Introduction to rigorous numerics in dynamics: general functional analytic setup and an example that forces chaos

Jan Bouwe van den Berg

October 22, 2017

Abstract

In this paper some basic concepts of rigorous computing in a dynamical systems context are outlined. We often simulate dynamics on a computer, or calculate a numerical solution to a partial differential equation. This gives very detailed, stimulating information. However, mathematical insight and impact would be much improved if we can be sure that what we see on the screen genuinely represents a solution of the problem. In particular, rigorous validation of the computations allows such objects to be used as ingredients of theorems.

The past few decades have seen enormous advances in the development of computer-assisted proofs in dynamics. One approach is based on a functional analytic setup. The goal of this paper is to introduce the ideas underlying this rigorous computational method. As the central example we use the problem of finding a particular periodic orbit in a nonlinear ordinary differential equation that describes pattern formation in fluid dynamics. This simple setting keeps technicalities to a minimum. Nevertheless, the rigorous computation of this single periodic orbit implies chaotic behavior via topological arguments (in a sense very similar to "period 3 implies chaos" for interval maps).

1 Introduction

This paper contains slightly extended lecture notes that accompanied the introductory lecture of the AMS short course on *Rigorous Numerics in Dynamics* at the Joint Mathematics Meetings 2016 in Seattle. We begin with a few sections that are meant to provide motivation and general background information, and these are therefore of a rather global nature (there is a vast amount of introductory texts on nonlinear dynamics; some examples are [26, 33, 17]). In §2 we turn to a precise mathematical formulation of the general analytic-computational approach. In §3 we introduce an example for which the approach is subsequently worked out in complete detail in §4.

The core of the material presented here is a synopsis of the state-of-the-art. In §1.2 we give several references to recent work which forms the foundation of the method described in this paper. However, it is by no means an exhaustive list. Additional references can be found in the papers mentioned. The main underlying references for the detailed calculations in this paper are [40] and [18]. In particular, the bulk of §3 is summarized from [40, §1]. We note that in the details of the computer-assisted proof we deviate substantially from the original approach. The new estimates presented in §4.5 and §4.6 are simpler to explain and lead to a computationally cheaper proof. Furthermore, the setup in §4.3 with *two* so-called radii polynomials depending on *two* radii (which characterize a rectangle in a the product space of unknowns) is more systematic than the ad-hoc choices of weights that were used in [40]. We believe that this leads to improved understanding and has the potential for generalizations far beyond the relatively simple example chosen here to illustrate the method.

1.1 Background: dynamical systems

Informally, a dynamical system is a system that evolves in time according to a deterministic law. The time variable $t \in \mathcal{T}$ may be either discrete or continuous. The time may flow forward ($\mathcal{T} = \mathbb{N}$ or $\mathcal{T} = \mathbb{R}^+ = [0, \infty)$) or forward-and-backward ($\mathcal{T} = \mathbb{Z}$ or $\mathcal{T} = \mathbb{R}$). It may happen that an evolution stops after some finite time (e.g. because it reaches infinity in finite time), but we will not be considering such details of a technical nature here. Mathematically, what we need are a topological space $\mathcal{X}$, the phase space, and a continuous function, the flow, $\varphi : \mathcal{X} \times \mathcal{T} \to \mathcal{X}$ with the properties

$$\varphi(x; 0) = x \qquad \text{for all } x \in \mathcal{X}, \tag{1a}$$

$$\varphi(x; t_1 + t_2) = \varphi(\varphi(x; t_1); t_2) \qquad \text{for all } x \in \mathcal{X} \text{ and } t_1, t_2 \in \mathcal{T}. \tag{1b}$$

The interpretation is that after a time t the state x has moved to $\varphi(x; t)$. The identity (1b) expresses that flowing for some time t_1 to a point $x_1 = \varphi(x; t_1)$ and then flowing another t_2 from x_1, gives the same result as flowing for a time $t_1 + t_2$ directly.

One usually writes $x(t) = \varphi(x(0); t)$ if it is clear which flow is meant. The fundamental problem of dynamical systems is that in all but the simplest cases we do not know φ, i.e., we have no useful explicit description of the flow. However, there are many cases where a "law of motion" (for which we have an explicit description) together with initial data determine the entire future dynamics deterministically, hence we know there *is* a flow φ, i.e., it is a dynamical system.

Indeed, dynamical systems appear in all branches of sciences. To give a few examples:

- Physics: Newton's law of motion in classical mechanics
- Chemistry: the Schrödinger equation
- Fluid dynamics: the Navier-Stokes equations
- Economics: the Black-Scholes equation for option pricing
- Meteorology: coupled ocean-atmosphere models in weather forecasting
- Systems biology: reaction-diffusion equations for reaction networks
- Astronomy: planet-, star-, galaxy- and cluster-formation and their evolution
- Neuroscience: dynamic models for bursting neurons

Dynamical systems do not just occur in applications. Indeed, they are interesting from a mathematical perspective in their own right. Moreover, they appear in many branches of mathematics, such as in the Calculus of Variations (gradient flows), in differential geometry (e.g. the Ricci flow, leading, eventually, to the resolution of the Poincaré conjecture), in evolutionary and stochastic partial differential equations, and in symplectic geometry (pseudoholomorphic curves).

The modern theory of dynamical systems dates back to Poincaré and his study of the (in)stability of the solar system and its ultimate fate. More generally, the central question in dynamical systems is the following: start at some point $x(0)$ and let it evolve under the flow φ; *now what happens?* And how do the individual orbits fit together to give a global description of the behavior of the system as a whole? Additionally, one would like to know not just how the ultimate fate depends on the initial position, but also on parameters in the system, and on the "type" of system (e.g. Hamiltonian, gradient, symmetric, network structure).

Clearly, no single method can answer all these questions, especially since there are so many types of dynamical systems (linear, nonlinear, ordinary differential equations (ODEs), partial differential equations (PDEs), delay equations, finite and infinite dimensional maps).

Linear dynamical systems can often be analyzed by hand using linear algebra, functional analysis and special functions. In this paper we are interested in *nonlinear* dynamics, for which the analytic tools are much coarser. Topological approaches (e.g. fixed point theorems, Conley index) and variational methods (e.g. Morse theory) give some information about existence of solutions

for large classes of systems. However, such robust but coarse results often lack information about multiplicity of solutions for specific nonlinear dynamical systems. Moreover, they usually provide very little qualitative (let alone quantitative) information about the shape of the solutions.

In practice, a topological approach is often complemented by computer simulations. While the latter give numerical evidence for the behavior of solutions, there is unquantified uncertainty in this information due to the absence of explicit error bounds on the outcome of numerics. These errors originate from rounding errors and, more importantly, truncation and discretization errors.

The goal of the paper is to introduce a framework in which the quantitative but tentative information from computer simulations can be turned into mathematically rigorous statements (theorems). Numerical computations are naturally limited to a relatively small range of parameter values and a restricted part of phase space. In particular, our methods focus on rigorously (in the mathematical sense) finding solutions in nonlinear dynamical systems that represent *special features*. One may think of equilibria, cycles, connecting orbits (homoclinic or heteroclinic), horseshoes ("chaos"), invariant tori, etc.

What the special solutions of most interest are will depend on the dynamical system. Nevertheless, some useful global statements can be made. In particular, in dissipative systems all dynamics will evolve towards an *attractor*, and hence studying this attractor is of most interest. More generally, the fundamental theorem of dynamical systems, due to Conley [9], states that any flow (on a compact metric space) decomposes into a *chain recurrent* part and a *gradient-like* part. The recurrent components consist, roughly speaking, of points that come back arbitrarily close to themselves under the evolution of the flow, whereas points in the gradient part do not. The dichotomy is thus that points are either recurrent, or they run "downhill" towards a recurrent component (and if we can go back in time: they come from *another* recurrent component). It is thus interesting to study recurrent components as well as the connecting (heteroclinic) orbits between them. As an example, in a gradient flow the principle interest is in the (hyperbolic) equilibria and the heteroclinic orbits that connect equilibria with Morse index difference one.

1.2 Rigorous numerics

computer-assisted proofs have a relatively short history in mathematics. Nevertheless, the proof of the four color theorem [1, 27] and of the densest sphere packing (Kepler problem) [16] belong to the classics in the field. Regarding the role of computers in dynamical systems, the start of the systematic study of nonlinear dynamical systems coincided with the advent of the computer (Feigenbaum diagram, Mandelbrot set). Although some qualitative properties were already sketched by Poincaré and contemporaries, only with the help of numerical simulations were people able to imagine and map the variety of dynamics exhibited by nonlinear systems. Ever since, computers have been a crucial instrument in the mathematical analysis of dynamical systems, providing both inspiration and a way to test (and reject) conjectures. Moreover, due to the very limited availability of tools for doing quantitative nonlinear analysis by hand, computers have been used as proof-assistants very early on. Prime examples are the proof of the universality of the Feigenbaum constant [21] and the existence of the (numerically "obvious") chaotic attractor in the Lorenz system [36].

By now there is a variety of computational frameworks that aid in the study of nonlinear dynamics. Each of these involves a combination of topological (or analytic-topological) arguments with (interval arithmetic) computations. Here we delve into the wealth of possibilities of one approach. Nevertheless, it is important to stress that there is a range of partly overlapping methodologies, each with their own strengths and relative weaknesses. To name a few (very briefly; we refer the reader to the references for examples and more details):

- Piotr Zgliczyński and collaborators have developed a method based on integrating the flow using rigorous enclosures and Taylor methods, combined with topological arguments (Conley index, covering relations) and derivative information (through so-called cone conditions) [43, 44]. An extensive C++ package CAPD (computer-assisted Proofs in Dynamics) is available for supporting these proofs [7].

- Hans Koch, in collaboration with Gianni Arioli, has developed a general functional analytic

approach to computer-assisted proofs in nonlinear analysis (e.g. [3, 4]). That work has provided the crucial underlying ideas for the techniques discussed in this paper.

- In a spirit similar to the above method, techniques have been developed by Mitsuhiro Nakao, Michael Plum and collaborators for studying elliptic problems by combining a Newton-like fixed point map with computer-assisted computations with a variable mix of functional analytic and computational effort (e.g. [6, 25]). This forms another stepping stone for the techniques discussed below.

- Konstantin Mischaikow and collaborators have developed a "database" approach to representing the global behavior of parameterised families of nonlinear dynamical systems [2, 20]. The database encodes the minimal Morse-Conley graphs and their dependence on parameters. This framework focuses on global dynamics while working up to a finite resolution. It hence complements the local but "infinitesimal" results that form the core of the method discussed in the current paper.

The method described in this paper has been developed in the past decade by a large number of people, and we refer to [8, 11, 12, 14, 23, 37, 39] and the references therein for a limited sample of the contributions.

1.3 Forcing theorems

Forcing theorems in dynamical systems are statements of the form: if there exists a solution (orbit) of type A, then there must be orbits of types B, C and D. In the most interesting cases, the list of implied orbits is infinitely long. Such forcing theorems are strong tools in the study of nonlinear dynamics. The main obstacle to applying them in concrete situations is that it is usually very difficult to show that an orbit of type A exists, even if numerical simulations may be tentatively convincing. This is the perfect setting for showcasing the power of rigorous computations, as they may be used as a tool to prove the existence of a type A orbit in concrete problems.

Examples of forcing theorems are (roughly in increasing order of sophistication)

- In scalar autonomous ordinary differential equations: if there are two equilibria then there exists a heteroclinic orbit (non necessarily between these equilibria), or (in very degenerate situations) a continuum of equilibria.

- In a planar system of ODEs: a periodic orbit implies the existence of an equilibrium (located inside the orbit). This is a classical result, where the forcing is via index theory.

- Interval maps: "period 3 implies chaos" [24]. If there exists a periodic orbit of (minimal) period 3, then there are periodic orbits of arbitrary period (this generalizes to the Sharkovsky ordering [31]) and there is an invariant set on which the dynamics is chaotic (has positive entropy with respect to a suitably defined distance).

- Three dimensional systems of ODEs: a Shilnikov homoclinic orbit (to an equilibrium with a one-dimensional unstable manifold and a two-dimensional 'spiral' stable manifold) implies chaos if the expansion rate of the unstable manifold is larger than the contraction rate in the stable manifold [32].

- In a system of ODEs with three or more degrees of freedom: the existence of a transversal homoclinic orbit to a cycle implies that there is a Smale horseshoe in the return map, hence chaos. Similarly, any hyperbolic periodic point of a discrete time dynamical system that has its stable and unstable manifolds intersecting transversally, implies chaos.

- Maps on two-dimensional manifolds: a periodic orbit whose associated braid (after suspension) is of pseudo-Anosov type implies chaotic dynamics for the map [35].

- Second order Lagrangian system: certain periodic orbits (to be made explicit in Section 3) imply chaos [15].

- In variational problems: the *Morse-Floer homology* encodes forcing relations between equilibria with certain (Morse-Floer) indices and connecting (heteroclinic) orbits between such equilibria [29].
- More generally for dynamical systems without a variational structure but with "sufficient compactness properties": as already mentioned, the *Conley index* encodes relations between invariant set and connecting orbits.

We note that in most of these examples a combination of topological arguments and rigorous computation leads to the strongest results.

2 The general setup

2.1 Motivation: Newton's method

If we numerically try to find a solution of $f(x) = 0$, where $f : \mathbb{R}^N \to \mathbb{R}^N$ is some nonlinear problem, then, starting from with some initial guess x_0, we may iterate the Newton scheme

$$x_{n+1} = \widetilde{T}(x_n) \stackrel{\text{def}}{=} x_n - Df(x_n)^{-1} \cdot f(x_n)$$

to obtain a sequence that hopefully approaches a zero of f quickly. Since

$$D\widetilde{T}(x)y = y - (D[Df(x)^{-1}]y) \cdot f(x) - Df(x)^{-1} \cdot Df(x) \cdot y$$

we find that $D\widetilde{T}(\hat{x})y = 0$ for any $\hat{x}$ such that $f(\hat{x}) = 0$, provided $Df(\hat{x})$ is invertible. This implies that if $Df(\hat{x})$ is invertible, then $\|D\widetilde{T}(x)\|$ is small near the fixed point $\hat{x}$, hence $\widetilde{T}$ is a contraction mapping with very strong contraction rate. If we merely want a (moderately strong) contraction mapping (e.g. because we rely on analytic arguments in addition to computer calculations), then there is thus quite a lot of room to alter the definition of the map $\widetilde{T}$ and still be successful. For example, one may try replacing $Df(x)^{-1}$ by $Df(x_0)^{-1}$ so that there is no need to recompute the inverse after every iteration.

2.2 Setup

Our approach starts by recasting the dynamic problem of interest as a zero finding problem $F(x) = 0$ in some *infinite* dimensional space. To fix ideas, let X and X' be Banach spaces, and $F : X \to X'$ be a (Fréchet) differentiable map (in practice: $F \in C^1(X, X')$ or $C^2(X, X')$). Assume that $\overline{x} \in X$ is an approximate zero of F, i.e. $\|F(\overline{x})\|_{X'} \approx 0$, and that we have a left inverse $DF(\overline{x})^{-1}$ of $DF(\overline{x})$. Then we can define the Newton-like map

$$\widehat{T}(x) = x - DF(\overline{x})^{-1}F(x).$$

If, as is explicitly assumed, $DF(\overline{x})$ is invertible, then fixed points of $\widehat{T}$ are in one-to-one correspondence with zeros of F. In this case we can attempt to show that $\widehat{T}$ is a contraction mapping in a neighborhood of $\overline{x}$. However, in practice inverting $DF(\overline{x})$ is "too difficult" in this infinite dimensional setting.

Thus we make use of an approximate inverse, typically obtained in part by computing a numerical inverse. We approximate $DF(\overline{x})$ by a (simpler) linear operator $A^\dagger : X \to X'$ and choose $A \in L(X', X)$ to be an approximate left inverse of $A^\dagger$.

In our computational framework the choice of the linear operator A is based on a computer calculation (a non-rigorous one). Namely, we choose (Galerkin) projections π_N and π'_N of X and X' onto N-dimensional subspaces X_N and X'_N, respectively, for some $N \in \mathbb{N}$. The natural embedding of X_N into X is denoted by ι. The finite dimensional truncated problem is given by

$$0 = F_N(x_N) \stackrel{\text{def}}{=} \pi'_N F(\iota x_N),$$

where $F_N : X_N \to X'_N$ and $x_N \in X_N$. We then seek a numerical approximation $\overline{x}_N$ of a solution to $F_N(x_N) = 0$, for example using a (finite dimensional) Newton iteration scheme. The idea is that $\overline{x} \stackrel{\text{def}}{=} \iota\overline{x}_N$ is an approximate solution of $F(x) = 0$.

Next we compute the Jacobian $DF_N(\overline{x}_N)$ and determine a numerical inverse $A_N \approx DF_N(\overline{x}_N)^{-1}$. We now need to "extend" the operator $A_N : X'_N \to X_N$ (or $\iota A_N : X'_N \to X$) to an *injective* operator $A : X' \to X$. How one goes about making this extension depends on the problem. It should be simple enough to allow for explicit analysis, while accurate enough to be an "approximate" left inverse of $DF(\iota\overline{x}_N)$ in a sense that will be made precise in §2.3.

Finally we are ready to define the Newton-like operator

$$T(x) \stackrel{\text{def}}{=} x - AF(x). \tag{2}$$

Observe that if A is injective, then fixed points of T correspond to zeros of F. Our goal is to show that T is a contraction map on a ball around $\overline{x}$. This ball should not be chosen too small (since then it won't contain a zero of F), but neither should it be chosen too large (since one cannot prove that T is a contraction on such larger balls (in fact it might not *be* a contraction there)).

Remark 1. *The space X' has no significance. The only important requirement is that we choose a linear operator A such that $AF(x) \in X$ for all $x \in X$.*

Remark 2. *To have correspondence between fixed points of T and zeros of F, we need that $AF(x) = 0$ implies $F(x) = 0$, i.e., A is injective on the range of F. In fact, one only needs that A is injective on the range of F when the domain of F is restricted to the fixed point space of T (e.g., one may have a priori knowledge about this due to symmetries).*

2.3 The theorem

The following theorem provides conditions under which we can guarantee the existence of a fixed point of T (and hence a zero of F provided A is injective).

Let X be a Banach space. Let $B_r(\overline{x})$ denote the *closed* ball of radius r around $\overline{x} \in X$.

Theorem 3. *Let T be a Fréchet differentiable map from a Banach space X to itself, such that for all $r > 0$*

$$\|T(\overline{x}) - \overline{x}\|_X \leq Y \tag{3a}$$

$$\|DT(x)\|_{B(X)} \leq Z(r) \qquad \text{for all } x \in B_r(\overline{x}), \tag{3b}$$

for some $Y \in \mathbb{R}^+$ and some function $Z : \mathbb{R}^+ \to \mathbb{R}^+$. If there exists an $\hat{r} > 0$ such that

$$Y + \hat{r}Z(\hat{r}) < \hat{r}, \tag{4}$$

then T has a unique fixed point in $B_{\hat{r}}(\overline{x})$.

The proof of the theorem is given in §2.5. We note that this is an abstract theorem in the sense that does not immediately reveal the context of §2.2. However, one should keep in mind that the bounds Y and $Z(r)$ will be given by complicated but explicit formulas that depend on numerically obtained values, see §4 for a detailed example. Indeed, both A (which appears in the definition (2) of T) and $\overline{x}$ are determined (largely) numerically. The inequality (4) is then checked with the help of a computer by using interval arithmetic. In practice Z usually depends polynomially on r. We then denote the so-called *radii polynomial* by

$$p(r) \stackrel{\text{def}}{=} Y + r[Z(r) - 1],$$

and we look for an $\hat{r} > 0$ such that $p(\hat{r}) < 0$. We note that we keep r as a variable parameter at the cost of having to work with a functional form for $Z = Z(r)$, which requires careful bookkeeping. The advantage is that we have more flexibility, not just enhancing the chance of proving that a fixed point of T exists, but also allowing better bounds both on the location of the fixed point (by

determining the smallest $\hat{r}$ for which $p(\hat{r})$ is negative) and on the uniqueness range (by determining the largest $\hat{r}$ for which $p(\hat{r})$ is negative).

One may interpret (3a) as an estimate of the residue (but observe that the residue is "preconditioned" with the linear operator A), while (3b) represents both a bound on how well A approximates $DF(\overline{x})^{-1}$ and an estimate of how nonlinear the problem is.

On the one hand, one may view Theorem 3 as a mathematically rigorous form of an explicit a-posteriori error analysis. On the other hand, it represents the widely used approach in analysis of setting up a fixed point problem on some ball around a guess for the solution, using a "suitable choice" of a linear operator (denoted by A in our case). In our context, the formulation is adjusted to a situation where both the (center of the) ball and the linear operator depend heavily on numerically obtained values.

Since the theorem gives uniqueness, it is clear that one can only attack problems that are formulated in such a way that they have isolated solutions (e.g. one needs to quotient out any continuous symmetries). Moreover, since the solution is obtained from a contraction argument (see §2.5), the solution is stable under small perturbations of the problem $F(x) = 0$, hence only robust solutions can be found, and only problems with robust solutions can be analyzed. Often this requirement can be achieved by carefully reformulating the problem, after understanding the sources of non-uniqueness and/or non-robustness. This robustness feature should also be viewed as an advantage, since if one finds a solution then it is automatically "in general position" (e.g. hyperbolic or transversal), a property that is often required when using it as an ingredient in a larger mathematical context, such as in forcing theorems. Moreover, the setup is perfect for performing continuation in parameters, which is of crucial importance in most applications of nonlinear dynamical systems.

2.4 Alternative formulations

Since $T(x) = x - AF(x)$, the conditions (3a) and (3b) may be reformulated in terms of F. Furthermore, in practice the estimate on DF is split into three components with the help of an operator $A^\dagger$ that approximates $DF(\overline{x})$. The conditions (3) are then replaced by

$$\|AF(\overline{x})\|_X \leq Y \tag{5a}$$

$$\|I - AA^\dagger\|_{B(X)} \leq Z^0 \tag{5b}$$

$$\|A[A^\dagger - DF(\overline{x})]\|_{B(X)} \leq Z^1 \tag{5c}$$

$$\|A[DF(x) - DF(\overline{x})]\|_{B(X)} \leq Z^2(r) \qquad \text{for all } x \in B_r(\overline{x}), \tag{5d}$$

for some Y, Z^0, $Z^1 \in \mathbb{R}^+$ and $Z^2 : \mathbb{R}^+ \to \mathbb{R}^+$, and one sets $Z(r) = Z^0 + Z^1 + Z^2(r)$. The final estimate (5d) is often replaced by a Lipschitz bound

$$\|A[DF(x) - DF(x')]\|_{B(X)} \leq \widehat{Z}^2 \|x - x'\|_X \qquad \text{for all } x, x' \in B_{r^*}(\overline{x}),$$

where $r^* > 0$ is some a priori bound on the radius, and $Z^2(r) = \widehat{Z}^2 r$. The Lipschitz bound may, in turn, be supplanted by a bound on the second derivative:

$$\|AD^2F(x)[v, w]\|_X \leq \widehat{Z}^2 \qquad \text{for all } v, w \in B_1(\overline{x}) \text{ and all } x \in B_{r^*}(\overline{x}).$$

In both cases one needs an a posteriori check that the radius $\hat{r}$ for which $p(\hat{r}) < 0$ also satisfies $\hat{r} \leq r^*$. In this notation the radii polynomial is written as

$$p(r) = Y + r[Z^0 + Z^1 - 1] + r^2 \widehat{Z}^2 \qquad \text{for } r \in [0, r^*].$$

When we find a radius $\hat{r} \leq r^*$ such that $p(\hat{r}) < 0$, and if we know that A is injective, then we conclude that F has a unique zero in $B_{\hat{r}}(\overline{x})$. As mentioned before, the smallest such $\hat{r}$ gives the best bound on where the solution is, whereas the biggest such $\hat{r}$ gives the strongest uniqueness result.

2.5 The proof of Theorem 3

The goal is to show that T is a uniform contraction on $B_{\hat{r}}(\overline{x})$. First, we check that $B_{\hat{r}}(\overline{x})$ gets mapped into itself. Let $x \in B_{\hat{r}}(\overline{x})$ and apply the Mean Value Theorem for Banach spaces to obtain

$$\begin{aligned}\|T(x) - \overline{x}\|_X &\leq \|T(x) - T(\overline{x})\|_X + \|T(\overline{x}) - \overline{x}\|_X \\ &\leq \sup_{x' \in B_{\hat{r}}(\overline{x})} \|DT(x')\|_{B(X)} \|x - \overline{x}\|_X + Y \\ &\leq Z(\hat{r})\hat{r} + Y.\end{aligned}$$

Since $Z(\hat{r})\hat{r} + Y < \hat{r}$ by assumption, it follows that that T maps $B_{\hat{r}}(\overline{x})$ into itself. Second, to see that T is a contraction on $B_{\hat{r}}(\overline{x})$, let $x_1, x_2 \in B_{\hat{r}}(\overline{x})$ be arbitrary, and use the estimate (again applying the Mean Value Theorem)

$$\|T(x_1) - T(x_2)\|_X \leq \sup_{x' \in B_{\hat{r}}(\overline{x})} \|DT(x')\|_{B(X)} \|x_1 - x_2\|_X \leq Z(\hat{r})\|x_1 - x_2\|_X. \tag{6}$$

It follows readily from $Y + Z(\hat{r})\hat{r} < \hat{r}$ with $Y \geq 0$ that the contraction rate $Z(\hat{r})$ in (6) is less than unity. Applying the Banach contraction mapping theorem finishes the proof.

3 Example of a rigorous computation that forces chaos

To illustrate how one can apply the general setup from §2, we consider as our key example a problem from pattern formation. The following description (in §3) is a condensed version of the main result in the paper "Chaotic braided solutions via rigorous numerics: chaos in the Swift-Hohenberg equation" by Jan Bouwe van den Berg and Jean-Philippe Lessard [40].

We first sketch the context. The *Swift-Hohenberg* equation [10, 34] is the fourth order parabolic partial differential equation (PDE)

$$\frac{\partial U}{\partial T} = -\left(\frac{\partial^2}{\partial X^2} + 1\right)^2 U + \alpha U - U^3. \tag{7}$$

It is a model equation for pattern formation due to a finite wavelength instability (e.g. in Rayleigh-Bénard convection). The onset of instability is at parameter value $\alpha = 0$. Time-independent solutions to the PDE (7) satisfy the *ODE*

$$-U'''' - 2U'' + (\alpha - 1)U - U^3 = 0. \tag{8}$$

The latter has a conserved quantity, the *energy*:

$$E = U'''U' - \frac{1}{2}U''^2 + U'^2 - \frac{\alpha - 1}{2}U^2 + \frac{1}{4}U^4 + \frac{(\alpha - 1)^2}{4}.$$

For $\alpha > 1$, the energy level $E = 0$ contains two constant solutions $U = \pm\sqrt{\alpha - 1}$. For $\alpha > \frac{3}{2}$, these are stable equilibria of the PDE (7) and saddle-foci for the ODE (8). It is well known that saddle-foci can act as organizing centers for complicated dynamics [13, 19]. By combining a topological forcing result with rigorous numerics (to obtain the "seed" of the forcing), we can indeed establish chaos for a large range of parameter values, as explained below.

Proposition 4 (Proposition 1 in [40])**.** *The dynamics of the Swift-Hohenberg ODE* (8) *on the energy level* $E = 0$ *is chaotic for all* $\alpha \geq 2$.

We first perform a change of coordinates that compactifies the parameter range:

$$y = \frac{X}{\sqrt[4]{\alpha - 1}}, \qquad u(y) = \frac{U(X)}{\sqrt{\alpha - 1}}, \qquad \xi = \frac{2}{\sqrt{\alpha - 1}}. \tag{9}$$

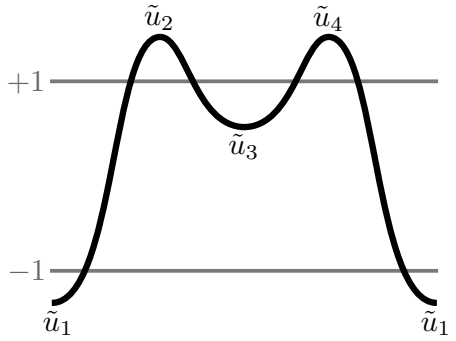

Figure 1: Sketch of a periodic profile $\tilde{u}$ satisfying the geometric properties $\mathcal{H}$.

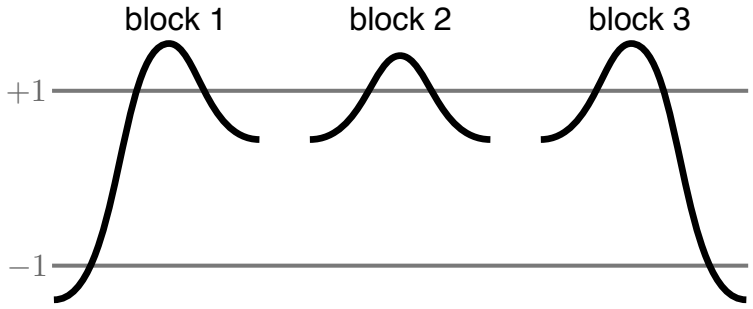

Figure 2: The shape of the building blocks that lead to positive entropy in Theorem 5. It should be intuitive that block 3 can be followed by block 1 only, while blocks 1 and 2 can be followed by either block 2 or block 3.

The parameter range $\alpha \geq 2$ in Proposition 4 corresponds to $\xi \in (0, 2]$. The differential equation now becomes

$$-u'''' - \xi u'' + u - u^3 = 0, \tag{10}$$

with energy

$$E = u'''u' - \frac{1}{2}u''^2 + \frac{\xi}{2}u'^2 + \frac{1}{4}(u^2 - 1)^2. \tag{11}$$

Proposition 4 is proved by showing that chaos is forced by a *single* periodic solution $\tilde{u}$ with the following geometric properties (see also Figure 1):

$$\mathcal{H} \begin{cases} (H_1) & \tilde{u} \text{ has exactly four monotone laps and extrema } \{\tilde{u}_i\}_{i=1}^4; \\ (H_2) & \tilde{u}_1 \text{ and } \tilde{u}_3 \text{ are minima, and } \tilde{u}_2 \text{ and } \tilde{u}_4 \text{ are maxima;} \\ (H_3) & \tilde{u}_1 < -1 < \tilde{u}_3 < 1 < \tilde{u}_2, \tilde{u}_4; \\ (H_4) & \tilde{u} \text{ is symmetric in its minima } \tilde{u}_1 \text{ and } \tilde{u}_3. \end{cases}$$

The forcing theorem is formulated below.

Theorem 5 (forcing; Theorem 2 in [40])**.** *Let $\xi \in [0, \sqrt{8})$, and suppose there exists a periodic solution $\tilde{u}$ of* (10) *at the energy level $E = 0$, satisfying the geometric conditions $\mathcal{H}$. Then* (10) *is chaotic on the energy level $E = 0$ in the sense that there exists a two-dimensional Poincaré return map which has a compact invariant set on which the topological entropy is positive.*

The set of solutions of (10) that implies chaotic dynamics is obtained by combining the three building blocks in Figure 2 in any order that obeys intuitive geometric restrictions. The final technical step is then to establish a semi-conjugacy to a subshift of finite type, see [40, §2].

The solutions to the ODE (10) thus constructed from the three building blocks, correspond to stationary profiles of the PDE (7). It turns out that all these profiles are *stable* under the evolution of the PDE, illustrating the pattern forming tendencies of the Swift-Hohenberg model.

Since Theorem 5 is a forcing theorem, to establish chaos we only need to find, in our case via rigorous numerics, a *single* periodic solution $\tilde{u}$ satisfying $\mathcal{H}$.

Theorem 6 (rigorous computation; Theorem 3 in [40])**.** *For every $\xi \in [0,2]$ the ODE (10) has a periodic solution $\tilde{u}$ at energy level $E = 0$ satisfying the geometric properties $\mathcal{H}$.*

The change of variables (9) directly converts Theorems 5 and 6 into Proposition 4.

In this paper we present a new, streamlined proof of Theorem 6. In particular, we will discuss all details of how to prove that for *fixed* parameter value ξ equation (10) has a periodic solution at energy level $E = 0$, see §4. Since the method provides a *quantitative* description of the solution with a small *explicit* error bound, it is relatively easy to check the geometric properties $\mathcal{H}$. In §4.7 we briefly discuss this issue, and we refer to [40, §4] for more details.

One can also prove the result for a *continuous* range of ξ-values. Indeed, in §4.7 we explain how one may use interval arithmetic computations to perform "brute force" continuation (by extending from single parameter values to small intervals of parameter values). Again we refer to [40] for more details. Moreover, a discussion of proper continuation techniques can be found elsewhere in this volume [22, 42], as well as in [41, 5].

Finally, we will not discuss the proof of the forcing Theorem 5 as it requires methods that are disjoint from the aims of the current paper; we refer the interested reader to [40, §2].

4 The computational proof

We will first convert the problem of finding a periodic solution to Fourier space (§4.1). Then we introduce a convenient norm on the space Fourier coefficients. The brief description in §4.2 is similar to the one in [18]. It diverges from [40] for the purpose of simplifying the exposition. The full problem consists of the ODE (10) and the energy constraint $E = 0$, with the Fourier coefficients and the frequency as variables. In §4.3 we reformulate this as a fixed point problem of the type (2), although viewed from a non-standard perspective, namely situated in a product space (of Fourier coefficients and frequency). In particular, we define finite dimensional projections and the linear operator A. The two radii polynomials for the problem are defined in §4.4. The formulation and proof of Theorem 11, which is a variation on Theorem 3 adapted to the product space setting (with two radii polynomials depending on two radii), are novel. Theorem 12 explains that injectivity of the linear operator A is automatic if one finds a pair of radii for which both radii polynomials are negative. Explicit expressions for the necessary bounds Y and Z are derived in §4.5 and §4.6, respectively. Finally, the Matlab code that clinches the proof is discussed in §4.7.

4.1 Fourier transform

We are going to restrict our attention to symmetric periodic solutions u satisfying $\mathcal{H}$, hence we expand u as a cosine series:

$$u(y) = a_0 + 2\sum_{m=1}^{\infty} a_m \cos(mqy), \tag{12}$$

with $q > 0$ an *a priori unknown* variable ($\frac{2\pi}{q}$ is the period) and $a_m \in \mathbb{R}$ for $m \geq 0$. One may also think of this as

$$u(y) = \sum_{m\in\mathbb{Z}} a_{|m|} e^{imqy}, \tag{13}$$

i.e., the Fourier transform with symmetry constraint $a_{-m} = a_m$.

Since $u'(0) = 0$, and since the energy (11) is a conserved quantity along the orbits of (10), we get that

$$\begin{aligned} E &= u'''(0)u'(0) - \frac{1}{2}u''(0)^2 + \frac{\xi}{2}u'(0)^2 + \frac{1}{4}(u(0)^2-1)^2 \\ &= -\frac{1}{2}\left[u''(0) - \frac{1}{\sqrt{2}}\big(u(0)^2-1\big)\right]\left[u''(0) + \frac{1}{\sqrt{2}}\big(u(0)^2-1\big)\right]. \end{aligned}$$

We look for u such that $E = 0$, $u(0) < -1$ and $u''(0) > 0$, hence the energy condition boils down to

$$u''(0) - \frac{1}{\sqrt{2}}[u(0)^2 - 1] = 0. \tag{14}$$

Substituting the expansion (12) for $u(y)$ in (14), we obtain

$$f_q(q,a) \stackrel{\text{def}}{=} -2q^2 \sum_{m=1}^{\infty} m^2 a_m - \frac{1}{\sqrt{2}} \left[a_0 + 2 \sum_{m=1}^{\infty} a_m \right]^2 + \frac{1}{\sqrt{2}} = 0. \tag{15}$$

In Fourier space the ODE (10) converts into an infinite sequence of algebraic equations:

$$(f_a)_k(q,a) \stackrel{\text{def}}{=} -\lambda_k(q) a_k - \sum_{\substack{k_1+k_2+k_3=k \\ k_i \in \mathbb{Z}}} a_{|k_1|} a_{|k_2|} a_{|k_3|} = 0 \qquad \text{for all } k \geq 0, \tag{16}$$

where

$$\lambda_k(q) \stackrel{\text{def}}{=} k^4 q^4 - \xi k^2 q^2 - 1 = (k^2 q^2 - \xi/2)^2 - (1 + \xi^2/4).$$

Equations (15) and (16) together constitute $F = (f_q, f_a)$.

4.2 The norm on the space of Fourier coefficients

Since the differential equation is analytic, the solution is going to be analytic as well and the coefficients will decay geometrically. This motivates us to work with the norm

$$\|a\|_{\ell^1_\nu} \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} |a_k| \, \omega_k(\nu),$$

where the exponential weights are

$$\omega_k(\nu) = \omega_k \stackrel{\text{def}}{=} \begin{cases} 1 & k = 0 \\ 2\nu^k & k \geq 1, \end{cases} \tag{17}$$

for some $\nu > 1$ to be chosen when doing the final computer-assisted step of the proof (we suppress the dependence on ν in the notation for ω_k). The factor 2 in (17) is related to our choice of cosine series; it facilitates the simple expression for the estimate (18) below. We denote the corresponding Banach space by ℓ^1_ν. In view of (13), this norm may also be written as

$$\|a\|_{\ell^1_\nu} = \sum_{k \in \mathbb{Z}} |a_{|k|}| \, \nu^{|k|}.$$

We endow the one-sided ℓ^1_ν with the following discrete convolution product: given $a, b \in \ell^1_\nu$ the convolution product $a * b \in \ell^1_\nu$ has components

$$(a * b)_k \stackrel{\text{def}}{=} \sum_{k' \in \mathbb{Z}} a_{|k'|} b_{|k-k'|}.$$

This product satisfies the Banach algebra property

$$\|a * b\|_{\ell^1_\nu} \leq \|a\|_{\ell^1_\nu} \|b\|_{\ell^1_\nu}, \tag{18}$$

as can be checked directly (applying the triangle inequality):

$$\begin{aligned} \|a * b\|_{\ell^1_\nu} &= \sum_{k \in \mathbb{Z}} |(a*b)_{|k|}| \nu^{|k|} \leq \sum_{k,k' \in \mathbb{Z}} |a_{|k'|}| |b_{|k-k'|}| \nu^{|k|} \\ &\leq \sum_{k,k-k' \in \mathbb{Z}} |a_{|k'|}| \nu^{|k'|} |b_{|k-k'|}| \nu^{|k-k'|} = \|a\|_{\ell^1_\nu} \|b\|_{\ell^1_\nu}. \end{aligned}$$

In this notation the expression for $(f_a)_k$ in (16) reduces to

$$(f_a)_k = -\lambda_k(q) - (a * a * a)_k.$$

The dual of ℓ^1_ν is $(\ell^1_\nu)^* \cong \ell^\infty_{\nu^{-1}} = \{(b_m)_{m=0}^\infty : \sup_{m\in\mathbb{N}} |b_m\omega_m^{-1}| < \infty\}$. Namely, for $b \in (\ell^1_\nu)^*$ we may use linearity to write

$$b(a) = \sum_{m\in\mathbb{N}} b(\delta^m)a_m,$$

where δ is the usual Knonecker delta:

$$\delta_k^m \stackrel{\text{def}}{=} \begin{cases} 1 & k = m \\ 0 & k \neq m. \end{cases}$$

Hence, writing $b_m \stackrel{\text{def}}{=} b(\delta^m)$ we obtain

$$|b(a)| = \left|\sum_{m=0}^\infty b_m a_m\right| = \left|\sum_{m=0}^\infty (b_m\omega_m^{-1})(a_m\omega_m)\right| \leq \sup_{m\in\mathbb{N}} |b_m\omega_m^{-1}| \sum_{m\in\mathbb{N}} |a_m|\omega_m.$$

We will identify $(\ell^1_\nu)^*$ with $l^\infty_{\nu^{-1}}$ and use

$$\|b\|_{(\ell^1_\nu)^*} \stackrel{\text{def}}{=} \sup_{m\in\mathbb{N}} |b(\delta^m)|\omega_m^{-1} = \sup_{m\in\mathbb{N}} |b_m|\omega_m^{-1}$$

Remark 7. *If $G \in B(\ell^1_\nu, \ell^1_\nu)$, then by linearity*

$$\begin{aligned}\|G(a)\|_{\ell^1_\nu} &= \sum_{k\in\mathbb{N}} |G(a)_k|\omega_k = \sum_{k,m\in\mathbb{N}} |G(\delta^m)_k a_m|\omega_k \\ &\leq \sum_{m\in\mathbb{N}} |a_m|\omega_m \sup_{m\in\mathbb{N}}\Big(\omega_m^{-1}\sum_{k\in\mathbb{N}} |G(\delta^m)_k|\omega_k\Big) = \|a\|_{\ell^1_\nu} \sup_{m\in\mathbb{N}} \omega_m^{-1}\|G(\delta^m)\|_{\ell^1_\nu}.\end{aligned}$$

Hence, writing $G_{km} \stackrel{\text{def}}{=} G(\delta^m)_k$, we obtain the explicit expression

$$\|G\|_{B(\ell^1_\nu,\ell^1_\nu)} = \sup_{m\in\mathbb{N}} \omega_m^{-1} \sum_{k\in\mathbb{N}} |G_{km}|\,\omega_k.$$

4.3 The product space

Since our unknowns are $x = (q, a)$, we use the product space $X = \mathbb{R} \times \ell^1_\nu$. We will use projections π_q and π_a onto $\mathbb{R}$ and ℓ^1_ν, respectively. The trivial embedding of $\mathbb{R}$ and ℓ^1_ν into X are both denoted by ι, since this will never cause confusion.

We fix a computational parameter $N \in \mathbb{N}$ (to be chosen when we do the final computer-assisted step of the proof, see §4.7). Given $x = (q, a) \in X$, we define the finite dimensional projection $\pi_N x = (q, a_0, \ldots, a_N) \in X_N \cong \mathbb{R} \times \mathbb{R}^{N+1} \cong \mathbb{R}^{N+2}$. We slightly abuse notation to write π_q and π_a for the projections from X_N onto $\mathbb{R}$ and $\mathbb{R}^{N+1}$, respectively: $\pi_q x_N = q$ and $\pi_a x_N = (a_k)_{k=0}^N$. The embedding of X_N into X by extending with zeros is again denoted by ι:

$$\pi_N \iota x_N = x_N \qquad \text{and} \qquad (\pi_a \iota x_N)_k = 0 \quad \text{for } k > N.$$

The truncated system, or Galerkin projection, is given by

$$F_N : X_N \to X_N, \qquad F_N(x_N) = \pi_N F(\iota x_N).$$

We solve $F_N = 0$ numerically (using a good initial guess and Newton's iteration method) to obtain an approximate zero $\overline{x}_N = (\overline{q}, \overline{a}_0, \ldots, \overline{a}_N)$ of F_N, and a corresponding approximate zero $\overline{x} = \iota\overline{x}_N$ of F. The Jacobian matrix

$$A_N^\dagger \stackrel{\text{def}}{=} DF_N(\overline{x}_N)$$

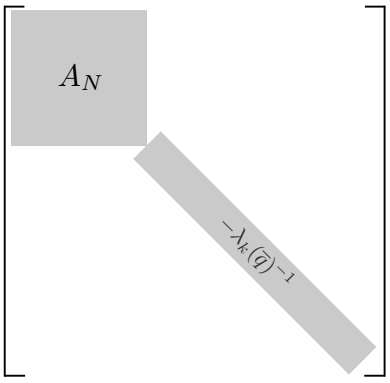

Figure 3: The shape of the linear operator A interpreted as an infinite matrix. The finite matrix A_N has size $(N+2)\times(N+2)$. The shape of $A^\dagger$ is identical (but of course with different values for the nonzero entries).

is inverted *numerically* to obtain

$$A_N \approx (A_N^\dagger)^{-1}.$$

Based on our expectation that the linear part in (16) will dominate for large k, the linear operators $A^\dagger$ and A are then built up as follows:

$$\pi_N(A^\dagger x) = A_N^\dagger \pi_N x \quad \text{and} \quad (\pi_a(A^\dagger x))_k = -\lambda_k(\overline{q})(\pi_a x)_k \qquad \text{for } k > N \tag{19a}$$

$$\pi_N(Ax) = A_N \pi_N x \quad \text{and} \quad (\pi_a(Ax))_k = -\lambda_k(\overline{q})^{-1}(\pi_a x)_k \quad \text{for } k > N. \tag{19b}$$

The shape of the operator A (and $A^\dagger$) as an "infinite matrix" is illustrated in Figure 3. We are now ready to define

$$T(x) \stackrel{\text{def}}{=} x - AF(x).$$

Remark 8. *Let Γ be a linear operator on X, then it can be decomposed into*

$$\Gamma = \begin{bmatrix} \Gamma_{qq} & \Gamma_{qa} \\ \Gamma_{aq} & \Gamma_{aa} \end{bmatrix}$$

with $\Gamma_{qq} \in \mathbb{R}$, $\Gamma_{qa} \in (\ell^1_\nu)^*$, $\Gamma_{aq} \in \ell^1_\nu$ and $\Gamma_{aa} \in B(\ell^1_\nu, \ell^1_\nu)$. Formally:

$$\pi_q \Gamma x = \Gamma_{qq}\pi_q x + \Gamma_{qa}\pi_a x$$
$$\pi_a \Gamma x = \Gamma_{aq}\pi_q x + \Gamma_{aa}\pi_a x.$$

We will use the same notation for bounded linear operators on X_N.

Remark 9. *Moreover, suppose $\Gamma \in L(X,X)$ is of the form*

$$\pi_N \Gamma x = \Gamma_N \pi_N x \qquad \text{and} \qquad (\pi_a \Gamma x)_k = \gamma_k x_k \quad \text{for } k > N,$$

with $\sup_{k>N}|\gamma_k| = \hat{\gamma} < \infty$ given. Then $(\Gamma_{qa})_k$ and $(\Gamma_{aq})_k$ vanish for $k > N$, hence their norms are computable, and

$$\|\Gamma_{aa}\|_{B(\ell^1_\nu,\ell^1_\nu)} = \max\left\{\hat{\gamma}\,,\ \max_{0\le m\le N}\frac{1}{\omega_m}\sum_{k=0}^{N}|(\Gamma_{aa})_{km}|\omega_k\right\}$$

is also computable.

Remark 10. *We choose the truncation dimension $N \geq \widehat{N}(\overline{q}, \xi)$, where*

$$\widehat{N}(\overline{q}, \xi) \stackrel{\text{def}}{=} \frac{((1+\xi^2/4)^{1/2} + \xi/2)^{1/2}}{\overline{q}},$$

so that

$$0 < \lambda_{N+1}(\overline{q}) \leq \lambda_k(\overline{q}) \quad \text{for all } k \geq N+1. \tag{20}$$

This ensures both that the factor $\lambda_k(\overline{q})^{-1}$ in (19b) is well-defined for all $k > N$, and that A is of the form discussed in Remark 9 with $\hat{\gamma} = \lambda_{N+1}(\overline{q})^{-1}$.

4.4 The radii polynomials

Since we (decide to) work on a product space, we are going to deviate from the setup in §2.3 and [40]. For $r = (r_1, r_2)$ with $r_1, r_2 > 0$ we define the *rectangle*

$$\mathcal{B}_r(\overline{x}) = \{x \in X : |\pi_q(x - \overline{x})| \leq r_1, \|\pi_a(x - \overline{x})\|_{\ell^1_\nu} \leq r_2\}.$$

We will establish bounds (with D_q and D_a denoting partial derivatives)

$$|\pi_q[T(\overline{x}) - \overline{x}]| \leq Y_1 \tag{21a}$$
$$\|\pi_a[T(\overline{x}) - \overline{x}]\|_{\ell^1_\nu} \leq Y_2 \tag{21b}$$
$$\|D_q \pi_q T(x)\|_{B(\mathbb{R},\mathbb{R})} \leq Z_{11}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}), \tag{21c}$$
$$\|D_a \pi_q T(x)\|_{B(\ell^1_\nu,\mathbb{R})} \leq Z_{12}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}), \tag{21d}$$
$$\|D_q \pi_a T(x)\|_{B(\mathbb{R},\ell^1_\nu)} \leq Z_{21}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}), \tag{21e}$$
$$\|D_a \pi_a T(x)\|_{B(\ell^1_\nu,\ell^1_\nu)} \leq Z_{22}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}), \tag{21f}$$

for some $Y_1, Y_2 \in \mathbb{R}^+$ and $Z_{ij} : (\mathbb{R}^+)^2 \to \mathbb{R}^+$ for $i, j \in \{1, 2\}$. Clearly in (21c)-(21e) one should read $B(\mathbb{R}, \mathbb{R}) = \mathbb{R}$, $B(\ell^1_\nu, \mathbb{R}) = (\ell^1_\nu)^*$ and $B(\mathbb{R}, \ell^1_\nu) = \ell^1_\nu$. Then the *two radii polynomials* for our problem are

$$p_1(r_1, r_2) = Y_1 + r_1[Z_{11}(r_1, r_2) - 1] + r_2 Z_{12}(r_1, r_2) \tag{22a}$$
$$p_2(r_1, r_2) = Y_2 + r_1 Z_{21}(r_1, r_2) + r_2[Z_{22}(r_1, r_2) - 1]. \tag{22b}$$

We need to find an $\hat{r} = (\hat{r}_1, \hat{r}_2) > 0$ such that $p(\hat{r}) < 0$ component-wise, i.e.,

$$p_1(\hat{r}_1, \hat{r}_2) < 0 \qquad \text{and} \qquad p_2(\hat{r}_1, \hat{r}_2) < 0. \tag{23}$$

The geometry of these sublevel sets is illustrated in Figure 4.

Theorem 11. *Assume that Y_i and Z_{ij} satisfy the bounds (21) for $i, j \in \{1, 2\}$. Let p_1 and p_2 be the radii polynomials defined in (22). If $\hat{r}_1, \hat{r}_2 > 0$ are such that the inequalities (23) are satisfied, then T has a unique fixed point in $\mathcal{B}_{\hat{r}}(\overline{x})$.*

Proof. To show that T maps $\mathcal{B}_{\hat{r}}(\overline{x})$ into itself, we use arguments similar to the ones in §2.5. To simplify notation, define the projected balls

$$B^q \stackrel{\text{def}}{=} \{q \in \mathbb{R} : |q - \overline{q}| \leq \hat{r}_1\} \qquad \text{and} \qquad B^a \stackrel{\text{def}}{=} \{a \in \ell^1_\nu : \|a - \overline{a}\|_{\ell^1_\nu} \leq \hat{r}_2\},$$

so that $\mathcal{B}_{\hat{r}}(\overline{x}) = B^q \times B^a$. Let $x = (q, a) \in \mathcal{B}_{\hat{r}}(\overline{q}, \overline{a})$. Then, by applying the triangle inequality and the intermediate value theorem, we obtain

$$\begin{aligned}
\left|\pi_q T(q,a) - \overline{q}\right| &\leq \left|\pi_q[T(q,a) - T(q,\overline{a})]\right| \\
&\qquad + \left|\pi_q[T(q,\overline{a}) - T(\overline{q},\overline{a})]\right| + \left|\pi_q T(\overline{q},\overline{a}) - \overline{q}\right| \\
&\leq \sup_{a' \in B^a} \|D_a \pi_q T(q,a')\|_{(\ell^1_\nu)^*} \|a' - \overline{a}\|_{\ell^1_\nu} \\
&\qquad + \sup_{q' \in B^q} |D_q \pi_q T(q',\overline{a})| \, |q' - q| \; + \; Y_1 \\
&\leq Z_{12}(\hat{r})\hat{r}_2 + Z_{11}(\hat{r})\hat{r}_1 + Y_1.
\end{aligned}$$

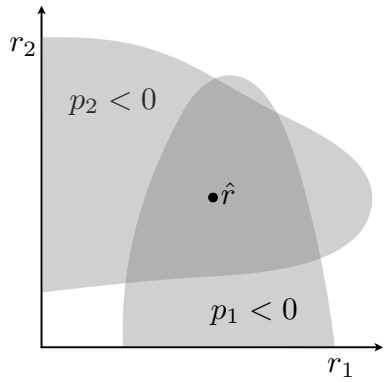

Figure 4: An illustration of a nonempty intersection of $\{p_1(r_1, r_2) < 0\}$ and $\{p_2(r_1, r_2) < 0\}$. Note that p_1 is monotone in r_2, while p_2 is monotone in r_1. Hence the set $\{p_1 = 0\}$ is a graph over r_1, while the set $\{p_2 = 0\}$ is a graph over r_2.

Since $p_1(\hat{r}) < 0$ by assumption, it follows that $|\pi_q T(q, a) - \overline{q}| < \hat{r}_1$. By an analogous argument, $\|\pi_a T(q, a) - \overline{a}\|_{\ell^1_\nu} < \hat{r}_2$. Hence T maps $\mathcal{B}_{\hat{r}}(\overline{x})$ into itself.

To show that T is a contraction mapping on $\mathcal{B}_{\hat{r}}(\overline{x})$ we need to choose a norm on the product space $X = \mathbb{R} \times \ell^1_\nu$. It turns out that T contracts with respect to the *weighted norm*

$$\|(q, a)\|_X \stackrel{\text{def}}{=} \max\{|q|/\hat{r}_1, \|a\|_{\ell^1_\nu}/\hat{r}_2\}. \tag{24}$$

Indeed, let $x_1, x_2 \in B_{\hat{r}}(\overline{x})$, then, by applying the mean value theorem and the triangle inequality,

$$\begin{aligned} |\pi_q[T(x_1) - T(x_2)]| &\leq \sup_{x' \in B_{\hat{r}}(\overline{x})} |D_q \pi_q T(x')| \, |\pi_q(x_1 - x_2)| \\ &\qquad + \sup_{x' \in B_{\hat{r}}(\overline{x})} \|D_a \pi_q T(x')\|_{(\ell^1_\nu)^*} \|\pi_a(x_1 - x_2)\|_{\ell^1_\nu} \\ &\leq [Z_{11}(\hat{r})\hat{r}_1 + Z_{12}(\hat{r})\hat{r}_2] \, \|x_1 - x_2\|_X, \end{aligned}$$

and analogously

$$\|\pi_a[T(x_1) - T(x_2)]\|_{\ell^1_\nu} \leq [Z_{21}(\hat{r})\hat{r}_1 + Z_{22}(\hat{r})\hat{r}_2] \, \|x_1 - x_2\|_X.$$

From the above estimates we conclude that

$$\|T(x_1) - T(x_2)\|_X \leq \max\big\{Z_{11}(\hat{r}) + Z_{12}(\hat{r})\hat{r}_2/\hat{r}_1, Z_{21}(\hat{r})\hat{r}_1/\hat{r}_2 + Z_{22}(\hat{r})\big\}\|x_1 - x_2\|_X. \tag{25}$$

It follows from the negativity of both radii polynomials (22) that the contraction rate

$$\max\big\{Z_{11}(\hat{r}) + Z_{12}(\hat{r})\hat{r}_2/\hat{r}_1, Z_{21}(\hat{r})\hat{r}_1/\hat{r}_2 + Z_{22}(\hat{r})\big\} \tag{26}$$

is less than unity. □

Theorem 12. *Suppose the assumptions of Theorem 11 are met. Then the unique fixed point $\hat{x} = (\hat{q}, \hat{a})$ obtained in Theorem 11 corresponds to a zero of F, hence $\hat{a}$ corresponds, via* (12)*, to a $2\pi/\hat{q}$-periodic solution of* (10) *with energy $E = 0$.*

Proof. The only thing left is to prove is injectivity of A, which is *automatic* from the negativity of the radii polynomials. Indeed, since $\lambda_k(\overline{q}) \neq 0$ for $k > N$ by (20), injectivity of A is equivalent to injectivity of A_N. The latter could be checked numerically, but that is not needed, since by the arguments in the proof of Theorem 11, negativity of the radii polynomials for $r = \hat{r}$ implies that $\|(I - ADF(\overline{x}))v\|_X < \|v\|_X$ for all $v \in X$, where the norm on X is given by (24). Namely, it follows from (25) that the operator norm $\|I - ADF(\overline{x})\|_{B(X,X)}$ is bounded by the expression (26), which is less than unity. Specializing to $v = \iota v_N$ one finds, for the induced finite dimensional norm $\|x_N\|_{X_N} = \|\iota x_N\|_X$, that $\|(I_N - A_N A_N^\dagger)v_N\|_{X_N} < \|v_N\|_{X_N}$ for all $v_N \in X_N$. This implies that A_N is invertible, hence A is injective. □

4.5 The estimates Y

Recalling (21a) and (21b), in this section we establish the bounds

$$|\pi_q A F(\overline{x})| \le Y_1,$$
$$\|\pi_q A F(\overline{x})\|_{\ell^1_\nu} \le Y_2.$$

We have

$$f_q(\overline{x}) = -2\overline{q}^2 \sum_{m=1}^{N} m^2 \overline{a}_m - \frac{1}{\sqrt{2}} \left[\overline{a}_0 + 2 \sum_{m=1}^{N} \overline{a}_m \right]^2 + \frac{1}{\sqrt{2}} = 0,$$

$$(f_a)_k(\overline{x}) = \begin{cases} -\lambda_k(\overline{q})\overline{a}_k - (\overline{a} * \overline{a} * \overline{a})_k & k = 0, \dots, N \\ -(\overline{a} * \overline{a} * \overline{a})_k & k = N+1, \dots, 3N \\ 0 & k > 3N. \end{cases}$$

There are thus only finitely many non-vanishing terms. Hence we set

$$Y_1 = \left| \pi_q A_N \pi_N F(\overline{x}) \right|,$$
$$Y_2 = \sum_{k=0}^{N} \left| (\pi_a A_N \pi_N F(\overline{x}))_k \right| \omega_k + \sum_{k=N+1}^{3N} \lambda_k(\overline{q})^{-1} |(\overline{a} * \overline{a} * \overline{a})_k| \omega_k,$$

computed using interval arithmetic to obtain rigorous upper bounds.

4.6 The estimates Z

In this section we construct bounds, see (21),

$$|D_q \pi_q T(x)| \le Z_{11}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}), \tag{27a}$$
$$\|D_a \pi_q T(x)\|_{(\ell^1_\nu)^*} \le Z_{12}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}), \tag{27b}$$
$$\|D_q \pi_a T(x)\|_{\ell^1_\nu} \le Z_{21}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}), \tag{27c}$$
$$\|D_a \pi_a T(x)\|_{B(\ell^1_\nu, \ell^1_\nu)} \le Z_{22}(r) \qquad \text{for all } x \in \mathcal{B}_r(\overline{x}). \tag{27d}$$

We recall that

$$\mathcal{B}_r(0) = \{(q,a) \in X : |q| \le r_1, \|a\|_{\ell^1_\nu} \le r_2\}.$$

As already mentioned in §2.4, for each of the four derivatives in (27), we split $DT(\overline{x}+w)v$, where $w = (w_q, w_a) \in \mathcal{B}_r(0)$ and $v = (v_q, v_a) \in \mathcal{B}_{1,1}(0)$ are arbitrary, into three pieces:

$$DT(\overline{x}+w)v = [I - AA^\dagger]v - A[DF(\overline{x}) - A^\dagger]v - A[DF(\overline{x}+w) - DF(\overline{x})]v. \tag{28}$$

By bounding these three terms separately, we will find bounds $Z_{ij}(r) = Z^0_{ij} + Z^1_{ij} + Z^2_{ij}(r)$, $i,j \in \{1,2\}$. We note that the first two terms are independent of w, hence only the third term depends on $r = (r_1, r_2)$.

4.6.1 The bounds Z^0

We infer from the definitions of A and $A^\dagger$ in (19) that the first term in (28) reduces to a finite dimensional operator:

$$[I - AA^\dagger]v = \iota[I_N - A_N A_N^\dagger]\pi_N v.$$

Let us write $\Gamma = I - AA^\dagger$, which is of the form discussed in Remark 9 with $\hat{\gamma} = 0$. With the notation introduced in Remark 8 it follows from Remark 9 that the computable numbers

$$Z^0_{11} = |\Gamma_{qq}|$$
$$Z^0_{12} = \|\Gamma_{qa}\|_{(\ell^1_\nu)^*}$$
$$Z^0_{21} = \|\Gamma_{aq}\|_{\ell^1_\nu}$$
$$Z^0_{22} = \|\Gamma_{aa}\|_{B(\ell^1_\nu, \ell^1_\nu)},$$

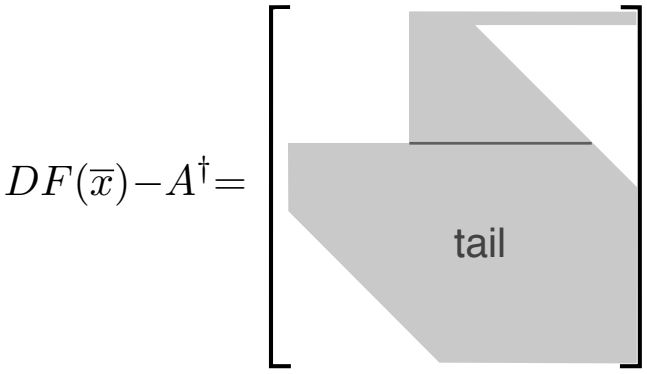

Figure 5: A sketch of the shape of the linear operator $DF(\overline{x}) - A^\dagger$ interpreted as an infinite matrix. Note that the first column vanishes, whereas the top row has infinitely many nonzero elements. The size of the block of zeros at the top left is $(N+2) \times (N+2)$. The width of the infinite non-vanishing "diagonal" strip (the "tail") is $4N+1$.

bound the respective components of $I - AA^\dagger$.

4.6.2 The bounds Z^1

For the second term in (28) we first compute, recalling that $A_N^\dagger = DF_N(\pi_N \overline{x})$,

$$[Df_q(\overline{x}) - \pi_q A^\dagger]v = -2\overline{q}^2 \sum_{m=N+1}^{\infty} m^2 (v_a)_m - \sqrt{2}\Big(\overline{a}_0 + 2\sum_{m=1}^{N} \overline{a}_m\Big)\Big(2\sum_{m=N+1}^{\infty} (v_a)_m\Big) \tag{29a}$$

and

$$([Df_a(\overline{x}) - \pi_a A^\dagger]v)_k = \begin{cases} -3\,(\overline{a} * \overline{a} * v_a^0)_k & \text{for } 0 \le k \le N \\ -3\,(\overline{a} * \overline{a} * v_a)_k & \text{for } k > N, \end{cases} \tag{29b}$$

where

$$(v_a^0)_k \stackrel{\text{def}}{=} \begin{cases} 0 & 0 \le k \le N \\ (v_a)_k & k > N. \end{cases}$$

Figure 5 illustrates the nonzero elements in the linear operator $Df(\overline{x}) - A^\dagger$ interpreted as an "infinite matrix". We note that both right-hand sides in (29) are independent of v_q, which implies that $D_q f(\overline{x})\pi_q - A^\dagger \iota \pi_q = 0$, hence

$$Z^1_{11} = 0 \qquad \text{and} \qquad Z^1_{21} = 0.$$

We then estimate for all $v \in \mathcal{B}_{1,1}(0)$, using the characterization of $(\ell^1_\nu)^*$ in §4.2,

$$\begin{aligned} \Big|\sum_{m=N+1}^{\infty} (v_a)_m\Big| &\le \sup_{m>N} \omega_m^{-1} \le \omega_{N+1}^{-1} \\ \Big|\sum_{m=N+1}^{\infty} m^2 (v_a)_m\Big| &\le \sup_{m>N} m^2 \omega_m^{-1} \le \tfrac{1}{2} Q(\nu, N), \end{aligned}$$

where

$$Q(\nu, N) \stackrel{\text{def}}{=} \begin{cases} \frac{4}{(e \ln \nu)^2} & N+1 \le 2/\ln(\nu) \\ 2(N+1)^2 \omega_{N+1}^{-1} & N+1 > 2/\ln(\nu). \end{cases}$$

We immediately see that ν should not be chosen too close to 1 as this would lead to a very large value for Q. We define bounds $d_N = (d_q, d_{a0}, \ldots, d_{aN})$ as follows:

$$d_q \stackrel{\text{def}}{=} \overline{q}^2 Q(\nu, N) + \sqrt{2}\,\omega_{N+1}^{-1} \left| \overline{a}_0 + 2\sum_{m=1}^{N} \overline{a}_m \right| \geq |[Df_q(\overline{x}) - \pi_q A^\dagger]v| \quad \text{for all } v \in \mathcal{B}_{1,1}(0),$$

and for $k = 0, \ldots, N$

$$d_{ak} \stackrel{\text{def}}{=} 3(|\overline{a} * \overline{a}| * \chi)_k \geq |([Df_a(\overline{x}) - \pi_a A^\dagger]v)_k| \quad \text{for all } v \in \mathcal{B}_{1,1}(0),$$

where

$$\chi_k = \begin{cases} 0 & 0 \leq k \leq N \\ \omega_k^{-1} & N+1 \leq k \leq 3N \\ 0 & k > 3N, \end{cases}$$

and where absolute values in $|\overline{a} * \overline{a}|$ are taken component-wise. For $k > N$ we note that

$$|(\pi_a A[Df(\overline{x}) - A^\dagger]v)_k| \leq \lambda_k(\overline{q})^{-1} |3(\overline{a} * \overline{a} * v_a)_k|,$$

hence we may estimate, by using (18) and (20),

$$\sum_{k=N+1}^{\infty} |(\pi_a A[Df(\overline{x}) - A^\dagger]v)_k| \omega_k \leq 3\lambda_{N+1}(\overline{q})^{-1} \|\overline{a} * \overline{a}\|_{\ell^1_\nu} \qquad \text{for all } v \in \mathcal{B}_{1,1}(0),$$

provided $N \geq \widehat{N}(\overline{q}, \xi)$. We thus set

$$\begin{aligned} Z_{21}^1 &= \pi_q |A_N| d_N \\ Z_{22}^1 &= \|\pi_a |A_N| d_N\|_{\ell^1_\nu} + 3\lambda_{N+1}(\overline{q})^{-1} \|\overline{a} * \overline{a}\|_{\ell^1_\nu}, \end{aligned}$$

where absolute values in the matrix $|A_N|$ are taken entry-wise.

4.6.3 The bounds Z^2

The final term in (28) is expanded in powers of r_1 and r_2 by writing $w = (r_1 \widetilde{w}_q, r_2 \widetilde{w}_a)$ with $\widetilde{w} \in \mathcal{B}_{1,1}(0)$:

$$[Df_q(\overline{x} + w) - Df_q(\overline{x})]v = v_q \sum_{i,j} c_q^{ij}(\overline{x}, \widetilde{w}) r_1^i r_2^j + \sum_{i,j} \widetilde{c}_q^{ij}(\overline{x}, v_a, \widetilde{w}) r_1^i r_2^j, \tag{30a}$$

$$\big([Df_a(\overline{x} + w) - Df_a(\overline{x})]v\big)_k = v_q \sum_{i,j} c_{ak}^{ij}(\overline{x}, \widetilde{w}) r_1^i r_2^j + \sum_{i,j} \widetilde{c}_{ak}^{ij}(\overline{x}, v_a, \widetilde{w}) r_1^i r_2^j, \tag{30b}$$

where we write $(c_a^{ij})_k = c_{ak}^{ij}$ for convenience. All sums are finite sums; the non-vanishing coefficients c_q^{ij} and c_{ak}^{ij} are listed in Table 1.

We now compute uniform bounds

$$C_q^{ij}(\overline{x}) \geq |c_q^{ij}(\overline{x}, \widetilde{w})| \qquad \text{for all } \widetilde{w} \in \mathcal{B}_{1,1}(0), \tag{31a}$$

$$\widetilde{C}_q^{ij}(\overline{x}) \geq |\widetilde{c}_q^{ij}(\overline{x}, v_a, \widetilde{w})| \qquad \text{for all } \widetilde{w} \in \mathcal{B}_{1,1}(0), \|v_a\|_{\ell^1_\nu} \leq 1, \tag{31b}$$

$$C_{ak}^{ij}(\overline{x}) \geq |c_{ak}^{ij}(\overline{x}, \widetilde{w})| \qquad \text{for all } \widetilde{w} \in \mathcal{B}_{1,1}(0), \quad 0 \leq k \leq N, \tag{31c}$$

$$\widetilde{C}_{ak}^{ij}(\overline{x}) \geq |\widetilde{c}_{ak}^{ij}(\overline{x}, v_a, \widetilde{w})| \qquad \text{for all } \widetilde{w} \in \mathcal{B}_{1,1}(0), \|v_a\|_{\ell^1_\nu} \leq 1, \quad 0 \leq k \leq N. \tag{31d}$$

These are summarized in Table 2, using the notation

$$Q_\nu \stackrel{\text{def}}{=} \frac{4}{(e \ln \nu)^2} \geq 2 \sum_{m \in \mathbb{N}} m^2 |(v_a)_m| \qquad \text{for all } \|v_a\|_{\ell^1_\nu} \leq 1. \tag{32}$$

c_q^{10}	$-4\widetilde{w}_q \sum_{m=1}^{N} m^2 \overline{a}_m$	$\widetilde{c}_q^{10}$	$-4\overline{q}\widetilde{w}_q \sum_{m=1}^{\infty} m^2 (v_a)_m$
c_q^{01}	$-4\overline{q} \sum_{m=1}^{\infty} m^2 (\widetilde{w}_a)_m$	$\widetilde{c}_q^{01}$	$-\sqrt{2}\left((\widetilde{w}_a)_0 + 2\sum_{m=1}^{\infty}(\widetilde{w}_a)_m\right)\left((v_a)_0 + 2\sum_{m=1}^{\infty}(v_a)_m\right)$
		$\widetilde{c}_q^{20}$	$-2\widetilde{w}_q^2 \sum_{m=1}^{\infty} m^2 (v_a)_m$
c_q^{11}	$-4\widetilde{w}_q \sum_{m=1}^{\infty} m^2 (\widetilde{w}_a)_m$		
c_{ak}^{10}	$-2k^2\widetilde{w}_q\big(6k^2\overline{q}^2 - \xi\big)\overline{a}_k$	$\widetilde{c}_{ak}^{10}$	$-2k^2\widetilde{w}_q\big(2k^2\overline{q}^3 - \xi\overline{q}\big)(v_a)_k$
c_{ak}^{01}	$-2k^2\big(2k^2\overline{q}^3 - \xi\overline{q}\big)(\widetilde{w}_a)_k$	$\widetilde{c}_{ak}^{01}$	$-6\,(\overline{a} * \widetilde{w}_a * v_a)_k$
c_{ak}^{20}	$-12k^4\widetilde{w}_q^2\overline{q}\overline{a}_k$	$\widetilde{c}_{ak}^{20}$	$-k^2\widetilde{w}_q^2\big(6k^2\overline{q}^2 - \xi\big)(v_a)_k$
c_{ak}^{11}	$-2k^2\widetilde{w}_q\big(6k^2\overline{q}^2 - \xi\big)(\widetilde{w}_a)_k$		
		$\widetilde{c}_{ak}^{02}$	$-3\,(\widetilde{w}_a * \widetilde{w}_a * v_a)_k$
c_{ak}^{30}	$-4k^4\widetilde{w}_q^3\overline{a}_k$	$\widetilde{c}_{ak}^{30}$	$-4k^4\widetilde{w}_q^3\overline{q}(v_a)_k$
c_{ak}^{21}	$-12k^4\widetilde{w}_q^2\overline{q}(\widetilde{w}_a)_k$		
		$\widetilde{c}_{ak}^{40}$	$-k^4\widetilde{w}_q^4(v_a)_k$
c_{ak}^{31}	$-4k^4\widetilde{w}_q^3(\widetilde{w}_a)_k$		

Table 1: The non-zero coefficients in the expansions (30).

C_q^{10}	$4\big\lvert\sum_{m=1}^{N} m^2\overline{a}_m\big\rvert$	$\widetilde{C}_q^{10}$	$2\overline{q}Q_\nu$
C_q^{01}	$2\overline{q}Q_\nu$	$\widetilde{C}_q^{01}$	$\sqrt{2}$
		$\widetilde{C}_q^{20}$	Q_ν
C_q^{11}	$2Q_\nu$		
C_{ak}^{10}	$2k^2\big\lvert 6k^2\overline{q}^2 - \xi\big\rvert\lvert\overline{a}_k\rvert$	$\widetilde{C}_{ak}^{10}$	$2k^2\big\lvert 2k^2\overline{q}^3 - \xi\overline{q}\big\rvert\omega_k^{-1}$
C_{ak}^{01}	$2k^2\big\lvert 2k^2\overline{q}^3 - \xi\overline{q}\big\rvert\omega_k^{-1}$	$\widetilde{C}_{ak}^{01}$	$6\lVert\overline{a}\rVert_{\ell_\nu^1}\omega_k^{-1}$
C_{ak}^{20}	$12k^4\overline{q}\lvert\overline{a}_k\rvert$	$\widetilde{C}_{ak}^{20}$	$k^2\big\lvert 6k^2\overline{q}^2 - \xi\big\rvert\omega_k^{-1}$
C_{ak}^{11}	$2k^2\big\lvert 6k^2\overline{q}^2 - \xi\big\rvert\omega_k^{-1}$		
		$\widetilde{C}_{ak}^{02}$	$3\omega_k^{-1}$
C_{ak}^{30}	$4k^4\lvert\overline{a}_k\rvert$	$\widetilde{C}_{ak}^{30}$	$4k^4\overline{q}\omega_k^{-1}$
C_{ak}^{21}	$12k^4\overline{q}\omega_k^{-1}$		
		$\widetilde{C}_{ak}^{40}$	$k^4\omega_k^{-1}$
C_{ak}^{31}	$4k^4\omega_k^{-1}$		

Table 2: The uniform bounds $C(\overline{x})$ and $\widetilde{C}(\overline{x})$ on the coefficients $c(\overline{x}, \widetilde{w})$ and $\widetilde{c}(\overline{x}, v_a, \widetilde{w})$, see (31).

		$\widetilde{C}^{10}_{\text{tail}}$	$4\overline{q}^3\mu_{N+1}$
C^{01}_{tail}	$4\overline{q}^3\mu_{N+1}$	$\widetilde{C}^{01}_{\text{tail}}$	$6\lambda_{N+1}(\overline{q})^{-1}\|\overline{a}\|_{\ell^1_\nu}$
		$\widetilde{C}^{20}_{\text{tail}}$	$6\overline{q}^2\mu_{N+1}$
C^{11}_{tail}	$12\overline{q}^2\mu_{N+1}$		
		$\widetilde{C}^{02}_{\text{tail}}$	$3\lambda_{N+1}(\overline{q})^{-1}$
		$\widetilde{C}^{30}_{\text{tail}}$	$4\overline{q}\mu_{N+1}$
C^{21}_{tail}	$12\overline{q}\mu_{N+1}$		
		$\widetilde{C}^{40}_{\text{tail}}$	μ_{N+1}
C^{31}_{tail}	$4\mu_{N+1}$		

Table 3: The uniform norm bounds $C^{ij}_{\text{tail}}(\overline{x})$ and $\widetilde{C}^{ij}_{\text{tail}}(\overline{x})$ on the non-vanishing tail terms c^{ij}_{ak} and $\widetilde{c}^{ij}_{ak}$ for $k > N$, incorporating the left-multiplication by the diagonal part of A, see (33).

For $k > N$ we use that $N \geq \widehat{N}(\overline{q}, \xi)$, see Remark 10, hence we obtain the bound

$$\mu_{N+1} \stackrel{\text{def}}{=} \frac{(N+1)^4}{\lambda_{N+1}(\overline{q})} \geq \frac{k^4}{\lambda_k(\overline{q})} \qquad \text{for all } k \geq N+1.$$

This allows the uniform "tail" estimates

$$C^{ij}_{\text{tail}}(\overline{x}) \geq \sum_{k>N} \left|\lambda_k(\overline{q})^{-1} c^{ij}_{ak}(\overline{x}, \widetilde{w})\right| \omega_k \qquad \text{for all } \widetilde{w} \in \mathcal{B}_{1,1}(0), \tag{33a}$$

$$\widetilde{C}^{ij}_{\text{tail}}(\overline{x}) \geq \sum_{k>N} \left|\lambda_k(\overline{q})^{-1} \widetilde{c}^{ij}_{ak}(\overline{x}, v_a, \widetilde{w})\right| \omega_k \qquad \text{for all } \widetilde{w} \in \mathcal{B}_{1,1}(0) \text{ and } \|v_a\|_{\ell^1_\nu} \leq 1, \tag{33b}$$

where the non-zero C^{ij}_{tail} and $\widetilde{C}^{ij}_{\text{tail}}$ are listed in Table 3.

Finally, with the notation

$$C^{ij}_N = (C^{ij}_q, C^{ij}_{a0}, \ldots, C^{ij}_{aN}) \quad \text{and} \quad \widetilde{C}^{ij}_N = (\widetilde{C}^{ij}_q, \widetilde{C}^{ij}_{a0}, \ldots, \widetilde{C}^{ij}_{aN}),$$

we set

$$\begin{aligned}
Z^2_{11}(r_1, r_2) &= \sum_{i,j} \pi_q |A_N| C^{ij}_N \, r^i_1 r^j_2, \\
Z^2_{12}(r_1, r_2) &= \sum_{i,j} \pi_q |A_N| \widetilde{C}^{ij}_N \, r^i_1 r^j_2, \\
Z^2_{21}(r_1, r_2) &= \sum_{i,j} \big\|\pi_a |A_N| C^{ij}_N\big\|_{\ell^1_\nu} \, r^i_1 r^j_2 + \sum_{i,j} C^{ij}_{\text{tail}} \, r^i_1 r^j_2, \\
Z^2_{22}(r_1, r_2) &= \sum_{i,j} \big\|\pi_a |A_N| \widetilde{C}^{ij}_N\big\|_{\ell^1_\nu} \, r^i_1 r^j_2 + \sum_{i,j} \widetilde{C}^{ij}_{\text{tail}} \, r^i_1 r^j_2.
\end{aligned}$$

All sums are over (a subset of) $0 \leq i \leq 4$, $0 \leq j \leq 2$.

4.7 Success

The estimates from §4.5 and §4.6, as well as the radii polynomials (23) have been implemented in the Matlab code `SHproof.m`, which uses the interval arithmetic package Intlab [28]. All code can be found at [38]. Based on Theorems 11 and 12 the script `runproofpointwise.m` successfully proves the existence aspect (as discussed at the end of §3) of Theorem 6 for $\xi = 0.01n$ where $n = 0, 1, \ldots, 203$. We use $N = 20$ and $\nu = 1.5$. These choices were made after some experimentation with the code. The run time is less than a minute on a standard Macbook Pro.

The code also verifies, by using `checkgeometry.m`, that the geometric conditions $\mathcal{H}$ are satisfied. Indeed, first we use the cosine series (12) to transform the Fourier coefficients $\overline{a}$ back to physical

space to obtain the approximate solution

$$\overline{u}(y) = \overline{a}_0 + 2\sum_{m=1}^{N} \overline{a}_m \cos(m\hat{q}y).$$

Note that the value of $\hat{q}$ (which lies somewhere in the interval $[\overline{q} - \hat{r}_1, \overline{q} + \hat{r}_1]$) plays no role in the geometric conditions $\mathcal{H}$ and is thus not used in this part of the code. Denoting the true solution by

$$\hat{u}(y) = \hat{a}_0 + 2\sum_{m=1}^{\infty} \hat{a}_m \cos(m\hat{q}y),$$

the characterization of $(\ell^1_\nu)^*$ in §4.2 leads to the error estimate

$$\|\hat{u} - \overline{u}\|_\infty \le |\hat{a}_0 - \overline{a}_0| + 2\sum_{m=1}^{\infty} |\hat{a}_m - \overline{a}_m| \le \hat{r}_2.$$

Adding the uncertainty interval $[-\hat{r}_2, \hat{r}_2]$ to $\overline{u}$ thus leads to a rigorous interval arithmetic description of the solution:

$$\hat{u}(y) \in \overline{u}(y) + [-\hat{r}_2, \hat{r}_2].$$

Analogous constructions lead to explicit quantitative descriptions of the first and second derivative $\hat{u}'(y)$ and $\hat{u}''(y)$, with extra factors $\hat{q}/(e \ln \nu)$ and $4\hat{q}^2/(e \ln \nu)^2$, respectively, in the uncertainty intervals for these derivatives, cf. (32). Since our interest is in the sign of these derivatives only, the value of $\hat{q}$ again plays no role (and is "scaled out" in the code). Next, we split the interval $[0, \pi/\hat{q}]$ into 50 equal pieces. We note that $\hat{u}'(0) = \hat{u}'(\pi/\hat{q}) = 0$ by construction. To verify that there are exactly two monotone laps on $[0, \pi/\hat{q}]$ and that the local minima and maxima satisfy (H_3), we check that the second derivative is nonzero on the intervals where the first derivative is potentially zero, and that the value of $\hat{u}$ on these intervals must lie in the appropriate range. See the code `checkgeometry.m` and [40, §4] for additional details.

Finally, the script `runproofcontinuous.m` proves Theorem 11 for all $0 \le \xi \le 2.02$, i.e., for a continuous range of parameter values. The interval $[0, 2.02]$ is split into 202000 intervals of width 0.00001, and all bounds are derived uniformly for ξ in each of these small intervals (by using interval arithmetic). We use $N = 70$ and $\nu = 1.15 + 0.15\xi$, i.e., we need more modes than for the pointwise proof and the choice of ν is a little more subtle, namely linear in ξ. These choices were made after some experimentation with the code. Again, we also verify that the geometric conditions $\mathcal{H}$ are satisfied, as described above. This brute force continuation takes several hours (see [41, 5, 22, 30, 42] for more sophisticated approaches).

References

[1] Kenneth Appel and Wolfgang Haken, *Every planar map is four colorable*, Contemporary Mathematics, vol. 98, American Mathematical Society, Providence, RI, 1989, With the collaboration of J. Koch. MR 1025335

[2] Zin Arai, William Kalies, Hiroshi Kokubu, Konstantin Mischaikow, Hiroe Oka, and PawełPilarczyk, *A database schema for the analysis of global dynamics of multiparameter systems*, SIAM J. Appl. Dyn. Syst. **8** (2009), no. 3, 757–789. MR 2533624

[3] Gianni Arioli and Hans Koch, *Computer-assisted methods for the study of stationary solutions in dissipative systems, applied to the Kuramoto-Sivashinski equation*, Arch. Ration. Mech. Anal. **197** (2010), no. 3, 1033–1051. MR 2679365

[4] ———, *Integration of dissipative partial differential equations: a case study*, SIAM J. Appl. Dyn. Syst. **9** (2010), no. 3, 1119–1133. MR 2728184

[5] Maxime Breden, Jean-Philippe Lessard, and Matthieu Vanicat, *Global bifurcation diagrams of steady states of systems of PDEs via rigorous numerics: a 3-component reaction-diffusion system*, Acta Appl. Math. **128** (2013), 113–152. MR 3125637

[6] B. Breuer, J. Horák, P. J. McKenna, and M. Plum, *A computer-assisted existence and multiplicity proof for travelling waves in a nonlinearly supported beam*, J. Differential Equations **224** (2006), no. 1, 60–97. MR 2220064

[7] CAPD*: Computer Assisted Proofs in Dynamics, a Package for Rigorous Numerics*, `http://capd.ii.uj.edu.pl/`.

[8] Roberto Castelli and Jean-Philippe Lessard, *Rigorous numerics in Floquet theory: computing stable and unstable bundles of periodic orbits*, SIAM J. Appl. Dyn. Syst. **12** (2013), no. 1, 204–245. MR 3032858

[9] Charles Conley, *Isolated invariant sets and the Morse index*, CBMS Regional Conference Series in Mathematics, vol. 38, American Mathematical Society, Providence, R.I., 1978. MR 511133

[10] M.C. Cross and P.C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Mod. Phys. **65** (1993), no. 3, 851–1112.

[11] Sarah Day and William D. Kalies, *Rigorous computation of the global dynamics of integrodifference equations with smooth nonlinearities*, SIAM J. Numer. Anal. **51** (2013), no. 6, 2957–2983. MR 3124898

[12] Sarah Day, Jean-Philippe Lessard, and Konstantin Mischaikow, *Validated continuation for equilibria of PDEs*, SIAM J. Numer. Anal. **45** (2007), no. 4, 1398–1424. MR 2338393

[13] Robert L. Devaney, *Homoclinic orbits in Hamiltonian systems*, J. Differential Equations **21** (1976), no. 2, 431–438. MR 0442990

[14] Marcio Gameiro and Jean-Philippe Lessard, *Efficient rigorous numerics for higher-dimensional PDEs via one-dimensional estimates*, SIAM J. Numer. Anal. **51** (2013), no. 4, 2063–2087. MR 3077902

[15] R. W. Ghrist, J. B. Van den Berg, and R. C. Vandervorst, *Morse theory on spaces of braids and Lagrangian dynamics*, Invent. Math. **152** (2003), no. 2, 369–432. MR 1974892

[16] Thomas C. Hales, *A proof of the Kepler conjecture*, Ann. of Math. (2) **162** (2005), no. 3, 1065–1185. MR 2179728

[17] Morris W. Hirsch, Stephen Smale, and Robert L. Devaney, *Differential equations, dynamical systems, and an introduction to chaos*, third ed., Elsevier/Academic Press, Amsterdam, 2013. MR 3293130

[18] Allan Hungria, Jean-Philippe Lessard, and J. D. Mireles James, *Rigorous numerics for analytic solutions of differential equations: the radii polynomial approach*, Math. Comp. **85** (2016), no. 299, 1427–1459. MR 3454370

[19] W. D. Kalies, J. Kwapisz, and R. C. A. M. VanderVorst, *Homotopy classes for stable connections between Hamiltonian saddle-focus equilibria*, Comm. Math. Phys. **193** (1998), no. 2, 337–371. MR 1618147

[20] William D. Kalies, Konstantin Mischaikow, and Robert C. A. M. Vandervorst, *Lattice structures for attractors I*, J. Comput. Dyn. **1** (2014), no. 2, 307–338. MR 3415257

[21] Oscar E. Lanford, III, *A computer-assisted proof of the Feigenbaum conjectures*, Bull. Amer. Math. Soc. (N.S.) **6** (1982), no. 3, 427–434. MR 648529

[22] Jean-Philippe Lessard, *Continuation of Solutions and Studying Delay Differential Equations via Rigorous Numerics*, 2017, To appear.

[23] Jean-Philippe Lessard and Christian Reinhardt, *Rigorous numerics for nonlinear differential equations using Chebyshev series*, SIAM J. Numer. Anal. **52** (2014), no. 1, 1–22. MR 3148084

[24] T. Y. Li and James A. Yorke, *Period three implies chaos*, Amer. Math. Monthly **82** (1975), no. 10, 985–992. MR 0385028

[25] Mitsuhiro T. Nakao, *Numerical verification methods for solutions of ordinary and partial differential equations*, Numer. Funct. Anal. Optim. **22** (2001), no. 3-4, 321–356, International Workshops on Numerical Methods and Verification of Solutions, and on Numerical Function Analysis (Ehime/Shimane, 1999). MR 1849323

[26] Douglas E. Norton, *The fundamental theorem of dynamical systems*, Comment. Math. Univ. Carolin. **36** (1995), no. 3, 585–597. MR 1364499

[27] Neil Robertson, Daniel Sanders, Paul Seymour, and Robin Thomas, *The four-colour theorem*, J. Combin. Theory Ser. B **70** (1997), no. 1, 2–44. MR 1441258

[28] L.N. Rump, *INTLAB - INTerval LABoratory*, Developments in Reliable Computing (Tibor Csendes, ed.), Kluwer Academic Publishers, Dordrecht, 1999, `http://www.ti3.tuhh.de/rump/`, pp. 77–104.

[29] Dietmar Salamon, *Morse theory, the Conley index and Floer homology*, Bull. London Math. Soc. **22** (1990), no. 2, 113–140. MR 1045282

[30] Evelyn Sander and Thomas Wanner, *Validated saddle-node bifurcations and applications to lattice dynamical systems*, SIAM J. Appl. Dyn. Syst. **15** (2016), no. 3, 1690–1733. MR 3546337

[31] O. M. Sarkovs'kiĭ, *Co-existence of cycles of a continuous mapping of the line into itself*, Ukrain. Mat. Ž. **16** (1964), 61–71. MR 0159905

[32] L. P. Sil'nikov, *A case of the existence of a denumerable set of periodic motions*, Dokl. Akad. Nauk SSSR **160** (1965), 558–561. MR 0173047

[33] S.H. Strogatz, *Nonlinear dynamics and chaos : with applications to physics, biology, chemistry, and engineering*, Studies in nonlinearity, Westview Press, Cambridge (Mass.), 1994.

[34] J. Swift and P. C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. A **15** (1977), no. 1, 319–328.

[35] William P. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. (N.S.) **19** (1988), no. 2, 417–431. MR 956596

[36] Warwick Tucker, *A rigorous ODE solver and Smale's 14th problem*, Found. Comput. Math. **2** (2002), no. 1, 53–117. MR 1870856

[37] J. B. van den Berg, C. M. Groothedde, and J. F. Williams, *Rigorous computation of a radially symmetric localized solution in a Ginzburg-Landau problem*, SIAM J. Appl. Dyn. Syst. **14** (2015), no. 1, 423–447. MR 3323206

[38] Jan Bouwe van den Berg, *MATLAB code for "Introduction to Rigorous Numerics in Dynamics: General Functional Analytic Setup and an Example that Forces Chaos"*, 2017, `http://www.math.vu.nl/~janbouwe/code/introrignumdyn/`.

[39] Jan Bouwe van den Berg, Andréa Deschênes, Jean-Philippe Lessard, and Jason D. Mireles James, *Stationary coexistence of hexagons and rolls via rigorous computations*, SIAM J. Appl. Dyn. Syst. **14** (2015), no. 2, 942–979. MR 3353132

[40] Jan Bouwe van den Berg and Jean-Philippe Lessard, *Chaotic braided solutions via rigorous numerics: chaos in the Swift-Hohenberg equation*, SIAM J. Appl. Dyn. Syst. **7** (2008), no. 3, 988–1031. MR 2443030

[41] Jan Bouwe van den Berg, Jean-Philippe Lessard, and Konstantin Mischaikow, *Global smooth solution curves using rigorous branch following*, Math. Comp. **79** (2010), no. 271, 1565–1584. MR 2630003

[42] Thomas Wanner, *Computer-Assisted Bifurcation Diagram Validation and Applications in Materials Science*, 2017, To appear.

[43] Piotr Zgliczyński, *Rigorous numerics for dissipative partial differential equations. II. Periodic orbit for the Kuramoto-Sivashinsky PDE—a computer-assisted proof*, Found. Comput. Math. **4** (2004), no. 2, 157–185. MR 2049869

[44] ———, *Covering relations, cone conditions and the stable manifold theorem*, J. Differential Equations **246** (2009), no. 5, 1774–1819. MR 2494688