Validated Continuation Over Large Parameter Ranges for Equilibria of PDEs

Marcio Gameiro [*] Jean-Philippe Lessard[†]

Konstantin Mischaikow[‡]

Abstract

Validated continuation was introduced in [4] as means of checking that the classical continuation method applied to a Galerkin projection of a PDE provides a locally unique equilibrium to the PDE of interest. In this paper we extend the numerical technique to include a parameter that leads to better bounds on the errors associated with the Galerkin truncation. We test this method on the Swift-Hohenberg and Cahn-Hilliard equations on one dimensional domains. For the first equation, we find no numerical obstructions to the validated continuation technique. This is not the case for the Cahn-Hilliard equation.

1 Introduction

The question of finding zeros of a nonlinear function arises in many mathematical domains. In the setting of a one parameter family of nonlinear differential equations

$$u_t = f(u, \nu), \quad u \in H, \ \nu \in \mathbb{R} \tag{1}$$

defined on a Hilbert space H, the zeros correspond to equilibria. If f is a sufficiently smooth function, then the set of equilibria $\mathcal{E} := \{(u, \nu) \mid f(u, \nu) = 0\}$ is, at least locally, typically a smooth curve in $H \times \mathbb{R}$. In this case, continuation provides a particularly efficient manner of approximating $\mathcal{E}$. Recall, that this method involves a predictor and corrector step: given, within a prescribed

[*] Department of Mathematics, Rutgers University, 110 Frelinghusen Rd, Piscataway, NJ 08854-8019, USA (gameiro@math.rutgers.edu). The author was partially supported by grants from D.O.E. and DARPA.

[†] School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332-0160 USA (lessard@math.gatech.edu). The author was partially supported by FQRNT, NSF DMS 0511115 and grants from D.O.E. and DARPA. Currently visiting: Department of Mathematics, Rutgers University, 110 Frelinghusen Rd, Piscataway, NJ 08854-8019, USA

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

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.

Our interest is in the case where (1) is a partial differential equation, and hence, H is an infinite dimensional space. Thus, to perform the continuation described above requires a priori reducing the infinite dimensional problem to a finite dimensional system, this often involves a Galerkin projection. A fundamental implication of this is that while the continuation method may succeed, it is not clear that the solution closely approximates elements of $\mathcal{E}$. To address this problem, the concept of validated continuation was introduced in [4].

To review the essential elements of validated continuation assume that (1) takes the form

$$u_t = L(u,\nu) + \sum_{p=0}^{d} c_p(\nu) u^p \tag{2}$$

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 $L(\cdot,\nu)$. Using an appropriate Fourier series expansion of (2) results in a countable system of differential equations on the coefficients of the expanded solution, which for the sake of simplicity we assume takes the form

$$\dot{u}_k = f_k(u,\nu) := \mu_k u_k + \sum_{p=0}^{d} \sum_{\sum n_i = k} (c_p)_{n_0} u_{n_1} \cdots u_{n_p} \quad k = 0, 1, 2, \ldots \tag{3}$$

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.

The continuation method is applied to the m-dimensional system of ordinary differential equations

$$\dot{u}_k = \mu_k u_k + \sum_{p=0}^{d} \sum_{\substack{\sum n_i = k \\ |n_i| < m}} (c_p)_{n_0} u_{n_1} \cdots u_{n_p} \quad k = 0, 1, \ldots, m-1 \tag{4}$$

obtained from (3) via a Galerkin projection. The validation technique provides justification that the numerical approximations of equilibria of (4) are satisfactory approximations of the equilibria of (2).

The theoretical basis for the validation technique is provided by the Banach fixed point theorem: *A contraction mapping $T : X \to X$ on a complete metric space has a unique fixed point in X.* Observe that in a neighborhood of a hyperbolic fixed point the Newton operator is a contraction mapping that provides super linear convergence. Furthermore, the not insignificant computational cost of deriving a numerical Newton-like operator is performed in the continuation step. Thus, for the sake of efficiency the validation technique focuses on efficiently determining the set X on which a controlled perturbation of the numerical operator is a contraction mapping.

We apply the validated continuation technique to parabolic PDEs for which the equilibria are smooth. We exploit this smoothness and assume that we can express the set X in the form $W_{\bar{u}}(r)$ where $\bar{u}$ is a numerical zero obtained from the continuation method applied to (4) and

$$W_{\bar{u}}(r) = \bar{u} + \prod_{k=0}^{m-1} [-r, r] \times \prod_{k=m}^{\infty} \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s} \right]. \tag{5}$$

As the notation suggests, we think of s and A_s as constants and r, the *validation radius*, as a variable to be solved for. In particular, as is described in Section 2, r is taken to be a solution to a set of polynomial inequalities which encode the truncation errors associated with performing the continuation technique on the projection of $W_{\bar{u}}(r)$ to $\prod_{k=0}^{m-1}[-r, r]$.

Two important points of [4] are the following: (1) if the validation radius r exists, then the Banach fixed point theorem applies and hence there exists a unique equilibrium in the set $W_{\bar{u}}(r)$, and (2) the cost of validated continuation is comparable with that of the standard continuation applied to (4). In order to focus on the essential elements of the validated continuation technique, several obvious numerical improvements and questions were explicitly left undeveloped in [4, Section 7]. We turn to two of these issues in this paper.

1. Observe that if (2) has a polynomial nonlinearity of order d, then straightforward evaluation of the nonlinear term in (4) involves on the order of m^d operations. This computational cost can be reduced by making use of Fast Fourier Transform (FFT) techniques.

2. As is mentioned above, the truncation of $W_{\bar{u}}(r)$ to $\prod_{k=0}^{m-1}[-r, r]$ introduces errors that must be overcome in order to solve for a validation radius. The simple assumption that $|u_k| \leq \frac{A_s}{k^s}$ for all $k \geq m$ provides a computationally cheap, but large, bound on the error. Though computationally more expensive, it is shown in [4] that the bounds can be improved by using explicit constraints on $|u_k|$ for $k = m, \ldots, M$ for some $M \geq m$. For the sake of clarity the computations performed in [4] were restricted to $M = m$. In this paper we exploit the computational parameter M to carry out continuation for large ranges of parameter values.

This paper is organized as follows. Establishing that an operator T is a contraction mapping on a space X is a question of estimates. This is addressed in Section 2. In particular, after introducing some notation the results of [4] are recalled in Section 2.1. As is mentioned above, given m, M is a computational parameter that is used to control the size of truncation errors. In Section 2.2 we provide a lower bound on the choice of M. In Section 2.3, we indicate how the FFT can be used to compute the nonlinear sums. In Section 3, we apply our techniques to two model problems: the Swift-Hohenberg equation and the Cahn-Hilliard equation. Finally, in Section 4 we use the results of our numerical experiments to suggest future directions of research.

2 Essential Estimates

Throughout the paper, we will use the subscript $(\cdot)_F$ to denote the components $k \in \{0, \cdots, m-1\}$. Recall that following the expansion of the system in the appropriate basis, we have

$$\dot{u} = f(u, \nu) \tag{6}$$

whose component-wise expansion is given by (3). Let m be a fixed projection dimension and consider the following truncation of (6).

For $u_F := (u_0, \ldots, u_{m-1}) \in \mathbb{R}^m$, define $f^{(m)} : \mathbb{R}^m \to \mathbb{R}^m$ by $f^{(m)}(u_F) = (f_0^{(m)}(u_F), \ldots, f_{m-1}^{(m)}(u_F))$ where for $k = 0, \ldots, m-1$,

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

The *Galerkin projection* of (6) is given by

$$\dot{u}_F = f^{(m)}(u_F, \nu) \tag{7}$$

(compare with (4)). The numerical equilibrium of (7) that we want to *validate* is denoted by $\bar{u}_F \in \mathbb{R}^m$. To simplify some of the expressions in the next section, we let $J_{m\times m}$ represent the numerical inverse of $Df^{(m)}(\bar{u}_F)$. Finally, let $\bar{u} = (\bar{u}_F, 0) \in \mathbb{R}^\infty$.

2.1 The Radii Polynomials

The philosophy of the validated continuation is to construct, at every step of the predictor-corrector algorithm, a set of the form (5) centered at $\bar{u} = (\bar{u}_F, 0)$ that will contain a unique equilibrium solution of the original problem (6). The idea is to construct an operator T whose fixed points correspond to equilibrium solutions of (6) and show that T contracts a set of the form (5). In order to verify that T is a contraction on $W_{\bar{u}}$, we will have to verify a finite number of polynomial inequalities given by the *radii polynomials* defined below. In principle, the computational parameter M will provide a way to compute with the components $k \in \{0, \cdots, M-1\}$ of the set $W_{\bar{u}}$ and to compute with the M first components $f_0, \cdots, f_{M-1}$ of (3). Hence if we take M big enough, the a priori upper bound on the truncation error term involved in doing the continuation on a Galerkin projection of dimension m will significantly decrease. However, the tradeoff will be an increase in the computational cost.

Recalling the results of [4], we introduce constants depending on the projection dimension m, the numerical equilibrium $\bar{u}_F$, the decay rate $s \geq 2$ and the tail constant $A_s > 0$ that are necessary to define the radii polynomials. Let

$$\begin{array}{rcl} \alpha & := & \dfrac{2}{s-1} + 2 + 3.5 \cdot 2^s \\ \bar{A} & := & \max\limits_{1 \le k < m} \{|\bar{u}_0|, |\bar{u}_k||k|^s\} \\ C_p & := & \max\limits_{k} \{|(c_p)_0|, |(c_p)_k||k|^s\} \\ A & := & A_s. \end{array}$$

Turning to the definition of the radii polynomials, for the components $k \in \{0, \cdots, m-1\}$, set

$$C_F^Y := \left| J_{m\times m} f^{(m)}(\bar{u}_F) \right| \ , \tag{8}$$

where $|\cdot|$ represents component-wise absolute values. We now introduce the computational parameter $M \ge m$. For $k \in \{0, \cdots, M-1\}$, let

$$C_k(p,j,l,M) := \sum_{\substack{k_0+\cdots+k_p=k \\ m \le |k_{p-j+1}|\dots,|k_p|<M \\ |k_0|,|k_1|,\dots,|k_{p-j}|<m}} \left|(c_p)_{k_0} \bar{u}_{k_1} \cdots \bar{u}_{k_{p-l}}\right| \frac{A_s}{|k_{p-j+1}|^s} \cdots \frac{A_s}{|k_p|^s}$$

and

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

Setting

$$1_F := (1, \cdots, 1)^T \in \mathbb{R}^m$$

$$\begin{array}{rcl} C_F^K(0) & := & |J_{m\times m}| \left[\displaystyle\sum_{l=1}^{d} \sum_{p=l}^{d} l \binom{p}{l} C_F(p,l,l,M) \right] + |J_{m\times m}| \displaystyle\sum_{l=2}^{d} \sum_{p=l}^{d} l \binom{p}{l} \epsilon_F(p,l,M) \qquad (9) \\ C_F^K(1) & := & |J_{m\times m}| \left[\displaystyle\sum_{l=1}^{d} \sum_{p=l}^{d} l \binom{p}{l} l \cdot C_F(p,l-1,l,M) \right] + \left| I_{F\times F} - J_{m\times m} Df^{(m)}(\bar{u}_F) \right| 1_F \\ C_F^K(i) & := & |J_{m\times m}| \left[\displaystyle\sum_{l=i}^{d} \sum_{p=l}^{d} l \binom{p}{l} \binom{l}{i} C_F(p,l-i,l,M) \right], \quad i = 2, \cdots, d \end{array}$$

allows us to define the m *finite radii polynomials* $P_0, \cdots, P_{m-1}$ by

$$P_k(r) := \sum_{i=2}^{d} C_k^K(i) r^i + \left[C_k^K(1) - 1 \right] r + \left[C_k^K(0) + C_k^Y \right] \ . \tag{10}$$

Set

$$C_k^Y := \begin{cases} \frac{|f_k(\bar{u}_F)|}{|\mu_k|} & \text{if } m \le k \le d(m-1) \\ 0 & \text{if } k > d(m-1). \end{cases}$$

Similarly, for $k \in \{m, \cdots, M-1\}$, let

$$\begin{aligned} C_k^K(0) &:= \frac{1}{|\mu_k|}\left[\sum_{l=1}^{d}\sum_{p=l}^{d} l\binom{p}{l} C_k(p,l,l,M)\right] + \frac{1}{|\mu_k|}\sum_{l=1}^{d}\sum_{p=l}^{d} l\binom{p}{l}\epsilon_k(p,l,M) \\ C_k^K(1) &:= \frac{1}{|\mu_k|}\left[\sum_{l=1}^{d}\sum_{p=l}^{d} l\binom{p}{l} l\cdot C_k(p,l-1,l,M)\right] \\ C_k^K(i) &:= \frac{1}{|\mu_k|}\left[\sum_{l=i}^{d}\sum_{p=l}^{d} l\binom{p}{l}\binom{l}{i} C_k(p,l-i,l,M)\right] \quad , \quad i = 2, \cdots, d \end{aligned}$$

and define the $M-m$ *tail radii polynomials*, $P_m, \cdots, P_{M-1}$ by

$$P_k(r) := \sum_{i=1}^{d} C_k^K(i) r^i + \left[C_k^K(0) + C_k^Y - \frac{A_s}{k^s}\right] \ . \tag{11}$$

Finally, let

$$C(\bar{A}, A) := \sum_{l=1}^{d}\sum_{p=l}^{d} l\binom{p}{l}\alpha^p C_p \bar{A}^{p-l} A^l$$

and define the *tail term* by

$$\tilde{P}_M := \frac{C(\bar{A}, A)}{|\mu_M|} - A_s \ . \tag{12}$$

The following proposition provides a concise representation of [4, Section 6] that is sufficient for the purpose of this paper.

Proposition 2.1 *Let $m \in \mathbb{N}$, $s \ge 2$, $A_s > 0$ and $M > d(m-1)$ be fixed and suppose that $|\mu_k| \le |\mu_{k+1}|$ for $k \ge M$. Assume that there exists an $r > 0$ such that the following conditions are simultaneously satisfied:*

1. *$P_k(r) < 0$ for all $k \in \{0, \ldots, m-1\}$,*
2. *$r(m-1)^s < A_s$,*
3. *$P_k(r) < 0$ for all $k \in \{m, \cdots, M-1\}$,*
4. *$\tilde{P}_M < 0$.*

Then, there exists a unique equilibrium solution of (6) in the set

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

If the conditions of Proposition 2.1 are satisfied, then we say that the set $W_{\bar{u}}$ *validates* the numerical equilibrium $\bar{u}_F \in \mathbb{R}^m$ with *validation radius* $r > 0$.

Hence, we see that to validate the numerical equilibrium $\bar{u}_F$, we need to compute the coefficients defined in (10), (11) and (12). Clearly, the computational cost arises from the terms $C_k(p,j,l,M)$, for $k \in \{0, \cdots, M-1\}$ which we handle as follows. For sake of simplicity, assume that the coefficients $(c_p)_k$ in $C_k(p,j,l,M)$ are always 0 for $k \neq 0$. Define

$$\tilde{a} := (|\bar{u}_0|, \cdots, |\bar{u}_{m-1}|, 0, \cdots, 0)^T \ , \ \tilde{A} := \left(0, \cdots, 0, \frac{A_s}{m^s}, \cdots, \frac{A_s}{(M-1)^s}\right)^T \in \mathbb{R}^M \ .$$

Then,

$$\begin{aligned} C_k(p,j,l,M) \quad &:= \sum_{\substack{k_0+\cdots+k_p=k \\ m \leq |k_{p-j+1}|\cdots,|k_p|<M \\ |k_0|,|k_1|,\ldots,|k_{p-j}|<m}} \left|(c_p)_{k_0} \bar{u}_{k_1} \cdots \bar{u}_{k_{p-l}}\right| \frac{A_s}{|k_{p-j+1}|^s} \cdots \frac{A_s}{|k_p|^s} \\ &= \quad |(c_p)_0| \sum_{\substack{k_1+\cdots+k_p=k \\ |k_1|,\ldots,|k_p|<M}} \tilde{a}_{k_1} \cdots \tilde{a}_{k_{p-l}} \tilde{A}_{k_{p-j+1}} \cdots \tilde{A}_{k_p} \ . \end{aligned}$$

As is discussed in section 2.3, FTT can be used to quickly compute $C_F(p,j,l,M) \in \mathbb{R}^M$ given the indices p, j, l, M.

2.2 Lower Bounds for M

The reason why we can get an a priori lower bound for M comes from the fact that the *tail term* $\tilde{P}_M$ is independent of the *validation radius* $r > 0$. Indeed, supposing that $M \geq d(m-1)$ the tail term inequality is given by

$$\frac{C(\bar{A}, A)}{|\mu_M|} - A_s < 0 \ . \tag{13}$$

Rather than obscuring the point in an abstract computation, observe that in the context of the Swift-Hohenberg equation (22), we have

$$C(\bar{A}, A) = 3\alpha(s)^3 A(\bar{A} + A)^2$$

and

$$\mu_M = \nu - \left(1 - M^2 L^2\right)^2 \ .$$

Since $A = A_s$, (13) becomes

$$3\alpha(s)^3 A_s (\bar{A} + A_s)^2 < A_s \left|\nu - (1 - M^2 L^2)^2\right| \ .$$

Supposing that $(1 - M^2L^2)^2 > \nu$ and dividing on both sides by $A_s > 0$, we get that

$$(M^2L^2 - 1)^2 > 3\alpha(s)^3 (\bar{A} + A_s)^2 + \nu \ .$$

Finally, supposing $M^2L^2 > 1$, we get

$$M > \gamma(L,\nu,s,\bar{u}_F,A_s) := \frac{1}{L}\sqrt{1+\sqrt{\nu+3\alpha(s)^3(\bar{A}(\bar{u}_F)+A_s)^2}} \tag{14}$$

Note that this lower bound only depends on the a priori information. Indeed, before starting the validation, we get all the quantities : L_0, ν and $\bar{u}_F$ from the continuation and s and A_s a priori given. Hence, before starting the validation process, we fix M to be at least γ.

2.3 Computing Sums Using the Fast Fourier Transform

In this section, we address the use of the FFT algorithm to compute sums of the form

$$\sum_{\substack{l_1+\cdots+l_p=l \\ |l_1|,\cdots,|l_p|<M}} a^1_{l_1}\cdots a^p_{l_p}\ , \tag{15}$$

where $a^1 := (a^1_{-M+1},\cdots,a^1_{M-1}),\cdots,a^p := (a^p_{-M+1},\cdots,a^p_{M-1}) \in \mathbb{R}^{2M-1}$. Note that we are not the first to use the FFT to compute sums of the form (15). In [6], the authors gave an explicit way to compute (15) for the cases $p=3$ and $p=5$. Here, we present the theory for a general $p \in \mathbb{N}$.

Definition 2.2 Let $b=(b_0,\cdots,b_{2M-2}) \in \mathbb{R}^{2M-1}$. Its *Discrete Fourier Transform* $\mathcal{F}(b)$ is given by

$$a_l = \mathcal{F}(b)|_l := \sum_{j=0}^{2M-2} b_j e^{-2\pi\mathbf{i}\left(\frac{jl}{2M-1}\right)}\ , \qquad \text{for } l \in \{-M+1,\cdots,M-1\}$$

Definition 2.3 Let $a=(a_{-M+1},\cdots,a_{M-1}) \in \mathbb{R}^{2M-1}$. Its *Inverse Discrete Fourier Transform* $\mathcal{F}^{-1}(a)$ is given by

$$b_j = \mathcal{F}^{-1}(a)|_j := \sum_{l=-M+1}^{M-1} a_l e^{2\pi\mathbf{i}\left(\frac{jl}{2M-1}\right)}, \qquad \text{for } j \in \{0,\cdots,2M-2\}$$

Let $\delta := \frac{p+1}{2}$, if p is odd and $\delta := \frac{p+2}{2}$ if p is even. Given $a^i=(a^i_{-M+1},\cdots,a^i_{M-1}) \in \mathbb{R}^{2M-1}$, define $\tilde{a}^i \in \mathbb{R}^{2\delta M-1}$ by

$$\tilde{a}^i_j = \begin{cases} a^i_j & \text{for } -M<j<M \\ 0 & \text{for } -\delta M+1 \le j \le -M \ \text{ and } M \le j \le \delta M-1 \end{cases} \tag{16}$$

For $j \in \{0,\cdots,2\delta M-2\}$, set

$$\tilde{b}^i_j := \mathcal{F}^{-1}(\tilde{a}^i)|_j = \sum_{l=-\delta M+1}^{\delta M-1} \tilde{a}^i_l e^{2\pi\mathbf{i}\left(\frac{jl}{2\delta M-1}\right)}\ . \tag{17}$$

For $l = -\delta M + 1, \cdots, \delta M - 1$,

$$\begin{aligned}
\mathcal{F}(\tilde{b}^1 * \cdots * \tilde{b}^p)|_l &= \sum_{j=0}^{2\delta M-2} \tilde{b}^1_j \cdots \tilde{b}^p_j e^{-2\pi \mathbf{i}\left(\frac{jl}{2\delta M-1}\right)} \\
&= \sum_{j=0}^{2\delta M-2} \left[\sum_{l_1=-\delta M+1}^{\delta M-1} \tilde{a}^1_{l_1} e^{2\pi \mathbf{i}\left(\frac{jl_1}{2\delta M-1}\right)} \right] \cdots \left[\sum_{l_p=-\delta M+1}^{\delta M-1} \tilde{a}^p_{l_p} e^{2\pi \mathbf{i}\left(\frac{jl_p}{2\delta M-1}\right)} \right] e^{-2\pi \mathbf{i}\left(\frac{jl}{2\delta M-1}\right)}
\end{aligned}$$

where

$$(\tilde{b}^1 * \cdots * \tilde{b}^p)_j := \tilde{b}^1_j \cdots \tilde{b}^p_j \ . \tag{18}$$

Defining

$$\begin{aligned}
S(j) &:= \prod_{i=1}^{p} \left[\sum_{l_i=-\delta M+1}^{\delta M-1} \tilde{a}^i_{l_i} e^{2\pi \mathbf{i}\left(\frac{jl_i}{2\delta M-1}\right)} \right] e^{-2\pi \mathbf{i}\left(\frac{jl}{2\delta M-1}\right)} \\
&= \sum_{\substack{l_1+\cdots+l_p=l \\ |l_1|,\cdots,|l_p|<M}} a^1_{l_1} \cdots a^p_{l_p} \ + \ \sum_{k=1}^{p} \left(\sum_{\substack{l_1+\cdots+l_p=l\pm k(2\delta M-1) \\ |l_1|,\cdots,|l_p|<M}} a^1_{l_1} \cdots a^p_{l_p} \right) \\
&\quad + \sum_{\substack{l_1+\cdots+l_p \notin \{l\pm k(2\delta M-1)|k=0,\cdots,p\} \\ |l_1|,\cdots|l_p|<M}} a^1_{l_1} \cdots a^p_{l_p} e^{2\pi \mathbf{i}\left(\frac{l_1+\cdots+l_p-l}{2\delta M-1}\right)j} \ ,
\end{aligned}$$

we obtain

$$\begin{aligned}
\mathcal{F}(\tilde{b}^1 * \cdots * \tilde{b}^p)|_l &= \sum_{j=0}^{2\delta M-2} S(j) \\
&= (2\delta M - 1) \sum_{\substack{l_1+\cdots+l_p=l \\ |l_1|,\cdots,|l_p|<M}} a^1_{l_1} \cdots a^p_{l_p} \\
&\quad +(2\delta M - 1) \sum_{k=1}^{p} \left(\sum_{\substack{l_1+\cdots+l_p=l\pm k(2\delta M-1) \\ |l_1|,\cdots,|l_p|<M}} a^1_{l_1} \cdots a^p_{l_p} \right) \\
&\quad + \sum_{\substack{l_1+\cdots+l_p \ \notin \ \{l\pm k(2\delta M-1)|k=0,\cdots,p\} \\ |l_1|,\cdots,|l_p|<M}} a^1_{l_1} \cdots a^p_{l_p} \left[\sum_{j=0}^{2\delta M-2} e^{2\pi \mathbf{i}\left(\frac{l_1+\cdots+l_p-l}{2\delta M-1}\right)j} \right] \ .
\end{aligned} \tag{19}$$

Euler's formula gives that for $l_1 + \cdots + l_p - l \not\equiv 0 \ mod \ (2\delta M - 1)$,

$$\sum_{j=0}^{2\delta M-1} e^{2\pi \mathbf{i}\left(\frac{l_1+\cdots+l_p-l}{2\delta M-1}\right)j} = 0 \ .$$

Hence, the third sum in (19) is zero. Turning to the second sum in (19), observe that $|l_1|, \cdots, |l_p| < M$ and $l \in \{0, \cdots, M-1\}$ implies that

$$l_1 + \cdots + l_p - l \in \{-(p+1)(M-1), \cdots, p(M-1)\} \; .$$

Hence, given the above mentioned choice of δ, the second sum of (19) is zero. Therefore, we can conclude that

$$\sum_{\substack{l_1+\cdots+l_p=l \\ |l_1|,\cdots,|l_p|<M}} a^1_{l_1} \cdots a^p_{l_p} = \frac{1}{2\delta M - 1} \cdot \mathcal{F}(\tilde{b}^1 * \cdots * \tilde{b}^p)|_l \; . \tag{20}$$

The discrete Fourier transforms required in the computations of (17) and (20) are computed using the FFT algorithm (e.g. see [1]).

3 Results

In this section we present some computations for the one-dimensional Swift-Hohenberg and the one-dimensional Cahn-Hilliard equations. This is meant both to show the practicality of the method of validated continuation and to highlight its current limitations.

The starting point for our computations is the trivial solution, $u_0 \equiv 0$, at a particular value of the continuation parameter, and an arbitrarily chosen Galerkin projection dimension.

The iteration of validated continuation proceeds as follows. As is indicated in the Introduction, we use a standard predictor-corrector numerical method to find a numerical solution at the next parameter value. That is, given a numerical zero of the Galerkin projection at ν_0, we find a new numerical zero $\bar{u}_F$ at the parameter value $\nu_1 = \nu_0 + \Delta\nu$. We then proceed with the validation step. We choose M to be the smallest integer satisfying

$$M \geq \max\{d(m-1), 2\gamma\} \; , \tag{21}$$

where γ is given by (14), and check the inequalities of Proposition 2.1. If the inequalities are satisfied, then Proposition 2.1 applies, we have validated the solution $\bar{u}_F$ at ν_1, and we proceed to the next step; that is, we increment ν and repeat the process. If validation fails we increase m by 2, recompute the numerical zero $\bar{u}_F$ at ν_1 and try to validate it. This procedure is repeated until the numerical zero $\bar{u}_F$ at ν_1 is validated or a maximum number of trials is reached. We remark for future reference that for Swift-Hohenberg our procedure always resulted in validation of the numerical zero.

At each step we monitor the determinant of the Jacobian to detect bifurcations. So starting with the trivial branch ($u \equiv 0$) we find branches that bifurcate from it, and then find branches that bifurcate from the newly found branches, and so on. In the case of Swift-Hohenberg we followed multiple branches. In each case we started with a low dimensional Galerkin projection, $m = 7$, and allowed the validation procedure to determine an appropriate value for m.

It is important to mention that we do not compute continuous branches of equilibria. The dots on Figures 1, 2, and 6, represent the points were we computed and validated equilibrium solutions. Notice also that the step size from one step to the next is not constant, but changes along each branch according to the formula $\Delta\nu := 2^{(4-k)/3}\Delta\nu$, where k is the number of iterations needed for the Newton method during the continuation step.

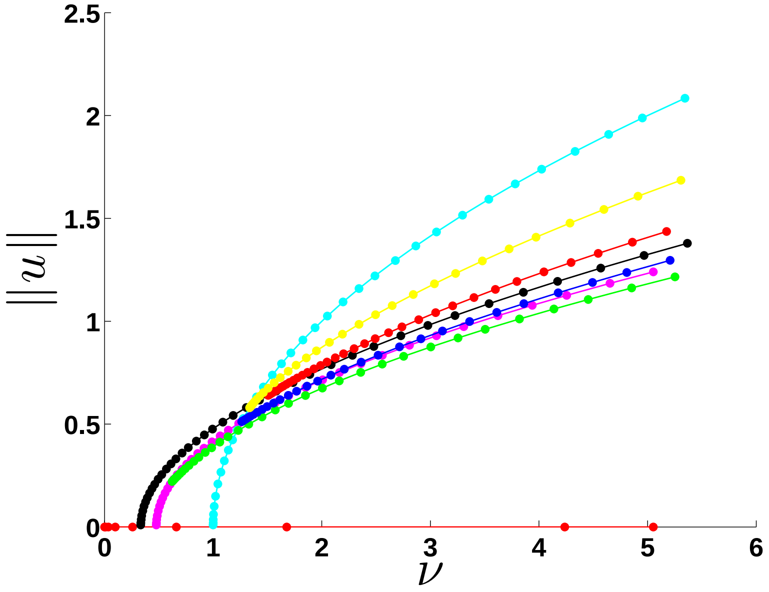

Figure 1: Bifurcation diagram for the Swift-Hohenberg equation (22) for $0 \leq \nu \leq 5$. The dots indicate the points at which a numerical zero was validated.

3.1 Swift-Hohenberg

Consider the Swift-Hohenberg equation [8]

$$u_t = \left(\nu - \left(1 + \frac{\partial^2}{\partial x^2}\right)^2\right) u - u^3, \quad x \in [0, 2\pi/L] \tag{22}$$

with $u(x,t) = u(x+2\pi/L, t)$ and $u(-x,t) = u(x,t)$. Expanding u in the Fourier basis $\{\cos(kLx) \mid k = 0, 1, 2, \ldots\}$ gives

$$u(x,t) = u_0(t) + 2\sum_{k=1}^{\infty} u_k(t)\cos(kLx).$$

Then (22) takes the form

$$\dot{u_k} = \mu_k u_k - \sum_{k_1+k_2+k_3=k} u_{k_1}u_{k_2}u_{k_3},$$

where

$$\mu_k = \nu - \left(1 - k^2 L^2\right)^2 , \tag{23}$$

is the eigenvalue of the linear part of (22). We fix $L = 0.65$, and use ν as the continuation parameter.

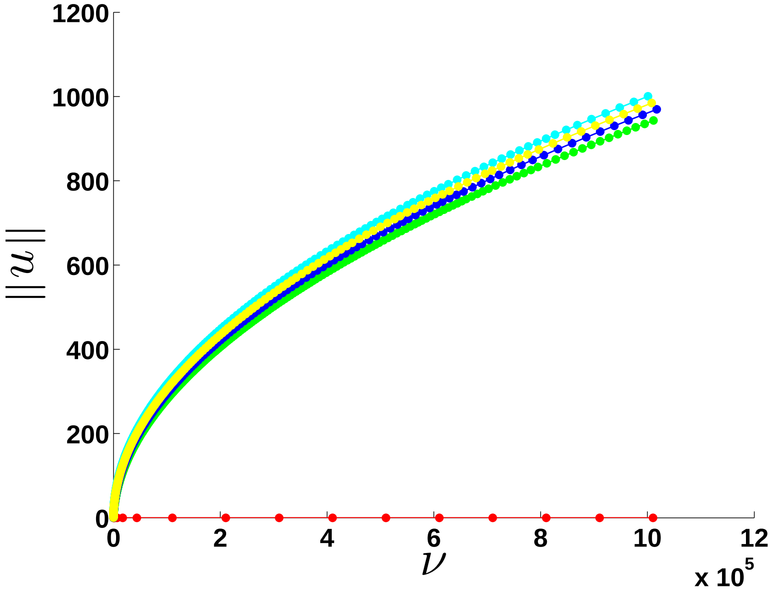

Figure 2: Some of the branches of equilibria of (22) for $0 \le \nu \le 10^6$. The dots indicate the points at which a numerical zero was validated. For the values $0 \le \nu \le 10^4$ the validation was done using interval arithmetic and hence at these points we have a mathematical proof of the existence and uniqueness of these solutions in the sets $W_{\bar{u}}(r)$. The color coding of the branches in this figure matches that of Figure 1.

As is indicated in the Introduction, we view the set $W_{\bar{u}}(r)$ (5) as a function of r. This implies that s and A_s are considered to be constants. For (22) we set $s = 4$ and $A_s = 1$. We discuss the choice of these values in Section 4.

We computed what we believe are all the branches of equilibria for $0 \le \nu \le 5$ and followed some of the branches up to $\nu \approx 10^6$. The diagrams are shown on Figure 1 and Figure 2. We validated all the branches up to $\nu \approx 10^4$ in Figure 2 using interval arithmetic to control floating point errors and thus rigorously verified that the inequalities of Proposition 2.1 are satisfied. This implies that we have mathematically proven the existence and uniqueness within the sets $W_{\bar{u}}(r)$ of the equilibria for Swift-Hohenberg at those values of $\nu \le 10^4$ indicated by the dots in Figure 2.

To describe some of the details and implications of these computations we focus on a particular branch of solutions. Given the title of this paper it is natural to consider a branch from Figure 2. We choose the blue one and note that the results for the other branches are similar. Plots of some of the solutions along the blue branch are presented in Figure 3. The computational cost of validating these branches are determined by m and M. Observe that m plays a significant role in the cost of the continuation step - the Newton step requires an approximation of the inverse of the Jacobian. The use of the FFT implies that the size of M determines the cost of the computation of the coefficients of the radii polynomials. Figure 4 indicates how m and M varies as a function of ν, though the reader should recall that in this setting given m, M is chosen according to (21).

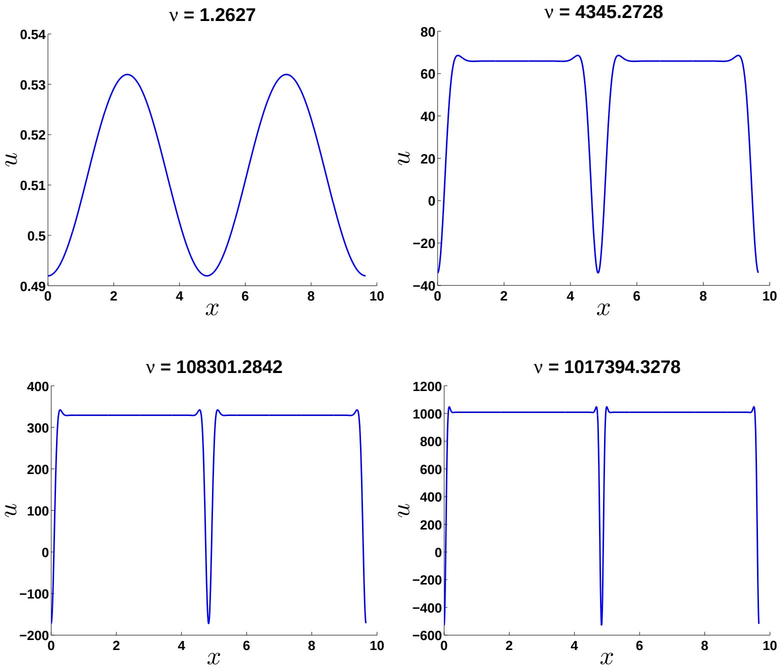

Figure 3: Some solutions along the blue branch of the diagram on Figure 2.

At the risk of being redundant, what Figure 2 indicates are the points in parameter space at which we have found a set of the form

$$W_{\bar{u}}(r) = \bar{u} + \prod_{k=0}^{m-1} [-r, r] \times \prod_{k=m}^{\infty} \left[-\frac{1}{k^4}, \frac{1}{k^4} \right]$$

in which there exists a unique equilibrium of (22). $\bar{u}$ is determined by the continuation method. m as a function of ν is given in Figure 4 and r as a function of ν is given in Figure 5. Observe that the knowledge that the equilibrium lies inside of $W_{\bar{u}}(r)$ gives very tight bounds. In particular, the true equilibrium of

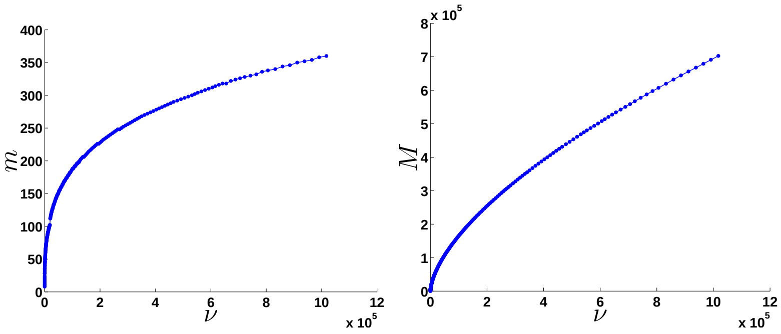

Figure 4: Plots of m and M along the blue branch of the diagram on Figure 2.

(22) at $\nu = 1017394.3278$ differs from that shown in Figure 3 by less than 10^{-10} in the L^2 norm. Thus, the peaks in the solution are not numerical artifacts.

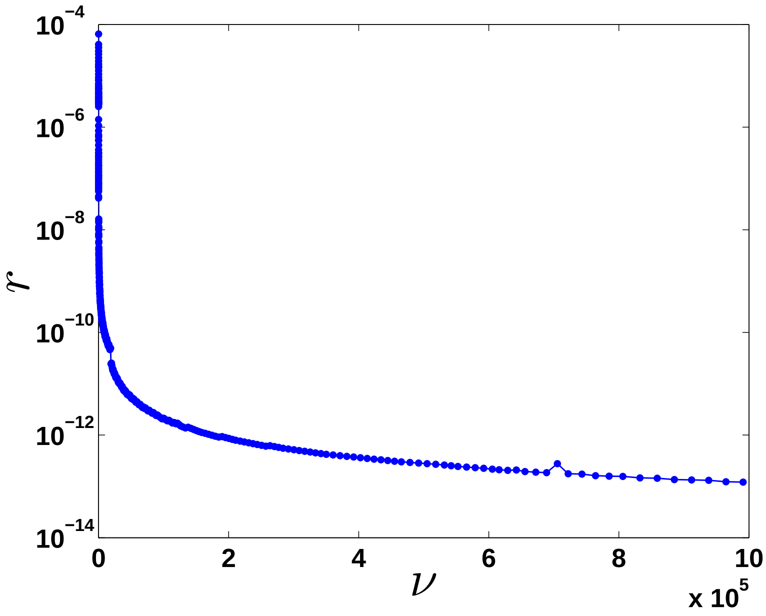

Figure 5: Plot of r along the blue branch of the diagram on Figure 2.

The computation time for the blue branch for ν up to $\nu \approx 10^4$ was 6.5 minutes without interval arithmetic and 9.19 hours with interval arithmetic. The computation for the whole branch (up to $\nu \approx 10^6$) was 11.67 hours without interval arithmetic. The computation times for the other branches were similar.

3.2 Cahn-Hilliard

We now turn our attention to the Cahn-Hilliard equation [2]

$$\begin{array}{l} u_t = -(\epsilon^2 u_{xx} + u - u^3)_{xx}, \quad x \in \Omega = [0,1] \\ u_x = u_{xxx} = 0, \quad \text{on} \quad \partial\Omega = \{0,1\} \end{array} \tag{24}$$

We consider the case of mass equal zero, that is,

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

In this case, to compute the equilibria of (24), is it sufficient to work with the Cahn-Allen equation

$$\begin{array}{l} u_t = \epsilon^2 u_{xx} + u - u^3, \quad x \in \Omega = [0,1] \\ u_x = 0, \quad \text{on} \quad \partial\Omega = \{0,1\} \end{array} \tag{25}$$

For this equation we use $\lambda = 1/\epsilon^2$ as the continuation parameter. For (25) we use the Fourier basis $\{\cos(k\pi x) \mid k = 0, 1, 2, \ldots\}$, then

$$u(x,t) = u_0(t) + 2\sum_{k=1}^{\infty} u_k(t)\cos(k\pi x).$$

So (25) takes the form

$$\dot{u_k} = \mu_k u_k - \sum_{k_1+k_2+k_3=k} u_{k_1} u_{k_2} u_{k_3},$$

where

$$\mu_k = 1 - \frac{k^2}{\lambda}, \tag{26}$$

is the eigenvalue of the linear operator in (25).

We computed using the procedure described at the beginning of Section 3. Choosing $s = 3$ and $A_s = 0.01$ led to the branches indicated in Figure 6. In particular, equilibria associated with the black branch are indicated in Figure 7.

The branches in Figure 6 terminate because the above mentioned procedure failed. To be more precise, we declare that our method fails when validation fails for 40 consecutive times at the same value of λ (recall that each time validation fails we increase m by 2, recompute the equilibrium and try to validate it again). Figure 8 indicates the rapid increase in m as a function of λ for the black branch in Figure 6. Observe that trying to validate a solution for 40 consecutive times is equivalent to increasing the dimension of the Galerkin projection by 80, recomputing the equilibrium and trying to validate it. In all the cases the reason for failure was that we were unable to find an r satisfying condition (1) of Proposition 2.1. In fact, it appears that the failure is due to the fact that at least one of the finite radii polynomials (10) fails to have any

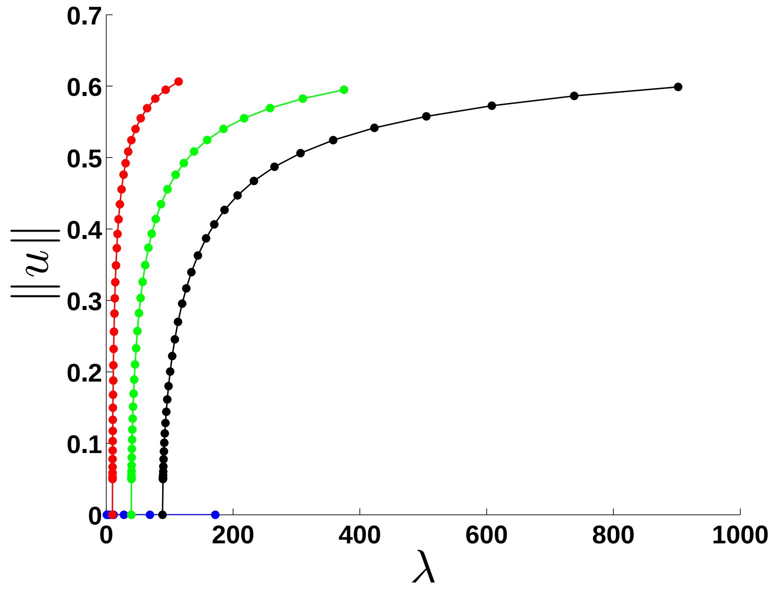

Figure 6: Bifurcation diagram for (24).

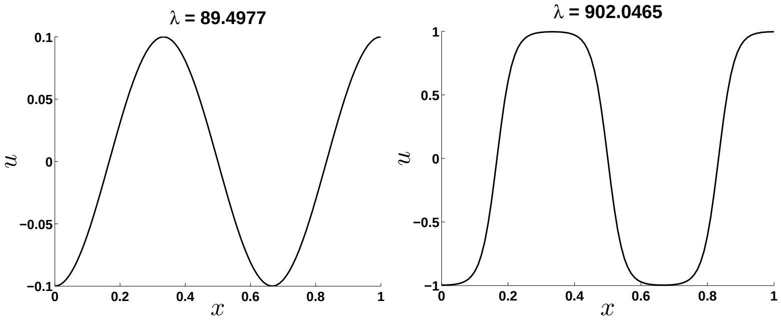

Figure 7: Solutions along the lower branch of the diagram on Figure 6.

positive roots. Since $P_k(0) > 0$, this implies that there is no positive solution to $P_k(r) < 0$.

As is indicated at the beginning of Section 2.1, there are only a few free constants involved in the definition of the radii polynomials: m, the dimensional of the Galerkin projection; M, a computational parameter; s, the decay rate; and A_s an a priori bound on the size of the Fourier coefficients. As is described above, failure of the procedure implies that m has been increased by 80. As one may expect and as the results in Figure 8 corroborates, this implies values of u_k for k close to m are essentially zero. Thus, further increase of the Galerkin projection at this point has little effect on the validation procedure.

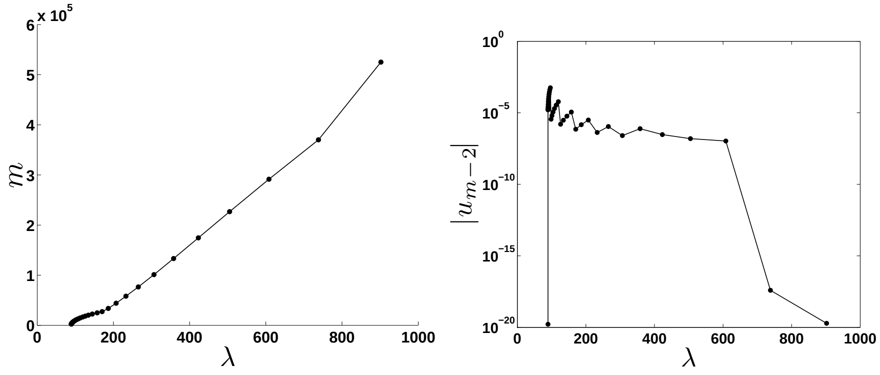

Figure 8: (Left) The dimension of the Galerkin projection m as a function of λ along the lower branch of the diagram on Figure 6. (Right) The value of $|u_{m-2}|$ as a function of λ along the same branch.

We tried to increase the value of M, since this results in better control on the tail errors. In particular, all the results indicated in Figure 6 were obtained using M equals twice the lower bound given by (14). We tried the same computations, from the beginning, using M equals four, six and ten times the lower bound in (14). In each case we were able to continue the branches in Figure 6 a bit further. However, in each case the procedure failed in the same way as before; there was no positive solution to the finite radii polynomial inequalities. This suggests that just increasing M does not provide an adequate solution to the problem.

We have no good heuristics for the choice of s and A_s. Random choices did not produce any significantly better results than $s = 3$ and $A_s = 0.01$.

4 Conclusion

The dramatic contrast between the Swift-Hohenberg equation, where validation computations even at extreme parameter values succeeded, and Cahn-Hilliard equation, where all computations eventually failed at fairly small parameter values calls for an explanation and suggests future directions of research for improvements in the technique of validated continuation.

The starting point for this discussion are the finite radii polynomials because, as is indicated in Section 3.2, it is the first condition of Proposition 2.1 that fails. Since both equations have cubic nonlinearities, the condition on the finite radii polynomials take the form

$$P_k(r) := C_k^K(3)r^3 + C_k^K(2)r^2 + \left[C_k^K(1) - 1\right] r + \left[C_k^K(0) + C_k^Y\right] < 0.$$

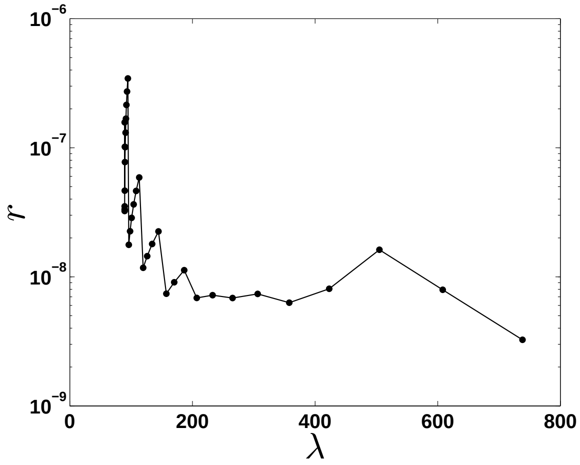

Figure 9: Plot of r as a function of λ along the lower branch of the diagram on Figure 6 .

Finding a positive solution requires that $C_k^K(0)$ and C_k^Y be sufficiently small. As is indicated in Figure 9, even though near the parameter value of failure m and M are large, the validation radius r is large as compared to the values for Swift-Hohenberg (Figure 5). This is further evidence that $C_k^K(0) + C_k^Y$ is not small.

To simplify the argument assume that $Df^{(m)}(\bar{u}_F)$ is diagonal. Then the diagonal terms of $J_{m\times m}$, the inverse of $Df^{(m)}(\bar{u}_F)$, become smaller as the eigenvalues μ_k become larger. By (9), $C_k^K(0)$ is proportional to $|J_{m\times m}|$ and by (8) C_k^Y is proportional to the tolerance of the numerical Newton method and is inversely proportional to μ_k. Comparing the eigenvalues of Swift-Hohenberg (23) against those of Cahn-Allen (26) we have

$$\mu_k = \nu - (1 - L^2k^2)^2 \quad \text{vs.} \quad \mu_k = 1 - \frac{k^2}{\lambda} = 1 - \epsilon^2 k^2.$$

Clearly, the magnitudes of the eigenvalues of Swift-Hohenberg are getting large at a much faster rate than those of Cahn-Hilliard.

This analysis provides some justification for the success in Swift-Hohenberg as opposed to Cahn-Hilliard. It also suggests necessary directions to improve the validated continuation methods.

The term C_k^Y essentially depends on the chosen tolerance of the numerical scheme. Thus, the failure of the first condition of Proposition 2.1 in Cahn-Hilliard is likely due to $C_k^K(0)$ which as presented in (9) is the sum of two terms. In the first term, the roles of s and A_s are clear. However, as is remarked in [4, Section 7] being able to optimize our choice of these constants remains an open question. The second term involves $\epsilon_k(p, l, M)$ which arises from the estimates of [4, Lemma 6.1]. The advantage of these estimates is that they can be straightforwardly applied directly to any polynomial nonlinearity. This simplicity comes at a price. For Swift-Hohenberg, with eigenvalues whose magnitudes grow at a fourth order rate, these estimate suffice. For Cahn-Hilliard we need better estimates. Work on this is in progress [5, 7].

References

[1] Brigham, E., *Fast Fourier Transform and its Applications*, Prentice Hall, 1st edition, 1988.

[2] 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.

[3] S. Day, Y. Hiraoka, K. Mischaikow, and T. Ogawa, *Rigorous numerics for global dynamics: A study of the Swift-Hohenberg equation*, SIAM Journal on Applied Dynamical Systems, 4 (2005), pp. 1–31.

[4] S. Day, J.-P. Lessard, and Mischaikow K, *Validated Continuation for Equilibria of PDEs*, to appear in SIAM Journal on Numerical Analysis, 2007.

[5] S. Day, J.-P. Lessard, and Mischaikow K, work in progress.

[6] Y. Hiraoka, and T. Ogawa, *An efficient estimate based on FFT in topological verification method*, J. Comput. Appl. Math. 199 (2007), no. 2, 238–244.

[7] J.-P. Lessard, *Validated Continuation for Infinite Dimensional Problems*, Ph.D Thesis, Georgia Institute of Technology.

[8] J.B. Swift and P.C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Physical Review A, 15:319, 1977.

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

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