)^s \leq A_s$.

Here, for the sake of simplicity, we set $M = m$. If the truncation error bounds prove too large for the computations, then M should be increased as described in Remark 6.3 in Section 6. Finally, we add a condition which reduces the check of the tail polynomials $P_k(r) < 0$, $k > m$ to a finite number of computations. The following procedure outlines this approach.

Procedure 3.5. *Suppose that the eigenvalues μ_k are such that $|\mu_k| \to \infty$. Suppose further that we may choose $m, m_+, \bar{m} \in \mathbb{N}$, $\bar{m} \geq m_+ \geq m$, and $\bar{\mu} > 0$ such that*

1. *m is the Galerkin projection dimension used for numerical continuation,*
2. *m_+ is the parameter used in the computation of $C_+(\bar{A}, A)$, and*
3. *$\bar{m}$ measures where the tail terms are bounded from below by $\bar{\mu}$ as follows: for all $k \geq \bar{m}$, $|\mu_k| \geq |\bar{\mu}|$.*

Set $M = m$.

Remark: m should be chosen to give the expected nonzero modes along the bifurcation branch under study and $\bar{m} = m_+ = (2d+1)(m-1)+1$ if $(c_p)_n = 0$ for all $n \neq 0$ and the eigenvalues, μ_k are monotonically increasing in magnitude after $k = (2d+1)(m-1)$.

Fix the decay constants

$$s \geq 2 \quad \textit{and} \quad A_s > 0. \tag{3.15}$$

Remark: In practice, A_s and s should be determined by regularity properties of the equation.

Set $A := A_s$. Using the finite radii polynomials given in Definition 3.1*, for $k = 0, \cdots, m-1$, numerically compute $I_k := \{r > 0 \mid P_k(r) < 0\}$ and*

$$\mathcal{I} := \bigcap_{k=0}^{m-1} I_k\ . \tag{3.16}$$

Check that $\mathcal{I} \neq \emptyset$.

Remark: If $\mathcal{I} = \emptyset$, begin the procedure again either by choosing m larger or by choosing s larger and/or A_s smaller in (3.15).

Check that there exists $\bar{r} \in \mathcal{I}$ such that

$$\bar{r} \leq \frac{A_s}{(m-1)^s} \ . \tag{3.17}$$

Remark: If such an $\bar{r}$ exists, then $A = A_s = \max\{A_s, \bar{r}(m-1)^s\}$. This in turn implies that component-wise $P_F(\bar{r}) < 0$. If $\bar{r}$ does not exist, then begin the procedure again either by choosing m larger or by choosing s larger and/or A_s smaller in (3.15).

Check the inequalities

$$P_m(\bar{r}) < 0, \ \cdots, \ P_{\bar{m}-1}(\bar{r}) < 0 \ \textit{and} \ \frac{C(\bar{A}, A)}{|\bar{\mu}|} - A_s < 0.$$

Remark: If any of these inequalities fails, begin the procedure again either by choosing m larger or by choosing s larger and/or A_s smaller in (3.15).

Observe that if Procedure 3.5 is successful, the hypotheses of Theorem 3.4 are satisfied with validation radius $\bar{r}$.

4. Computational cost. We now provide a rough comparison of the cost of continuation with the cost of validated continuation for PDEs of the form

$$u_t = L(u, \nu) - u^3 \ . \tag{4.1}$$

Since the degree of the polynomial nonlinearity in (4.1) is cubic and we use a Newton-like operator in the continuation procedure, the most expensive terms of the computation involve m^3 operations, where m is the number of modes used in the Galerkin projection

$$f_k^{(m)}(u_F, \nu) = \mu_k(\nu) u_k - \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} u_{n_1} u_{n_2} u_{n_3}, \qquad k = 0, \ldots, m-1. \tag{4.2}$$

With this in mind we count the number of m^3 operations for both approaches to obtain an estimate for the asymptotic costs and conclude with statistics obtained from calculations for the Swift-Hohenberg and Cahn-Hilliard equations.

4.1. Cost of continuation. We decompose the analysis of the cost of continuation into four steps, assuming that we begin with an approximate zero x_0 at ν_0.

Step 1. In order to get the Euler predictor (3.4), we need to evaluate the vector $-f_x^{(m)}(x_0, \nu_0)^{-1} f_\nu^{(m)}(x_0, \nu_0)$. This requires computing the m by m matrix $f_x^{(m)}(x_0^{(0)}, \nu_0)$, where for $0 \leq i, j < m$,

$$\left[f_x^{(m)}(x_0^{(0)}, \nu_0)\right]_{i+1,j+1} = \delta_{i,j}\mu_i - 3\Big(\sum_{\substack{n_1+n_2+j=i \\ |n_i|<m}} [x_0^{(0)}]_{|n_1|}[x_0^{(0)}]_{|n_2|} + \sum_{\substack{n_1+n_2-j=i \\ |n_i|<m}} [x_0^{(0)}]_{|n_1|}[x_0^{(0)}]_{|n_2|}\Big).$$

This involves the evaluation of $2m^2$ sums demanding $2m-1$ multiplications and $2m-2$ additions each. Therefore, determining $f_x^{(m)}(x_0^{(0)}, \nu_0)$ requires $8m^3$ operations.

Next, we compute the LU decomposition of $f_x^{(m)}(x_0^{(0)}, \nu_0)$ in order to compute the action of its inverse on $f_\nu^{(m)}(x_0, \nu_0)$. This involves $\frac{2}{3}m^3$ operations. In our case, $f_\nu^{(m)}(x_0, \nu_0) = x_0$, requiring no additional cost. The predictor is then

$$\begin{cases} x_1^{(0)} &= x_0 - \Delta\nu {f_x^{(m)}(x_0, \nu_0)}^{-1} x_0 \\ \nu_1 &= \nu_0 + \Delta\nu. \end{cases}$$

Step 2. We now start the corrector. To construct the quasi-Newton operator (3.5), we need the action of the inverse of $f_x^{(m)}$ at the predictor $(x_1^{(0)}, \nu_1)$. As seen before, it costs $8m^3$ to evaluate $f_x^{(m)}(x_1^{(0)}, \nu_1)$ and $\frac{2}{3}m^3$ to compute its inverse using LU decomposition. Note that we need to compute the LU decomposition only at the first step.

Step 3. At the j^{th} iteration of (3.5), we need to evaluate $f^{(m)}(x_1^{(j-1)}, \nu_1)$. Its i^{th} component is

$$\begin{aligned} [f^{(m)}(x_1^{(j-1)}, \nu_1)]_i = \mu_i(\nu_1)[x_1^{(j-1)}]_i \\ - \sum_{\substack{n_1+n_2+n_3=i \\ |n_i|<m}} [x_1^{(j-1)}]_{|n_1|}[x_1^{(j-1)}]_{|n_2|}[x_1^{(j-1)}]_{|n_3|} \end{aligned}$$

which requires at least $3m^2$ operations to evaluate. Since $f^{(m)}$ has m components, we get a total of $3m^3$. If k is the total number of iterations of the corrector, then this step requires $3km^3$ operations.

Step 4. The corrector ends when $||f^{(m)}(x_1^{(k)}, \nu_1)|| <$ `tolerance`. Let $\bar{a}_F := x_1^{(k)}$. Evaluating the function at $(\bar{u}_F, \nu_1)$ is another $3m^3$. Now, note that we have to compute the action of the inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$ to get the predictor for the next step. Recall $J_{F\times F}$ is the numerical inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$ computed as before using an LU decomposition. Explicitly computing all the coefficients in $f_x^{(m)}(\bar{u}_F, \nu_1)$ requires an extra $2m^3$ operations. We do not count the m^3 involved to get the next predictor, since that is part of the next predictor-corrector step.

Combining the costs of the four above mentioned steps suggests that the cost of one application of the predictor-corrector algorithm is on the order of $(20 + 3k)m^3$, where k is the number of iterations in the quasi-Newton corrector.

4.2. Cost of validation. We now show that the extra cost of performing validation for a cubic function ($d = 3$) with constant function coefficients is of the order of $6m^3$ operations where m is the projection dimension used for continuation. The additional cost comes primarily from computing the coefficients of the radii polynomials. In the following, we construct $m_+ = d(m-1)+1 = 3m-2$ polynomials $P_0, \ldots, P_{3m-3}$ using Procedure 3.5 and calculate the associated computational cost. Both to simplify the presentation and because this is what is used to perform the computations presented in Section 5, we set $\bar{m} = m_+ = d(m-1)+1$, with $|\mu_k| \geq |\mu_{\bar{m}}|$ for all $k \geq \bar{m}$, and $M = m$. As described in Procedure 3.5, $A = A_s$ and we consider fixed $s > 2$ and $A_s > 0$.

The only nonlinear term of (4.1) is a monomial of degree 3. Thus, if $p \neq 3$, then $C_k(p, j, l, M) = 0$. In addition, we have set $M = m$. Hence, if $j \neq 0$, then $C_k(p, j, l, M) = 0$ (see Remark 6.3). Therefore, the only nonzero terms of this form are

$$C_k(3,0,l,m) = \left| \sum_{\substack{n_1+n_2+n_3=k \\ |n_1|,|n_2|,|n_3|<m}} \bar{u}_{n_1} \cdots \bar{u}_{n_{3-l}} \right|. \tag{4.3}$$

Hence, by (3.13) we set

$$C_k^K(0) \geq (|J_{F\times F}|\epsilon_F)_k + V_k^{(0)} \tag{4.4}$$

for $0 \leq k < m$ and $|\cdot|$ denotes the component-wise absolute value. Note that it is possible to get an analytic upper bound on $V_k^{(0)}$ using Lemma 6.2 in which case computing $V_k^{(0)}$ doesn't require any m^3 operations. Hence, all necessary computations for $C_F^K(0)$ are of order less than m^3. Using (3.13),

$$C_k^K(1) \geq V_k^{(1)}$$

for $0 \leq k < m$ and evaluating $V_F^{(1)}$ does not require any m^3 operations.

Finally, combining (3.13) and (4.3)

$$C_F^K(2) \geq 6|J_{F\times F}|C_F(3,0,2,m)$$

where $C_n(3,0,2,m) = |\bar{u}_n|$ and

$$C_F^K(3) \geq 3|J_{F\times F}|C_F(3,0,3,m)$$

where $C_n(3,0,3,m) = 1$.

The last coefficient to compute to get all the finite radii polynomials (3.14) is

$$C_F^Y \geq |J_{F\times F} f_F(\bar{u})|$$

where again $|\cdot|$ denotes the component-wise absolute value. This comes with no extra m^3 cost since $f_F(\bar{u}) = f^{(m)}(\bar{u}_F, \nu_1)$ was computed in Step 4 of the predictor-corrector algorithm.

The next step in Procedure 3.5 is checking for the existence of a validation radius $r > 0$. This requires finding the numerical zeros of each of the cubic polynomials $P_0, \cdots, P_{m-1}$, constructing $I_0, \cdots, I_{m-1}$ where I_k are closed intervals such that $I_k \subsetneq \{r > 0 | P_k(r) < 0\}$, and finally checking for a non-empty intersection $\mathcal{I} = \cap_{k=0}^{m-1} I_k$. All of these steps are of order less than m^3.

Assuming there exists a positive $\bar{r} \in \mathcal{I}$ such that $\bar{r}(m-1)^s \leq A_s$, we construct and evaluate the tail radii polynomials $P_m, \cdots, P_{3m-1}$ at $\bar{r}$. We compute Y_k using (3.7) which requires $6m^3$ operations since we need to evaluate $f_k(\bar{u})$ for $k = m, \cdots, 3m-3$.

Using Definition 3.2 and the assumption that $A = A_s$ we compute

$$C(\bar{A}, A) = \sum_{l=1}^{3} l \binom{3}{l} \alpha^3 \bar{A}^{3-l} A^l = 3\alpha^3 A_s (\bar{A} + A_s)^2.$$

This latter step and the remaining computations for Procedure 3.5 are all of order less than m^3.

In summary, the m^3 cost of computing the coefficients of the radii polynomials is $6m^3$. Thus the additional cost of validation is on the order of $6m^3$ operations.

4.3. Relative cost. Combining the results of Sections 4.1 and 4.2 suggests that asymptotically the ratio of the cost of validated continuation to the cost of traditional continuation is

$$\frac{26+3k}{20+3k}.$$

where k is the number of iterations performed in the corrector step. We tested this hypothesis again two fourth order partial differential equations with cubic nonlinearities, Swift-Hohenberg and Cahn-Hilliard. The results are discussed in greater detail in Section 5. For the moment we are only interested in the relative times of computation.

We performed validated continuation for 46 predictor-corrector steps involving a total of 90 quasi-Newton iterations for the cubic Swift-Hohenberg equation. We repeated the computations without validation. The ratio of elapsed time for validated continuation to the time used for continuation alone was ≈ 1.156. Given that we had an average of 90/46 iterations per predictor-corrector step, this is close to the rough estimate of $\frac{26+3\cdot 90/46}{20+3\cdot 90/46} \approx 1.232$ given by the above arguments.

Similarly, we performed validated continuation for 15 predictor-corrector steps involving a total of 37 quasi-Newton iterations for Cahn-Hilliard. Again, we repeated the computations without validation. The ratio of elapsed time for validated continuation to the time used for continuation alone was ≈ 1.173. Given that we had an average of 37/15 iterations per predictor-corrector step, the asymptotic ratio is $\frac{26+3\cdot 37/15}{20+3\cdot 37/15} \approx 1.219$.

The results of these computations are summarize in Figure 4.1.

PDE	m	$\frac{\#\ \text{iterations}}{\#\ \text{steps}}$	Experimental Ratio	Estimated Ratio $\frac{26+3k}{20+2k}$
S-H	27	1.96	1.156	1.232
C-H	60	1.65	1.173	1.219

FIG. 4.1. *Comparison of the asymptotic ratios.*

5. Sample results.

To demonstrate the practical applicability of validated continuation we turn to two model problems, Cahn-Hilliard and Swift-Hohenberg. In both cases we follow a branch of equilibria and validate at each parameter value of the continuation. In the case of Swift-Hohenberg we also use interval arithmetic to evaluate the radii polynomials, thus allowing us to rigorously verify the existence and uniqueness of the equilibria.

5.1. Cahn-Hilliard.

The Cahn-Hilliard equation was introduced in [1] as a model for the process of phase separation of a binary alloy at a fixed temperature. On a one-dimensional domain it takes the form

$$\begin{aligned} u_t &= -(\frac{1}{\nu}u_{xx} + u - u^3)_{xx}\ , \quad x \in [0,1] \\ u_x &= u_{xxx} = 0\ , \quad \text{at } x = 0,1. \end{aligned} \tag{5.1}$$

The assumption of an equal concentration of both alloys is formulated as

$$\int_0^1 u(x,\cdot)\,dx = 0$$

Note that when looking for the equilibrium solutions of (5.1), it is sufficient to work with the Allen-Cahn equation

$$\begin{aligned} &\frac{1}{\nu}u_{xx} + u - u^3 = 0 \\ &u_x = 0 \quad \text{at } x = 0, 1. \end{aligned} \tag{5.2}$$

Re-writing (5.2) in the form of (1.2), the linear operator is $L(\cdot,\nu) = \frac{1}{\nu}\frac{\partial^2}{\partial x^2} + 1$ and the polynomial nonlinearity is of degree $d = 3$ with coefficient functions

$$(c_p)_n = \begin{cases} -1 & p = 3 \text{ and } n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Applying Procedure 3.5 with $M = m = 60$, $s = 3$, and $A_s = 0.01$, results in the branch of equilibria indicated in Figure 5.1 where each point represents the center of the infinite dimensional validation set of the form $\bar{u} + W(\bar{r})$, containing a unique equilibrium of (5.1). These are the points used to obtain the cost estimates presented in Figure 4.1. To avoid drowning the reader in large lists of numbers, we only provide the detailed numerical output at one parameter value.

Validated Result 5.1. *Let* $\nu = 43.57415358799057$. *Then,*

$$\bar{r} = 4.846104201261526 \times 10^{-8}$$

is a validation radius for the numerical zero $\bar{u}_F$ *given in Figure* 5.2. *Thus, there exists a unique equilibrium for* (5.1) *in the validation set*

$$(\bar{u}_F, 0) + \prod_{k=0}^{59}[-\bar{r}, \bar{r}] \times \prod_{k=60}^{\infty}\left[-\frac{0.01}{k^3}, \frac{0.01}{k^3}\right].$$

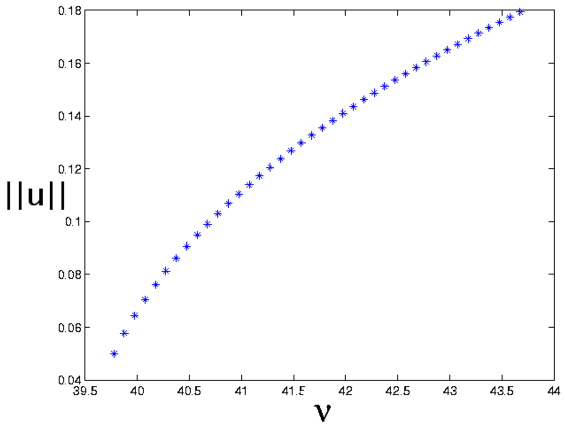

Fig. 5.1. *Validated continuation in* ν *for the Cahn-Hilliard equation on* $[0, 1]$.

5.2. Swift-Hohenberg. The Swift-Hohenberg equation

$$u_t = f(u,\nu) = \left\{\nu - \left(1 + \frac{\partial^2}{\partial x^2}\right)^2\right\} u - u^3, \qquad u(\cdot,t) \in L^2\left(0, \frac{2\pi}{L_0}\right),$$

$$u(x,t) = u\left(x + \frac{2\pi}{L_0}, t\right), \qquad u(-x,t) = u(x,t), \qquad \nu > 0, \tag{5.3}$$

k	$\bar{u}_k$
1	$1.773844149032812 \times 10^{-1}$
3	$-7.601617928785714 \times 10^{-4}$
5	$3.271672072176762 \times 10^{-6}$
7	$-1.408100160017936 \times 10^{-8}$
9	$6.060344382471457 \times 10^{-11}$
11	$-2.608320515803233 \times 10^{-13}$
13	$1.122598345048980 \times 10^{-15}$
15	$-4.831561184682242 \times 10^{-18}$
17	$2.079457485469691 \times 10^{-20}$
19	$-8.949770271275235 \times 10^{-23}$
21	$3.851880360024139 \times 10^{-25}$
23	$-1.657801422354123 \times 10^{-27}$
25	$7.134947464114615 \times 10^{-30}$
27	$-3.070770234245256 \times 10^{-32}$
29	$1.321605495419571 \times 10^{-34}$
31	$-5.687926883858248 \times 10^{-37}$
33	$2.447955395983479 \times 10^{-39}$
35	$-1.053537452697732 \times 10^{-41}$
37	$4.534120813401209 \times 10^{-44}$
39	$-1.951337823193323 \times 10^{-46}$
41	$8.397842606319005 \times 10^{-49}$
43	$-3.614086242431264 \times 10^{-51}$
45	$1.555336697148314 \times 10^{-53}$
47	$-6.693373497802139 \times 10^{-56}$
49	$2.880447985844179 \times 10^{-58}$
51	$-1.239563989182517 \times 10^{-60}$
53	$5.334225825486573 \times 10^{-63}$
55	$-2.295445428599939 \times 10^{-65}$
57	$9.877687199770852 \times 10^{-68}$
59	$-4.250458946966345 \times 10^{-70}$
$\geq$60	0

FIG. 5.2. *The numerical zero $\bar{u}_F$ obtained by continuation for the Cahn-Hilliard equation at $\nu = 43.57415358799057$. Note that all even coefficients are* 0.

was originally introduced to describe the onset of Rayleigh-Bénard heat convection [8], where L_0 is a fundamental wave number for the system size $2\pi/L_0$. The parameter ν corresponds to the Rayleigh number and its increase is associated with the appearance of multiple solutions that exhibit complicated patterns. For the computations presented here we fixed $L_0 = 0.65$.

Re-writing (5.3) in the form of (1.2), the linear operator is $L(\cdot, \nu) = \nu - (1 + \frac{\partial^2}{\partial x^2})^2$ and the polynomial nonlinearity is of degree $d = 3$ with coefficient functions

$$(c_p)_n = \begin{cases} -1 & p = 3 \text{ and } n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Applying Procedure 3.5 with $M = m = 27$, $s = 4$, and $A_s = 0.002$, results in the branch of equilibria indicated in Figure 5.3 where each point represents the center of the infinite dimensional validation set of the form $\bar{u} + W(\bar{r})$, containing a unique equilibrium of (5.3). Again, these are the points used to obtain the cost estimates presented in Figure 4.1.

As in the case of the Cahn-Hilliard equation, we only include the output at one point on the branch of the Figure 5.3.

VALIDATED RESULT 5.2. *Let $\nu = .6674701641462312$. Then*

$$\bar{r} = 1.998167170445973 \ \times \ 10^{-9}$$

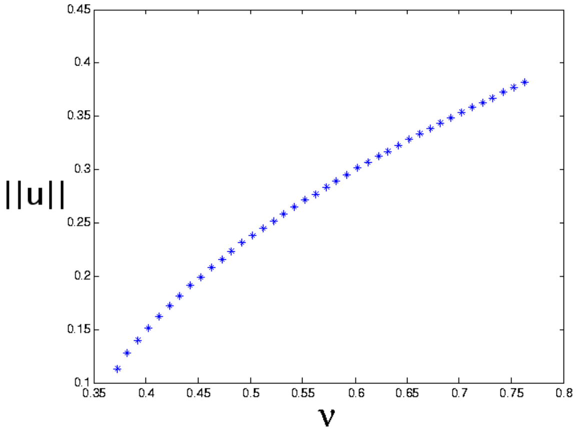

FIG. 5.3. *Validated continuation in ν for the Swift-Hohenberg equation at $L_0 = 0.65$.*

k	$\bar{u}_k$
1	$-3.359998711939212 \times 10^{-1}$
3	$4.824376413178060 \times 10^{-3}$
5	$-1.761066797314072 \times 10^{-5}$
7	$7.535865329757206 \times 10^{-8}$
9	$-2.790895103063484 \times 10^{-10}$
11	$9.411109491227775 \times 10^{-13}$
13	$-3.113936321690645 \times 10^{-15}$
15	$1.007016979585499 \times 10^{-17}$
17	$-3.200410295859874 \times 10^{-20}$
19	$1.003878817132397 \times 10^{-22}$
21	$-3.114244522738206 \times 10^{-25}$
23	$9.573156964813860 \times 10^{-28}$
25	$-2.920394630491221 \times 10^{-30}$
$\geq$26	0

FIG. 5.4. *The numerical zero $\bar{u}_F$ obtained by continuation for the Swift-Hohenberg equation at $\nu = .6674701641462312$ and $L_0 = 0.65$. All even coefficients are 0.*

is a validation radius for the numerical zero $\bar{u}_F$ whose coefficient values are indicated in Figure 5.4. Thus, there exists a unique equilibrium solution for (5.3) in the validation set

$$(\bar{u}_F, 0) + \prod_{k=0}^{26} [-\bar{r}, \bar{r}] \times \prod_{k=27}^{\infty} \left[-\frac{0.002}{k^4}, \frac{0.002}{k^4} \right].$$

Observe that in all the above mentioned calculations floating point round-off errors have not been controlled, thus at this point one cannot claim that the validation results presented above are rigorous. However, with additional computational effort a computer-assisted proof can be obtain. To be more precise, our technique relies on the existence of a validation radius $\bar{r}$ making all radii polynomials strictly negative. Hence, rigorous validation follows if the inequalities are satisfied when one includes bounds to control the possible of floating point errors. The first step in checking these inequalities on this level is to obtain floating point outer bounds for the coefficients of the polynomials. This can be done by defining each entry of

$$\bar{u}_F, \; f^{(m)}(\bar{u}_F, \nu), \; J_{F \times F}, \; f_x^{(m)}(\bar{u}_F, \nu), \; \mu_k(\nu), \; A_s, \text{ and } s$$

to be an interval and then computing (3.13), (3.14) and the quantities in Definition 3.2 using interval arithmetic. The resulting radii polynomials, which we denote by $\tilde{P}_k$, have interval coefficients. Let $\bar{r}$ be the smallest representable number such that using interval arithmetic, the corresponding finite radii polynomials may be shown to be strictly contained in $(-\infty, 0)$. Assume such an $\bar{r}$ exists. If, again using interval arithmetic, $\bar{r}(m-1) - A_s \subset (-\infty, 0)$ and the intervals obtained from evaluating tail radii polynomials at $\bar{r}$ are strictly contained in $(-\infty, 0)$, i.e. $\tilde{P}_k(\bar{r}) \subset (-\infty, 0)$ for all $k \geq m$, then the hypotheses of Theorem 3.4 are satisfied and we obtain a proof.

The above mentioned computations were performed using the interval arithmetic package in *Matlab*. Thus, we can state the following theorem.

THEOREM 5.3. *Each point in Figure* 5.3 *represents the center of an infinite dimensional set of the form*

$$\bar{u}_F + \prod_{k=0}^{26}[-\bar{r}, \bar{r}] \times \prod_{k=27}^{\infty}\left[-\frac{0.002}{k^4}, \frac{0.002}{k^4}\right]$$

containing a unique equilibrium to (5.3).

The actual values for the various numerical zeros and validation radii are of limited interest and thus not presented. Of greater interest is understanding how large are the errors induced by the floating point computations as opposed to the magnitudes of the floating point computations of $P_k(\bar{r})$, $k \geq 0$, where $\bar{r}$ is the validation radius.

Let us restrict our attention to the equilibrium described by Validated Result 5.2. Following Procedure 3.5 at this parameter value, beginning using radii polynomials with interval coefficients and performing the computations with interval arithmetic leads to an interval of potential validation radii

$$\mathcal{I} = [3.373873850437414 \times 10^{-9}, 9.003755731999980 \times 10^{-4}].$$

Hence, we choose $\bar{r} = 3.373873850437415 \times 10^{-9}$. There are 53 inclusions that need to be satisfied, those arising from the $2m - 2 = 52$ tail radii polynomials with interval coefficients and the one associated with the inequality (3.17). The fact that the inclusions are satisfied leads to the conclusion of Theorem 5.3 at this parameter value. Again, rather than listing all 53 inclusions let us focus on the two extremes, the interval closest to 0

$$\tilde{P}_{27}(\bar{r}) = -3.191484496597115 \times 10^{-11} \pm 7.037497555236307 \times 10^{-24}$$

and the interval the farthest from 0

$$-1.973098298147102 \times 10^{-3} \pm 8.673617379884037 \times 10^{-19}$$

corresponding to the inequality (3.17). Observe that in both cases, the width of the interval induced by the floating point errors is more than ten orders of magnitude smaller than the value of the center. Furthermore, this behavior is typical for all the validation computations that were performed. This suggests that it is reasonably safe to assume that a validated equilibrium is a true equilibrium.

6. Justification of radii polynomials. In this section, we describe the construction of the radii polynomials that were defined in Section 3.2 and encode the bounds required for Theorem 2.1. We begin by computing the required bounds Y_n and K_n in (2.5) for the Newton-like operator constructed in (3.6).

Using a Taylor expansion of the Newton-like operator $T(u) = u - Jf(u)$ around the numerical equilibrium $\bar{u} = (\bar{u}_F, 0, 0, \dots)$ leads to

$$\begin{aligned} DT(\bar{u}+w')w &= [I - J\cdot Df(\bar{u}+w')]\, w \\ &= \left(I - J\left(Df(\bar{u}) + D^2 f(\bar{u})(w') + \cdots + \frac{D^l f(\bar{u})}{(l-1)!}(w')^{l-1} + \cdots + \frac{D^d f(\bar{u})}{(d-1)!}(w')^{d-1}\right)\right) w \\ &= [I - J\cdot Df(\bar{u})]w - J\left(\sum_{l=2}^{d} \frac{D^l f(\bar{u})}{(l-1)!}(w')^{l-1}\right) w \\ &= [I - J\cdot Df(\bar{u})]\, w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} \frac{p! c_p \bar{u}^{p-l}(w')^{l-1}}{(l-1)!(p-l)!}\right) w \\ &= [I - J\cdot Df(\bar{u})]\, w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l} c_p \bar{u}^{p-l}(w')^{l-1}\right) w\ . \end{aligned}$$

In the rest of the section, we will make use of the discrete convolution of bi-infinite vectors i.e. considering two bi-infinite vectors $(a_j)_{j\in\mathbb{Z}}$, $(b_j)_{j\in\mathbb{Z}}$, we define their convolution by

$$(a*b)_k = \sum_{n=-\infty}^{\infty} a_n b_{k-n} = \sum_{\substack{k_1+k_2=k \\ k_i\in\mathbb{Z}}} a_{k_1} b_{k_2} \quad , \quad k \in \mathbb{Z}\ .$$

Expanding into Fourier modes, we can write the nonlinear part in terms of convolution

$$\begin{aligned} DT(\bar{u}+w')w &= [I - J\cdot Df(\bar{u})]\, w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l} c_p \bar{u}^{p-l}(w')^{l-1}\right) * w \\ &= [I - J\cdot Df(\bar{u})]\, w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l} (c_p \bar{u}^{p-l}) * (w')^{l-1} * w\right)\ . \end{aligned} \tag{6.1}$$

Thus,

$$(c_p\bar{u}^{p-l})*((w')^{l-1})*w = \left[\sum_{\bar{n}} \left(\sum_{\sum n_i = \bar{n}} (c_p)_{n_0}\bar{u}_{n_1}\cdots \bar{u}_{n_{p-l}}\right)\left(\sum_{\sum n_i = n-\bar{n}} w'_{n_1}\cdots w'_{n_{l-1}} w_{n_l}\right)\right]_n .$$

Here, $[\cdot]_n$ denotes the bi-infinite vector indexed by $n \in \mathbb{Z}$ and $(\cdot)_k$ denotes the entry at index k.

We use this expansion to compute the bounds

$$\begin{aligned} K_k &\geq \max |(DT(\bar{u}+W)W)_k| \\ &\geq \max \left| [I - J\cdot Df(\bar{u})]\tilde{w} - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l}(c_p\bar{u}^{p-l}) * \tilde{w}^l\right)\right| \end{aligned}$$

where, as in Section 2, W has the form (2.1).

The block-diagonal structure of J allows us to decompose (6.1) into a finite, m-dimensional piece and the infinite dimensional tail terms. For the following, we adopt the notation $[\cdot]_F$ to denote the m-vector whose nth entry is computed at index value $n-1$ for $1 \leq n \leq m$, the subscript $\tilde{F}$ to denote the bi-infinite vector in which the kth entries for $|k| \geq m$ are set equal to 0, and the subscript $\tilde{I}$ to denote the bi-infinite vector in which the kth entries for $|k| < m$ are set equal to 0. We begin with the following decomposition of the finite part of the linear term.

$$\begin{aligned}
\{[I - J\cdot Df(\bar{u})]w\}_F &= w_F - [J\cdot Df(\bar{u})w]_F \\
&= w_F - J_{F\times F}\,[Df(\bar{u})w]_F \\
&= w_F - J_{F\times F}\cdot Df_F(\bar{u})w \\
&= w_F - J_{F\times F}\cdot \left[Df^{(m)}(\bar{u}_F)w_F + R_F(\bar{u}, w)\right] \\
&= \left[I_{F\times F} - J_{F\times F}\cdot Df^{(m)}(\bar{u}_F)\right] w_F - J_{F\times F}\cdot R_F(\bar{u}, w)\ , \qquad (6.2)
\end{aligned}$$

where for $k \in \{0, \cdots, m-1\}$,

$$\begin{aligned}
R_k(\bar{u}, w) &:= \sum_{i=m}^{\infty} \frac{\partial f_k}{\partial u_i}(\bar{u}) w_i \\
&= \sum_{\substack{\bar{n}=-\infty \\ |k-\bar{n}|\geq m}}^{\infty} \left| \sum_{p=1}^{d} p \sum_{\sum n_i = \bar{n}} (c_p)_{n_0} \bar{u}_{n_1} \ldots \bar{u}_{n_{p-1}} \right| \frac{A_s}{|k-\bar{n}|^s}\ . \qquad (6.3)
\end{aligned}$$

It follows that

$$\begin{aligned}
[DT(\bar{u}+W)W]_F \subseteq & \left[I_{F\times F} - J_{F\times F}\cdot Df^{(m)}(\bar{u}_F)\right] \tilde{w}_F - J_{F\times F}\cdot R_F(\bar{u}, w) \\
& - \left(J_{F\times F} \sum_{l=2}^{d} \sum_{p=l}^{d} l \binom{p}{l} [(c_p \bar{u}^{p-l}) * \tilde{w}^l]_F \right). \qquad (6.4)
\end{aligned}$$

For $k \geq m$,

$$(DT(\bar{u}+W)W)_k \subseteq -J(k,k) \sum_{l=1}^{d} \sum_{p=l}^{d} l \binom{p}{l} ((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k. \qquad (6.5)$$

We now focus on finding bounds on the terms given in (6.4) and (6.5). First consider

$$((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k = \sum_{\bar{n}} \left(\sum_{\sum n_i = \bar{n}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right) \left(\sum_{\sum n_i + \bar{n} = k} \tilde{w}_{n_1} \cdots \tilde{w}_{n_l} \right) \qquad (6.6)$$

where p is the degree of the original monomial term of f and $l \in \{1, \ldots, p\}$ is the order of the derivative being taken. One upper bound for (6.6) is given in the following lemma, which uses asymptotic bounds first listed in Section 3.2.

Lemma 6.1. *Let* $\alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$, $\bar{u}_k \in \frac{\bar{A}}{k^s}[-1,1]$, $(c_p)_k \in \frac{C_p}{k^s}[-1,1]$, *and* $\tilde{w}_k \subset \frac{A}{k^s}[-1,1]$ *for all* k. *Then*

$$((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k \subseteq \begin{cases} \frac{\alpha^p C_p \bar{A}^{p-l} A^l}{|k|^s}[-1,1] & k \neq 0 \\ \\ \alpha^p C_p \bar{A}^{p-l} A^l [-1,1] & k = 0. \end{cases}$$

Proof. Note that

$$\begin{aligned}\sum_{\bar n}\left(\sum_{\sum n_i=\bar n}(c_p)_{n_0}\bar u_{n_1}\cdots\bar u_{n_{p-l}}\right)&\left(\sum_{\sum n_i+\bar n=k}\tilde w_{n_1}\cdots\tilde w_{n_l}\right)\\ &\subseteq\sum_{\sum n_i=k}(c_p)_{n_0}\bar u_{n_1}\cdots\bar u_{n_{p-l}}\tilde w_{n_{p-l+1}}\cdots\tilde w_{n_p}\\ &\subseteq\sum_{\sum n_i=k}\frac{C_p}{|n_0|^s}\frac{\bar A}{|n_1|^s}\cdots\frac{\bar A}{|n_{p-l}|^s}\frac{A}{|n_{p-l+1}|^s}\cdots\frac{A}{|n_p|^s}[-1,1]\end{aligned}$$

The remainder of the proof is a modification of [2, Lemma 5.8]. □

In most cases, especially when l is small relative to p, this bound will be too large to use for the low modes. In particular, $\bar u$ may be far from zero, resulting in a large constant $\bar A$. By taking k sufficiently large, the contraction given by $J(k,k)\approx{\mu_k}^{-1}$ will overcome the large bound. A more practical approach for obtaining bounds for the low modes is given by the following lemma. For flexibility in balancing numerical computations (requiring a finite number of operations) with analysis (to obtain truncation bounds), we choose $M\geq m$ to be the dimension used to split these sums.

Lemma 6.2. *For $M\geq m$,*

$$((c_p\bar u^{p-l})*\tilde w^l)_k\subseteq\left(\sum_{j=0}^{l}\binom{l}{j}C_k(p,j,l,M)r^{l-j}+\epsilon_k(p,l,M)\right)[-1,1].$$

Proof. This lemma is a modification of [2, Lemma 5.10] combined with Lemma 6.1. In [2, Lemma 5.10], the bound is split into finite sums and the tail term, bounded by

$$\frac{p\alpha^{p-1}C_p\bar A^{p-l}A^l}{(M-1)^{s-1}(s-1)}\left[\frac{1}{(M-k)^s}+\frac{1}{(M+k)^s}\right][-1,1].$$

We obtain a polynomial in r by rewriting the finite sums as follows:

$$\begin{aligned}&\sum_{\bar n}\Big(\sum_{\substack{\sum n_i=\bar n\\|n_i|<M}}(c_p)_{n_0}\bar u_{n_1}\cdots\bar u_{n_{p-l}}\Big)\Big(\sum_{\substack{\sum n_i+\bar n=k\\|n_i|<M}}\tilde w_{n_1}\cdots\tilde w_{n_l}\Big)\\ &\quad=\sum_{\bar n}\Big(\sum_{\substack{\sum n_i=\bar n\\|n_1|,\dots,|n_{p-l}|<m\\|n_0|<M}}(c_p)_{n_0}\bar u_{n_1}\cdots\bar u_{n_{p-l}}\Big)\Big(\sum_{\substack{\sum n_i+\bar n=k\\|n_i|<M}}\tilde w_{n_1}\cdots\tilde w_{n_l}\Big)\\ &\quad=\sum_{\bar n}\left(\sum_{\substack{\sum n_i=\bar n\\|n_1|,\dots,|n_{p-l}|<m\\|n_0|<M}}(c_p)_{n_0}\bar u_{n_1}\cdots\bar u_{n_{p-l}}\right)\left(\sum_{j=0}^{l}\binom{l}{j}\sum_{\substack{\sum n_i+\bar n=k\\m\leq|n_1|,\dots,|n_j|<M\\|n_{j+1}|,\dots,|n_l|<m}}\tilde w_{n_1}\cdots\tilde w_{n_l}\right)\\ &\quad=\sum_{\bar n}\left(\sum_{\substack{\sum n_i=\bar n\\|n_1|,\dots,|n_{p-l}|<m\\|n_0|<M}}(c_p)_{n_0}\bar u_{n_1}\cdots\bar u_{n_{p-l}}\right)\left(\sum_{j=0}^{l}\binom{l}{j}r^{l-j}[-1,1]\sum_{\substack{\sum n_i+\bar n=k\\m\leq|n_1|,\dots,|n_j|<M\\|n_{j+1}|,\dots,|n_l|<m}}\frac{A_s^j}{|n_1|^s\cdots|n_j|^s}\right)\end{aligned}$$

$$= \sum_{j=0}^{l} \binom{l}{j} r^{l-j} \sum_{|\bar{n}|<(p-l)(m-1)+M} [-1,1] \left| \sum_{\substack{\sum n_i = \bar{n} \\ |n_1|,\dots,|n_{p-l}|<m \\ |n_0|<M}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right| \left(\sum_{\substack{\sum n_i + \bar{n} = k \\ m \le |n_1|,\dots,|n_j| < M \\ |n_{j+1}|,\dots,|n_l|<m}} \frac{A_s^j}{|n_1|^s \cdots |n_j|^s} \right). \qquad \square$$

Remark 6.3. *Note that in Lemma 6.2, $C_k(p,j,l,M)$ captures the contribution to the $(l-j)$th polynomial coefficient from the l-th derivative of the p-th monomial term of f in the Taylor expansion. If $M = m$, then $C_k(p,j,l,M) = 0$ for all $j > 0$ and*

$$C_k(p,0,l,m) = \left| \sum_{\substack{n_0 + \cdots + n_{p-l} = k \\ |n_0|,\dots,|n_{p-l}|<m}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right| .$$

For $M > m$ there is also a (small) contribution to the coefficients of higher degrees of r in the polynomials, while simultaneously decreasing the ϵ_k term. This offers a method for using additional computations to decrease the bound ϵ_k if this bound proves to be too large for the validation procedure.

For notational purposes, set ϵ_F, $C_F(p,j,l,M)$, $V_F^{(0)}$ and $V_F^{(1)}$ to be the m-vectors as defined in Section 3.2. For $0 \le k < m$, we substitute the bounds from Lemma 6.2 into (6.4),

$$\begin{aligned}
(DT(\bar{u}+W)W)_k &\subseteq rV_k^{(1)}[-1,1] + V_k^{(0)}[-1,1] \\
&\quad + \left(-J_{F\times F} \sum_{l=2}^{d} \sum_{p=l}^{d} l \binom{p}{l} \left(\sum_{j=0}^{l} \binom{l}{j} (C_F(p,j,l,M) r^{l-j} + \epsilon_F(p,l,M)) \right) [-1,1] \right)_k \\
&= (|J_{F\times F}|\epsilon_F)_k [-1,1] + rV_k^{(1)}[-1,1] + V_k^{(0)}[-1,1] \\
&\quad + \left(\sum_{l=2}^{d} \sum_{p=l}^{d} \sum_{j=0}^{l} r^{l-j} l \binom{p}{l} \binom{l}{j} |J_{F\times F}| C_F(p,j,l,M) \right)_k [-1,1] \\
&= \left(|J_{F\times F}|\epsilon_F + V_F^{(0)} \right)_k [-1,1] + rV_k^{(1)}[-1,1] \\
&\quad + \left(\sum_{i=0}^{d} r^i \sum_{l=\max\{2,i\}}^{d} \sum_{p=l}^{d} l \binom{p}{l} \binom{l}{i} |J_{F\times F}| C_F(p,l-i,l,M) \right)_k [-1,1]
\end{aligned}$$

where $|\cdot|$ denotes entry-wise absolute value. For $0 \le k < m$, set K_k to be

$$K_k := \sum_{i=0}^{d} C_k^K(i) r^i \ge |(DT(\bar{u}+W)W)_k|$$

where $C_k^K(i)$ satisfies (3.13).

Recall that our goal is to find a polynomial bound for $Y_k + K_k - |\tilde{w}_k|$ for Theorem 2.1. This requires also computing the bounds for Y_k satisfying the following equation.

$$Y_k \ge |(T(\bar{u}) - \bar{u})_k|$$

$$\begin{aligned} &= |[-Jf(\bar u)]_k| \\ &= \left| \left(-J \left[\mu_n \bar u_n + \sum_{p=0}^{d} \sum_{\substack{n_0+\cdots+n_p=n \\ |n_1|,\dots,|n_p|<m}} (c_p)_{n_0} \bar u_{n_1} \cdots \bar u_{n_p} \right]_n \right)_k \right|. \qquad (6.7) \end{aligned}$$

Therefore, for $k < m$, set $Y_k = C_k^Y$ where C_F^Y is given by (3.14). Note that these terms involve the Galerkin projection of f at $\bar u$ onto the first m modes and, therefore, are expected to be small.

For $0 \leq k < m$, we now combine our bounds for Y_k with the bounds for K_k to compute the coefficients of the polynomials $P_k(r)$ giving the bounds $Y_k + K_k - |\tilde w_k|$. This leads us to the definition of the finite radii polynomials presented in Definition 3.1.

In modes $k \geq m$, we use Lemma 6.1 and (6.5) to obtain

$$\begin{aligned} (6.8) \qquad (DT(\bar u + W)W)_k &\subseteq -J(k,k) \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l \binom{p}{l} ((c_p \bar u^{p-l}) * \tilde w^l)_k \\ &\subseteq \frac{1}{|\mu_k| k^s} \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l \binom{p}{l} \alpha^p C_p \bar A^{p-l} A^l [-1,1]. \end{aligned}$$

Therefore, set K_k, $k \geq m$, such that

$$K_k \geq \frac{C(\bar A, A)}{|\mu_k| k^s}. \qquad (6.9)$$

Recall (6.7). For $k \geq m$, choose Y_k (Compare with (3.7)) such that

$$\begin{aligned} Y_k &\geq |(T(\bar u) - \bar u)_k| \\ &= |-J(k,k)(f_k(\bar u))| \\ &= \frac{|\sum_{p=2}^{d} (c_p \bar u^p)_k|}{|\mu_k|}. \qquad (6.10) \end{aligned}$$

Using Lemma 6.1,

$$\frac{|\sum_{p=2}^{d} (c_p \bar u^p)_k|}{|\mu_k|} \subseteq \sum_{p=2}^{d} \frac{\alpha C_p \bar A^p}{|\mu_k| |k|^s} [-1,1] \ . \qquad (6.11)$$

These bounds are overestimates and should only be used for large k. In fact, if the coefficient functions c_p have finite Fourier expansions (as in the examples we consider in Section 5) then $Y_k = 0$ for k sufficiently large.

We may now define the polynomial bounds for $Y_k + K_k - |\tilde w_k|$ in the tail modes. Suppose the bounds Y_k are numerically or analytically computed for $m \leq k < m_+$. Then for $k \geq m$, the tail radii polynomial (see Definition 3.2) satisfies

$$\begin{aligned} P_k(r) &= Y_k + K_k(r) - \frac{A_s}{k^s} \\ &= \begin{cases} \frac{|\sum_{p=2}^{d} (c_p \bar u^p)_k|}{|\mu_k|} + \frac{C(\bar A, A)}{|\mu_k| k^s} - \frac{A_s}{k^s} & m \leq k < m_+ \\ \frac{C_+(\bar A, A)}{|\mu_k| k^s} - \frac{A_s}{k^s} & k \geq m_+. \end{cases} \end{aligned}$$

Checking that $P_k < 0$ for $k \geq m$ reduces to checking the inequalities $P_m < 0, \ldots, P_{m_+ - 1} < 0$ and, by rearranging terms,

$$C_+(\bar{A}, A) < |\mu_k| A_s. \tag{6.12}$$

Therefore, the assumption that $|\mu_k|$ is growing in k ensures that (6.12) may be verified for all $k \geq m$ with only a finite number of checks. More explicitly, computing a lower bound on $|\mu_k|$, $k \geq m_+$ would allow us to verify all inequalities of type (6.12), $k \geq m_+$, in one step. Indeed, since $\frac{C_+(\bar{A},A)}{|\bar{\mu}|} - A_s < 0$ and $f_k(\bar{u}) = 0$ and $|\mu_k| \geq |\bar{\mu}|$ for all $k \geq \bar{m} \geq m_+$,

$$\begin{aligned} P_k(\bar{r}) &= Y_k + K_k - \frac{A_s}{k^s} \\ &= \frac{C_+(\bar{A}, A)}{|\mu_k| k^s} - \frac{A_s}{k^s} \\ &\leq \frac{C_+(\bar{A}, A)}{|\bar{\mu}| k^s} - \frac{A_s}{k^s} \\ &< 0. \end{aligned}$$

We have now constructed the radii polynomials to give the bounds required for Theorem 2.1.

Recall that $r > 0$ is a *validation radius* if $P_k(r) < 0$ for all radii polynomials P_k as defined in Definitions 3.1 and 3.2. We may now prove Theorem 3.4 from Section 3.2.

THEOREM 3.4. *If there exists a validation radius $r > 0$ and the eigenvalues μ_k satisfy $|\mu_k| \to \infty$, then there exists a unique equilibrium solution of* (3.1) *in $\bar{u} + W(r)$.*

Proof. The radii polynomials have been constructed so that $P_k(r) < 0$ for all k ensures that the first condition of Theorem 2.1 is satisfied. Since the first condition is satisfied, we also have that $\frac{K_k}{|\tilde{w}_k|} < 1$ for all k. Finally, since $|\mu_k| \to \infty$,

$$\frac{K_k}{|\tilde{w}_k|} = \frac{\frac{C_+(\bar{A},A)}{|\mu_k| k^s}}{\frac{A_s}{k^s}} = \frac{C_+(\bar{A}, A)}{A_s |\mu_k|} \to 0$$

Therefore, $K := \sup \left\{ \frac{K_k}{|\tilde{w}_k|} \right\} < 1$ and the second and final hypothesis in Theorem 2.1 is also satisfied. □

7. Concluding remarks. As is indicated in the Introduction, the purpose of this paper is to communicate the essential ideas of our proposed validation method. As such we have presented it in a somewhat limited setting. Thus, we conclude with a range of comments, beginning with obvious generalizations, describing ongoing work, and ending with some open questions.

The particular choice of the abstract expression for the expansion of the partial differential equation (1.3) was chosen because it was appropriate for the application to Cahn-Hilliard (5.1) and Swift-Hohenberg (5.3). Hopefully it is clear that a different choice of boundary conditions or symmetries does not affect the essential estimates. It is expected, but remains to be checked, that the form of the estimates can be lifted to parabolic PDEs on rectangular domains (see [6] where similar estimates were used to study the equilibria of the Cahn-Hilliard equation on the unit square) and to systems of such PDEs. We also believe that generalizing this technique to pseudo-arclength continuation should be fairly straightforward. Furthermore, treating the parameter ν

as an interval allows us to prove the existence and uniqueness of a branch of solutions over the interval $\tilde{\nu}$. By adapting the predictor step length, this approach may be used to prove existence and uniqueness along continuous, finite branches of equilibria.

While there are numerous directions in which our validation technique can be expanded or improved we focus on the following four.

- Observe that if (1.2) has a polynomial nonlinearity of order d, then straightforward evaluation of the nonlinear term in (1.4) involves on the order of m^d operations. In a forthcoming work [5], this computational cost is reduced by the use of the fast Fourier transform.
- For the computations presented in this paper, we fixed $M = m$. This was done for the sake of simplicity of presentation. Clearly, the success of validation strongly depends on upper bounds presented in Lemma 6.2. In general, for fixed m choosing $M > m$ increases the computational cost, but provides a smaller bound for the truncation error ϵ_k. Improved bounds should facilitate validated continuation with a smaller projection dimension m, which decreases the computational cost. The exact trade off is currently being explored.
- The computational strategy adopted for this work is to fix A_s and s throughout the continuation procedure. In particular, in the Swift-Hohenberg example we obtained 40 successful predictor-corrector steps with $A_s = 0.002$ and $s = 4$ held constant over a parameter range of length 0.4. We were able to do this because we chose a projection dimension $m = 27$ which is unnecessarily large. For example, with $m = 11$, $A_s = 0.002$ and $s = 4.52$ we were able to perform a validated continuation over a parameter range of length 0.01. In this case, we obtained $s = 4.52$ by fixing A_s and seeking a successful s by trial and error. This suggests that it is worthwhile to develop a method for choosing A_s and s adaptively during the validated continuation procedure.
- As is pointed out in Section 5, the floating point errors are many orders of magnitude smaller than the magnitude of the radii polynomials evaluated at the validation radius. This suggests that it might be possible to compute a priori bounds on the floating point errors from which one could conclude that the validation computations are in fact rigorous computations. The techniques in [7] might prove useful for this purpose.

Acknowledgments. The authors would like to thank L. Dieci for numerous helpful conversations concerning continuation methods and M. Gameiro, R. Beardmore, and the referees for helpful comments on the layout of the paper.

REFERENCES

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

[2] S. Day, *A Rigorous Numerical Method in Infinite Dimensions*, Ph.D. thesis, Georgia Institute of Technology, 2003.

[3] 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), pp. 1–31.

[4] Z. Galias and P. Zgliczyński, *An Interval Method for Finding Fixed Points and Periodic Orbits of Infinite Dimensional Discrete Dynamical Systems*, preprint.

[5] M. Gameiro, J.-P. Lessard, and K. Mischaikow, *Rigorous Continuation over Long Parameter Ranges for Equilibria of PDEs*, in preparation.

[6] S. Maier-Paape, K. Mischaikow, and T. Wanner, *Structure of the attractor of the Cahn-Hilliard equation on a square*, RWTH Aachen, 5 (2005), pp. 1–68.

[7] M. Mrozek, *Rigorous error analysis of numerical algorithms via symbolic computations*, J. Symbolic Comput., 22 (1996), pp. 435–458.

[8] J. B. Swift and P. C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. A (3), 15 (1977), p. 319.

[9] N. Yamamoto, *A numerical verification method for solutions of boundary value problems with local uniqueness by Banach's fixed-point theorem*, SIAM J. Numer. Anal., 35 (1998), pp. 2004–2013.

[10] P. Zgliczyński and K. Mischaikow, *Rigorous numerics for partial differential equations: The Kuramoto-Sivashinsky equation*, Found. Comp. Math., 1 (2001), pp. 255–288.