Validated numerical solutions for some semilinear elliptic equations on the disk

Gianni Arioli [1] and Hans Koch [2]

Abstract. Starting with approximate solutions of the equation $-\Delta u = wu^3$ on the disk, with zero boundary conditions, we prove that there exist true solutions nearby. One of the challenges here lies in the fact that we need simultaneous and accurate control of both the (inverse) Dirichlet Laplacean and nonlinearities. We achieve this with the aid of a computer, using a Banach algebra of real analytic functions, based on Zernike polynomials. Besides proving existence, and symmetry properties, we also determine the Morse index of the solutions.

1. Introduction

In this paper we consider semilinear elliptic equations of the form

$$-\Delta u = w f'(u)\,, \qquad u\,\big|_{\partial\Omega} = 0\,, \tag{1.1}$$

where Ω is the unit disk in $\mathbb{R}^2$, w is a nonnegative function on Ω, and f' is the derivative of a regular function f on $\mathbb{R}$. Our primary goal is to develop techniques that can be used to prove the existence of solutions in a constructive way, with the help of a computer. In the concrete cases considered here, w is always radial (invariant under rotations) and $f'(u) = u^3$. But it will be clear from our description that the same methods work for other choices of w and f. In fact, similar techniques should apply to other types of equations, and to other radially symmetric domains in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Before giving more details, let us state a result that will help to set the stage.

Theorem 1.1. *There exists a positive radial polynomial w on Ω, such that the equation (1.1) with $f'(u) = u^3$ admits a real analytic solution $u = u_w$ that has Morse index 2, with the property that $|u_w|$ is not invariant under any nontrivial rotation.*

The weight function w and the solution u_w are shown in Figure 1. A precise definition of w is given in [25]. We note that u_w is symmetric under a reflection. This is one symmetry that solutions cannot avoid [3]. Our goal was to find an index-2 solution that has no other symmetries.

Concerning the Morse index, recall that solutions of equation (1.1) are critical points of the functional J on $\mathrm{H}^1_0(\Omega)$,

$$J(u) = \int_\Omega \Big[\tfrac{1}{2}\big|\nabla u\big|^2 - w f(u)\Big]\, dxdy\,, \tag{1.2}$$

assuming that f satisfies some growth and regularity conditions. The Morse index of a critical point u is the number of descending directions of J at u.

One of the difficulties with proving Theorem 1.1 is that Ω is a disk. For a square domain, an analogous result was proved in [11]. And for the disk, it is possible [12] to

[1] Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano.

[2] Department of Mathematics, University of Texas at Austin, Austin, TX 78712

obtain an accurate numerical "solution" that looks as shown in Figure 1. But we have hitherto been unable to prove that there exists a true solution nearby.

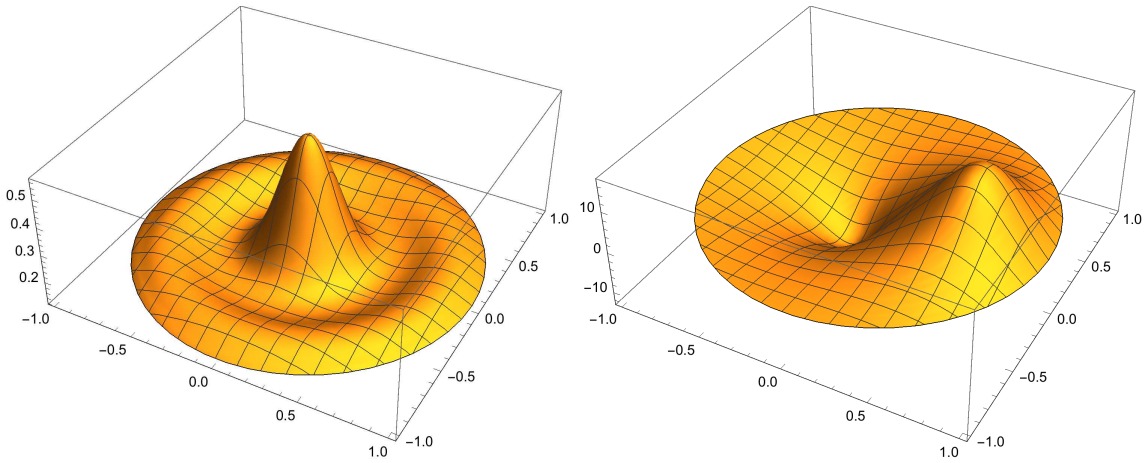

Figure 1. The weight function w and solution u_w described in Theorem 1.1.

Before describing our approach in more detail, let us state two other results that can be proved in a similar way. The first results concern again "minimally symmetric solutions to a highly symmetric problem". While the weight w in Theorem 1.1 had to be chosen carefully to obtain a minimally symmetric solution of index 2, a standard Hénon weight $w(r,\vartheta) = r^\alpha$ suffices in the index-1 case. Here, and in what follows, (r,ϑ) denote the standard polar coordinates on Ω.

Theorem 1.2. *For $\alpha = 2, 4, 6$, the equation (1.1), with $w = r^\alpha$ and $f'(u) = u^3$, admits a real analytic solution $u = u_\alpha > 0$. This solution has Morse index 1 and is not invariant under any nontrivial rotation.*

The solutions u_2, u_4, and u_6 are shown in Figure 2.

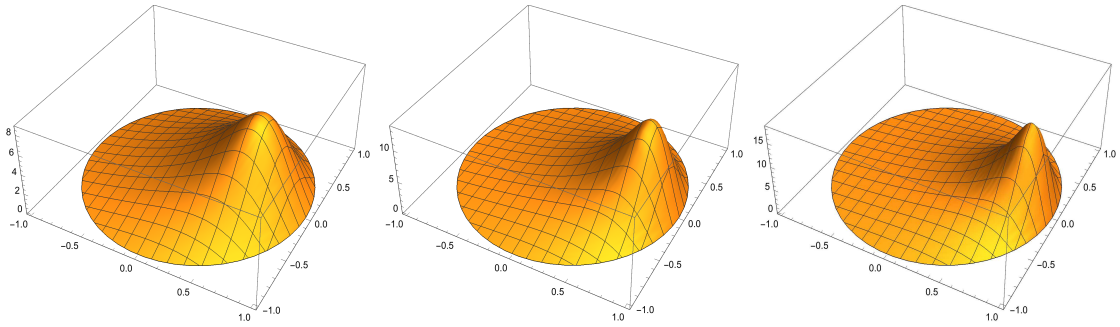

Figure 2. The solutions u_2, u_4, and u_6 described in Theorem 1.2.

The same result, but without the statement about the lack of symmetry, is easy to prove: minimizing J on the Nehari manifold $\mathfrak{N} = \{u \in \mathrm{H}^1_0(\Omega) : DJ(u)u = 0\,, u \neq 0\}$ shows that index-1 solutions exist and that they do not vanish anywhere on Ω. Intuitively, the asymmetry of the positive minimizers u_α stems from fact that the term $-wf(u)$ in the

integral (1.2) rewards u for concentrating off-center. Indeed, an analogue of Theorem 1.2 can be proved by variational methods for sufficiently large values of α [2]; see also [4] and references therein. Numerical results on a number of nonlinear elliptic equation can be found in [1]. They include positive non-radial solutions u_α as described in Theorem 1.2, but for $\alpha = 1, 9$. Given that the positive solution for $\alpha = 0$ is radial, one expects that there is a symmetry-breaking bifurcation as α is increased from 0 to 1.

Our method of proof is not limited to solutions of index 1 or 2, although the computations become impractical at high index. In the next theorem, we consider two solutions that are close to sums of index-1 solutions,

$$u_{\alpha,n} \approx \sum_{m=1}^{2n} (S_n)^m u_\alpha \,, \qquad (S_n u)(r,\vartheta) = -u(r,\vartheta+\pi/n)\,. \tag{1.3}$$

If u_α is one of the solutions described in Theorem 1.2, then the functions in the above sum are solutions of the same equation; and if n is not too large, then most of their mass is contained in mutually disjoint sectors of the disk. Thus, the sum in (1.3) is an approximate solution of the Hénon equation, and we expect to find a true solution nearby. Furthermore, this solution should have index $2n$. Indeed, this holds in the two cases considered here:

Theorem 1.3. *For $n = 1, 2$, the equation (1.1), with $w = r^2$ and $f'(u) = u^3$, admits a nontrivial real analytic solution $u = u_{2,n}$ that is invariant under S_n and has index $2n$.*

The functions $u_{2,1}$ and $u_{2,2}$ are shown in Figure 3. We expect that solutions of the type (1.3) exist for any given $n > 0$, provided that α is sufficiently large.

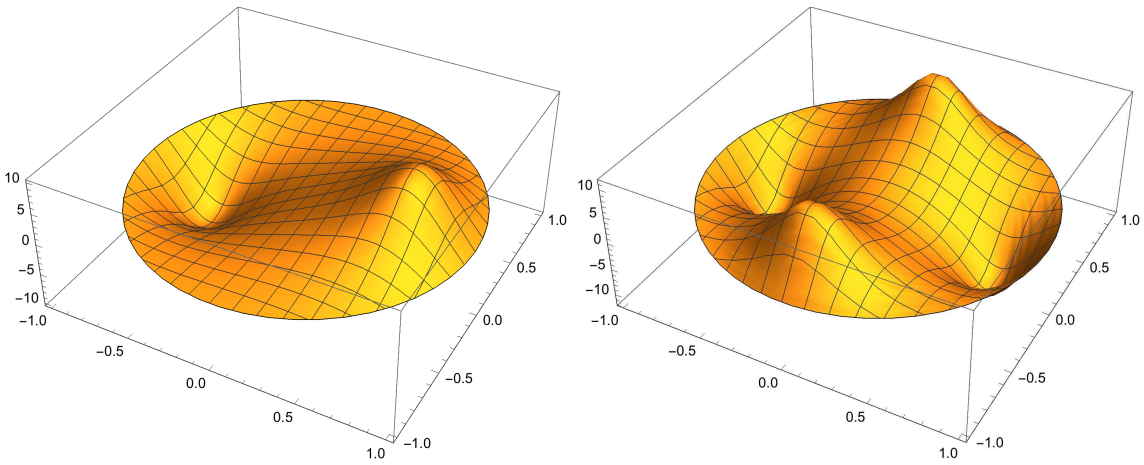

Figure 3. The solutions $u_{2,1}$ and $u_{2,2}$ described in Theorem 1.3.

As indicated earlier, the three theorems stated above are proved with the aid of a computer. In addition to the properties described in these theorems, we obtain accurate bounds on the difference $u - \bar{u}$ between the true solution u and a numerical approximation $\bar{u}$. The accuracy of the result is limited only by the computational resources available. To give a rough idea: our estimates on the solution $u = u_w$ described in Theorem 1.1, carried out on a standard desktop machine, yield an upper bound less than 2^{-35} on the norm (defined later) of $u - \bar{u}$ relative to the norm of $\bar{u}$.

Following a strategy that has been successful in many other computer-assisted proofs in analysis [11-20], we start by converting the given equation (1.1) to a fixed point equation for a suitable operator G. As in [11], we use

$$G(u) = -\Delta^{-1}\big[wf'(u)\big]\,, \tag{1.4}$$

where Δ^{-1} is the inverse Dirichlet Laplacean on Ω. Then we consider a Newton-type map $\mathcal{N}$ associated with G and prove that $\mathcal{N}$ is a contraction in a small neighborhood of an approximate fixed point $\bar{u}$.

Clearly, approximations play a crucial role. Without loss of generality, we are looking for a representation $u = \sum_k c_k \Psi_k$, where approximation corresponds to truncating the series to a finite sum. So it is desirable to work with a space $\mathcal{B}$ and basis function Ψ_k that

(a) *are well adapted to the operators involved*,

(b) *have useful algebraic properties and*

(c) *good approximation properties*.

The same criteria apply to most computer-assisted proofs in analysis [11-20]. In problems that involve a Laplacean with a compact inverse, the ideal way to satisfy (a) is to take for Ψ_k the k-th eigenvector of Δ^{-1}. This works well for rectangular domains and Fourier series. In this case, the basis functions Ψ_k also have a simple product expansion $\Psi_i\Psi_j = \sum_k c_{i,j,k}\Psi_k$. This is a desired property (b) in problems that include nonlinearities such as the term $f'(u)$ in (1.4). Ideally, $\mathcal{B}$ is a Banach algebra. Concerning (c), the expected solution u should have coefficients c_k that decrease more rapidly than those for a typical function in $\mathcal{B}$.

In the problems considered here, the eigenfunctions of the Dirichlet Laplacean (on the disk) are of the form $\Psi_k(r,\vartheta) = \psi_k(r)e^{im_k\vartheta}$, where ψ_k is an appropriately scaled Bessel function. Unfortunately, there is no convenient product expansion for these Bessel functions. Thus, in earlier attempts to prove a result like Theorem 1.1, we used various approximations or alternatives for the Bessel functions. Some choices worked well for numerical computations, as described in [12], but they never led to a successful proof. We had also considered using Zernike polynomials, but obviously not carefully enough.

As it turns out, the Zernike polynomials are close to ideal for the type of problems considered here. There are good reasons for this, as we will explain below. The Zernike polynomials R_n^m are widely used in optics. But despite the vast literature on (the use of) these polynomials, we found no evidence that would have justified trying yet another approach. That is, until we became aware of the references [7,10].

In [7] it is shown that the Zernike functions $V_n^m(r,\vartheta) = R_n^m(r)e^{im\vartheta}$ have a product expansion whose coefficients are the squares of certain Clebsch-Gordan coefficients. This property is obtained by relating the Zernike functions to generalized spherical harmonics, using the azimuthal projection of the sphere to the disk. So in essence, one works indirectly with functions on SO(3), whence the nice behavior under multiplication (product representations). But the Laplacean on the sphere does not map to the Laplacean on the disk, which creates a potential conflict between the desired properties (a) and (b). Surprisingly, this problem is quite harmless: both ΔV_n^m and $\Delta^{-1}V_n^m$ are linear combinations of at most three Zernike functions [10].

This motivates the following expansion for our functions on the unit disk:

$$u(r,\vartheta) = \sum_{m,l=0}^{\infty} R^m_{m+2l}(r)\Big[a_{m,l}\cos(m\vartheta) + b_{m,l}\sin(m\vartheta)\Big], \tag{1.5}$$

with $b_{0,l} = 0$ for all l. To be more precise, the solutions of (1.1) described in the preceding theorems are symmetric under the reflection $\vartheta \mapsto -\vartheta$, so their coefficients $b_{m,l}$ all vanish. The function spaces used in our analysis are the following. Given $\rho > 1$, we define $\mathcal{B}_\rho$ to be the space of all functions (1.5) that have a finite norm

$$\|u\|_\rho = \sum_{m,l=0}^{\infty} \big(|a_{m,l}| + |b_{m,l}|\big)\rho^{m+2l}\,. \tag{1.6}$$

Using the relationship between the Zernike functions and the generalized spherical harmonics, one immediately gets complex bounds on the Zernike functions. This, together with the Clebsch-Gordan series for products can be used to show that $\mathcal{B}_\rho$ is in fact a Banach algebra of real analytic functions on Ω.

Clearly, if $u \in \mathcal{B}_\varrho$ with $\varrho > \rho$, then u is well approximated in $\mathcal{B}_\rho$ by truncated sums (1.5). This allows us to obtain highly accurate numerical approximations to the solutions described in Theorems 1.1, 1.2, and 1.3. The limiting factor here is the computation or storage of Clebsch-Gordan coefficients. The computation is quite costly, so we compute the coefficients beforehand and store them in an array. To reduce symmetries and thus storage space, we use ideas described in [23].

2. Zernike functions

Here we introduce the Zernike functions and describe the properties that we need in our analysis. Our need for the product expansion and complex bounds favors the following approach.

Consider unitary representations of SU(2), with Hermitian generators L_1, L_2, and L_3 satisfying $[L_2, L_3] = iL_1$, $[L_3, L_1] = iL_2$, and $[L_1, L_2] = iL_3$. The eigenvalues of L_3 are commonly referred to as weights. Each irreducible representation is characterized uniquely (up to unitary equivalence) by the value ν of the largest weight, which is a nonnegative half-integer. In such a representation, the spectrum of each generator L_j is the set $W_\nu = \{-\nu, -\nu+1, \ldots, \nu-1, \nu\}$, and all the eigenvalues are simple. Furthermore, the $d = 2\nu + 1$ eigenvectors of L_3 constitute an orthogonal basis for the underlying space. Using the bra-ket notation that is common in physics, the normalized eigenvector of L_3 with eigenvalue $\mu \in W_\nu$ will be denoted by $|\nu\mu\rangle$.

Without loss of generality, we may identify each operator R in our representation by the $d \times d$ matrix whose elements are $R_{\mu,\mu'} = \langle \nu\mu|R|\nu\mu'\rangle$. Here we consider the matrices $e^{-i\alpha L_3}e^{-i\beta L_2}e^{-i\gamma L_3}$, also known as the Wigner D-matrices. The angles (α, β, γ) describe the orientation of a coordinate frame in $\mathbb{R}^3$ with respect to a fixed frame. We will restrict to $\gamma = 0$, which suffices to describe the direction of the rotated 3-axis. Then the matrix elements

$$D^\nu_{\mu,\mu'}(\alpha,\beta) \stackrel{\text{def}}{=} \big\langle \nu\mu\big|e^{-i\alpha L_3}e^{-i\beta L_2}\big|\nu\mu'\big\rangle = e^{-i\mu\alpha}\big\langle \nu\mu\big|e^{-i\beta L_2}\big|\nu\mu'\big\rangle \tag{2.1}$$

define functions on the unit sphere, also referred to as generalized spherical harmonics. The functions $D^{\nu}_{\mu,0}$ are in essence the ordinary spherical harmonics.

Consider now the diagonal elements $D^{\nu}_{\mu,\mu}$. They are even functions of β. This follows e.g. from the identity $(-1)^{\nu-L_3}L_2(-1)^{\nu-L_3} = -L_2$, which can be derived by elementary computations. Using the spectral decomposition $L_2 = \sum_\kappa \kappa P_\kappa$ of the operator L_2, we may write

$$D^{\nu}_{\mu,\mu}(\alpha,\beta) = e^{-i\mu\alpha} \sum_{\kappa\in W_\nu} \langle \nu\mu|P_\kappa|\nu\mu\rangle \cos(\kappa\beta)\,. \tag{2.2}$$

Notice that the sum in this equation is a polynomial of degree 2ν in the variable $\cos(\beta/2)$. Let $n = 2\nu$ and $m = 2\mu$. Then we have

$$V^m_n(r,\vartheta) \stackrel{\text{def}}{=} D^{\nu}_{\mu,\mu}(-2\vartheta,\beta) = R^m_n(r)e^{im\vartheta}\,, \qquad r = \cos(\beta/2)\,, \tag{2.3}$$

where R^m_n is a polynomial of degree n. This equation will serve as our definition of the Zernike functions V^m_n and the Zernike polynomials R^m_n.

Next, we describe a few properties of the functions V^m_n that will be needed in our analysis. Using the well-known Clebsch-Gordan series for the product of Wigner functions $D^{\nu}_{\mu,\mu'}$, one obtains directly the product rule

$$V^{m_1}_{n_1}V^{m_2}_{n_2} = \sum_{n_3} \big|\langle \nu_1\mu_1\nu_2\mu_2|\nu_3\mu_3\rangle\big|^2 V^{m_3}_{n_3}\,, \qquad m_3 = m_1 + m_2\,. \tag{2.4}$$

Here, $\langle \nu_1\mu_1\nu_2\mu_2|\nu_3\mu_3\rangle$ are the so-called Clebsch-Gordan coefficients, with $\nu_j = m_j/2$ and $\mu_j = m_j/2$. These coefficients will be discussed in Section 5.

Another immediate consequence of the definition (2.3) is the following complex bound. Given that $-\nu \le L_2 \le \nu$ in a representation of highest weight ν, the Wigner functions (2.2) extend to entire analytic functions, and we have the bound

$$\big|V^m_n(\cos(\beta/2),\vartheta)\big| \le e^{|\mathrm{Im}(\beta/2)|n}e^{|\mathrm{Im}(\vartheta)||m|}\,, \qquad \beta,\vartheta\in\mathbb{C}\,. \tag{2.5}$$

The generalized spherical harmonics are known to be directly related to the Jacobi polynomials $P^{(a,b)}_l$. In particular, $V^m_{m+2l}(r,\vartheta) = z^m P^{(0,m)}_l\big(2|z|^2-1\big)$ for $m,l\ge 0$, where $z = re^{i\vartheta}$. Using the Rodrigues formula for the Jacobi polynomials, one finds the following Rodrigues formula for the Zernike functions [5]. Let $k,l\ge 0$. Then

$$V^m_n(r,\vartheta) = \mathcal{V}^m_n\big(re^{i\vartheta}, re^{-i\vartheta}\big)\,, \qquad n = k+l\,, \quad m = k-l\,, \tag{2.6}$$

where

$$\mathcal{V}^m_n = \partial^l_z\partial^k_{\bar z}\mathcal{P}_n\,, \qquad \mathcal{P}_n(z,\bar z) = \frac{1}{n!}(z\bar z - 1)^n\,. \tag{2.7}$$

Here ∂_z and $\partial_{\bar z}$ denote the partial derivatives with respect to the (independent) variables z and $\bar z$, respectively. The identity (2.7) can be used e.g. to give a simple proof of the following lemma. Denote by Δ the Dirichlet Laplacean on the disk Ω.

Lemma 2.1. [10] *Let $n \geq m \geq 0$ with $n - m$ even. Then*

$$\begin{aligned} \Delta^{-1} V_n^m &= c_2 V_{n+2}^m + c_1 V_n^m + c_0 V_{n-2}^m \qquad \text{if } n > m\,, \\ \Delta^{-1} V_n^m &= c_2 V_{n+2}^m - c_2 V_n^m \qquad\qquad\quad \text{if } n = m\,, \end{aligned} \tag{2.8}$$

where

$$c_2 = \frac{1}{4(n+2)(n+1)}\,, \quad c_1 = -\frac{1}{2n(n+2)}\,, \quad c_0 = \frac{1}{4n(n+1)}\,. \tag{2.9}$$

Proof. A trivial computation shows that

$$\partial_z \partial_{\bar z} \mathcal{P}_k = k \mathcal{P}_{k-1} + \mathcal{P}_{k-2}\,, \qquad k \geq 2\,. \tag{2.10}$$

We claim that there exist constants c_2, c_1, and c_0, such that

$$c_2 \big(\partial_z \partial_{\bar z}\big)^2 \mathcal{P}_{n+2} + c_1 \big(\partial_z \partial_{\bar z}\big) \mathcal{P}_n + c_0 \mathcal{P}_{n-2} = \tfrac{1}{4} \mathcal{P}_n\,. \tag{2.11}$$

Indeed, each term on the left hand side of this equation is a linear combination of $\mathcal{P}_n$, $\mathcal{P}_{n-1}$, and $\mathcal{P}_{n-2}$. So we have 3 linear equations with 3 unknowns. A straightforward computation yields the solution (2.9). Applying $4\partial_z^l \bar\partial_{\bar z}^k$ to both sides of the equation (2.11), we obtain

$$4 \partial_z \partial_{\bar z} \big[c_2 \mathcal{V}_{n+2}^m + c_1 \mathcal{V}_n^m + c_0 \mathcal{V}_{n-2}^m \big] = \mathcal{V}_n^m\,. \tag{2.12}$$

Notice that $c_0 + c_1 + c_2 = 0$. Thus, the function $[\ldots]$ in the equation (2.12) vanishes for $z\bar z = 1$. Taking $z = x + iy$ and $\bar z = x - iy$, we have $4\partial_z \partial_{\bar z} = \partial_x^2 + \partial_y^2 = \Delta$, and the first identity in (2.8) follows. The second identity is verified similarly. **QED**

To conclude this section, we note that the change of variables $r = \cos(\beta/2)$ used in the definition (2.3) is far from ad hoc. It defines the azimuthal projection $(\beta, \vartheta) \mapsto (r, \vartheta)$ from the sphere to the disk. This projection preserves area (up to a trivial factor). Using the orthogonality properties of the generalized spherical harmonics, one finds that the Zernike functions V_n^m constitute a complete orthogonal set for $\mathrm{L}^2(\Omega)$. Alternatively, one can use the orthogonality properties of the Jacobi polynomials [9].

3. Real analytic functions on the disk

In this section we prove that $\mathcal{B}_\rho$ is a Banach algebra under pointwise multiplication of functions. In addition, we give a bound on the inverse Dirichlet Laplacean, and we introduce some notation that will be needed later on. Unless specified otherwise, the domain of a function $u = u(r, \vartheta)$ is assumed to be the cylinder $[0, 1] \times \mathbb{S}^1$. But we still regard u as a function on the disk Ω.

For every integer m, define $N_m = \{|m|, |m| + 2, |m| + 4, \ldots\}$. Given a real number $\rho \geq 1$, denote by $\mathcal{A}_\rho$ the real vector space of all functions u,

$$u = \sum_{m \in \mathbb{Z}} u_m\,, \qquad u_m = \sum_{n \in N_m} u_{m,n} V_n^m\,, \qquad u_{m,n} \in \mathbb{C}\,, \tag{3.1}$$

that have a finite norm

$$\|u\|_\rho = \sum_{m\in\mathbb{Z}} \|u_m\|_\rho\,, \qquad \|u_m\|_\rho = \sum_{n\in N_m} |u_{m,n}|_1\,\rho^n\,. \tag{3.2}$$

Here $|x+iy|_1 = |x| + |y|$ for $x, y \in \mathbb{R}$. When equipped with this norm, $\mathcal{A}_\rho$ is a Banach space over $\mathbb{R}$. A real-valued function $u \in \mathcal{A}_\rho$ can be written in the form

$$u = \sum_{n\in N_0} A_{0,n} V_n^0 + \frac{1}{2}\sum_{m\neq 0}\sum_{n\in N_m} \big[A_{m,n} - iB_{m,n}\big] V_n^m\,, \tag{3.3}$$

with real coefficients $A_{m,n}$ and $B_{m,n}$ satisfying $A_{-m,n} = A_{m,n}$ and $B_{-m,n} = -B_{m,n}$, respectively, for all integers m and all $n \in N_m$. Let us relabel the coefficients by setting $a_{m,l} = A_{m,n}$ and $b_{m,l} = B_{m,n}$, where $l = (n - |m|)/2$. Then a short computation shows that u agrees with the function (1.5), and that the norm (3.2) of u is given by (1.6). In other words, $\mathcal{B}_\rho$ is the subspace of real-valued functions $u \in \mathcal{A}_\rho$.

Lemma 3.1. *$\mathcal{A}_\rho$ is a Banach algebra under pointwise multiplication. If $\rho > 1$ then the functions in $\mathcal{A}_\rho$ extend analytically to some complex open neighborhood of Ω.*

Proof. Consider first fixed integers m_1, m_2, and define $m_3 = m_1 + m_2$. Recall from (2.4) that

$$V_{n_1}^{m_1} V_{n_2}^{m_2} = \sum_{n_3} C_{n_1,n_2,n_3}^{m_1,m_2,m_3}\, V_{n_3}^{m_3}\,, \qquad n_1 \in N_{m_1}\,, \quad n_2 \in N_{m_2}\,, \tag{3.4}$$

where $C_{n_1,n_2,n_3}^{m_1,m_2,m_3}$ is the square of a Clebsch-Gordan coefficient and thus nonnegative. These coefficients vanish whenever $n_j \notin N_{m_j}$ for some j. They also vanish if $n_3 > n_1 + n_2$, as we will describe later. In addition, we have $\sum_{n_3} C_{n_1,n_2,n_3}^{m_1,m_2,m_3} = 1$. This follows from unitarity, but it can be seen also from (3.4) by noting that $R_n^m(1) = 1$ whenever $n \in N_m$. As a result,

$$\|V_{n_1}^{m_1} V_{n_2}^{m_2}\|_\rho \le \sum_{n_3} C_{n_1,n_2,n_3}^{m_1,m_2,m_3} \rho^{n_3} \le \rho^{n_1+n_2}\,. \tag{3.5}$$

Let now u and v be two functions in $\mathcal{A}_\rho$. To simplify notation, we define $u_{m,n} = 0$ and $v_{m,n} = 0$ whenever $n \notin N_m$. By using the bound (3.5), we immediately get

$$\begin{aligned} \|uv\|_\rho &\le \sum_{m_1,n_1,m_2,n_2} |u_{m_1,n_1} v_{m_2,n_2}|_1 \|V_{n_1}^{m_1} V_{n_2}^{m_2}\|_\rho \\ &\le \sum_{m_1,n_1,m_2,n_2} |u_{m_1,n_1}|_1 |v_{m_2,n_2}|_1 \rho^{n_1+n_2} = \|u\|_\rho \|v\|_\rho\,. \end{aligned} \tag{3.6}$$

This shows that $\mathcal{A}_\rho$ is a Banach algebra, as claimed.

Consider now $\rho > 1$. From the bound (2.5), is is clear that a function $u \in \mathcal{A}_\rho$ extends analytically to a complex open neighborhood A of $[0,1] \times \mathbb{S}^1$ in the variables (r, ϑ). So the series (3.1) for u converges uniformly on compact subsets of A. Changing to Cartesian variables (x, y), this translates into uniform convergence on compact subsets of some open

neighborhood Ω_ρ of Ω. But the Zernike functions V_n^m are polynomials in (x, y), as can be seen e.g. from (2.6). Thus, being a locally uniform limit of analytic functions, u is analytic on Ω_ρ. **QED**

We note that Banach algebras of disk polynomials have been considered before in [8].

For our computer-assisted error estimates, we approximate a function by truncating its Zernike series. Given $N \geq 0$, define the projection $\mathbb{P}_N : \mathcal{A}_\rho \to \mathcal{A}_\rho$ as follows. For every $u \in \mathcal{A}_\rho$, the function $\mathbb{P}_N u$ is obtained from u by truncating the Zernike series (3.1) of u to terms with index $n < N$.

Proposition 3.2. *Consider the inverse Dirichlet Laplacean Δ^{-1} as a linear operator on $\mathcal{A}_\rho$. Then Δ^{-1} is compact, and for every $N \geq 1$, the operator norm of $\Delta^{-1}(\mathbb{I} - \mathbb{P}_N)$ is bounded by*

$$\left\|\Delta^{-1}(\mathbb{I} - \mathbb{P}_N)\right\| \leq \frac{\left(\rho + \rho^{-1}\right)^2}{4N(N+2)}\,. \tag{3.7}$$

The bound (3.7) is an immediate consequence of Lemma 2.1. It implies in particular that Δ^{-1} is a uniform limit of finite rank operators, and thus compact.

4. Main steps in the proof

In this section we describe how Theorems 1.1, 1.2, and 1.3 can be proved by verifying the assumptions of five technical lemmas. The estimates that are used to verify these assumptions will be discussed in Section 6.

As mentioned in the introduction, we find solutions of the equation (1.1) by solving the fixed point problem for the map G given by (1.4). Here we follow the approach used in [11]. We always assume that $f'(u) = u^3$. Let ρ be a real number larger than 1, to be specified later. Assuming that w belongs to $\mathcal{B}_\rho$, G defines a smooth compact map on $\mathcal{B}_\rho$. This follows from the fact that $\mathcal{B}_\rho$ is a Banach algebra, and from Proposition 3.2.

In this paper, we are interested only in solutions $u = u(r, \vartheta)$ that are even functions of ϑ. The even subspace of $\mathcal{B}_\rho$ will be denoted by $\mathcal{B}_\rho^0$. Given $r > 0$ and $u \in \mathcal{B}_\rho^0$, denote by $B_r(u)$ the close ball in $\mathcal{B}_\rho^0$ of radius r, centered at u.

Given a function $\bar{u} \in \mathcal{B}_\rho^0$ and a bounded linear operator M on $\mathcal{B}_\rho^0$, define

$$\mathcal{N}(h) = G(\bar{u} + Ah) - \bar{u} + Mh\,, \qquad A = \mathbb{I} - M\,, \tag{4.1}$$

for every $h \in \mathcal{B}_\rho^0$. Clearly, if h is a fixed point of $\mathcal{N}$ then $\bar{u} + Ah$ is a fixed point of G. Furthermore, if the operator $\mathbb{I} - DG(\bar{u})$ is invertible, and if A sufficiently close to its inverse, then $\mathcal{N}$ is a contraction near $\bar{u}$.

Our goal is to apply the contraction mapping theorem to the map $\mathcal{N}$, on some small ball $B_r(0)$. Thus $\bar{u}$ is chosen to be an approximate fixed point of G. For practical reasons, $\bar{u} = \mathbb{P}_N \bar{u}$ for some N. For the same reasons, we choose M to satisfy $M = \mathbb{P}_N M \mathbb{P}_N$ for some N. So M is in essence a matrix.

To guarantee the existence of a true fixed point of G near $\bar{u}$, it suffices to prove the hypotheses of the following lemma.

Lemma 4.1. *Let $\rho > 1$ and $w \in \mathcal{B}^0_\rho$. Assume that there exists a function $\bar{u} \in \mathcal{B}^0_\rho$, a finite-rank operator $M : \mathcal{B}^0_\rho \to \mathcal{B}^0_\rho$, and a real number $\delta > 0$, such that the map map $\mathcal{N}$ defined by (4.1) admits bounds*

$$\varepsilon \geq \|\mathcal{N}(0)\|_\rho\,, \qquad K \geq \|D\mathcal{N}(h)\|\,, \quad \forall h \in B_\delta(0)\,, \tag{4.2}$$

with ε and K satisfying $\varepsilon + K\delta < \delta$. Then the equation (1.1) has a solution $u_* \in \mathcal{B}_\rho$ within a distance $\|A\|\delta$ of $\bar{u}$. Furthermore, if M has no eigenvalue 1, then this solution u_* is locally unique.

Proof. By the contraction mapping principle, the given bounds imply that $\mathcal{N}$ has a unique fixed point h_* in the ball $B_\delta(0)$. In fact, h_* lies in the interior of $B_\delta(0)$, since the inequality $\varepsilon + K\delta < \delta$ is strict. Clearly $u_* = \bar{u} + Ah_*$ is a fixed point of G. If M has no eigenvalue 1, then this fixed point is locally unique, since the fixed point h_* of $\mathcal{N}$ is locally unique and $A = \mathbb{I} - M$ is invertible. The distance of u_* from $\bar{u}$ is $\|u_* - \bar{u}\|_\rho = \|Ah_*\|_\rho \leq \|A\|\delta$. **QED**

With u_* as above, consider the possibility that $|u_*|$ is invariant under some nontrivial rotation. Then the function $u = u_*^2$ is invariant under a rotation by $2\pi/k$ for some integer $k \neq 1$. So the component u_m in the representation (3.1) vanishes, unless m is a multiple of k. This proves the following.

Lemma 4.2. *If some coefficient $u_{1,l}$ in the Zernike expansion (1.5) for $u = u_*^2$ is nonzero, then $|u_*|$ is not invariant under any nontrivial rotation.*

This fact is used to verify that the solutions described in Theorems 1.1 and 1.2 have no rotation symmetries. The property $u_\alpha > 0$ mentioned in Theorem 1.2 follows from the well-known fact that index-1 solutions do not vanish anywhere on Ω, if $w > 0$. In order to prove the symmetry properties of the solutions described in Theorem 1.3, we use the following. We assume that w is radial.

Lemma 4.3. *Under the assumptions of Lemma 4.1, if $\bar{u}$ is invariant under S_n, and if M commutes with S_n, then the solution u_* described in (the proof of) Lemma 4.1 is invariant under S_n.*

This claim follows from the fact that, under the given assumptions, if $h \in \mathcal{B}^0_\rho$ is invariant under S_n, then so is $\mathcal{N}(h)$. Thus, the limit $h_* = \lim_{k\to\infty} \mathcal{N}^k(0)$ and the function $u_* = \bar{u} + h_* - Mh_*$ are invariant under S_n.

Next, we consider the problem of determining the Morse index of a solution u of the equation (1.1). As in [11] we use the identity

$$D^2J(u)(v \times v) = \int_\Omega \Big[|\nabla v|^2 - 3wu^2v^2\Big]\,dxdy = \big\langle v, [\mathrm{I} - DG(u)]v\big\rangle_{\mathrm{H}^1}\,, \tag{4.3}$$

which relates the second derivative of the functional J defined in (1.2) to the first derivative of the map G defined in (1.4). Here, it is assumed that v belongs to $\mathrm{H}^1_0 = \mathrm{H}^1_0(\Omega)$. Notice that, if W is a bounded linear operator on $\mathrm{L}^2 = \mathrm{L}^2(\Omega)$, then $(-\Delta)^{-1}W$ is a bounded linear operator from L^2 to H^1_0, and

$$\big\langle (-\Delta)^{-1}Wv, h\big\rangle_{\mathrm{H}^1} = \langle Wv, h\rangle_{\mathrm{L}^2}\,, \qquad v, h \in \mathrm{H}^1_0\,. \tag{4.4}$$

Clearly, if W is self-adjoint on L^2, then the restriction of $(-\Delta)^{-1}W$ to H^1_0 is self-adjoint.

These observations explain much of the following.

Proposition 4.4. *Assume that wu^2 is of class C^1 and nonnegative. Then $DG(u)$ is a compact positive self-adjoint operator on H^1_0. If wu^2 is positive almost everywhere on Ω, then all eigenvalues of $DG(u)$ are positive. Assume now that wu^2 belong to $\mathcal{B}_\rho$. Then every eigenvector of $DG(u)$ with nonzero eigenvalue belongs to $\mathcal{B}_\rho$. If in addition u solves the equation (1.1), then the Morse index of u equals the number of eigenvalues of $DG(u)$ that exceed* 1.

This proposition was proved in [11] for a square domain. The same arguments apply in the case of a disk, using that Δ^{-1} defines a compact linear operator on $\mathcal{B}_\rho$ by Proposition 3.2, and that $\mathcal{B}_\rho$ is dense in H^1. The density of $\mathcal{B}_\rho$ in H^1 follows e.g. from the fact that the Zernike functions V_n^m constitute a complete orthogonal set for L^2.

Assume now that $u \in \mathcal{B}_\rho$ is a nontrivial fixed point of G, with $w \in \mathcal{B}_\rho$. By Proposition 4.4, the Morse index of u agrees with the number of eigenvalues of $DG(u)$ that are larger than 1. And it suffices to consider $DG(u)$ as a linear operator on $\mathcal{B}_\rho$.

Notice that $\mathcal{B}_\rho^0$ is an invariant subspace of $DG(u)$. Another invariant subspace of $DG(u)$ is the space $\mathcal{B}_\rho^1$ of all functions $u = u(r,\vartheta)$ in $\mathcal{B}_\rho$ that are odd functions of ϑ. We refer to $\mathcal{B}_\rho^1$ as the odd subspace of $\mathcal{B}_\rho$. Clearly, every eigenvalue of $DG(u)$ has an eigenfunction that belongs to one of these two subspaces. Two eigenvalues are known explicitly: $u \in \mathcal{B}_\rho^0$ is an eigenvector of $DG(u)$ with eigenvalue 3, and $\partial_\vartheta u \in \mathcal{B}_\rho^1$ is an eigenvector of $DG(u)$ with eigenvalue 1. This follows from the fact that nonlinearity $f'(u)$ in the equation (1.4) is cubic, and that $(r,\vartheta) \mapsto u(r,\vartheta+t)$ is a fixed point of G for every real number t, respectively.

Consider now the restriction of $DG(u)$ to one of the subspaces $\mathcal{B}_\rho^\sigma$ of fixed parity $\sigma \in \{0,1\}$. Our goal is to determine the number of eigenvalues of $DG(u)$ that are larger than 1. In order simplify our description, let $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be the eigenvalues of $DG(u)$, listed with their multiplicities, and let $u_1, u_2, \ldots$ be the corresponding eigenvectors. We may assume that these eigenvectors are mutually orthogonal. If $\lambda_n < \theta < \lambda_{n+1}$, then the operator

$$DG(u) - \mathcal{K}\,, \qquad \mathcal{K} = \sum_{j=1}^{n} \lambda_j \frac{\langle h, u_j\rangle_{\mathrm{H}^1}}{\langle u_j, u_j\rangle_{\mathrm{H}^1}} u_j\,, \tag{4.5}$$

has a spectral radius less than θ. We are interested in obtaining a similar conclusion by using only approximate eigenvalues and eigenvectors.

The is possible by using the following fact.

Lemma 4.5. *Let A and K be bounded linear operators on a Hilbert space. Assume that A is normal, that K is of finite rank n, and that $\|A-K\| < \theta$. Then A has at most n eigenvalues λ_j (counting multiplicities) satisfying $|\lambda_j| \geq \theta$.*

Proof. As a rank n operator, K admits a representation $Kh = \sum_{i=1}^n a_i \langle h, w_i\rangle v_i$. Assume for contradiction that A admits an orthonormal set $\{u_1, u_2, \ldots, u_{n+1}\}$ of $n+1$ eigenvectors with eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_{n+1}$ satisfying $|\lambda_j| \geq \theta$. Then some nontrivial linear combination $h = \sum_{j=1}^{n+1} c_j u_j$ is orthogonal to each of the n vectors w_i. It satisfies $Kh = 0$, and

thus

$$\big\|(A-K)h\big\|^2 = \|Ah\|^2 = \sum_{j=1}^{n+1} |\lambda_j c_j|^2 \geq \theta \sum_{j=1}^{n+1} |c_j|^2 = \theta \|h\|^2 \,. \tag{4.6}$$

This contradicts the assumption that $\|A-K\| < \theta$. **QED**

In our application, $\theta < 0.981$, and A is the restriction of $DG(u)$ to the subspace $\mathcal{B}^\sigma_\rho$. To be more specific: for the solutions $u = u_w, u_2, u_4, u_6, u_{2,2}, u_{2,4}$ described in Theorems 1.1, 1.2, and 1.3, we verify the assumptions of this proposition with $n = 2, 1, 1, 1, 2, 3$, respectively, on the even subspace, and with $n = 1, 1, 1, 1, 1, 2$, respectively, on the odd subspace.

Lemma 4.5 shows that n is an upper bound on the number of eigenvalues of A in the interval $[\theta, \infty)$. We need a lower bound as well; but only if $n > 1$, since A has an eigenvalue 3 or 1, as described earlier. Such a bound can obtained by using the following lemma, for some real number $a > 1$.

Lemma 4.6. *Let A be a compact self-adjoint linear operator on a Hilbert space $\mathcal{H}$. Let $\{v_1, v_2, \ldots, v_m\}$ be an orthonormal set in $\mathcal{H}$, and assume that*

$$A_{j,j} - \sum_{i \neq j} |A_{i,j}| > a \,, \qquad A_{i,j} = \langle v_i, A v_j \rangle \,, \quad 1 \leq i, j \leq m \,. \tag{4.7}$$

Then A has at least m eigenvalues (counting multiplicities) in the interval $[a, \infty)$.

This lemma is an immediate consequence of the Gershgorin circle theorem and Courant's min-max principle. We verify the hypotheses with $m = 2, 1, 1, 1, 2, 3$, respectively, for the solutions $u = u_w, u_2, u_4, u_6, u_{2,2}, u_{2,4}$ on the even subspace, and with $m = 1$ for the solution $u = u_{2,4}$ on the odd subspace. The value $a = 3$ works in all cases. More accurate eigenvalue bounds can be found in [25].

We remark that this method for determining the Morse index is significantly simpler, and more efficient, than the method used in [11].

5. Clebsch-Gordan coefficients

In this section we give a brief description of the identities and algorithms that we use to compute and index Clebsch-Gordan coefficients. For details we refer to the Ada packages `Regges` and `CG` in [25].

The Clebsch-Gordan coefficients $\langle \nu_1 \mu_1 \nu_2 \mu_2 | \nu_3 \mu_3 \rangle$ that appear in the product expansion (2.4) vanish unless the angular momenta ν_j and μ_j satisfy certain constraints. These constraints, as well as symmetries, are most conveniently described in terms of the Regge symbol

$$\begin{vmatrix} \nu_2+\nu_3-\nu_1 & \nu_3+\nu_1-\nu_2 & \nu_1+\nu_2-\nu_3 \\ \nu_1-\mu_1 & \nu_2-\mu_2 & \nu_3-\mu_3 \\ \nu_1+\mu_1 & \nu_2+\mu_2 & \nu_3+\mu_3 \end{vmatrix} = \frac{(-1)^{\nu_1-\nu_2+\mu_3}}{\sqrt{2\nu_3+1}} \big\langle \nu_1 \mu_1 \nu_2 \mu_2 | \nu_3 (-\mu_3) \big\rangle \,. \tag{5.1}$$

A Regge symbol $\boxed{R}$ vanishes unless its entries R_{ij} are all nonnegative integers, and unless the row sums $R_{i1}+R_{i2}+R_{i3}$ and column sums $R_{1j}+R_{2j}+R_{3j}$ all have the same value. Furthermore, its absolute value is invariant under interchanges of rows, interchanges of columns, and transposition [21]. Under odd row or column permutations of the matrix R, the symbol acquires a factor $(-1)^J$, where $J=\nu_1+\nu_2+\nu_3$.

A commonly used formula [22] for Wigner's $3j$ symbol, expressed here in terms of the Regge symbol, is

$$\boxed{R} = (-1)^{R_{12}-R_{33}} \sqrt{\frac{R_{11}!R_{12}!\cdots R_{2,3}!R_{3,3}!}{(J+1)!}} \sum_k \frac{(-1)^k}{Q_k(R)}\,, \tag{5.2}$$

where

$$Q_k(R) = k!(R_{12}-R_{21}+k)!(R_{11}-R_{32}+k)!(R_{31}-k)!(R_{21}-k)!(R_{32}-k)!\,. \tag{5.3}$$

The sum in (5.2) runs over all integers k such that the arguments of all factorials in (5.3) are nonnegative.

As can be seen from (5.2), the square of a Regge symbol is a rational number P/Q. In our programs, we compute P and Q exactly, following roughly a procedure described in [24]. The summands $(-1)^k/Q_k$ in the equation (5.2) are multiplied first by their least common multiple, so the sum becomes a sum of integers. Factorials and their products are computed in terms of their prime factorization.

This computation is too costly to be repeated whenever we need the value of a Regge symbol. Thus, we compute the necessary values beforehand and store them in a linear array. Due to the above-mentioned symmetries, it suffices to index Regge matrices of the form

$$\boxed{\mathcal{R}} = \boxed{\begin{array}{ccc} S & L & X+B-T \\ X & B & S+L-T \\ L+B-T & S+X-T & T \end{array}}\,, \tag{5.4}$$

with $L\ge X\ge T\ge B\ge S\ge 0$. As was shown in [23], any Regge matrix R with nonzero symbol can be transformed to such a normal form $\mathcal{R}$ via row and column permutations, and possibly a transposition. This eliminates half of the 72 symmetries.

Here, we eliminate much of the other half as well by requiring that $T\le (L+S)/2$. This can be achieved by exchanging the last two rows of $\mathcal{R}$, if necessary.

We index such Regge matrices by enumerating the set

$$\mathcal{S} = \big\{(l,s,t,x,b):\ \ l\ge x\ge t\ge b\ge s\ge 0\,,\ \ t\le (l+s)/2\,\big\}\,, \tag{5.5}$$

using the following lexicographical order: recursively, define $(u,v,w,\ldots)<(U,V,W,\ldots)$ to mean that either $u<U$, or else $u=U$ and $(v,w,\ldots)<(V,W,\ldots)$. Assuming that the quintuple (L,S,T,X,B) belongs to $\mathcal{S}$, the index of the Regge matrix (5.4) is defined to be the number of quintuples in $\mathcal{S}$ that are less than or equal to (L,S,T,X,B). This index can be expressed as

$$\mathcal{I}(\mathcal{R}) = \sum_{l\le L}\sum_{s\le S}\sum_{t<T}\sum_{x,b} \chi_{\mathcal{S}}(l,s,t,x,b) + C_{S,T,X,B}\,, \tag{5.6}$$

where $C_{S,T,X,B}$ is the sum over all $x \le X$ and $b \le B$ of the numbers $\chi_{\mathcal{S}}(L,S,T,x,b)$. Here $\chi_{\mathcal{S}}$ denotes the indicator function of the set $\mathcal{S}$. Notice that the five-fold sum in (5.6) defines a function of the three variables (L,S,T). In our programs, the values of this function are determined simply by having the computer carry out the five-fold sum. The values are then stored in a three-dimensional array. As for the values $C_{S,T,X,B}$, a straightforward computation shows that

$$C_{S,T,X,B} = (X-T)(T-S+1) + (B-S+1)\,. \tag{5.7}$$

6. Computer estimates

In order to complete our proof of Theorems 1.1, 1.2, and 1.3, we need to verify the assumptions of the lemmas in Section 4. This is done with the aid of a computer. For each of the six models considered, we have chosen $\rho = 65/64$.

To fix ideas, consider Lemma 4.1 for some given choice of $w \in \mathcal{B}_\rho$. To verify the assumptions of this lemma, we first determine an approximate fixed point $\bar{u}$ of G and an approximation M for the operator $\mathbb{I} - [\mathbb{I} - DG(\bar{u})]^{-1}$. These numerical data are included with the source code of our programs in [25]. The remaining steps are rigorous: First, we compute an upper bound ε on the norm of $\mathcal{N}(0)$. Using this bound, we define an increasing function d on the interval $[0,3/4]$ such that $d(K) > \varepsilon/(1-K)$ on this interval. Now we compute an upper bound K on the operator norm of $D\mathcal{N}(h)$ that holds for all h of norm $d(3/4)$ or less. After verifying that $K \le 3/4$, we set $\delta = d(K)$. This guarantees that $\varepsilon + K\delta < \delta$. We also verify that M has no eigenvalue 1.

The rigorous part is still numerical, but instead of truncating series and ignoring rounding errors, it produces guaranteed enclosures at every step along the computation. Our choice of enclosures and associated data types will be described below.

The above-mentioned steps are analogous to those used in the proof of Theorem 4.1 in [11]. The main difference is that [11] uses functions on the square and data of type `Fourier2`, while here we use functions on the disk and data of type `Zernike`. To avoid undue repetition, we will focus here on those aspects of the proof where the differences are relevant.

We will also describe our computation of the Morse index, which amounts to verifying the assumptions of the Lemmas 4.5 and 4.6. But any description given here is necessarily incomplete. For precise definitions and other details, the ultimate reference is the source code of our programs [25]. This code is written in the programming language Ada [26].

One of the basic data type in our programs is the type `Ball` that we use to define enclosures for real numbers. A data item of type `Ball` is a pair `B=(B.C,B.R)`, where `B.C` is a representable number (type `Rep`), and where `B.R` a nonnegative representable number (type `Radius`). The corresponding subset of $\mathbb{R}$ is the interval $\mathtt{B}^\flat = \{b \in \mathbb{R} : |b - \mathtt{B.C}| \le \mathtt{B.R}\}$. Using controlled rounding, it is trivial to implement e.g. a "`function Sum(A,B: Ball) return Ball`" with the property that $\mathtt{Sum}(\mathtt{A},\mathtt{B})^\flat$ contains $a+b$ whenever $a \in \mathtt{A}^\flat$ and $b \in \mathtt{B}^\flat$. Similarly for other elementary operations involving real numbers.

Next, we describe our enclosures for functions in $\mathcal{B}_\rho$ that belong to the even subspace $\mathcal{B}_\rho^0$ or to the odd subspace $\mathcal{B}_\rho^1$. The enclosures depend on the choice of a positive integer

`Size` which we denote here by S. Define $D = \lfloor S/2 \rfloor$. We start by considering functions $f : [0,1] \to \mathbb{R}$ with the property that

$$(J_\rho^{\pm m} f)(r,\vartheta) \stackrel{\text{def}}{=} f(r) e^{\pm i m \vartheta} \tag{6.1}$$

defines a function in $\mathcal{B}_\rho$. In this step, $\rho \geq 1$ and $m \geq 0$ are considered fixed, with $m \leq D$. Our enclosures for such functions f are associated with a data type `Radial`. A data item of type `Radial` is (in essence) a triple `F=(F.M,F.C,F.E)`, where `F.C` is an `array(0 .. D) of Ball`, `F.E` is an `array(0 .. D+1) of Radius`, and `F.M` $= m$. The corresponding set `F`$^\flat$ is the set of all function $f : [0,1] \to \mathbb{R}$ that admit a representation

$$f(r) = \sum_{j=0}^{D_m} C_j R^m_{m+2j}(r) + \sum_{j=0}^{D_m+1} E_j(r)\,, \qquad D_m = \lfloor (D-m)/2 \rfloor\,, \tag{6.2}$$

with $C_j \in$ `F.C(j)`$^\flat$ for $j \leq D_m$, and $\|J_\rho^m E_j\|_\rho \leq$ `F.E(j)` for $j \leq D_m + 1$. In addition, we require that the Zernike series for the functions $J_\rho^m E_j$ include only modes $V_{n'}^m$ with $n' \geq n$, where $n = m + 2j$. Notice that the coefficient array `F.C` specifies a set of polynomials of degree $\leq m + 2D_m \leq D$. The numbers in `F.E` represent error bounds.

An item of type `Zernike` is a quadruple `U=(U.R,U.P,U.C,U.E)`, where `U.R` ≥ 1 is of type `Radius`, `U.C` is an `array(0 .. S) of Radial` with `U.C(m).M=m` fixed for each `m`, `U.E` is an `array(0 .. 2*S) of Radius`, and `U.P` is either 0 or 1. If `U.P` $= 1$ then `U` defines a subset `U`$^\flat$ of $\mathcal{B}^1_\rho$ with $\rho =$ `U.R`. Consider now `U.P` $= 0$. In this case, `U` defines a subset `U`$^\flat$ of $\mathcal{B}^0_\rho$ with $\rho =$ `U.R`. This set consists of all functions

$$u = \sum_{m=0}^{S} \frac{1}{2}\big(J_\rho^m + J_\rho^{-m}\big) f_m + \sum_{m=0}^{2S} E_m\,, \qquad E_m \in \mathcal{B}^0_\rho\,, \tag{6.3}$$

with $f_m \in$ `U.C(m)`$^\flat$ for $0 \leq m \leq S$, and $\|E_m\|_\rho \leq$ `U.E(m)` for $0 \leq m \leq 2S$. In addition, we require that the Zernike series for E_m include only modes $V_n^{m'}$ with $m' \geq m$. Our enclosures for $\mathcal{B}^1_\rho$ are defined analogously.

`Zernike`-type sets `U`$^\flat$ play the same role for functions in $\mathcal{B}^0_\rho \cup \mathcal{B}^1_\rho$ as `Ball`-type sets `B`$^\flat$ play for real numbers. It is trivial to implement e.g. a "`function Sum(U,V: Zernike) return Zernike`" with the property that `Sum(U,V)`$^\flat$ contains $u + v$ whenever $u \in$ `U`$^\flat$ and $v \in$ `V`$^\flat$, provided that `U.P=V.P`. Implementing a bound on the product of two such functions is a bit more involved. Here we use the Banach algebra property of $\mathcal{B}_\rho$ and enclosures for the Clebsch-Gordan coefficients. For details we refer to the Ada package `Zernikes` in [25]. This package also implements a bound `InvNegLap` on the operator $(-\Delta)^{-1}$, using estimates of the type (3.7) for the error terms in (6.2) and (6.3).

More problem-specific operations are defined in `Zernikes.GFix`, including bounds `GMap` and `DGMap` on the map G and its derivative $DG(u)$, respectively. Our proof of Lemma 4.1 is organized by the procedure `ContrFix`, using `DContrNorm` to obtain a bound on the operator norm of $D\mathcal{N}(h)$. The steps are as described at the beginning of this section. This applies to each of the solutions u_w, u_α, and $u_{2,n}$, described in Theorems

1.1, 1.2, and 1.3, respectively. For the solutions u_w and u_α we also verify the assumptions of Lemma 4.2, and for $u_{2,n}$ we verify the assumptions of Lemma 4.3. The details can be found in [25].

What remains to be discussed is the computation of the Morse index. Using Lemmas 4.5 and 4.6, with A being the restriction of $DG(u)$ to $\mathcal{B}^\sigma_\rho$, this task is relatively straightforward. The computations for $\sigma = 0$ and for $\sigma = 1$ are carried out separately. And this is done for each of the six models being considered.

Among the data included in [25] are approximate eigenvectors of A. They define a self-adjoint approximation K for the operator $\mathcal{K}$ described in (4.5). A bound on the the map $\mathcal{L} = A - K$ is implemented by the procedure `LLMap`. In order to estimate the spectral radius of $\mathcal{L}$, as required by Lemma 4.5, we first construct an enclosure `L` for the operator $\mathcal{L} : \mathcal{B}^\sigma_\rho \to \mathcal{B}^\sigma_\rho$, iterate $\mathtt{L} \mapsto \mathtt{L}^2$ several times, and then estimate the operator norm of the result. The inequalities (4.7) needed for Lemma 4.6 are verified in the procedure `KBound`. This procedure first orthonormalizes the approximate eigenvectors that were used to define the operator K.

In order to construct operator enclosures, we use some data types and procedures from `Zernikes` that we have not yet described. Notice that a `Zernike U` can be viewed as a collection of "coefficient modes" `U.C(m).C(j)` and "error modes" `U.C(m).E(j)` or `U.E(m)`. Coefficient modes represent one-dimensional subspaces of $\mathcal{B}^\sigma_\rho$, while error modes represent infinite-dimensional subspaces. To specify individual modes we use a data type `ZMode`. We are interested mostly in finite collections of modes whose subspaces Z_i define a partition of $\mathcal{B}^\sigma_\rho$, in the sense that $\bigoplus_i Z_i = \mathcal{B}^\sigma_\rho$, and that $Z_i \cap Z_j = \{0\}$ for $i \neq j$. Such a "partition" is specified by our data type `ZModes`. Our linear operator $\mathcal{L} : \mathcal{B}^\sigma_\rho \to \mathcal{B}^\sigma_\rho$ now defines a "matrix" of operators $\mathcal{L}_{i,j} : Z_j \to Z_i$. By an enclosure for $\mathcal{L}$ we mean a corresponding matrix of bounds, with each element being a `Ball`. To be more precise, we restrict to `ZModes` that allow a `Zernike` to be distributed efficiently over the individual modes, using the procedure `Extract`. Then a bound $\mathtt{L}_{i,j}$ on $\mathcal{L}_{i,j}$ is obtained in essence by applying `LLMap` to the j-th `ZMode` and extracting the i-th `ZMode` from the result.

All major steps that are used to verify the assumptions of the five lemmas in Section 4 are implemented in the procedures described above. They are combined in the proper order, and invoked with the appropriate parameters, by the main program `Run_All`. Instructions on how to compile and run this program are in a file `README` that is included with the source code of our programs in [25]. The programs `Find_Fix` and `Find_Eigen` that were used to compute our numerical data are included as well.

The parameter `Size` that determines the size of our `Zernike`-type data ranges from 70 to 140, depending on the computation. For the set of representable numbers (`Rep`) we choose standard extended floating-point numbers [28] that support controlled rounding, and for bounds on non-elementary `Rep`-operations we use the open source MPFR library [29]. Our programs were run successfully on a standard desktop machine, using a public version of the gcc/gnat compiler [27].

References

[1] G. Chen, W.-M. Ni, J. Zhou, *Algorithms and visualization for solutions of nonlinear elliptic equations*, Int. J. Bifurc. Chaos **10**, 1565–1612 (2000).

[2] D. Smets, J. Su, M. Willem, *Non radial ground states for the Hénon equation*, Comm. Contemp. Math. **4** 467–480 (2002).
[3] F. Pacella, T. Weth, *Symmetry of solutions to semilinear elliptic equations via Morse index*, Proc. Amer. Math. Soc. **135**, 1753–1762 (2007).
[4] M. Badiale, G. Cappa, *Non radial solutions for non homogeneous Hénon equation*, Nonlinear Analysis **109**, 45–55 (2014).
[5] T. Koornwinder, *Two-variable analogues of the classical orthogonal polynomials*, in: Theory and application of special functions, R.A. Askey (ed.), Academic Press, 1975, pp. 435-495.
[6] E.C. Kintner, *On the mathematical properties of the Zernike Polynomials*. Opt. Acta. **23** 679–680 (1976).
[7] W.J. Tango, *The Circle Polynomials of Zernike and Their Application in Optics*, Appl. Phys. **13** 327–332 (1977).
[8] Y. Kanjin, *Banach algebra related to disk polynomials*, Tôhoku Math. Journ. **37** 395–404 (1985).
[9] A. Wünsche, *Generalized Zernike or disc polynomials*, J. Comput. Appl. Math. **174**, 135–163 (2005).
[10] A.J.E.M. Jansen, *Zernike expansion of derivatives and Laplacians of the Zernike circle polynomials*, J. Opt. Soc. Am. A **31**, 1604–1613 (2014).
[11] G. Arioli, H. Koch, *Non-symmetric low-index solutions for a symmetric boundary value problem*, J. Diff. Equations **252**, 448–458 (2012).
[12] G. Arioli, H. Koch, *Some symmetric boundary value problems and non-symmetric solutions*, J. Diff. Equations **259**, 796–816 (2015).
[13] J. Cyranka, P. Zgliczyński, *Existence of globally attracting solutions for one-dimensional viscous Burgers equation with nonautonomous forcing - a computer assisted proof*, SIAM J. Appl. Dyn. Syst. **14**, 787–821 (2015).
[14] Y. Watanabe, M.T. Nakao, *A numerical verification method for nonlinear functional equations based on infinite-dimensional Newton-like iteration*, Appl. Math. Comput. **276** 239–251 (2016).
[15] R. Castelli, J.-P. Lessard, J.D. Mireles-James, *Parameterization of invariant manifolds for periodic orbits (II): A posteriori analysis and computer assisted error bounds*, J. Dyn. Diff. Equat. (2017). `https://doi.org/10.1007/s10884-017-9609-z`
[16] J.L. Figueras, A. Haro, *Rigorous computer assisted application of KAM theory: a modern approach*, A. Found. Comput. Math. **17**, 1123–1193 (2017).
[17] J.-L. Figueras and R. de la Llave, *Numerical computations and computer assisted proofs of periodic orbits of the Kuramoto-Sivashinsky equation*, SIAM J. Appl. Dyn. Syst. **16**, 834–852 (2017).
[18] I. Balázs, J.B. van den Berg, J. Courtois, J. Dudás, J.-P. Lessard, A. Vörös-Kiss, J.F. Williams, X.Y. Yin, Computer-assisted proofs for radially symmetric solutions of PDEs, preprint (2017)
[19] G. Arioli, H. Koch, *Spectral stability for the wave equation with periodic forcing*, Preprint `mp_arc 17-23`.
[20] For earlier work, see references in [11-19].
[21] T. Regge, *Symmetry properties of Clebsch-Gordans coefficients*, Nuovo Cimento **10**,544–545 (1958).
[22] M. Rotenberg, R. Bivins, N. Metropolis, and J.K. Wooten, *The 3j and 6j symbols*, Cambridge, MA: MIT Press, 1959.
[23] J. Rasch and A.C.H. Yu, *Efficient storage scheme for precalculated Wigner* $3J$*,* $6J$ *and Gaunt coefficients*, Siam J. Sci. Comput **25**, 1416–1428 (2003).

[24] H.T. Johansson and C. Forssén, *Fast and accurate evaluation of Wigner* $3j$*,* $6j$*, and* $9j$ *symbols using prime factorisation and multi-word integer arithmetic*, SIAM J. Sci. Comput. **38**, A376A384 (2016)
[25] G. Arioli, H. Koch, The computer programs and data files are available at `http://www.ma.utexas.edu/users/koch/papers/zerni/`
[26] Ada Reference Manual, ISO/IEC 8652:2012(E), available e.g. at `http://www.ada-auth.org/arm.html`
[27] A free-software compiler for the Ada programming language, which is part of the GNU Compiler Collection; see `http://gnu.org/software/gnat/`
[28] The Institute of Electrical and Electronics Engineers, Inc., *IEEE Standard for Binary Floating–Point Arithmetic*, ANSI/IEEE Std 754–2008.
[29] The MPFR library for multiple-precision floating-point computations with correct rounding; see `http://www.mpfr.org/`.