

Validated Continuation for Equilibria of PDEs

Sarah Day* Jean-Philippe Lessard†

Konstantin Mischaikow‡

November 22, 2005

1 Introduction

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

$$u_t = E(u, \nu) \tag{1}$$

on a Hilbert space is to identify the set of equilibria $\mathcal{E} := \{(u, \nu) \mid E(u, \nu) = 0\}$. 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 [2], [7] and references therein). As mathematicians we are willing to argue forcefully for the importance 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

*Department of Mathematics, Cornell University, 310 Malott Hall, Ithaca, NY 14853-4201 (sday@math.cornell.edu).

†School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332 USA (lessard@math.gatech.edu).

‡School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332 USA (mischaik@math.gatech.edu).

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 an example 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*.

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) takes the form

$$u_t = L_\nu(u) + \sum_{p=0}^d c_p(\nu)u^p \quad (2)$$

where L_ν is a linear operator at parameter value ν and d is the degree of the polynomial nonlinearity. Expanding (2) using an orthonormal basis, chosen appropriately in terms of the eigenfunctions of L_ν , 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

$$\dot{a}_k = \mu_k a_k + \sum_{p=0}^d \sum_{\sum n_i=k} (c_p)_{n_0} a_{n_1} \cdots a_{n_p} \quad k = 0, 1, 2, \dots \quad (3)$$

where $\mu_k = \mu_k(\nu)$ are the parameter dependent eigenvalues of L_ν and $\{a_n\}$ and $\{(c_p)_n\}$ are the coefficients of the corresponding expansions of the functions u and $c_p(\nu)$ respectively with $a_n = a_{-n}$ and $(c_p)_n = (c_p)_{-n}$ for all n .

In particular, this is the case for the Swift-Hohenberg equation

$$\begin{aligned} u_t = E(u) &= \left\{ \nu - \left(1 + \frac{\partial^2}{\partial x^2} \right)^2 \right\} u - u^3, & u(\cdot, t) \in L^2 \left(0, \frac{2\pi}{L_0} \right), \\ u(x, t) &= u \left(x + \frac{2\pi}{L_0}, t \right), & u(-x, t) = u(x, t), \quad \nu > 0, \end{aligned} \quad (4)$$

considered in Section 4.

The continuation method is applied to the m -dimensional system of ODEs of the form

$$\dot{a}_k = \mu_k a_k + \sum_{p=0}^d \sum_{\substack{\sum n_i=k \\ |n_i|<m}} (c_p)_{n_0} a_{n_1} \cdots a_{n_p} \quad k = 0, 1, \dots, m-1. \quad (5)$$

obtained by performing a Galerkin projection on (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 $\{a_n\}$. 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 $L_\nu(u)$.

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 [6, Theorem 2.1]. A similar formulation can also be found in [3]. Recall that to apply the Banach fixed point theorem one must have a contraction mapping $g : X \rightarrow X$. 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, in the orthonormal basis used to obtain (3) consider the neighborhood $\bar{a} + W(r)$ of \bar{a} 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*, $\{P_k(r)\}_{k=0,1,\dots}$, whose coefficients are given explicitly in terms of the constants A_s , s , and (5). Theorem 3.6 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 (2) in the neighborhood $X = \bar{a} + W(r)$ of the numerical equilibria \bar{a} produced by the continuation procedure. It is important to remark that the conditions of Theorem 3.6 can be checked with a finite number of calculations.

To demonstrate the efficacy of this approach, we apply it to the Swift-Hohenberg equation (4) in Section 4. The decision to use (4) as the model equation was not made arbitrarily. The fact that it is a fourth order parabolic equation implies that a normalized Fourier series provides an appropriate orthonormal basis. Furthermore, the high regularity of the equilibria implies that any sufficiently high decay rate for the Fourier coefficients is, at least theoretically, acceptable. Finally, because of earlier work by S. Day, et. al. [5] the existence of the desired equilibria has been rigorously proven (though at a much higher computational cost). We emphasize this last point because we do not include a rigorous proof in this paper. In practice, proving the existence of a unique equilibrium using Theorem 3.6 requires generating and evaluating a finite set of the radii polynomials using interval arithmetic to account for the floating point computations. There exists a variety of software packages which perform such operations. However, since our focus is on comparing the cost

of continuation with the cost of continuation including validation we believe there are two good reasons for not including the final interval arithmetic computations. First, we have no control over the implementation of the interval arithmetic packages and hence over their computational cost. Second, if the researcher decides that the computational cost of applying an interval arithmetic package is too high, then the validity of the computation must be checked in some fashion. Our computations suggest that the ratio of the cost of validated continuation to continuation is significantly less than 1.5 (see Section 4.1 for the precise figures). Thus our method provides an alternative form of validation which is considerably cheaper than repeating the continuation algorithm at a reduced step size or with an increased number of modes included in the Galerkin projection.

As is indicated above, we see the results of this paper as a first step in the direction of combining continuation methods with rigorous computations. We conclude the paper in Section 5 with a discussion of open questions and on going work.

2 Computer-Assisted Proofs for Equilibria

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

$$\tilde{w}_k = \begin{cases} [-r, r] & 0 \leq k < m \\ [-\frac{A_s}{k^s}, \frac{A_s}{k^s}] & k \geq m \end{cases} \quad (6)$$

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

A particularly nice norm to use for this set (similar to the one used by Yamamoto in [6]) is the normalized sup norm

$$\|a\|_W := \sup_k \left\{ \frac{|a_k|}{|\tilde{w}_k|} \right\}$$

where $|\tilde{w}_k| := \max_{x \in \tilde{w}_k} |x|$. In this norm, W and $\bar{a} + W$ are compact, $W = B(0, 1)$, the unit ball around 0, and $\bar{a} + W = B(\bar{a}, 1)$, the unit ball around \bar{a} .

We will now reformulate our problem of studying equilibria for (3) by establishing an equivalent fixed point problem on $\bar{a} + W$. Suppose J is an invertible (operator) matrix. Then a is an equilibrium solution of (3) if and only if a is a fixed point of

$$g(a) = a - Jf(a). \quad (7)$$

In practice, g is constructed to be a contraction (Newton-like) operator with $J \approx (Df(\bar{a}))^{-1}$ 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 g is a contraction, we first consider the following Lipschitz condition on $\bar{a} + W$:

$$\|g(x) - g(y)\|_W \leq K \|x - y\|_W \quad \text{for } x, y \in \bar{a} + W. \quad (8)$$

Since by construction g is continuous and $\bar{a} + W$ is compact, such a finite constant K exists. The question now becomes whether we can compute a contraction constant $K < 1$ satisfying (8). We begin by computing Lipschitz constants, K_n , for the component functions g_n on $\bar{a} + W$ satisfying the following

$$|g_n(x) - g_n(y)| \leq K_n \|x - y\|_W \quad \text{for } x, y \in \bar{a} + W. \quad (9)$$

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

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

A constant K_n computed in this manner satisfies (9) by the following argument. Since $\bar{a} + W$ is convex, the Mean Value Theorem states that there exists $z \in \bar{a} + W$ such that

$$\begin{aligned} |g_n(x) - g_n(y)| &= |Dg_n(z)(x - y)| \\ &= \left| Dg_n(z) \frac{x - y}{\|x - y\|_W} \right| \|x - y\|_W. \end{aligned}$$

Finally, 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 it satisfies (8)

$$\begin{aligned} \|g(x) - g(y)\|_W &= \sup_n \frac{|g_n(x) - g_n(y)|}{|\tilde{w}_n|} \\ &\leq \sup_n \frac{K_n}{|\tilde{w}_n|} \|x - y\|_W \\ &= K \|x - y\|_W. \end{aligned}$$

Theorem 2.1 (Existence and Uniqueness) *If for all n there exist bounds $Y_n \geq |g_n(\bar{a}) - \bar{a}_n|$ and K_n satisfying (9) such that*

$$Y_n + K_n - |\tilde{w}_n| < 0 \quad (10)$$

and

$$K := \sup_n \frac{K_n}{|\tilde{w}_n|} < 1 \quad (11)$$

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

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

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

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

by assumption (10). Therefore, $g(\bar{a} + W) \subset \bar{a} + W$. The second inequality guarantees that g is also a contraction. Thus, the result follows from Banach's fixed point theorem. \blacksquare

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 (10) 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 non-linearity.) Furthermore, regularity of the equilibria may also be used to show that $\frac{K_n}{|\bar{w}_n|}$ becomes a decreasing sequence. Therefore, (11) follows automatically from (10).

Perhaps an even more important point to make for our intended algorithmic approach in this paper is that $Y_n + K_n - |\bar{w}_n|$ 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{a} will amount to showing that it is possible to simultaneously solve a (finite) list of polynomial inequalities in r .

3 Continuation and the Newton-like Operator

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 (7) onto the appropriate m -dimensional subspace is an intrinsic element of the continuation algorithm. In this section, we discuss exploiting this relationship to produce an automated, 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

$$\dot{a} = f(a, \nu) \tag{12}$$

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

$$\dot{a}_k = f_k(a) = \mu_k a_k + \sum_{p=0}^d \sum_{\sum n_i=k} (c_p)_{n_0} a_{n_1} \cdots a_{n_p} \quad (13)$$

A first approach 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 (13).

For $a_F := (a_0, \dots, a_{m-1}) \in \mathbb{R}^m$, define $f^{(m)} : \mathbb{R}^m \rightarrow \mathbb{R}^m$ by $f^{(m)}(a_F) = (f_0^{(m)}(a_F), \dots, f_{m-1}^{(m)}(a_F))$ where for $k = 0, \dots, m-1$,

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

The corresponding Galerkin projection of the original system (12) is then

$$\dot{a}_F = f^{(m)}(a_F, \nu) \quad (14)$$

This is the m -dimensional system to be studied numerically. Intuitively, we expect that if m is sufficiently large, (14) will capture the essential dynamics for the original system (12). In particular, given an equilibrium \bar{a}_F for (14) we expect that there is a small neighborhood around $\bar{a} := (\bar{a}_F, 0, \dots)$ which contains a unique equilibrium solution for (12). Our approach is to study this relationship via the tools outlined in Section 2.

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 (14), 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

$$\dot{x}_0 = -f_x^{(m)}(x_0, \nu_0)^{-1} f_\nu^{(m)}(x_0, \nu_0). \quad (15)$$

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

$$x_1^{(n+1)} = x_1^{(n)} - f_x^{(m)}(x_1^{(n)}, \nu_1)^{-1} f^{(m)}(x_1^{(n)}, \nu_1) \quad (16)$$

If k is the total number of iterations of (16), then $\bar{a}_F := x_1^{(k)}$ and $f^{(m)}(\bar{a}_F, \nu_1) \approx 0$.

As before, define the corresponding point $\bar{a} = (\bar{a}_F, 0, \dots)$ in the infinite dimensional space. We now use the information required for the next predictor

step, the numerical inverse of $f_x^{(m)}(\bar{a}_F, \nu_1)$, to construct a Newton-like operator near \bar{a} at the parameter value ν_1 . Let $J_{m \times m}$ be the numerical inverse of $f_x^{(m)}(\bar{a}_F, \nu_1)$ and define T by

$$T(a) = a - Jf(a) \tag{17}$$

where

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

is the block diagonal matrix which we expect to be close to $(Df(\bar{a}, \nu_1))^{-1}$. 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 $\bar{a} + W$ where W has the form (6). 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.

3.1 Computing the bounds

We now focus on computing the required bounds Y_n and K_n in (10) for the Newton-like operator constructed in (17).

Taking a Taylor expansion around the point \bar{a} ,

$$T(a) = T(\bar{a}) + T'(\bar{a})(a - \bar{a}) + \dots + \frac{T^{(p)}(\bar{a})}{p!}(a - \bar{a})^p + \dots + \frac{T^{(d)}(\bar{a})}{d!}(a - \bar{a})^d$$

and

$$T'(a) = T'(\bar{a}) + T''(\bar{a})(a - \bar{a}) + \dots + \frac{T^{(p)}(\bar{a})}{(p-1)!}(a - \bar{a})^{p-1} + \dots + \frac{T^{(d)}(\bar{a})}{(d-1)!}(a - \bar{a})^{d-1}$$

We use this expansion to compute the bounds

$$K_k \geq \max |[T'(\bar{a} + W)W]_k| \geq \sup_{z, y \in (\bar{a} + W)} \{[T'(z)(\bar{a} - y)]_k\}$$

where, as in Section 2, W has the form (6).

Formally,

$$\begin{aligned}
T'(\bar{a} + W)W &= T'(\bar{a})W + T''(\bar{a})W^2 + \cdots + \frac{T^{(l)}(\bar{a})W^l}{(l-1)!} + \cdots + \frac{T^{(d)}(\bar{a})W^d}{(d-1)!} \\
&= (I - Jf'(\bar{a}))W + (-Jf''(\bar{a}))W^2 + \cdots + \frac{(-Jf^{(l)}(\bar{a}))W^l}{(l-1)!} + \\
&\quad \cdots + \frac{(-Jf^{(d)}(\bar{a}))W^d}{(d-1)!} \\
&= (W - JL(W)) - J \left(\sum_{l=1}^d \frac{1}{(l-1)!} \left(\frac{d^l}{d\bar{a}^l} \Big|_{\bar{a}=\bar{a}} \sum_{p=0}^d c_p \bar{a}^p \right) W^l \right) \\
&= (W - JL(W)) - J \left(\sum_{l=1}^d \frac{1}{(l-1)!} \left(\sum_{p=l}^d \frac{p!}{(p-l)!} c_p \bar{a}^{p-l} \right) W^l \right) \\
&= (W - JL(W)) - J \sum_{l=1}^d \sum_{p=l}^d l \binom{p}{l} c_p \bar{a}^{p-l} W^l
\end{aligned}$$

Expanding into Fourier modes (or another suitable basis), we get

$$\begin{aligned}
T'(\bar{a} + W)W &= ([\tilde{w}_n]_n - J[L(W)]_n) \\
&\quad - J \sum_{l=1}^d \sum_{p=l}^d l \binom{p}{l} \left[\sum_{\sum n_i=n} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \right]_n
\end{aligned} \tag{18}$$

For flexibility in balancing numerical computations (requiring a finite number of operations) with analysis (to obtain truncation bounds), we now choose $M \geq m$ to be the dimension used to split these sums. The block-diagonal structure of J allows us to decompose (18) into a finite, M -dimensional piece and the infinite dimensional tail terms as follows:

$$\begin{aligned}
[T'(\bar{a} + W)W]_F &= (\tilde{w}_F - J_{F \times F} L(W)_F) \\
&\quad - J_{F \times F} \sum_{l=1}^d \sum_{p=l}^d l \binom{p}{l} \left[\sum_{\sum n_i=n} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \right]_F
\end{aligned} \tag{19}$$

where the subscript F denotes the vector consisting of modes 0 through $M-1$. For $k \geq M$,

$$\begin{aligned}
[T'(\bar{a} + W)W]_k &= (\tilde{w}_k - J(k, k) L(W)_k) \\
&\quad - J(k, k) \sum_{l=1}^d \sum_{p=l}^d l \binom{p}{l} \sum_{\sum n_i=k} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p}
\end{aligned} \tag{20}$$

Equation (18) and its decomposition into Equations (19) and (20) contain several finite sums determined by the degree of the polynomial nonlinearity with

terms consisting of infinite sums arising from derivatives of the monomial terms. We will now focus on computing bounds for the infinite sums since the finite sums will then be handled by the computer.

The infinite sums in (18) come in the form

$$\sum_{\sum n_i=k} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \quad (21)$$

where p is the degree of the original monomial term of f and $l \in \{1, 2, \dots, p\}$ is the order of the derivative being taken.

In one case, when $l = 1$ and k is a fixed number, we may exploit the fact that \bar{a} has only finitely many nonzero Fourier coefficients to show that the corresponding sum is actually a finite sum. In this case, it may be reasonable to compute a bound for the sum numerically. In all other cases, further analysis must be used to compute a bound for the sum.

First, let us establish asymptotic bounds for the Fourier expansions of the coefficient function c_p , the numerical zero \bar{a} , and the set W . Define \bar{A} , C_p and A by

$$\begin{aligned} \bar{A} &= \max_{0 \leq k < m} \{|\bar{a}_0|, |\bar{a}_k| |k|^s\} \\ C_p &= \max_k \{|(c_p)_0|, |(c_p)_k| |k|^s\} \\ A &= \max\{A_s, r(m-1)^s\}. \end{aligned}$$

We have now established the asymptotic bounds $\bar{a}_k \in \frac{\bar{A}}{k^s} [-1, 1]$, $(c_p)_k \in \frac{C_p}{k^s} [-1, 1]$, and $\tilde{w} \subset \frac{A}{k^s} [-1, 1]$ for all k . Note that these bounds are related to the norms of the corresponding sets.

One upper bound for (21) is given in the following lemma.

Lemma 3.1 *Let $\alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$. Then*

$$\sum_{\sum n_i=k} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \subseteq \begin{cases} \frac{\alpha^p C_p \bar{A}^{p-l} A^l}{|k|^s} [-1, 1] & k \neq 0 \\ \alpha^p C_p \bar{A}^{p-l} A^l [-1, 1] & k = 0. \end{cases}$$

Proof. This lemma is a modification of [1, 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{a} may be far from zero, resulting in a large constant \bar{A} . By taking k sufficiently large, the contraction given by $J(k, k)$ will overcome the large bound. A more practical approach for obtaining bounds for the low modes is given by the following lemma.

Lemma 3.2

$$\sum_{\sum n_i=k} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \subseteq \sum_{j=0}^l \binom{l}{j} C_k(p, j, l, M) r^{l-j} + \epsilon_k(p, l, M)$$

where

$$C_k(p, j, l, M) := \sum_{\substack{|n_0| < M \\ m \leq |n_{p-j+1}|, \dots, |n_p| < M}} \frac{A_s^j}{|n_{p-j+1}|^s \cdots |n_p|^s} \sum_{\substack{n_0 + \dots + n_p = k \\ |n_1|, \dots, |n_p| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \quad (22)$$

and

$$\epsilon_k(p, l, M) := \min \left\{ \frac{p\alpha^{p-1} C_p \bar{A}^{p-l} A^l}{(M-1)^{s-1} (s-1)} \left[\frac{1}{(M-k)^s} + \frac{1}{(M+k)^s} \right], \frac{\alpha^p C_p \bar{A}^{p-l} A^l}{k^s} \right\} [-1, 1] \quad (23)$$

Proof. This lemma is a modification of [1, Lemma 5.10] combined with Lemma 3.1. In [1, Lemma 5.10], the bound is split into a finite sum and the tail term, bounded by $\frac{p\alpha^{p-1} C_p \bar{A}^{p-l} A^l}{(M-1)^{s-1} (s-1)} \left[\frac{1}{(M-k)^s} + \frac{1}{(M+k)^s} \right]$.

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

$$\begin{aligned} & \sum_{\substack{n_0 + \dots + n_p = k \\ |n_0|, \dots, |n_p| < M}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \\ &= \sum_{\substack{n_0 + \dots + n_p = k \\ |n_0|, |n_{p-l+1}|, \dots, |n_p| < M \\ |n_1|, \dots, |n_{p-l}| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \\ &= \sum_{j=0}^l \binom{l}{j} \sum_{\substack{n_0 + \dots + n_p = k \\ m \leq |n_0|, |n_{p-j+1}|, \dots, |n_p| < M \\ |n_1|, \dots, |n_{p-j}| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \\ &= \sum_{j=0}^l \binom{l}{j} r^{l-j} \sum_{\substack{n_0 + \dots + n_p = k \\ m \leq |n_0|, |n_{p-j+1}|, \dots, |n_p| < M \\ |n_1|, \dots, |n_{p-j}| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-j+1}} \cdots \tilde{w}_{n_p} \\ &= \sum_{j=0}^l \binom{l}{j} r^{l-j} \sum_{\substack{n_0 + \dots + n_p = k \\ m \leq |n_0|, |n_{p-j+1}|, \dots, |n_p| < M \\ |n_1|, \dots, |n_{p-j}| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \frac{A_s}{|n_{p-j+1}|^s} \cdots \frac{A_s}{|n_p|^s} \\ &= \sum_{j=0}^l \binom{l}{j} r^{l-j} \sum_{\substack{|n_0| < M \\ m \leq |n_{p-j+1}|, \dots, |n_p| < M}} \frac{A_s^j}{|n_{p-j+1}|^s \cdots |n_p|^s} \sum_{\substack{n_0 + \dots + n_p = k \\ |n_1|, \dots, |n_p| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \end{aligned}$$

■

Remark 3.3 Note that $C_k(p, j, l, M)$ 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

$$C_k(p, 0, l, m) = \sum_{\substack{n_0 + \dots + n_p = k \\ |n_1|, \dots, |n_p| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-1}}$$

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

For $0 \leq k < M$, we substitute the bound from Lemma 3.2 into (19),

$$\begin{aligned} T'_k(\bar{a} + W)W &\subseteq (\tilde{w}_k - [J_{F \times F} L(W)]_k) \\ &+ \left[-J_{F \times F} \sum_{l=1}^d \sum_{p=l}^d l \binom{p}{l} \left[\sum_{j=0}^l \binom{l}{j} C_n(p, j, l, M) r^{l-j} + \epsilon_n(p, l, M) \right]_F \right]_k \\ &= (-J_{F \times F} [\epsilon_n(p, l, M)]_F)_k + r [\mathbb{I}_F - J_{F \times F} L(\mathbb{I})_F]_k \\ &+ \sum_{i=0}^d r^i \left(-J_{F \times F} \sum_{l=i}^d \sum_{p=l}^d l \binom{p}{l} \binom{l}{i} C_n(p, l - i, l, M)_F \right)_k \end{aligned}$$

where

$$\begin{aligned} \mathbb{I} &:= (1, 1, \dots)^T, \quad L(\mathbb{I})_k = \mu_k \\ \epsilon_n &:= \sum_{l=1}^d \sum_{p=l}^d l \binom{p}{l} \epsilon_n(p, l, M). \end{aligned} \quad (24)$$

For $k < M$, set K_k to be

$$K_k := \sum_{i=0}^d C_k^K(i) r^i \geq |T'_k(\bar{a} + W)W|$$

where

$$\begin{aligned} C_k^K(i) &\geq \left| \left[-J_{F \times F} \left[\sum_{l=i}^d \sum_{p=l}^d l \binom{p}{l} \binom{l}{i} C_n(p, l - i, l, M) \right]_F \right. \right. \\ &\quad \left. \left. + \begin{cases} -J_{F \times F} \epsilon_F & i = 0 \\ \mathbb{I}_F - J_{F \times F} L(\mathbb{I})_F & i = 1 \\ 0 & \text{otherwise} \end{cases} \right]_k \right|. \end{aligned} \quad (25)$$

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.

$$\begin{aligned}
Y_k &\geq |[T(\bar{a}) - \bar{a}]_k| \\
&= |[-Jf(\bar{a})]_k| \\
&= \left| \left(-J \left[\mu_n \bar{a}_n + \sum_{p=0}^d \sum_{\substack{n_0+\dots+n_p=n \\ |n_1|, \dots, |n_p| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_p} \right] \right)_n \right|_k
\end{aligned} \tag{26}$$

Therefore, for $k < M$, set $Y_k = C_k^Y$ where

$$C_k^Y \geq \left| \left[-J_{F \times F} \left[\mu_n \bar{a}_n + \sum_{p=0}^d \sum_{\substack{n_0+\dots+n_p=n \\ |n_1|, \dots, |n_p| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_p} \right] \right]_n \right|_k. \tag{27}$$

Note that these terms involve the Galerkin projection of f at \bar{a} onto the first m modes and, therefore, are expected to be small. Furthermore, since $\bar{a}_k = 0$ for all $k \geq m$, computing each C_k^Y involves only a finite number of computations. In particular, $f_k(\bar{a}) = 0$ for $k > d(m-1)$ so only finitely many of the bounds C_k^Y require numerical computations. Hence, for $0 \leq k < m$, we 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 $Y_k + K_k - |\tilde{w}_k|$.

Definition 3.4 The M finite radii polynomials, P_0, \dots, P_{M-1} , are given in vector notation as

$$P_F(r) := \sum_{n=0}^d C_F(n) r^n \tag{28}$$

where

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

and the terms C_F^Y and $C_F^K(n)$ are computed component-wise as previously defined.

In the tail modes, $k \geq M$ we have

$$\begin{aligned}
T'_k(\bar{a} + W)W &\subseteq (\tilde{w}_k - J(k, k)L(W)_k) & (29) \\
&\quad - J(k, k) \sum_{l=2}^d \sum_{p=l}^d l \binom{p}{l} \sum_{\sum n_i = k} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p} \\
&\subseteq (\tilde{w}_k - \mu_k^{-1} \mu_k \tilde{w}_k) \\
&\quad - \mu_k^{-1} \sum_{l=2}^d \sum_{p=l}^d l \binom{p}{l} \frac{\alpha^p C_p \bar{A}^{p-l} A^l}{k^s} [-1, 1] \\
&= \frac{-1}{\mu_k k^s} \left(\sum_{l=2}^d \sum_{p=l}^d l \binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A^l \right) [-1, 1] \\
&= \frac{-1}{\mu_k k^s} C(\bar{A}, A) [-1, 1]
\end{aligned}$$

where

$$C(\bar{A}, A) := \sum_{l=2}^d \sum_{p=l}^d l \binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A^l \quad (30)$$

Therefore, for $k \geq M$, set K_k to be such that

$$K_k \geq \frac{C(\bar{A}, A)}{|\mu_k| k^s} \geq |T'_k(\bar{a} + W)W|, \quad (31)$$

with $A := \max\{A_s, r(m-1)^s\}$. Recall (26). Then for $k \geq m$, choose Y_k such that

$$\begin{aligned}
Y_k &\geq |[T(\bar{a}) - \bar{a}]_k| \\
&= |-J(k, k)f_k(\bar{a})| \\
&= \left| -\mu_k^{-1} \left(\mu_k \bar{a}_k + \sum_{p=0}^d \sum_{\substack{n_0 + \cdots + n_p = k \\ |n_1|, \dots, |n_p| < m}} (c_p)_{n_0} \bar{a}_{n_1} \cdots \bar{a}_{n_p} \right) \right| \\
&= \frac{|f_k(\bar{a})|}{|\mu_k|}. \quad (32)
\end{aligned}$$

We may now define the polynomial bounds for $Y_k + K_k - |\tilde{w}_k|$ in the tail modes.

Definition 3.5 For $k \geq M$, the *tail radii polynomial*, P_k , is

$$P_k(r) := Y_k + K_k(r) - \frac{A_s}{k^s} \quad (33)$$

By considering the inequality $P_k < 0$, $k \geq M$, required for Theorem 2.1 and rearranging terms we get

$$k^s |f_k(\bar{a})| + C(\bar{A}, \max\{A_s, r(m-1)^s\}) < |\mu_k| A_s. \quad (34)$$

Since $f_k(\bar{a}) = 0$ for k sufficiently large, a regularity assumption that $|\mu_k|$ is growing in k ensures that (34) may be verified for all $k \geq M$ with only a finite number of checks. More explicitly, computing an upper bound for $k^s |f_k(\bar{a})|$, $M \leq k \leq d(m-1)$, and a lower bound on $|\mu_k|$, $k > d(m-1)$ would allow us to verify all inequalities of type (34) in one step. We have now constructed the radii polynomials to give the bounds required for Theorem 2.1.

Theorem 3.6 *If there exists a validation radius $r > 0$ such that $P_k(r) < 0$ for all P_k as defined in Definitions 3.4 and 3.5 and the eigenvalues μ_k satisfy $|\mu_k| \rightarrow \infty$ then there exists a unique equilibrium solution of (12) in $\bar{a} + W(r)$.*

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 $\frac{K_k}{|\bar{w}_k|} < 1$ for all k . Finally, since $|\mu_k| \rightarrow \infty$,

$$\frac{K_k}{|\bar{w}_k|} = \frac{C(\bar{A}, A)}{|\mu_k| k^s} = \frac{C(\bar{A}, A)}{A_s |\mu_k|} \rightarrow 0$$

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

We now present a slightly weaker version of Theorem 3.6 which provides a natural order for defining the constants A_s , s , and A . To motivate this corollary, we make the observation that $A := \max\{A_s, r(m-1)^s\}$ depends explicitly on the variable r . From a computational perspective, we would like to find $r > 0$ solving $P_0(r), \dots, P_{M-1}(r) < 0$ without having to simultaneously study the influence of r in the polynomials P_k , $k \geq M$. The influence of r in these tail polynomials comes in the definition of the constant A as $\max\{r(m-1)^s, A_s\}$. A practical way to overcome this problem is to set $A = A_s$ at the beginning of the procedure and then check in the end that a solution $r > 0$ to $P_0(r), \dots, P_{M-1}(r) < 0$ also satisfies $r(m-1)^s \leq A_s$. Finally, we add a condition which simplifies the check of the tail polynomials $P_k < 0$, $k > d(m-1)$ and also set $M = m$. The following follows from Theorem 3.6.

Corollary 3.7 *Suppose that the eigenvalues μ_k are such that $|\mu_k| \rightarrow \infty$. Let $m \in \mathbb{N}$ be such that $|\mu_k| \geq |\mu_{d(m-1)+1}|$ for all $k \geq d(m-1) + 1$ and set $M = m$. Let*

$$s \geq 2, \quad A_s > 0, \quad A = A_s \quad (35)$$

and let P_0, \dots, P_{m-1} be the finite radii polynomials (28) depending on these constants. For $k = 0, \dots, m-1$, let $I_k := \{r > 0 \mid P_k(r) < 0\}$ and define

$$\mathcal{I} := \bigcap_{k=0}^{m-1} I_k. \quad (36)$$

Let $P_m, \dots, P_{d(m-1)}$ be the tail radii polynomials defined by (33). Note that they are now independent of r . Finally, let

$$\tilde{P}_{d(m-1)+1} := \frac{C(\bar{A}, A)}{|\mu_{d(m-1)+1}|} - A_s. \quad (37)$$

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

$$\bar{r}(m-1)^s \leq A_s \quad (38)$$

and if $P_k < 0$ for $k = m, \dots, d(m-1)$ and $\tilde{P}_{d(m-1)+1} < 0$ then there exists a unique equilibrium solution of (12) in the set $\bar{a} + W(\bar{r})$, where

$$W(\bar{r}) = \prod_{k=0}^{m-1} [-\bar{r}, \bar{r}] \times \prod_{k=m}^{\infty} \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s} \right]. \quad (39)$$

Proof. Since $\tilde{P}_{d(m-1)+1} := \frac{C(\bar{A}, A)}{|\mu_{d(m-1)+1}|} - A_s < 0$ and $|\mu_k| \geq |\mu_{d(m-1)+1}|$ for all $k \geq d(m-1) + 1$,

$$\begin{aligned} Y_k + K_k &= K_k = \frac{C(\bar{A}, A)}{|\mu_k|k^s} \\ &\leq \frac{C(\bar{A}, A)}{|\mu_{d(m-1)+1}|k^s} \\ &< \frac{A_s}{k^s} \end{aligned}$$

for all $k \geq d(m-1) + 1$. Hence, the hypotheses of Theorem 3.6 are satisfied. ■

4 Sample Results for Swift-Hohenberg

The Swift-Hohenberg equation (4) was originally introduced to describe the onset of Rayleigh-Bénard heat convection [5], 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.

In the notation of Section 3.1, (4) has linear operator $L = \nu - (1 + \frac{\partial^2}{\partial x^2})^2$ with eigenvalues $\mu_k := L(\mathbb{I})_k = \nu - (1 - k^2 L_0^2)^2$, and polynomial nonlinearity of degree $d = 3$ with associated coefficient function c_3 with Fourier expansion

$$(c_3)_n = \begin{cases} -1 & n = 0 \\ 0 & \text{otherwise} \end{cases}$$

For a projection dimension m , the associated Galerkin projection is

$$f_k^{(m)}(a_F, \nu) = \mu_k(\nu)a_k - \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} a_{n_1}a_{n_2}a_{n_3}, \quad k = 0, \dots, m-1. \quad (40)$$

For validated continuation, we need to compute the coefficients of the $3m-1$ radii polynomials P_0, \dots, P_{3m-2} of Corollary 3.7 at the end of each predictor-corrector step. Let \bar{a}_F be the numerical zero coming from the last quasi-Newton iteration of the corrector at ν and let $J_{m \times m}$ the numerical inverse of $Df^{(m)}(\bar{a}_F, \nu)$. To simplify the presentation, we consider the specific case $M = m$.

We now compute the formulas for the radii polynomial coefficients for (4) given the constants $m = M$, $s > 2$, and $A_s > 0$ and the constant vector \bar{a}_F . Since the only nonlinear term is a monomial of degree $p = 3$ and we have set $M = m$, $C_k(p, j, l, M) = 0$ whenever $p < 3$ or $j > 0$ (see Remark 3.3). The only nonzero terms of this form are

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

Before computing the truncation errors ϵ_k , we set $A = A_s$ as in Corollary 3.7 and the constants

$$\bar{A} = \max_{0 \leq k < m} \{|\bar{a}_0|, |\bar{a}_k| \cdot |k|^s\} \quad \text{and} \quad \alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$$

as described in Section 3.1. By (24),

$$\epsilon_k = \sum_{l=1}^3 l \binom{3}{l} \epsilon_k(3, l, m) = 3\epsilon_k(3, 1, m) + 6\epsilon_k(3, 2, m) + 3\epsilon_k(3, 3, m)$$

where, by (23)

$$\epsilon_k(3, l, m) = \min \left\{ \frac{3\alpha^2 \bar{A}^{3-l} A^l}{(m-1)^{s-1}(s-1)} \left[\frac{1}{(m-k)^s} + \frac{1}{(m+k)^s} \right], \frac{\alpha^2 \bar{A}^{3-l} A^l}{k^s} \right\}.$$

Using (25), we set

$$C_F^K(0) = | -J_{F \times F} \epsilon_F | \quad (41)$$

where $|\cdot|$ denotes the component-wise absolute value. Now, from (25), we get

that the nonzero terms in $C_F^K(1)$ occur when $l = 1$ and $p = 3$. Hence, let

$$\begin{aligned}
C_F^K(1) &= \left| \mathbb{I}_F - J_{F \times F} \left(L(\mathbb{I})_F + \left[\sum_{l=1}^3 \sum_{p=l}^3 l \binom{p}{l} \binom{l}{1} C_k(p, l-1, l, m) \right]_k \right) \right| \\
&= \left| \mathbb{I}_F - J_{F \times F} \left(L(\mathbb{I})_F + 1 \binom{3}{1} \binom{1}{1} [C_k(3, 0, 1, m)]_k \right) \right| \\
&= \left| \mathbb{I}_F - J_{F \times F} \left(L(\mathbb{I})_F - 3 \left[\sum_{n_1+n_2+n_3=k} \bar{a}_{n_1} \bar{a}_{n_2} \right]_k \right) \right| \\
&= \left| \mathbb{I}_F - J_{F \times F} Df^{(m)}(\bar{a}_F) \mathbb{I}_F \right|.
\end{aligned}$$

By construction, $J_{F \times F} \approx (Df^{(m)}(\bar{a}_F), \nu)^{-1}$ so this term should be small.

Following the same ideas, we let

$$C_F^K(2) = 2 \binom{3}{2} \binom{2}{2} |J_{F \times F} C_F(3, 0, 2, m-1)| = 6 |J_{F \times F} \left[\sum_{\substack{n_1+n_2+n_3=k \\ |n_i| < m}} \bar{a}_{n_1} \right]_k|$$

and

$$C_k^K(3) = 3 \binom{3}{3} \binom{3}{3} |J_{F \times F} C_F(3, 0, 3, m-1)| = 3 |J_{F \times F} \left[\sum_{\substack{n_1+n_2+n_3=k \\ |n_i| < m}} 1 \right]_k|$$

The last coefficient to compute for the finite radii polynomials (28) is C_F^Y . From (27), $C_F^Y = |J_{F \times F} f^{(m)}(\bar{a}_F, \nu)|$.

At this point, we are now ready to check for the existence of an $r > 0$ which satisfies $P_F(r) = \sum_{n=0}^3 C_F(n) r^n < 0$. To do this we find the numerical zeros of each of the cubic polynomials P_0, \dots, P_{m-1} , construct I_0, \dots, I_{m-1} where $I_k = \{r > 0 | P_k(r) < 0\}$, and finally check for a non-empty intersection $\mathcal{I} = \bigcap_{k=0}^{m-1} I_k$ representing a solution to $P_F(r) < 0$. If we can get a positive $\bar{r} \in \mathcal{I}$ such that $\bar{r}(m-1)^s \leq A_s$, then we build the tail radii polynomials P_m, \dots, P_{3m-1} , and \tilde{P}_{3m-2} and attempt to verify the remaining hypotheses of Corollary 3.7.

We now construct the tail polynomials given in (33) and \tilde{P}_{3m-2} in (37). For $k = m, \dots, 3m-1$,

$$P_k = Y_k + K_k - \frac{A_s}{k^s}$$

where

$$Y_k = \left| \frac{1}{\mu_k} \sum_{\substack{n_1+n_2+n_3=k \\ |n_i| < m}} \bar{a}_{n_1} \bar{a}_{n_2} \bar{a}_{n_3} \right|$$

and

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

Using (30) with $A = A_s$,

$$C(\bar{A}, A) = \sum_{l=2}^2 l \binom{3}{l} \alpha^3 \bar{A}^{3-l} A^l = 6\alpha^3 \bar{A} A_s^2 + 3\alpha^3 A_s^3. \quad (42)$$

The final polynomial from (37) is

$$\tilde{P}_{3m-2} = \frac{C(\bar{A}, A)}{|\mu_{3m-2}|} - A_s$$

The two remaining steps in verifying the conditions of Corollary 3.7 are to check the inequalities for the tail polynomials $P_m, \dots, P_{3m-3}, \tilde{P}_{3m-2} < 0$ constructed above and the condition on the eigenvalues $\mu_k, k \geq m$. This last condition on the eigenvalues, that $|\mu_k| \geq |\mu_{3m-2}|$ for all $k \geq 3m - 2$ and $|\mu_k| \rightarrow \infty$ follow directly from the form of these values for (4) and the constants chosen below for the sample result.

Using $m = 27$, $L_0 = 0.65$, $s = 4$ and $A = A_s = 0.002$, we ran the validation procedure just outlined for (4). Figure 1 shows a branch of equilibria where each point represents the center of the infinite dimensional validation set of the form $\bar{a} + W(\bar{r})$, containing a unique equilibrium of (4). For example, the last point on the right of the branch is \bar{a} given in Figure 6, the approximate zero for (4) at $\nu = 0.7617$, and has validation set

$$(\bar{a}_F, 0) + \prod_{k=0}^{26} [-4.1395e^{-9}, 4.1395e^{-9}] \times \prod_{k=27}^{\infty} \left[-\frac{0.002}{k^4}, \frac{0.002}{k^4} \right].$$

4.1 Computational Cost

We now compare the cost of the continuation with the cost of validated continuation for the Swift-Hohenberg model (4). Since the nonlinearity in (4) 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. We now provide a rough comparison of the number of m^3 operations for both approaches.

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 (15), we need to compute the vector $-f_x^{(m)}(x_0, \nu_0)^{-1} f_\nu^{(m)}(x_0, \nu_0)$. Hence, we first need to evaluate the m by

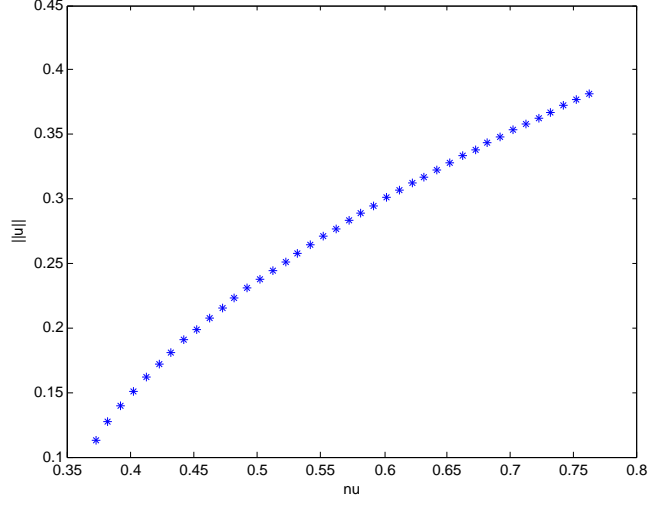


Figure 1: Validated continuation in ν for the Swift-Hohenberg equation at $L_0 = 0.65$.

m matrix $f_x^{(m)}(x_0^{(0)}, \nu_0)$. For $0 \leq i, j < m$, we have

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

This requires the evaluation of $2m^2$ sums demanding $2m - 1$ multiplications and $2m - 2$ additions each. Therefore, the evaluation of $f_x^{(m)}(x_0^{(0)}, \nu_0)$ 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. We then let $x_1^{(0)} = x_0 - \Delta \nu f_x^{(m)}(x_0, \nu_0)^{-1} x_0$ and $\nu_1 = \nu_0 + \Delta \nu$.

Step 2. To construct the quasi-Newton operator (16), we need the action of the inverse of $f_x^{(m)}(x_1^{(0)}, \nu_1)$. 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 (16), we need to evaluate $f^{(m)}(x_1^{(j-1)}, \nu_1)$. Its i^{th} component is

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

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

Step 4. The corrector ends when $\|f^{(m)}(x_1^{(k)}, \nu_1)\| < \textit{tolerance}$. Let $\bar{a}_F := x_1^{(k)}$. Evaluating the function at (\bar{a}_F, ν_1) is another $3m^3$. Now, note that we have to compute the action of the inverse of $f_x^{(m)}(\bar{a}_F, \nu_1)$ to get the predictor for the next step. Define $J_{F \times F}$ to be the numerical inverse of $f_x^{(m)}(\bar{a}_F, \nu_1)$ computed as before using an LU decomposition. Explicitly computing all the coefficients in $f_x^{(m)}(\bar{a}_F, \nu_1)$ 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.

We now examine the extra cost of performing the validation in this specific context. The additional work comes primarily from the fact that we need to build the coefficients of the $3m - 1$ polynomials $P_0, \dots, P_{3m-3}, \tilde{P}_{3m-2}$. From Step 4, computing $C_F^K(0)$ in (41) is not an m^3 operation. Computing $C_F^K(1)$ requires multiplying two m by m matrices at a cost of $2m^3$. Note that we used a combinatorial argument to compute $C_F^K(2)$ and $C_F^K(3)$. Therefore, computing these terms requires on the order of m^2 operations. Computing $C_k^Y = |J_{F \times F} f^{(m)}(\bar{a}_F, \nu_1)|$ comes with no extra m^3 cost since we already computed these terms in Step 4 of the predictor-corrector algorithm. In order to compute the coefficients of the tail radii polynomials, we need an extra $6m^3$ operations. Indeed, we need to evaluate $f_k(\bar{a})$ for $k = m, \dots, 3m - 3$. The remaining terms do not require computations of order m^3 . Hence, validated continuation requires on the order of an extra $10m^3$ operations over continuation alone.

In summary, if k is the number of iterations of the corrector, then the ratio of the cost of validated continuation to continuation is asymptotically roughly $\frac{30+3k}{20+3k}$. In the computations described in this section, we performed validated continuation for 40 predictor-corrector steps involving a total of 90 quasi-Newton iterations. We did the same without validation. The ratio of elapsed time for validated continuation to the time used for continuation alone was $\frac{56 \text{ seconds}}{44 \text{ seconds}} \approx 1.27$. Given that we had an average of 2.25 iterations per predictor-corrector

step, this is close to the rough estimate of $\frac{30+3\cdot 2.25}{20+3\cdot 2.25} \approx 1.37$ given by the above arguments.

5 Concluding remarks

In order to communicate the essential ideas of our proposed validation method, we presented it in a somewhat limited setting. Thus, we conclude with a range of comments, beginning with obvious generalizations and ending with future work.

As is indicated in the Introduction, the particular choice of the abstract expression for the expansion of the partial differential equation (3) was chosen because it was appropriate for the application to Swift-Hohenberg (4). 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 [4] 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.

As presented in the previous sections, our technique represents a method for validating the continuation results. With extra effort these ideas can be used to rigorously verify the continuation results. 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 even when one considers the possibility 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

$$\bar{a}_F, f^{(m)}(\bar{a}_F, \nu), J_{m \times m}, Df^{(m)}(\bar{a}_F, \nu), \mu_k(\nu), A_s, \text{ and } s$$

to be an interval of machine precision length and then computing (25), (27), (31) and (32) using exact floating point arithmetic. The resulting radii polynomials 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 tail radii polynomials are strictly contained in $(-\infty, 0)$, then the hypotheses of Corollary 3.7 are satisfied and we obtain a proof. Similarly 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 three.

- In the computations described above, for the sake of simplicity of presentation, we fixed $M = m$. The success of our technique strongly depends on upper bounds presented in Lemma 3.2. 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, however, may facilitate validated continuation with a smaller projection dimension m , which decreases the computational cost. At the moment we do not know how to choose M and m optimally.
- The computational strategy adopted for this work is to fix A_s and s throughout the continuation procedure. In particular, in our 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.
- It is important to observe that if (2) has a polynomial nonlinearity of order d , then straightforward evaluation of the nonlinear term in (5) involves on the order of m^d operations. This can be reduced by the use of the fast Fourier transform.

Acknowledgements

The authors would like to thank L. Dieci for numerous helpful conversations concerning continuation methods. S.D. was partially supported by NSF DMS 9983660. J.P.L. was partially supported by FQRNT. J.P.L. and K.M. were partially supported by NSF DMS 0511115, and grants from D.O.E. and DARPA.

References

- [1] S. Day. *A rigorous numerical method in infinite dimensions*. PhD thesis, Georgia Institute of Technology, 2003.
- [2] Sarah Day, Yasuaki Hiraoka, Konstantin Mischaikow, and Toshiyuki Ogawa. Rigorous numerics for global dynamics: a study of the Swift-Hohenberg equation. *SIAM J. Appl. Dyn. Syst.*, 4(1):1–31 (electronic), 2005.
- [3] Z. Galias and P. Zgliczyński. An interval method for finding fixed points and periodic orbits of infinite dimensional discrete dynamical systems. *preprint*.
- [4] S. Maier-Paape, K. Mischaikow, and T. Wanner. Structure of the Attractor of the Cahn-Hilliard Equation on a Square. *RWTH Aachen*, 5:1–68, 2005.

- [5] J.B. Swift and P.C. Hohenberg. Hydrodynamic fluctuations at the convective instability. *Phys. Rev. A*, 15:319, 1977.
- [6] Nobito 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(5):2004–2013 (electronic), 1998.
- [7] P. Zgliczyński and K. Mischaikow. Rigorous numerics for partial differential equations: the Kuramoto-Sivashinsky equation. *Found. Comp. Math.*, 1:255–288, 2001.

6 Appendix

Figure 2: Approximate stationary solution \bar{a} at $\nu = 0.7617$.

k	\bar{a}_k
0	0
1	0.38115715367932
2	0
3	-0.00695472240547
4	0
5	0.00003250545319
6	0
7	-0.00000017759055
8	0
9	0.00000000083957
10	0
11	-0.00000000000362
12	0
13	0.00000000000002
≥ 14	0