SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 1–17
© 2008 Society for Industrial and Applied Mathematics

J_2 Effect and Elliptic Inclined Periodic Orbits in the Collision Restricted Three-Body Problem*

E. Barrabés†, J. M. Cors‡, C. Pinyol§, and J. Soler†

Abstract. The existence of a new class of inclined periodic orbits of the collision restricted three-body problem is shown. The symmetric periodic solutions found are perturbations of elliptic Kepler orbits, and they exist only for special values of the inclination and are related to the motion of a satellite around an oblate planet.

Key words. collision restricted three-body problem, periodic orbits, symmetric orbits, critical inclination, continuation method

AMS subject classifications. 70F07, 70F15, 70H09, 70H12, 70M20

DOI. 10.1137/070683854

1. Introduction.

The launch of Sputnik in October 1957 opened the space age. The use of circular, elliptic, and synchronous orbits, combined with dynamical effects due to the Earth's equatorial bulge, gives rise to an array of orbits with specific properties to support various mission constraints. One example is the Molniya orbit, a highly elliptic 12-hour-period orbit the former USSR originally designed to observe the northern hemisphere. The orbital plane makes an angle of about 63 degrees with the equatorial plane of the Earth, and this is the only value that prevents the orbit itself from rotating slowly within its plane and around the focus.

In what follows we will introduce briefly a few common notions of orbital dynamics, together with the current terminology (sometimes a few centuries old), and state the aim of the paper.

The position of a body on a Keplerian elliptic orbit can be completely characterized by six parameters. One such set of parameters are the classical orbital elements. As the orbital plane is fixed in any inertial frame and passes through the origin, one should first give the position of this plane. In a Cartesian frame with axes xyz, this is given by the inclination i with respect to the xy-plane and the angle Ω from the positive x-axis to the intersection of the orbital plane with the xy-plane. In the classical terminology of astronomy this line is

*Received by the editors February 27, 2007; accepted for publication (in revised form) by J. Meiss May 26, 2007; published electronically January 16, 2008.
http://www.siam.org/journals/siads/7-1/68385.html

†Departament d'Informàtica i Matemàtica Aplicada, Universitat de Girona, Girona 17003, Spain (barrabes@ima.udg.edu, jaume.soler@ima.udg.edu). The first author was supported by grant MTM2006-05849/Consolider (including a FEDER contribution). The last author was supported by grant MTM2005-07660-C02-02.

‡Departament de Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona 08242, Spain (cors@epsem.upc.edu). This author was supported by MCYT grant MTM 2005-06098-C02-01 and by CIRIT grant 2001SGR00173.

§Departament d'Economia i Història Econòmica, Universitat Autònoma de Barcelona, Barcelona 08193, Spain (conxita.pinyol@uab.cat). This author was supported by grant SEJ2006/00712ECO.

known as the *line of nodes* (the nodes of the orbit being the two points of intersection with the xy-plane, and the ascending node that in which the body crosses from $z < 0$ to $z > 0$), and Ω is called the *longitude of the ascending node*.

Then we need the position of the ellipse on its plane. One focus is at the origin, and the line joining the pericenter and the apocenter (classically the *line of apsides*) forms an angle ω with the line of nodes which gives the position of the ellipse. Usually the half-line from the origin to the pericenter and the half-line from the origin to the ascending node are taken, and then we say that ω is the *longitude of the pericenter*.

The size and shape of the ellipse are given by the semimajor axis a (directly related to the energy) and the eccentricity e (related to the energy and the angular momentum).

Finally, the position of the body along the orbit is given either by the angle f (true anomaly) pericenter–origin–body, or by some other related angles such as E (eccentric anomaly) or M (mean anomaly). The three anomalies (a name already used in Greek astronomy) are related among themselves by the geometry and the dynamics of Keplerian motion. The position is actually a function of time, and the origin of time is called the *epoch* (see, for example, [2]).

Of course, Ω, ω, and f are not well defined for circular or zero-inclination orbits, a problem that can be solved in a variety of ways which go back to Laplace in the case of the classical elements and to Poincaré for the Hamiltonian formulation.

Thus the position and velocity of a point in space are completely characterized by the six orbital elements, which are constant (except f, or whichever anomaly is used) for a Keplerian orbit. This rather strange system of coordinates in phase space is useful because in most cases the non-Keplerian motion of a body subject to perturbations can be seen as a fast motion along a Keplerian orbit with slowly varying elements.

A set of variables closely related to the orbital elements are the *Delaunay elements*, which could be considered as the canonical (in the sense of Hamiltonian) version of the classical elements and will be defined in section 2.

Any small perturbation of the Keplerian motion has two kinds of effects on the motion: periodic and secular. An element subject to periodic perturbations simply oscillates around its central unperturbed value, while a secular perturbation is a steady, linear increase or decrease of its value. Of course, this is true only in a first order approximation, and it is a qualitative description, because a first order approximation is valid only on a finite interval of time, and the very concept of periodicity does not make sense. As for the secular effects we must remember that one of the major problems of the classical dynamical astronomy of the nineteenth century was the distinction between true secular effects and linearization of periodic effects of very long period, and that the whole matter has been settled only by the KAM theory.

In this sense, it is a result of classical astronomy that a, e, and i are subject to only periodic effects, while Ω, ω, and M display periodic and secular effects. In short, a perturbed Keplerian conic can be thought of, roughly speaking, as a conic which rotates slowly on its plane while the plane itself rotates around the z-axis.

The most common perturbations of the potential in celestial mechanics are due either to the presence of a third body or to lack of sphericity of the bodies. The latter can be dealt with by expanding the potential in spherical harmonics, so that if the body has axial symmetry,

the potential can be seen as that of an inverse square distance central force plus other terms:

$$V(r,\phi) = -Gm\frac{r_{eq}}{r}\left(1 - \sum_{k=2}^{\infty} J_k \left(\frac{r_{eq}}{r}\right)^k P_k(\cos\phi)\right),$$

where G is the gravitational constant, m is the mass of the body, r_{eq} is its equatorial radius, (r,θ,ϕ) are spherical coordinates (θ does not appear because of the axial symmetry), P_k is the kth Legendre polynomial, and J_k are the coefficients defining the expansion (see, for example, [10]).

The third body perturbation is quite a different matter because the equations of motion must be supplemented with those of the new body. In the restricted three-body problem, it is assumed that one of the bodies is so small that it does not affect the motion of the other two (the *primaries*), and then we usually have a Keplerian motion plus a nonautonomous perturbation. If, however, we consider the potential in a region far away from the primaries and normalize this distance to unity, the velocities of the primaries are very high and heuristically we can somehow average their mass along their whole orbits, so that dynamically we are again in the case of a nonspherical potential.

For Earth-orbit design, the main perturbation is that of the J_2 term in the expansion of the potential of an oblate ellipsoid. This term perturbs the orbit in the sense explained above, resulting in a precession both of the line of nodes and of the pericenter. It is apparent (see, for example, pages 503–504 of [2]) that there exists a critical inclination angle, $i \simeq 63^\circ$, such that the perigee is fixed in the first approximation because its secular terms are of opposite sign for inclinations above or under the critical value, irrespective of the eccentricity. The case of a prolate ellipsoid, though apparently not frequent in astronomy, could be treated in the same way, the only difference being that the J_2 term has the opposite sign, so that all the precessions are in the opposite direction.

The existence of a class of inclined periodic solutions of the circular three-body problem was shown by Jefferys in [5]. He showed the existence of families of elliptic orbits with inclination close to critical for any value of the eccentricity. His proof rests on a mirror theorem: in the rotating coordinate system of the restricted three-body problem any trajectory that hits twice perpendicularly a certain plane is a periodic solution. For an elliptic orbit, perpendicular crossing means that the body is at either the pericenter or the apocenter and the line of apsides lies on the mirror plane; for this situation to happen twice in time it is sufficient that the line of apsides does not have a secular motion, so the inclination must be near critical. Of course a precession of the line of nodes does exist, but it is hidden, as it were, in the rotating frame. It must be borne in mind that in celestial mechanics periodic usually means periodic in some rotating frame, and thus periodic or quasi-periodic in the inertial frame depending on whether the angle advanced by the rotating frame in a period is a rational multiple of π or not. The method used is the continuation method developed by Poincaré (see [8]), which is one of the most frequently used methods for proving the existence of periodic orbits.

The case dealt with in this paper is different from Jefferys's because the primaries move on an elliptic collision orbit along the z-axis. Heuristically speaking, however, it can be expected that far away from the primaries the potential will be similar to that of a very eccentric prolate

ellipsoid, so that a J_2 effect, with its critical inclination, will exist. We show the existence of periodic solutions of Jefferys type: large semiaxis compared to the that of the primaries, arbitrary eccentricity, and inclination close to critical.

The problem can be seen as a perturbed Kepler problem, where the small parameter is the semimajor axis of the primaries' orbit after rescaling. The perturbed problem is degenerate due to the fast motion of the primaries, and the equations are no longer analytic when the parameter equals zero, which precludes the use of a standard implicit function theorem. We overcome the difficulty by using Arenstorf's theorem, where weaker assumptions of differentiability are needed (see [1]). A planar configuration of this problem is studied in [6].

In our case, the problem has a rotational symmetry around the z-axis (which contains the colliding primaries). This symmetry would be lost if we considered elliptic noncollision orbits for the primaries. See [3] and [4], where the elliptic restricted three-body problem is considered. In those papers, the periodic orbits are perturbations of the circular solutions of the Kepler problem having large radii on a plane perpendicular to that of the primaries. Periodic orbits in the spatial elliptic restricted three-body problem are also studied using double averaging in [7].

The paper is organized as follows. Section 2 describes the general setting of the collision restricted three-body problem. Section 3 shows how its solutions can be approximated through successive corrections to Keplerian motion. Section 4 deals with the continuation problem. The main result is the existence of quasi-elliptic orbits for discrete values of the semimajor axis of the primaries, with arbitrary eccentricity and inclination close to critical. A number of technical computations are presented in section 5.

Figure 1 shows one of the orbits predicted by the main theorem, numerically computed with initial values $r_0 = 0.621114405$, $\phi_0 = 1.116457610$, $\theta_0 = 0$, $p_r = 0$, $p_\phi = 0$, $p_\theta = 0.4098780306$, and $\mu = 30^{-2/3}$. The equations of motion and the first variational equations were numerically integrated with a Runge–Kutta 7-8 routine, and the equations defining the initial conditions were solved with a Newton method starting with the Keplerian orbit with $a = 1$, $e = 0.4$, and $\cos i = 1/\sqrt{5}$ (critical inclination).

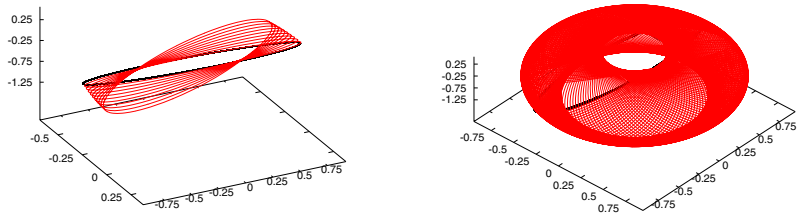


Figure 1. *Example of a quasi-periodic orbit in a Cartesian frame. The orbit is followed during* 6 *(plot on the left) and* 150 *(plot on the right) times the period of the primaries. The Keplerian orbit for* $\mu = 0$ *is plotted (black line). The primaries move along the vertical line passing through the origin.*

2. The collision restricted three-body problem.

The collision restricted three-body problem describes the motion of a massless particle under the attraction of two primaries with equal

masses, $m_1 = m_2 = 1/2$, moving on a collision elliptic orbit. In order to avoid a triple collision, we consider that the third body is far from the primaries compared to the distance between them. This fact can be introduced in the equations of motion by making the primaries very close to each other and looking for solutions of the massless particle at distance of order unity to the primaries.

Let μ be a small parameter. The distance between both primaries is given by

$$\rho = \mu(1 - \cos E_p(t)), \tag{2.1}$$

where $E_p = E_p(t)$ is the eccentric anomaly of m_1 and it is related to its mean anomaly ℓ_p through Kepler's equation

$$E_p - \sin E_p = \ell_p, \tag{2.2}$$

where $\ell_p = \mu^{-3/2}t$. The period of the motion of the primaries is $T_p = 2\pi\mu^{3/2}$, so that $E_p = k\pi$ when $t = \pi k\mu^{3/2}$.

Equation (2.2) is a particular case (for $e = 1$) of Kepler's equation $\ell = E - e\sin E$, where e is the eccentricity, ℓ is the mean anomaly (real time measured in such units that the period is 2π), and E is the eccentric anomaly, which is related to the angular position f of the body on the orbit (from the pericenter) through $\tan f/2 = \sqrt{(1+e)/(1-e)}\tan E/2$. The latter equation results from the geometry of elliptic orbits, and Kepler's equation is just the mathematical expression of the law of areas, i.e., the conservation of the angular momentum (see [10]).

We consider a fixed coordinate system (q_1, q_2, q_3) (see Figure 2) with origin at the center of mass of m_1 and m_2 in such a way that the primaries move along the q_3-axis. Their positions are given by $\mathbf{r}_1 = (0, 0, \frac{\mu}{2}(1 - \cos E_p))$ and $\mathbf{r}_2 = -\mathbf{r}_1$. Let $\mathbf{q} = (q_1, q_2, q_3)$ and $\mathbf{p} = (p_1, p_2, p_3) = (\dot q_1, \dot q_2, \dot q_3)$ be the position and momentum of the infinitesimal body m_3. The problem of describing its motion is known as the *three-dimensional collision restricted three-body problem*.

The equations of motion for the infinitesimal body can be written as a nonautonomous Hamiltonian system depending on the parameter μ as

$$\dot q_i = \frac{\partial \mathcal{H}}{\partial p_i}, \qquad \dot p_i = -\frac{\partial \mathcal{H}}{\partial q_i}, \qquad i = 1, 2, 3,$$

where

$$\mathcal{H} = \frac{1}{2}(p_1^2 + p_2^2 + p_3^3)^2 - \frac{1}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right), \tag{2.3}$$

and R_1 and R_2 are given by

$$\begin{aligned} R_1^2 &= q_1^2 + q_2^2 + \left(q_3 - \tfrac{\mu}{2}(1 - \cos E_p)\right)^2, \\ R_2^2 &= q_1^2 + q_2^2 + \left(q_3 + \tfrac{\mu}{2}(1 - \cos E_p)\right)^2. \end{aligned}$$

Let us introduce spherical coordinates (r, ϕ, θ), and (p_r, p_ϕ, p_θ) by means of the canonical change

$$\begin{aligned} q_1 &= r\cos\phi\cos\theta, & p_1 &= p_r\cos\phi\cos\theta - \tfrac{p_\phi}{r}\sin\phi\cos\theta - \tfrac{p_\theta}{r\cos\phi}\sin\theta, \\ q_2 &= r\cos\phi\sin\theta, & p_2 &= p_r\cos\phi\sin\theta - \tfrac{p_\phi}{r}\sin\phi\sin\theta + \tfrac{p_\theta}{r\cos\phi}\cos\theta, \\ q_3 &= r\sin\phi, & p_3 &= p_r\sin\phi + \tfrac{p_\phi}{r}\cos\phi. \end{aligned}$$

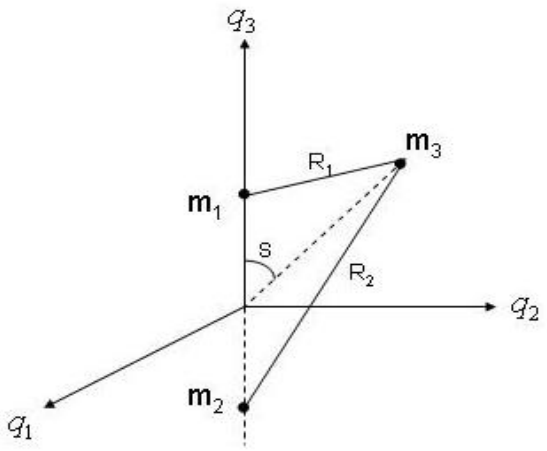


Figure 2. *Collision restricted three-body problem.*

In the new variables, the Hamiltonian (2.3) becomes

$$H = \frac{1}{2}\left(p_r^2 + \frac{p_\phi^2}{r^2} + \frac{p_\theta^2}{r^2\cos^2\phi}\right) - \frac{1}{2}\left(\frac{1}{R_1} + \frac{1}{R_2}\right), \tag{2.4}$$

with R_1 and R_2 given by

$$\begin{aligned} R_1^2 &= r^2 + \left(\tfrac{\mu}{2}\right)^2 (1-\cos E_p)^2 - r\mu(1-\cos E_p)\sin\phi, \\ R_2^2 &= r^2 + \left(\tfrac{\mu}{2}\right)^2 (1-\cos E_p)^2 + r\mu(1-\cos E_p)\sin\phi. \end{aligned} \tag{2.5}$$

Notice that E_p as given by (2.2) is a function of time t and μ, which is not defined for $\mu = 0$. So, neither the Hamiltonian (2.3) nor (2.4) is defined.

The equations of motion for the infinitesimal mass in spherical coordinates are

$$\begin{aligned} \dot r &= p_r, & \dot p_r &= -\tfrac{\partial \mathcal{H}}{\partial r}, \\ \dot\theta &= \tfrac{p_\theta}{r^2\cos^2\phi}, & \dot p_\theta &= 0, \\ \dot\phi &= \tfrac{p_\phi}{r^2}, & \dot p_\phi &= -\tfrac{\partial \mathcal{H}}{\partial \phi}. \end{aligned} \tag{2.6}$$

Since R_1 and R_2 do not depend on θ, $\dot p_\theta = 0$ and the angular momentum $p_\theta = \Theta$ is constant. Thus, it can be calculated from the initial conditions, and the equation for θ can be decoupled from the other equations. In this way, we can consider the system of equations

$$\begin{aligned} \dot r &= p_r, & \dot p_r &= -\tfrac{\partial \mathcal{H}}{\partial r}, \\ \dot\phi &= \tfrac{p_\phi}{r^2}, & \dot p_\phi &= -\tfrac{\partial \mathcal{H}}{\partial \phi}. \end{aligned} \tag{2.7}$$

Once r and ϕ are obtained, we will get θ from its equation in (2.6).

From now on, *reduced problem* means the problem given by (2.7), and *complete problem* means the whole set of equations (2.6). Our aim is to find periodic solutions of the reduced problem that will be periodic or quasi-periodic solutions of the complete problem.

It is easy to see that the equations of the reduced problem are invariant by the symmetry

$$S : (t, r, \phi, p_r, p_\phi, E_p) \quad \longrightarrow \quad (-t, r, \phi, -p_r, -p_\phi, -E_p),$$

so that, if

$$\gamma(t) = (r(t), \phi(t), p_r(t), p_\phi(t), E_p(t))$$

is a particular solution of (2.7), then so is

$$(r(-t), \phi(-t), -p_r(-t), -p_\phi(-t), -E_p(-t)),$$

and we have the following well-known result.

Proposition 2.1. *Let $\gamma(t) = (r(t), \phi(t), p_r(t), p_\phi(t), E_p(t))$ be a solution of the reduced problem given by (2.7). If $\gamma(t)$ satisfies $(p_r(0), p_\phi(0)) = (0, 0)$, $(p_r(T/2), p_\phi(T/2)) = (0, 0)$, and $E_p(T/2) = k\pi$, then $\gamma(t)$ is a periodic solution of period T.*

In order to find elliptic orbits we will introduce Delaunay variables (l, g, h) and (L, G, H), where

$$L = \sqrt{a}, \qquad H = G\cos i, \qquad G = \sqrt{a(1-e^2)},$$

a is the semimajor axis of the infinitesimal mass, G its angular momentum, i the inclination of its orbital plane with respect to the q_1q_2 reference plane, l the mean anomaly, g the argument of the pericenter measured from the ascending node, and h the longitude of the ascending node (see, for example, [9]).

We will use the symmetry conditions stated in Proposition 2.1 to obtain periodic solutions of the reduced problem. These conditions can be expressed in Delaunay variables as

$$l(t) = 0 \mod \pi, \qquad g(t) = \pi/2 \mod \pi \tag{2.8}$$

for epochs $t = 0$ and $t = T/2$, where $T = 2k\pi\mu^{3/2}$ in order to satisfy $E_p(T/2) = k\pi$.

3. Approximate solutions.

In this section we will show how those solutions of the three-dimensional collision elliptic restricted three-body problem in which the infinitesimal body keeps moving far away from the primaries can be approximated through successive corrections to the Keplerian motion. In section 4, we will use these approximations to continue some elliptic solutions of the Kepler problem to the case $\mu \neq 0$.

As the Hamiltonian (2.3) is not defined when $\mu = 0$, instead of expansions in power series (which are no longer available) we use asymptotic series. Using expressions (2.5) and (2.1), we can write

$$R_1 = r\sqrt{1 + \frac{\rho^2}{4r^2} - \frac{\rho}{r}\cos\mathcal{S}},$$

where $\mathcal{S} = \frac{\pi}{2} - \phi$ is the angle between the position vectors of m_1 and m_3 (see Figure 2). We assume that the distance from the origin to the primaries $(\mu/2)$ is small compared to the

distance from the origin to the infinitesimal body, so that $\rho \ll r$ and we can expand R_1^{-1} as a power series in $\frac{\rho}{2r}$ by using Legendre polynomials. Then

$$\frac{1}{R_1} = \frac{1}{r}\sum_{j=0}^{\infty} P_j(\cos\mathcal{S})\left(\frac{\rho}{2r}\right)^j = \frac{1}{r}\left[1 + \sum_{j=1}^{\infty} \mu^j P_j(\cos\mathcal{S})\left(\frac{1-\cos E_p}{2r}\right)^j\right],$$

where $P_j(\cos\mathcal{S})$ is the jth Legendre polynomial. Expanding R_2^{-1} in a similar way, the Hamiltonian (2.4) becomes

$$\mathcal{H}(\mathbf{q},\mathbf{p},t,\mu) = \mathcal{H}_0(\mathbf{q},\mathbf{p}) + \mu^2\mathcal{H}_1(\mathbf{q},\mathbf{p},t,\mu) + \mu^4\mathcal{H}_R(\mathbf{q},\mathbf{p},t,\mu), \tag{3.1}$$

where

$$\mathcal{H}_0(\mathbf{q},\mathbf{p}) = \frac{1}{2}\left(p_r^2 + \frac{p_\phi^2}{r^2} + \frac{p_\theta^2}{r^2\cos^2\phi}\right) - \frac{1}{r},$$

$$\mathcal{H}_1(\mathbf{q},\mathbf{p},t,\mu) = \frac{-1}{r}P_2(\cos S)\left(\frac{1-\cos E_p}{2r}\right)^2 = \frac{(1-\cos E_p)^2}{8r^3}\left(1-3\cos^2\mathcal{S}\right),$$

and

$$\mathcal{H}_R(\mathbf{q},\mathbf{p},t,\mu) = \frac{-1}{r}\sum_{k=2}^{\infty}\mu^{2(k-2)}P_{2k}(\cos\mathcal{S})\left(\frac{1-\cos E_p}{2r}\right)^{2k}.$$

The dependence on (t,μ) comes from the eccentric anomaly $E_p = E_p(t,\mu)$ given by (2.2). Notice that if $r \geq \delta$ for some fixed $\delta > 0$, then $\mathcal{H}_1(\mathbf{q},\mathbf{p},t,\mu)$ and $\mathcal{H}_R(\mathbf{q},\mathbf{p},t,\mu)$ are bounded. Thus, $\mu^2\mathcal{H}_1$ and $\mu^4\mathcal{H}_R$ are continuous at $\mu = 0$, although $\mathcal{H}_1$ and $\mathcal{H}_R$ are not so. This is the reason why expansions as power series in μ cannot be used.

Let us denote $\mathbf{z} = (l,g,h,L,G,H)$. Applying the corresponding symplectic change of variables, Hamiltonian (3.1) becomes

$$\mathcal{H}(\mathbf{z},t,\mu) = \mathcal{H}_0(\mathbf{z}) + \mu^2\mathcal{H}_1(\mathbf{z},t,\mu) + \mu^4\mathcal{H}_R(\mathbf{z},t,\mu), \tag{3.2}$$

where

$$\mathcal{H}_0(\mathbf{z}) = -\frac{1}{2L^2}, \tag{3.3}$$

$$\mathcal{H}_1(\mathbf{z},t,\mu) = \frac{(1-\cos E_p)^2}{8r^3}\left[1 - 3\left(1-\frac{H^2}{G^2}\right)\sin^2(g+f)\right]. \tag{3.4}$$

In (3.4) we have used the true anomaly f of the motion of the infinitesimal mass in order to write $q_3 = r\sin i\sin(f+g)$ and

$$\cos^2(\mathcal{S}) = \frac{q_3^2}{r^2} = \sin^2 i\sin^2(f+g) = \left(1-\frac{H^2}{G^2}\right)\sin^2(g+f).$$

Observe that $\mathcal{H}_0$ is the Hamiltonian of the Kepler problem and that, despite that Hamiltonian (3.2) is not defined for $\mu = 0$, the limit when $\mu \to 0$ exists and

$$\lim_{\mu\to 0}\mathcal{H}(\mathbf{z},t,\mu) = \mathcal{H}_0(\mathbf{z}).$$

Therefore, the equations of motion can be written as

$$\dot{\mathbf{z}} = \mathcal{F}(\mathbf{z}, t, \mu), \tag{3.5}$$

where $\mathcal{F} = J \cdot \nabla \mathcal{H}$, $J = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$, and I is the identity matrix of dimension 3×3. Using (3.2), the vector field $\mathcal{F}$ is given by

$$\mathcal{F}(\mathbf{z}, t, \mu) = \mathcal{F}_0(\mathbf{z}) + \mu^2 \mathcal{F}_1(\mathbf{z}, t, \mu) + \mu^4 \mathcal{F}_R(\mathbf{z}, t, \mu),$$

where

$$\begin{aligned} \mathcal{F}_0(\mathbf{z}) &= (L^{-3}, 0, 0, 0, 0, 0)^t, \\ \mathcal{F}_1(\mathbf{z}, t, \mu) &= J \cdot \nabla \mathcal{H}_1, \\ \mathcal{F}_R(\mathbf{z}, t, \mu) &= J \cdot \nabla \mathcal{H}_R, \end{aligned}$$

and $\mathcal{H}_1$ and $\mathcal{H}_R$ are the terms in (3.2).

The next lemma shows that the solutions of (3.5) can be written as the solutions of the Kepler problem plus terms of order μ^2, and the same is true of its partial derivatives with respect to the initial conditions.

Lemma 3.1. *Let $\mathbf{z}_0$ be an initial condition and $\mathbf{z}^{(0)}(t, \mathbf{z}_0)$ a solution of*

$$\dot{\mathbf{z}} = \mathcal{F}_0(\mathbf{z})$$

with $\mathbf{z}^{(0)}(0, \mathbf{z}_0) = \mathbf{z}_0$ such that it remains bounded and bounded away from the singularities of $\mathcal{F}(\mathbf{z}, t, \mu)$. Let $\mathbf{z}(t, \mathbf{z}_0, \mu)$ be a solution of (3.5) with the same initial condition $\mathbf{z}_0$. Then we can write

$$\mathbf{z}(t, \mathbf{z}_0, \mu) = \mathbf{z}^{(0)}(t, \mathbf{z}_0) + \mu^2 \mathbf{z}^{(1)}(t, \mathbf{z}_0, \mu) + \mathbf{z}_R(t, \mathbf{z}_0, \mu),$$

where $\mathbf{z}^{(1)}(t, \mathbf{z}_0, \mu)$ is the solution of

$$\dot{\mathbf{z}} = \mathcal{F}_1(\mathbf{z}^{(0)}(t, \mathbf{z}_0), t, \mu) + D\mathcal{F}_0(\mathbf{z}^{(0)}(t, \mathbf{z}_0))\, \mathbf{z}$$

with initial condition $\mathbf{z}^{(1)}(0, \mathbf{z}_0, \mu) = 0$, and $D\mathcal{F}$ is the matrix whose entries are the partial derivatives of $\mathcal{F}$ with respect to the $\mathbf{z}$ variables. Furthermore, $\mathbf{z}_R(t, \mathbf{z}_0, \mu)$ and $D_{\mathbf{z}_0}\mathbf{z}_R(t, \mathbf{z}_0, \mu)$ are $\mathcal{O}(\mu^4)$ in a finite interval of time.

These results can be obtained by using Taylor's expansions and Gronwall's inequality (see [4]). They are also valid for any initial conditions in a compact neighborhood of $\mathbf{z}_0$ satisfying the hypothesis of the lemma.

4. Continuation of symmetric periodic solutions. In this section we use the results of section 3 to show the existence of symmetric periodic solutions of the reduced problem.

Let us start by considering the Kepler problem given by Hamiltonian (3.3), whose solution with initial condition $\mathbf{z}_0$ is

$$\mathbf{z}^{(0)}(t, \mathbf{z}_0) = (l_0 + L_0^{-3} t, g_0, h_0, L_0, G_0, H_0).$$

Clearly, the orbit with initial conditions $\mathbf{z}_0^* = (0, \pi/2, h_0^*, 1, G_0^*, H_0^*)$ is symmetric and periodic of period $T = 2\pi$. We want to find periodic symmetric solutions close to $\mathbf{z}^{(0)}(t, \mathbf{z}_0^*)$.

From Lemma 3.1, any solution of the reduced problem can be written as the solution of the Kepler problem plus a perturbation, provided that $\mathbf{z}^{(0)}(t, \mathbf{z}_0^*)$ remains bounded and bounded away from the singularities. To ensure that $\mathbf{z}_0^*$ satisfies these conditions, it is sufficient that $G_0^* > 0$; that is, the eccentricity is less than 1. Also, notice that we are dealing with elliptic orbits, so $G_0^* < 1$.

Thus, we will look for initial conditions $\mathbf{z}_0 = (0, \pi/2, h_0, L_0, G_0, H_0)$ in a neighborhood of a fixed $\mathbf{z}_0^*$ with $0 < G_0^* < 1$, in such a way that the solution $\mathbf{z}(t, \mathbf{z}_0, \mu)$ of (3.5), with $\mu \neq 0$ small enough, is a symmetric periodic orbit of the reduced problem.

For a fixed $\mathbf{z}_0^*$, let $\mathcal{D}$ be a compact neighborhood of $\mathbf{z}_0^*$ where the conditions of Lemma 3.1 hold and $0 < G_0 < 1$.

Then, given $\mathbf{z}_0 \in \mathcal{D}$, we know that

$$\mathbf{z}(t, \mathbf{z}_0, \mu) = \mathbf{z}^{(0)}(t, \mathbf{z}_0) + \mu^2 \mathbf{z}^{(1)}(t, \mathbf{z}_0, \mu) + \mathcal{O}(\mu^4), \tag{4.1}$$

where

$$\mathbf{z}^{(0)}(t, \mathbf{z}_0) = (L_0^{-3} t, \pi/2, h_0, L_0, G_0, H_0), \tag{4.2}$$

$$\mathbf{z}^{(1)}(t, \mathbf{z}_0, \mu) = \mathcal{Z}(t, \mathbf{z}_0) \int_0^t \mathcal{Z}^{-1}(s, \mathbf{z}_0) \mathcal{F}_1(\mathbf{z}^{(0)}(s, \mathbf{z}_0), s, \mu)\, ds, \tag{4.3}$$

and

$$\mathcal{Z}(t, \mathbf{z}_0) = \left. \frac{\partial \mathbf{z}^{(0)}(t, \xi)}{\partial \xi} \right|_{\xi = \mathbf{z}_0}.$$

From (4.1),

$$\begin{aligned} l(t, \mathbf{z}_0, \mu) &= L_0^{-3} t + \mu^2 l^{(1)}(t, \mathbf{z}_0, \mu) + \mathcal{O}(\mu^4), \\ g(t, \mathbf{z}_0, \mu) &= \pi/2 + \mu^2 g^{(1)}(t, \mathbf{z}_0, \mu) + \mathcal{O}(\mu^4), \end{aligned}$$

where $l^{(1)}$ and $g^{(1)}$ are the first and second coordinates of $\mathbf{z}^{(1)}$, given by (4.3), respectively.

Obviously, the symmetry conditions given by (2.8) are fulfilled at $t = 0$. They also must be satisfied at $t = T/2 = k\pi\mu^{3/2}$ in order to have $E_p(T/2) = k\pi$. We consider k a natural number and $\mu > 0$ such that $\mu = k^{-2/3}$. Then, $T/2 = \pi$ and we have to find initial conditions $\mathbf{z}_0 \in \mathcal{D}$ satisfying

$$\begin{aligned} L_0^{-3}\pi - \pi + \mu^2 l^{(1)}(\pi, \mathbf{z}_0, \mu) + \mathcal{O}(\mu^4) &= 0, \\ g^{(1)}(\pi, \mathbf{z}_0, \mu) + \mathcal{O}(\mu^2) &= 0. \end{aligned} \tag{4.4}$$

A natural way to solve (4.4) is to find a solution for the case $\mu = 0$ and then to apply an implicit function theorem. The first handicap is that neither $l^{(1)}(\pi, \mathbf{z}_0, \mu)$ nor $g^{(1)}(\pi, \mathbf{z}_0, \mu)$ is defined for $\mu = 0$, and neither is the Hamiltonian $\mathcal{H}_1$. Moreover, they do not satisfy the differentiability conditions of the standard implicit function theorem. In order to overcome these difficulties, we will see first that both equations can be extended to the case $\mu = 0$. Second, we will use Arenstorf's theorem, which requires weaker conditions (see [1]).

Let us start by extending (4.4) to the case $\mu = 0$. Observe that, from (4.3), $l^{(1)}$ and $g^{(1)}$ are bounded. This fact will be sufficient to define the first equation in (4.4) for $\mu = 0$. As for the second one, we show that $g^{(1)}(\pi, \mathbf{z}_0, \mu)$ can be written in terms of L_0, H_0, and G_0 plus a term of order $\mu^{3/2}$. In order to prove this we need the following technical lemma.

Lemma 4.1. *Given $\mathbf{z}_0^*$, let $\mathbf{z}_0 = (0, \pi/2, h_0, L_0, G_0, H_0) \in \mathcal{D}$. Let be $\varphi(t, \mathbf{z}_0)$ be a function bounded in $\mathcal{D}$, such that $\frac{d\varphi}{dt}(t, \mathbf{z}_0)$ is also bounded in $\mathcal{D}$. Then,*

$$\int_0^\pi (1 - \cos E_p)^2 \varphi(t, \mathbf{z}_0)\, dt = \frac{5}{2} \int_0^\pi \varphi(t, \mathbf{z}_0)\, dt + \mathcal{R}(\mathbf{z}_0, \mu),$$

where $|\mathcal{R}(\mathbf{z}_0, \mu)| \leq K\mu^{3/2}$ for a certain constant K.

Proof. From (2.2), the function $(1 - \cos E_p)^2$ is even and 2π-periodic with respect to the variable ℓ_p. Its Fourier series is given by

$$(1 - \cos E_p)^2 = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\ell_p),$$

where

$$a_k = \frac{2}{\pi} \int_0^\pi (1 - \cos E_p)^2 \cos(k\ell_p)\, d\ell_p.$$

From the fact that $(1 - \cos E_p)\, dE_p = d\ell_p$, the zero coefficient is

$$a_0 = \frac{2}{\pi} \int_0^\pi (1 - \cos E_p)^3\, dE_p = 5.$$

Then, using that $\ell_p = \mu^{-3/2} t$,

$$\int_0^\pi (1 - \cos E_p)^2 \varphi(t, \mathbf{z}_0)\, dt = \frac{5}{2} \int_0^\pi \varphi(t, \mathbf{z}_0)\, dt + \underbrace{\sum_{k=1}^{\infty} a_k \int_0^\pi \varphi(t, \mathbf{z}_0) \cos(k\mu^{-3/2} t)\, dt}_{\mathcal{R}(\mathbf{z}_0, \mu)}.$$

Integrating by parts,

$$\int_0^\pi \varphi(t, \mathbf{z}_0) \cos(k\mu^{-3/2} t)\, dt = \frac{\mu^{3/2}}{k} \left(\varphi(\pi, \mathbf{z}_0) \sin(k\mu^{-3/2}\pi) - \int_0^\pi \frac{d\varphi}{dt}(t, \mathbf{z}_0) \sin(k\mu^{-3/2} t)\, dt \right).$$

Since $\varphi(t, \mathbf{z}_0)$, $\frac{d\varphi}{dt}(t, \mathbf{z}_0)$ are bounded for all $(t, \mathbf{z}_0)$ with $\mathbf{z}_0 \in \mathcal{D}$ (say, $|\varphi(t, \mathbf{z}_0)| \leq k_1$ and $|\frac{d\varphi}{dt}(t, \mathbf{z}_0)| \leq k_2$), we have that for a certain constant C

$$\left| \int_0^\pi \varphi(t, \mathbf{z}_0) \cos(k\mu^{-3/2} t)\, dt \right| \leq \frac{\mu^{3/2}}{k} \left(k_1 + 2k_2 \frac{\mu^{3/2}}{k} \right) = \frac{\mu^{3/2}}{k} C,$$

and

$$|\mathcal{R}(\mathbf{z}_0, \mu)| \leq \mu^{3/2} C \sum_{k=1}^{\infty} \frac{|a_k|}{k}.$$

The series on the right-hand side converges because a_k are Fourier coefficients of a regular function, and so the lemma is proved. ∎

Lemma 4.2. *Given $\mathbf{z}_0^*$, let $\mathbf{z}_0 = (0, \pi/2, h_0, L_0, G_0, H_0) \in \mathcal{D}$. Then, for $\mu > 0$ and small enough, $g^{(1)}(\pi, \mathbf{z}_0, \mu)$ is of type $\mathcal{C}^1$ in $\mathcal{D}$ and there exist functions $I_1(L_0, G_0)$ and $I_2(L_0, G_0)$ such that*

$$g^{(1)}(\pi, \mathbf{z}_0, \mu) = -\frac{15}{16}(I_1(L_0, G_0) - H_0^2 I_2(L_0, G_0)) + \mathcal{O}(\mu^{3/2}).$$

Furthermore, if $L_0 = 1$, then

$$g^{(1)}(\pi, \mathbf{z}_0, \mu) = -\frac{15\pi}{32 G_0^4}\left(5\frac{H_0^2}{G_0^2} - 1\right) + \mathcal{O}(\mu^{3/2}).$$

Proof. From (4.3), we have that

$$g^{(1)}(t, \mathbf{z}_0, \mu) = \int_0^t \left.\frac{\partial \mathcal{H}_1}{\partial G}\right|_{\mathbf{z}^{(0)}(s, \mathbf{z}_0)} ds,$$

where $\mathbf{z}^{(0)}(s, \mathbf{z}_0)$ is the function defined in (4.2). Then, using (3.4) we obtain that

$$g^{(1)}(\pi, \mathbf{z}_0, \mu) = \frac{-3}{8}\int_0^\pi (1 - \cos E_p)^2 \varphi(t, \mathbf{z}_0)\, dt,$$

where

$$\varphi(t, \mathbf{z}_0) = \frac{\partial}{\partial G}\left(\frac{1}{r^3}\left(\cos^2 f - \frac{1}{3}\right) - \frac{H^2 \cos^2 f}{r^3 G^2}\right)\Bigg|_{\mathbf{z}^{(0)}(t, \mathbf{z}_0)}. \tag{4.5}$$

Since $\mathbf{z}_0 \in \mathcal{D}$, it is clear that $g^{(1)}(\pi, \mathbf{z}_0, \mu)$ is of type $\mathcal{C}^1$ in $\mathcal{D}$.

From Lemma 4.1 and (4.5), we have that

$$\begin{aligned} g^{(1)}(\pi, \mathbf{z}_0, \mu) &= \frac{-15}{16}\int_0^\pi \varphi(t, \mathbf{z}_0)\, dt + \mathcal{O}(\mu^{3/2}) \\ &= \frac{-15}{16}\left(\int_0^\pi \frac{\partial}{\partial G}\left(\frac{1}{r^3}\left(\cos^2 f - \frac{1}{3}\right)\right)\Bigg|_{\mathbf{z}^{(0)}(t, \mathbf{z}_0)} dt \right. \\ &\qquad \left. - H_0^2 \int_0^\pi \frac{\partial}{\partial G}\left(\frac{\cos^2 f}{G^2 r^3}\right)\Bigg|_{\mathbf{z}^{(0)}(t, \mathbf{z}_0)} dt\right) + \mathcal{O}(\mu^{3/2}) \\ &= -\frac{15}{16}(I_1(L_0, G_0) - H_0^2 I_2(L_0, G_0)) + \mathcal{O}(\mu^{3/2}), \end{aligned}$$

where

$$I_1(L_0, G_0) = \frac{G_0}{e_0 L_0^5}\int_0^{E(e_0, L_0)} (e_0 - \cos E)\frac{5e_0^2 - 5 - 2e_0\cos E + (7 - 3e_0^2)\cos^2 E - 2e_0\cos^3 E}{(1 - e_0\cos E)^6}\, dE,$$

$$\begin{aligned} I_2(L_0, G_0) = \frac{1}{G_0 e_0 L_0^5}\left(\int_0^{E(e_0, L_0)} \frac{-2e_0}{1 - e_0^2}\frac{(\cos E - e_0)^2}{(1 - e_0\cos E)^4}\, dE \right. \\ \left. + \int_0^{E(e_0, L_0)} (\cos E - e_0)\frac{4 - 5e_0^2 + 4e_0\cos E + (2e_0^2 - 7)\cos^2 E + 2e_0\cos^3 E}{(1 - e_0\cos E)^6}\, dE\right), \end{aligned}$$

and $E(e_0, L_0)$ is the solution of the equation $\pi = L_0^3(E - e_0 \sin E)$ (see section 5 for the details). In particular, when $L_0 = 1$, both integrals can be calculated explicitly and

$$I_1(1, G_0) = \frac{-\pi}{2G_0^4}, \qquad I_2(1, G_0) = \frac{-5\pi}{2G_0^6},$$

which ensures the last statement of the lemma. ∎

The next lemma shows that derivatives of $g^{(1)}$ satisfy conditions similar to those in Lemma 4.2; i.e., they can be written in terms of L_0, H_0, and G_0 plus terms of order $\mu^{3/2}$.

Lemma 4.3. *Under the same hypothesis as in Lemma 4.2,*

$$\frac{\partial g^{(1)}}{\partial L_0}(\pi, \mathbf{z}_0, \mu) = -\frac{15}{16}\left(\frac{\partial I_1}{\partial L_0}(L_0, G_0) - H_0^2 \frac{\partial I_2}{\partial L_0}(L_0, G_0)\right) + \mathcal{O}(\mu^{3/2}),$$
$$\frac{\partial g^{(1)}}{\partial H_0}(\pi, \mathbf{z}_0, \mu) = \frac{15}{8} H_0 I_2(L_0, G_0) + \mathcal{O}(\mu^{3/2}).$$

Proof. The result is straightforward using arguments similar to those of Lemma 4.2. ∎

In order to extend the symmetry equation to $\mu = 0$, let

$$\Omega = \{(L_0, H_0);\, \mathbf{z}_0 = (0, \pi/2, h_0, L_0, G_0, H_0) \in \mathcal{D}\}.$$

We define the function

$$\Phi(\xi, \mu) = (\Phi_1(\xi, \mu), \Phi_2(\xi, \mu))$$

for $\xi = (L_0, H_0) \in \Omega$ and $\mu \geq 0$ as

$$\Phi_1(\xi, \mu) = \begin{cases} L_0^{-3}\pi - \pi + \mu^2 l^{(1)}(\pi, \mathbf{z}_0, \mu) + \mathcal{O}(\mu^4) & \text{if } \mu \neq 0, \\ L_0^{-3}\pi - \pi & \text{if } \mu = 0, \end{cases} \tag{4.6}$$

and

$$\Phi_2(\xi, \mu) = \begin{cases} g^{(1)}(\pi, \mathbf{z}_0, \mu) + \mathcal{O}(\mu^2) & \text{if } \mu \neq 0, \\ -\frac{15}{16}(I_1(L_0, G_0) - H_0^2 I_2(L_0, G_0)) & \text{if } \mu = 0, \end{cases} \tag{4.7}$$

where $I_1(L_0, G_0)$ and $I_2(L_0, G_0)$ are the functions stated in the proof of Lemma 4.2. Then, (4.4) can be written as

$$\Phi(\xi, \mu) = (0, 0)$$

for $\mu \geq 0$, and for $\mu = 0$, $L_0 = 1$, satisfies

$$\Phi_1(\xi, 0) = 0, \qquad \Phi_2(\xi, 0) = -\frac{15\pi}{32G_0^4}\left(5\frac{H_0^2}{G_0^2} - 1\right).$$

Thus, for each fixed value of G_0, $\Phi(\xi_0, 0) = (0, 0)$ if $\xi_0 = (1, G_0/\sqrt{5})$. Observe that $H_0^2/G_0^2 = \cos^2 i$, and so the solution for $\mu = 0$ corresponds to an inclination i with $\cos i = 1/\sqrt{5}$, which is the critical inclination angle obtained in the case of the problem of an Earth-centered orbit when the effects of J_2 are considered.

In order to show that there exist symmetric periodic solutions of the reduced problem for $\mu \neq 0$, we need to show that there exist solutions of $\Phi(\xi, \mu) = (0, 0)$. In this case, we will use the next proposition, proved in [4], which is a sufficient condition for Arenstorf's theorem.

Proposition 4.4. *Let U be an open domain in R^n, $I \subset R$ an open neighborhood of the origin, $f : U \times I \to R^n$ with $f(0, 0) = 0$, differentiable with respect to $x \in U$, and $D_x f(0, 0)$ nonsingular. Assume that there exist $c > 0$, $k > 0$ such that for $x \in U$, $\epsilon \in I$,*

1. $\|D_x f(x, \epsilon) - D_x f(0, 0)\| \leq c(\|x\| + \epsilon)$,
2. $\|f(0, \epsilon)\| \leq k\epsilon$.

Then there exists a function $x(\epsilon) \in U$, defined for $\epsilon \in I' \subset I$, such that $f(x(\epsilon), \epsilon) = 0$ and $x(0) = 0$.

In order to apply Proposition 4.4 we need to prove that the function Φ satisfies some properties.

Proposition 4.5. *Let G_0 be fixed and $\xi_0 = (1, G_0/\sqrt{5})$. For μ small enough, there exists η such that the function $\Phi(\xi, \mu)$ is differentiable with respect to ξ in $\mathcal{B} = \{\xi \in \Omega;\ \|\xi - \xi_0\| \leq \eta\}$ and satisfies the three properties*

(i) $\|\Phi(\xi_0, \mu)\| \leq C_0 \mu^{3/2}$,
(ii) $\|(D_\xi \Phi)^{-1}(\xi_0, 0)\| \leq M$,
(iii) $\|D_\xi \Phi(\xi, \mu) - D_\xi \Phi(\xi_0, 0)\| \leq C_1(\|\xi - \xi_0\| + \mu^{3/2})$,

where M, C_0, and C_1 are constants independent of μ and $D_\xi \Phi(\xi, \mu)$ denotes the Jacobi matrix of Φ with respect to the variables ξ.

Proof. Statement (i) is a direct consequence of the definition of Φ (see (4.6) and (4.7)), the fact that $l^{(1)}$ is a bounded function, and Lemma 4.2.

Using that the derivatives of $l^{(1)}$ are also bounded and Lemma 4.3, we have that

$$
\begin{aligned}
D_\xi \Phi(\xi, \mu) &= \begin{pmatrix} \frac{-3\pi}{L_0^4} + \mathcal{O}(\mu^2) & \mathcal{O}(\mu^2) \\ \mathcal{J}(L_0, G_0) + \mathcal{O}(\mu^{3/2}) & \frac{15}{8} H_0 I_2(L_0, G_0) + \mathcal{O}(\mu^{3/2}) \end{pmatrix}, \\
D_\xi \Phi(\xi, 0) &= \begin{pmatrix} \frac{-3\pi}{L_0^4} & 0 \\ \mathcal{J}(L_0, G_0) & \frac{15}{8} H_0 I_2(L_0, G_0) \end{pmatrix},
\end{aligned}
\tag{4.8}
$$

where $\mathcal{J}(L_0, G_0) = \frac{-15}{16} \frac{\partial (I_1 - H_0^2 I_2)}{\partial L_0}$. Then, as $I_2(1, G_0) = -5\pi/(2G_0^6) \neq 0$, $D_\xi \Phi(\xi_0, 0)$ can be inverted, and item (ii) is proved.

Let us prove (iii). First, we have that

$$
\|D_\xi \Phi(\xi, \mu) - D_\xi \Phi(\xi_0, 0)\| \leq \|D_\xi \Phi(\xi, \mu) - D_\xi \Phi(\xi, 0)\| + \|D_\xi \Phi(\xi, 0) - D_\xi \Phi(\xi_0, 0)\|. \tag{4.9}
$$

On one hand, as the components of $\Phi(\xi, 0)$ are of type $\mathcal{C}^1$ with respect to ξ, we get that

$$
\|D_\xi \Phi(\xi, 0) - D_\xi \Phi(\xi_0, 0)\| \leq c_0 \|\xi - \xi_0\|. \tag{4.10}
$$

On the other hand,

$$
\|D_\xi \Phi(\xi, \mu) - D_\xi \Phi(\xi, 0)\| \leq \sum_{i=1}^{2} \|D_\xi \Phi_i(\xi, \mu) - D_\xi \Phi_i(\xi, 0)\| \leq c_1 \mu^2 + c_2 \mu^{3/2} \tag{4.11}
$$

by expressions (4.8). Substituting (4.10) and (4.11) into (4.9), we prove item (iii). ■

Notice that, given h_0, G_0, and $\mathbf{z}_0^* = (0, \pi/2, h_0, 1, G_0, G_0/\sqrt{5})$, the solution $\mathbf{z}^{(0)}(t, \mathbf{z}_0^*)$ is a solution of the Kepler problem lying on a plane of the critical inclination i with $\cos i = 1/\sqrt{5}$. Finally, let us prove that there exist periodic symmetric solutions of the perturbed reduced problem close to $\mathbf{z}^{(0)}(t, \mathbf{z}_0^*)$.

Theorem 4.6. *Consider the three-dimensional collision restricted three-body problem with masses $m_1 = m_2 = 1/2$, and primaries' semimajor axis $\mu/2$. If $\mu = k^{-2/3}$, where k is a positive integer large enough, there exist initial conditions such that the infinitesimal body moves in a symmetric periodic orbit of the* reduced problem*, of period 2π, near a Keplerian elliptic orbit. The inclination of the orbit is close to the "critical value"* $\cos i = 1/\sqrt{5}$.

Proof. Let us consider initial values h_0, G_0, and $\xi_0 = (1, G_0/\sqrt{5})$. It is clear that $\Phi(\xi_0, 0) = (0, 0)$. Given $\xi \in \Omega$, we define $f(x, \mu) = \Phi(x + \xi_0, \mu)$, where $x = \xi - \xi_0$. From Proposition 4.5 we can easily prove that $f(x, \mu)$ is under the hypothesis of Proposition 4.4. Then there exists a function $x(\mu)$ such that $f(x(\mu), \mu) = (0, 0)$ and $x(0) = 0$.

This yields a continuum of solutions of system $\Phi(\xi, \mu) = (0, 0)$. These conditions must be satisfied simultaneously with $E_p(T/2) = k\pi$, which is equivalent to $T = 2k\pi\mu^{3/2}$. Thus, for each $\mu = k^{-2/3}$, k a large positive integer, a periodic solution of the reduced problem exists. ■

Remark. All the orbits found are on an integral resonance with the motion of the primaries; i.e., the primaries undergo k complete orbits in one orbit of the infinitesimal body. If $k = p/q$ is an irreducible rational, then similar arguments show that in q complete orbits of the infinitesimal the primaries undergo p complete orbits.

5. Appendix. Here we develop the calculations needed in the proof of Lemma 4.2. We want to compute

$$\begin{aligned}\int_0^\pi \varphi(t, \mathbf{z}_0)dt &= \int_0^\pi \left.\frac{\partial \Delta_1}{\partial G}\right|_{\mathbf{z}^{(0)}(t,\mathbf{z}_0)} dt - H_0^2 \int_0^\pi \left.\frac{\partial \Delta_2}{\partial G}\right|_{\mathbf{z}^{(0)}(t,\mathbf{z}_0)} dt \\ &= I_1(L_0, G_0) - H_0^2 I_2(L_0, G_0),\end{aligned}$$

where $\Delta_1 = \frac{\cos^2 f - 1/3}{r^3}$ and $\Delta_2 = \frac{\cos^2 f}{G^2 r^3}$.

We will introduce the change of variables given by $t = L_0^3(E - e_0 \sin E)$, where L_0^2 and e_0 correspond to the semimajor axis and the eccentricity of the Keplerian orbit $\mathbf{z}^{(0)}(t, \mathbf{z}_0)$, respectively. As the new variable to integrate will be E, we use the rule

$$\frac{\partial \Delta_i}{\partial G} = \frac{\partial \Delta_i}{\partial E}\frac{dE}{de}\frac{de}{dE} \tag{5.1}$$

for $i = 1, 2$. On one hand, from Kepler's equation $t = a^{3/2}(E - e \sin E)$, we have that

$$0 = a^{3/2}\left(\frac{dE}{de} - \sin E - e\cos E \frac{dE}{de}\right) \quad \text{and} \quad \frac{dE}{de} = \frac{\sin E}{1 - e\cos e}.$$

On the other hand, as $G^2 = a(1 - e^2)$, we have that $\frac{de}{dG} = \frac{-G}{ae}$. Substituting into (5.1), we have that

$$\frac{\partial \Delta_i}{\partial G} = \frac{-G \sin E}{ae(1 - e\cos E)}\frac{\partial \Delta_i}{\partial E}. \tag{5.2}$$

Next, as $r = a(1 - e\cos E) = \frac{a(1-e^2)}{1+e\cos f}$, we have that $\cos f = \frac{\cos E - e}{1 - e\cos E}$, and deriving both expressions we obtain

$$\frac{\partial r}{\partial E} = a\,\frac{e - \cos E}{\sin E}, \qquad -\sin f\,\frac{df}{dE} = \frac{\sin E(e^2 - 2 + e\cos E)}{(1 - e\cos E)^2}. \tag{5.3}$$

Thus, using the expressions (5.2) and (5.3), we have that

$$\begin{aligned}
\frac{\partial \Delta_1}{\partial G} &= \frac{G(e - \cos E)(5e^2 - 5 - 2\,e\cos E + (7 - 3\,e^2)\cos^2 E - 2\,e\cos^3 E)}{ea^4(1 - e\cos E)^7},\\
\frac{\partial \Delta_2}{\partial G} &= \frac{-2(\cos E - e)^2}{G^3a^3(1 - e\cos E)^5}\\
&\quad + \frac{(\cos E - e)}{Gea^4}\,\frac{4 - 5e^2 + 4e\cos E + (2e^2 - 7)\cos^2 E + 2e\cos^3 E}{(1 - e\cos E)^7}.
\end{aligned}$$

Finally, evaluating both expressions on the solution $\mathbf{z}^{(0)}(t, \mathbf{z}_0)$ of the Kepler problem, we obtain the expressions for the functions I_1 and I_2 as

$$\begin{aligned}
\int_0^\pi \left.\frac{\partial \Delta_1}{\partial G}\right|_{\mathbf{z}^{(0)}(t,\mathbf{z}_0)} dt &= \frac{G_0}{e_0L_0^5}\int_0^{E(e_0,L_0)} f_1(e_0, E)\,dE = I_1(L_0, G_0),\\
\int_0^\pi \left.\frac{\partial \Delta_2}{\partial G}\right|_{\mathbf{z}^{(0)}(t,\mathbf{z}_0)} dt &= \frac{1}{G_0e_0L_0^5}\int_0^{E(e_0,L_0)} f_2(e_0, E)\,dE = I_2(L_0, G_0),
\end{aligned}$$

where $E(e_0, L_0)$ is the solution of the equation $\pi = L_0^3(E - e_0\sin E)$ and

$$\begin{aligned}
f_1(e_0, E) &= (e_0 - \cos E)\frac{5e_0^2 - 5 - 2e_0\cos E + (7 - 3e_0^2)\cos^2 E - 2e_0\cos^3 E}{(1 - e_0\cos E)^6},\\
f_2(e_0, E) &= \frac{-2e_0}{1 - e_0^2}\,\frac{(\cos E - e_0)^2}{(1 - e_0\cos E)^4}\\
&\quad + (\cos E - e_0)\,\frac{4 - 5e_0^2 + 4e_0\cos E + (2e_0^2 - 7)\cos^2 E + 2e_0\cos^3 E}{(1 - e_0\cos E)^6}.
\end{aligned}$$

Furthermore, it is clear that for a fixed value of $e_0 < 1$ the functions f_1 and f_2 are continuous and differentiable with respect to E and e_0, and so $I_1(L_0, G_0)$ and $I_2(L_0, G_0)$ are functions of type $\mathcal{C}^1$ with respect to e_0. In particular, when $L_0 = 1$, $E(e_0, L_0) = \pi$, and both integrals can be calculated explicitly:

$$I_1(1, G_0) = \frac{-\pi}{2G_0^4}, \qquad I_2(1, G_0) = \frac{-5\pi}{2G_0^6}.$$

REFERENCES

[1] R. F. Arenstorf, *A new method of perturbation theory and its application to the satellite problem of celestial mechanics*, J. Reine Angew. Math., 221 (1966), pp. 113–145.

[2] R. H. Battin, *An Introduction to the Mathematics and Methods of Astrodynamics*, AIAA Education Series, American Institute of Aeronautics and Astronautics (AIAA), Washington, DC, 1987.

[3] J. M. Cors, C. Pinyol, and J. Soler, *Periodic solutions in the spatial elliptic restricted three-body problem*, Phys. D, 154 (2001), pp. 195–206.
[4] J. M. Cors, C. Pinyol, and J. Soler, *Analytic continuation in the case of non-regular dependency on a small parameter with an application to celestial mechanics*, J. Differential Equations, 219 (2005), pp. 1–19.
[5] W. H. Jefferys, *A new class of periodic solutions of the three dimensional restricted problem*, Astronom. J., 71 (1966), pp. 99–102.
[6] J. Llibre and D. Pasca, *Periodic orbits of the planar collision restricted* 3*-body problem*, Celestial Mech. Dynam. Astronom., 96 (2006), pp. 19–29.
[7] J. F. Palacian, P. Yanguas, S. Fernandez, and M. A. Nicotra, *Searching for periodic orbits of the spatial elliptic restricted three-body problem by double averaging*, Phys. D, 213 (2006), pp. 15–24.
[8] H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste*, Tome III, Gauthier-Vilars, Paris, 1899.
[9] E. L. Stiefel and G. Scheifele, *Linear and Regular Celestial Mechanics*, Grundlehren Math. Wiss. 174, Springer-Verlag, New York, 1971.
[10] L. G. Taff, *Celestial Mechanics. A Computational Guide for the Practitioner*, John Wiley & Sons, New York, 1985.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 18–38
© 2008 Society for Industrial and Applied Mathematics

Corner-Impact Bifurcations: A Novel Class of Discontinuity-Induced Bifurcations in Cam-Follower Systems*

Gustavo Osorio†, Mario di Bernardo‡, and Stefania Santini‡

Abstract. This paper is concerned with the analysis of a class of impacting systems of relevance in applications: cam-follower systems. We show that these systems, which can be modeled as discontinuously forced impact oscillators, can exhibit complex behavior due to the detachment at high rotational speeds between the follower and the cam. We propose that the observed phenomena can be explained in terms of a novel type of discontinuity-induced bifurcation, termed as corner-impact. We present a complete analysis of this bifurcation in the case of a nonautonomous impact oscillator and explain the transition to chaos observed in a representative cam-follower example. The theoretical findings are validated numerically.

Key words. discontinuity-induced bifurcation, piecewise-smooth systems, impact oscillator, cam-follower system

AMS subject classifications. 34C15, 34C28, 34C60, 37G15, 37E05, 70K50

DOI. 10.1137/060666433

1. Introduction.

Recently, much research effort has been spent to analyze the dynamics of piecewise-smooth dynamical systems with impacts [5, 41]. These systems arise in many areas of engineering and applied science. A typical example is that of mechanical systems characterized by structural components with displacement constraints. Examples include bouncing or hopping robots, systems with backlash or friction, gears, and vibro-impacting mechanical devices [5].

Cam-follower systems are a particularly important class of mechanical systems with displacement constraints widely used for the operation of various machines and mechanical devices [30]. Usually, their purpose is to actuate valves or other mechanisms through the movement of a follower forced by a rotating cam. For example, all types of automated production machines, including screw machines, spring winders, and assembly machines, rely heavily on this kind of system for their operation. One of the most common application is to the valve train of internal combustion engines (ICEs) [18], where the effectiveness of the ICE is based

*Received by the editors July 31, 2006; accepted for publication (in revised form) by W. Beyn June 6, 2007; published electronically January 16, 2008. This work was partially supported by the European Project SICONOS IST2001-37172. The authors also gratefully acknowledge support from the European Union (EU Project SICONOS - V Framework Programme, IST2001-37172) and the project MIUR-PRIN MACSI funded by the Italian Ministry for Research and University. The paper was completed during a research visit of the authors at the Centre de Recerca Matematica in Barcelona thanks to support from the Government of Catalunya.

http://www.siam.org/journals/siads/7-1/66643.html

†Corresponding author. Departamento de Ingeniería Eléctrica, Electrónica y Computación, Universidad Nacional de Colombia, Carrera 26 #64-60, Manizales, Colombia (gaosoriol@unal.edu.co), and Dipartimento di Informatica e Sistemistica, Universitá degli Studi di Napoli Federico II, Via Claudio 21, 80125, Napoli, Italia (gosorio@unina.it).

‡Dipartimento di Informatica e Sistemistica, Universitá degli Studi di Napoli Federico II, Via Claudio 21, 80125, Napoli, Italia (mario.dibernardo@unina.it, stefania.santini@unina.it).

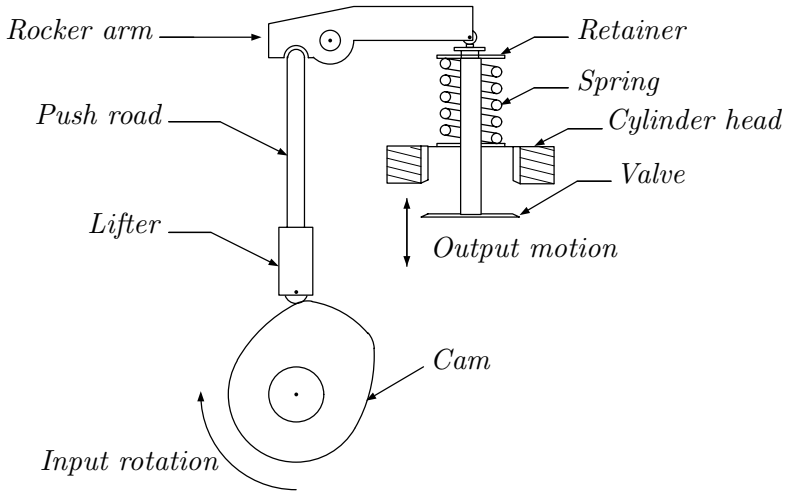


Figure 1. *Valve train configuration.*

on the proper working of a cam-follower system. A schematic of a single valve for a typical pushrod-type engine is presented in Figure 1. Here, the cam rotation results in a linear motion imparted to the valve. The valve spring in the system provides the restoring force necessary to maintain contact between the components.

To guarantee that the follower moves as required, it is important in applications to carefully design the cam profile. Different cam geometries are used in practice ranging from circular cams to highly complex cam profiles. In general, there is now a large variety of alternative methods to select the cam profile. For example, by using the constrained optimization algorithm, it is possible to use splines to obtain the cam geometry from the desired motion that the cam is required to impart on the follower (for examples see [9] and [16]). This often means that while the cam has a continuous displacement profile, it might have discontinuities in its acceleration [31].

It has been observed that, as the cam rotational speed increases, the follower can detach from the cam. This causes the onset of undesired behavior associated to impacts taking place between the follower and the cam. For example, in automotive engines this phenomenon can seriously deteriorate the engine performance as the valves can close with abnormally high velocity and even bounce off the seat (valve floating and bouncing) [21, 37, 10]. To avoid this phenomenon, a large spring force and preload are applied to the follower [34]. This often causes an increase in the contact force, which induces higher stresses possibly leading to early surface failure of the parts. The resulting high friction valve train reduces the efficiency of the engine system [39].

In general, cam-follower systems can be thought of as impact oscillators with moving boundaries [20, 30, 15, 40]. While the dynamics of impact oscillators with continuous forcing has been the subject of many papers in the existing literature (see, for example, [32, 17, 6, 7]), the possible intricate bifurcation behavior of impact oscillators with discontinuous forcing was discussed only recently, as, for example, in [8]. It was proposed that discontinuously forced oscillators can show a novel bifurcation phenomenon unique to their nature which was termed

as corner-impact bifurcation. Namely, in [8] the dynamics of an impact oscillator forced by a discontinuous sinusoidal forcing of the form $f(t) = A|\sin(\omega t)|$ are studied. It was shown that, under variation of the system parameters, abrupt changes of the system's qualitative behavior are observed when an impact occurs at a point where the forcing velocity is discontinuous (a corner-impact bifurcation point).

The observed behavior was explained in terms of appropriate local maps. In particular, by using the technique of discontinuity mappings recently proposed in [17] and [12], it was suggested that a corner-impact bifurcation of the oscillator corresponds to a border-collision of a fixed point of the associated Poincaré map. An important difference was highlighted between corner-impact bifurcations and other types of discontinuity-induced bifurcations [4] in impacting systems such as grazing of limit cycles [28, 38, 22, 26, 33, 24]. While the normal form map of a grazing bifurcation is typically characterized by a square root singularity [28], the local normal form map associated to a corner-impact bifurcation was shown to be a piecewise-linear map with a gap such as those studied in [19]. Hence, as explained in [8], an appropriate classification method needs to be used to investigate this novel class of bifurcations.

In [14], it was conjectured for the first time that corner-impact bifurcations are fundamental in organizing the complex behavior observed in cam-follower systems. It was shown that, as the cam rotational speed increases, these systems can exhibit sudden transitions from periodic solutions to chaos. Such transitions were conjectured to be due to corner-impact bifurcations.

In this paper, we present a careful analysis of corner-impact bifurcations in cam-follower systems. We analytically derive the normal form map associated to such a bifurcation in a representative example of interest where the cam profile is characterized by a discontinuous acceleration. In particular, we investigate the bifurcation behavior exhibited by this system under variations of the cam rotational speed. We find that following the detachment of the follower from the cam, the system can exhibit complex nonlinear phenomena involving chattering, period adding cascades, and the sudden transition from periodic attractors to chaos. We explain the sudden transition to chaos observed in the system in terms of a corner-impact bifurcation. Namely, we show that dramatic changes in the system's behavior are observed when, under parameter variation, one of the impacts characterizing the system's trajectory crosses one of the manifolds in phase space where the cam acceleration is discontinuous.

We prove that the normal form map of the corner-impact bifurcation in these systems is a piecewise-linear continuous map rather than discontinuous because of the higher degree of discontinuity of the forcing signal provided by the cam with respect to that of the forcing considered in [8]. We wish to emphasize that such a finding is generic for the wide class of impacting systems characterized by forcing terms with discontinuous acceleration.

As shown in the paper, the derivation of the mapping has an immediate practical relevance. In fact, the derivation of a piecewise-linear normal form map implies that the strategy to classify border-collisions in piecewise-linear continuous maps due to Feigin [13] can be used, under some circumstances, to classify corner-impact bifurcations in continuous-time impacting flows.

The rest of the paper is outlined as follows. In section 2, we present the modeling of the cam-follower system of our concern, where the cam profile has been assumed to be characterized by a discontinuous acceleration. Then in section 3 the numerical bifurcation analysis is

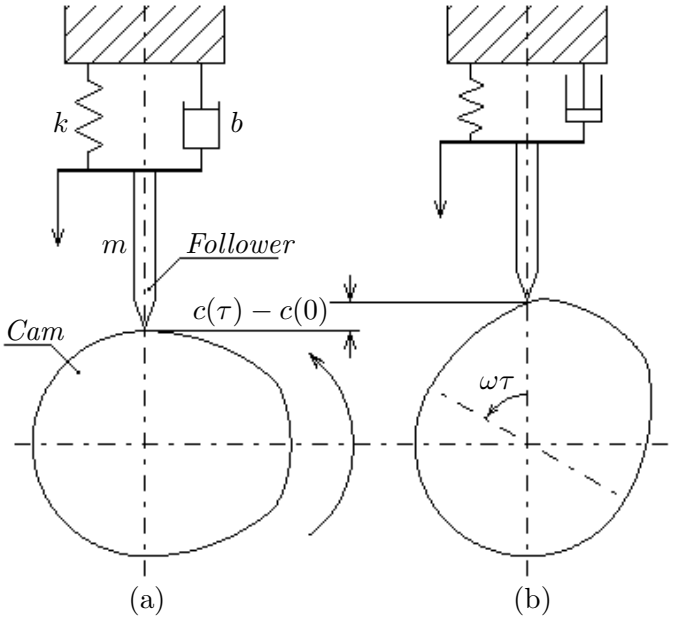


Figure 2. *Cam-follower schematics.* (a) $t = 0$. (b) $t = \tau$.

presented under variation of the cam rotational speed. In section 4 we present the analysis of the corner-impact bifurcation phenomenon detected in the system, and we classify the ensuing dynamics by using an appropriately derived local mapping. Finally, conclusions are drawn in section 5.

2. Modeling. The formulation of an appropriate model for a cam-follower system can be a challenging task for most applications. Various models with different degrees of complexity have been proposed and extensively studied. They range from simple models with one degree-of-freedom (DOF) such as that described in [20] to complex models characterized by many DOFs, as, for example, the 21 DOF model studied in [36], where additional effects of camshaft torsion and bending, backlashing, and squeezing of lubricant in bearings are included. Nevertheless, there is general agreement in the literature, confirmed by experiments, that a lumped parameter single DOF model is adequate to represent the main qualitative features of the dynamic behavior of the system of interest [3, 20, 1, 15].

The schematic diagram of the cam-follower system under investigation is shown in Figure 2. We consider the following second order equation to model the free body dynamics of the follower away from the cam:

$$\begin{aligned} mq''(t) + bq'(t) + kq(t) &= -mg \\ \text{if } q(t) &> c(t), \end{aligned} \tag{2.1}$$

where m, b, k, and g are constant positive parameters representing the follower mass, viscous damping, spring stiffness, and gravitational constant, respectively. The state of the follower is given by the position $q(t)$ and the velocity $q'(t)$. The cam position is given by $c(t)$, and we assume that the follower motion is constrained to the phase-space region where $q(t) > c(t)$.

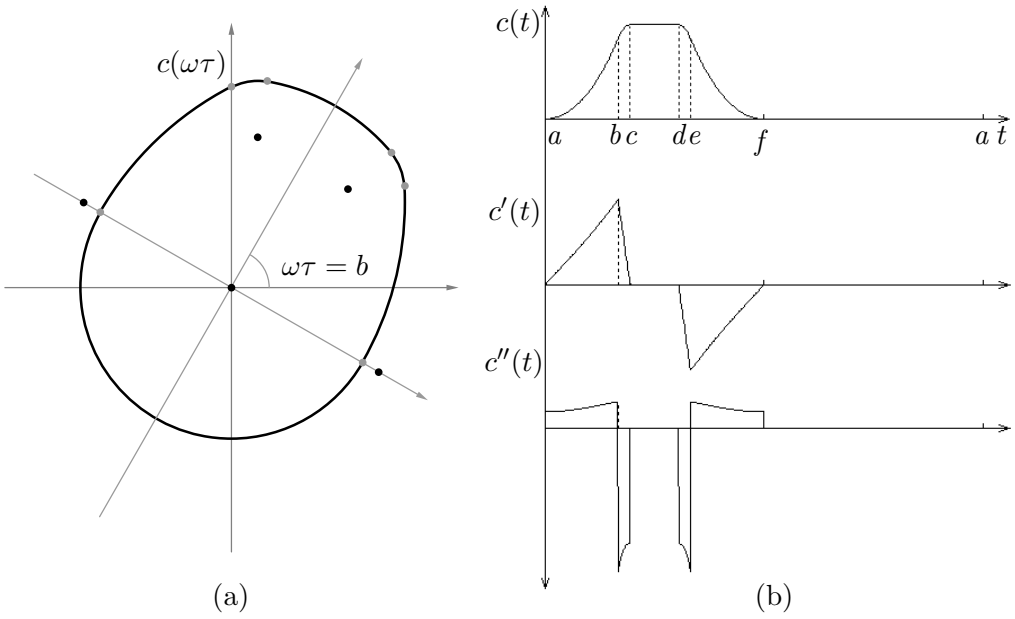


Figure 3. (a) *Cam profile.* (b) *Constraint position $c(t)$, velocity $c'(t)$, and acceleration $c''(t)$.*

The dynamic behavior when impact occurs (i.e., $q(t) = c(t)$) is modeled via a Newton restitution law as [5, 23, 29]

$$q'(t^+) = (1+r)c'(t) - rq'(t^-) \quad \text{if } q(t) = c(t), \tag{2.2}$$

where $q'(t^+)$ and $q'(t^-)$ are the post- and preimpact velocities, respectively, $c'(t)$ is the projection of the cam velocity vector at the contact point along the direction of the free movement of the follower, and $r \in [0, 1]$ is the coefficient of restitution used to model from plastic to elastic impacts.

An essential ingredient of the model is the choice of the cam profile, $c(t)$. The cam is assumed to be rotating at a constant angular velocity ω and can be interpreted as the "control action" acting on the follower state, as suggested in [30]. Therefore, $c(t)$ is carefully selected in applications as a trade-off between several optimality criteria dependent upon the specific device being considered and the unavoidable physical constraints on the system.

Typically, this results in a design process where the cam profile is selected by using splines and can contain several degrees of discontinuity. For example, the cam for a single overhead camshaft valve train is designed by using quadratic splines, and, as a consequence, discontinuities are present in its acceleration. In general, it is not uncommon in applications to find cam geometries characterized by continuous cam positions and velocities but a discontinuous second derivative [30].

In what follows, we assume the cam profile to be characterized by a discontinuous second derivative as shown in Figure 3. For the sake of brevity, the analytical expressions of the cam profile and its derivatives are reported in Appendix A. The case of a smooth cam profile with

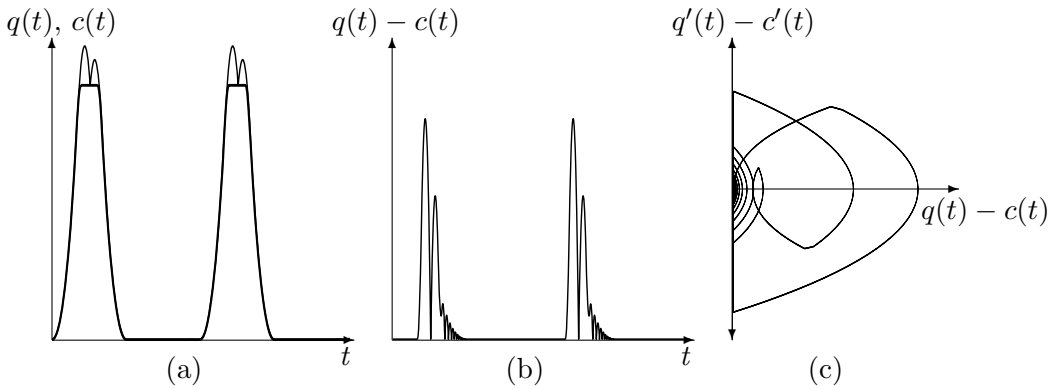


Figure 4. *Time simulation at $\omega = 183$ rpm.* (a) *Follower position, $q(t)$ (light); Cam position, $c(t)$ (dark).* (b) *Relative position, $q(t) - c(t)$.* (c) *Phase space, $q(t) - c(t)$ versus $q'(t) - c'(t)$.*

continuous first and second order derivatives is also of interest in applications and was studied experimentally in [2].

3. Numerical bifurcation analysis. The model represented by (2.1) and (2.2) was found to exhibit an intricate bifurcation behavior including the sudden transition to chaos under variation of the cam rotational speed, ω [14]. The presence of bifurcations and chaos was also confirmed by experiments, as described in [2].

Here we briefly summarize some of the most striking behavior exhibited by the system focusing on the abrupt transition from a one-periodic impacting solution to chaos observed when $\omega \approx 673.234445$ rpm.

In general, starting from low values of ω, the system exhibits solutions characterized by permanent contact between the cam and the follower. As ω increases, the follower is observed to detach from the cam during its evolution and then to impact with it. A typical periodic evolution with impacts is shown in Figure 4(a) when $\omega = 183$ rpm. We observe that the follower and the cam are in contact with zero relative velocity for part of the orbit (*sticking*) and then detach, giving rise to impacting behavior. As shown in Figure 4(b)–(c) a careful look at the follower evolution shows that a *chattering sequence* is present, where theoretically an infinite number of impacts accumulates in finite time. (Note that in practice chattering is associated to a large but finite number of impacts.)

Chattering can be associated to an intricate bifurcation structure. In Figure 5(a), the location of the impacts in the cam surface is depicted for each value of ω, characterizing the follower asymptotic solution. We see that following detachment at about 114 rpm, the follower immediately exhibits multi-impacting behavior and chattering (characterized by the accumulation of the impact lines in the diagram onto the darker areas corresponding to the chattering accumulation points). An interesting phenomenon is the appearance of resonant peaks associated to impact lines crossing the boundaries where the cam acceleration profile is discontinuous (represented by dotted lines in the figure). A detailed analysis of this bifurcation scenario is presented in [27].

This phenomenon can be classified as due to a *corner-impact bifurcation*, a type of discontinuity-induced bifurcation recently described in [8]. Namely, at certain values of ω, one of

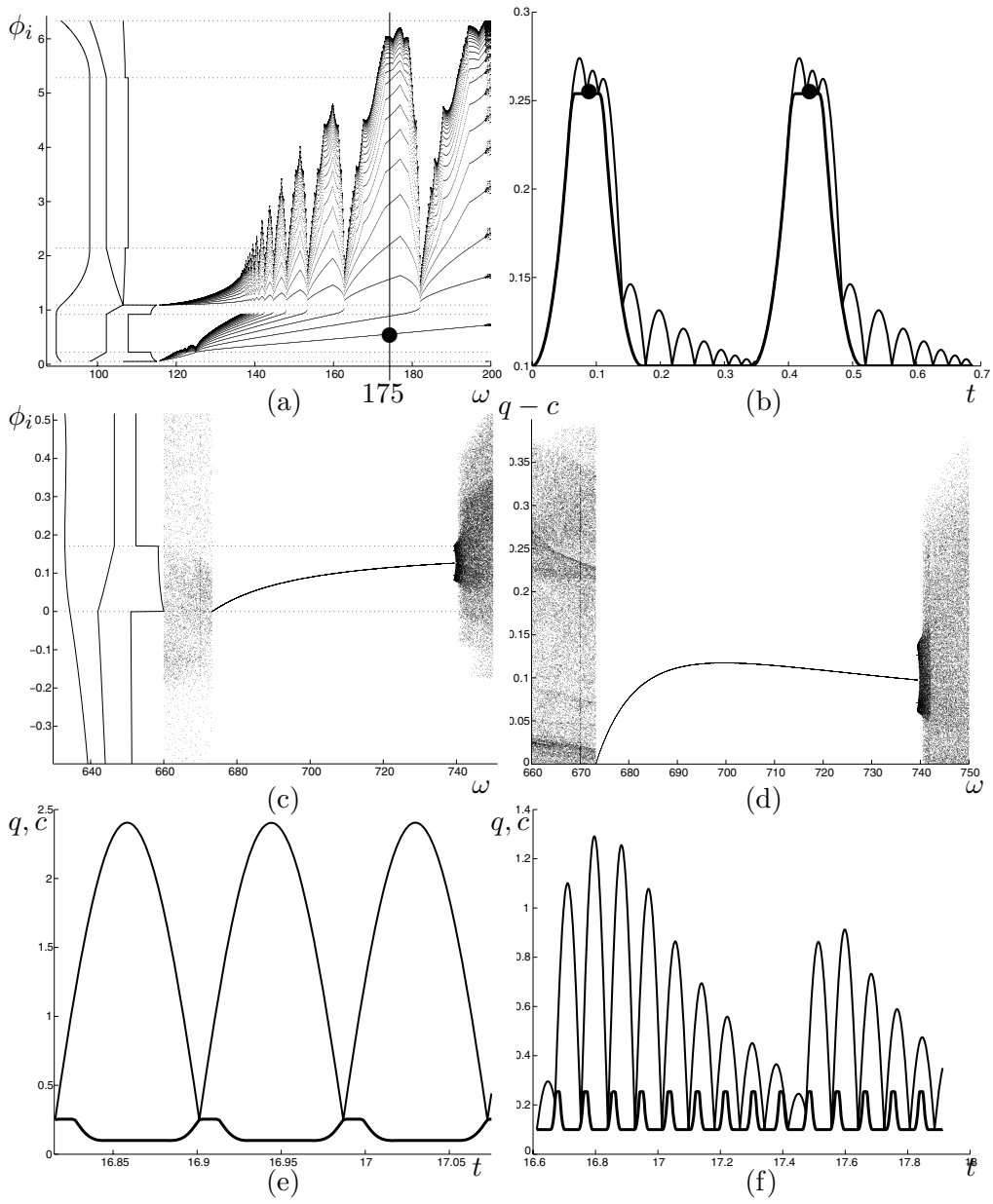


Figure 5. (a) *Impact bifurcation diagram for* $[115, 200]$ *rpm. The phase of an impact* ϕ_i *(rad) is plotted against* ω. (b) *Time evolution for* 175 *rpm.* (c) *Impact bifurcation diagram for* $\omega = [660, 750]rpm$. (d) *Stroboscopic bifurcation diagram for* $\omega = [660, 750]rpm$. (e) *Bifurcating orbit at the corner-impact point at* $\omega = 700rpm$. (f) *Chaotic evolution for* $\omega = 670rpm$. *Dotted and dashed lines represent phases where the cam profile is discontinuous. Vertical curves in panels* (a), (c) *show the cam position velocity and acceleration as function of the phase.*

the impacts characterizing the follower motion occurs at a point on the cam profile where the acceleration is discontinuous. We shall seek to analytically investigate this phenomenon and classify the behavior following the corner-impact event in the cam-follower system of interest. For the sake of simplicity, we focus on a different region of the system bifurcation diagram depicted in Figure 5(c). Here a one-periodic solution characterized by one impact per period exhibits sudden transitions to chaos as ω is decreased below 673.234445 rpm. A close look at the impact bifurcation diagram in Figure 5(c) and in the stroboscopic bifurcation diagram in Figure 5(d) shows that such transitions occur precisely when the impact characterizing the solution crosses the cam discontinuity boundaries (the dotted lines in Figure 5(c)). Specifically, the sudden transition to chaos is due to the corner-impact bifurcation of the periodic solution depicted in Figure 5(e). Past the corner-impact bifurcation point, the system exhibits chaotic behavior (see, for example, the trajectory reported in Figure 5(f) for $\omega \approx 670$ rpm). The rest of this paper is devoted to the analysis of this bifurcation scenario.

4. Corner-impact bifurcation analysis. The numerical observations reported above indicate that a corner-impact bifurcation is causing the transition to chaos observed in the cam-follower system. Specifically, we are interested in analyzing the occurrence of the corner-impact bifurcation depicted in Figure 5(c) when $\omega \approx 673.234445$ rpm. Numerically, we detected that the bifurcating orbit, shown in Figure 5(e), is a one-periodic orbit characterized by one impact per period. As the rotational speed of the cam is decreased, at the bifurcation point, the impact is observed to cross the point on the cam surface where the cam acceleration is discontinuous. To investigate this novel type of discontinuity-induced bifurcation we will analytically construct the Poincaré map of the system close to the bifurcation point. We will then study the bifurcations of the fixed point corresponding to the periodic solution of interest. A crucial point in the analysis is to assess whether the resulting map is piecewise-linear continuous or not. Indeed, only if this is the case, the theory of border-collision bifurcations (see [35, 13]) can be used to classify the possible solutions branching from the corner-impact bifurcation point [25].

We use the concept of discontinuity mapping (or normal form mapping) recently introduced in [17, 12] to analytically construct the Poincaré map associated to the bifurcating orbit of interest. We use the cam-follower system described in section 2 as a representative example to carry out the analytical derivations.

4.1. Poincaré map derivation. We are interested in the analysis of the period one orbit at the corner-impact bifurcation point. Such an orbit is sketched in Figure 6. Then, close to such a periodic orbit we define the stroboscopic map P as the mapping from the follower state $x_1 \in \Pi_1$ at a stroboscopic time instant t_1 to the next stroboscopic point $x_2 \in \Pi_2$. Without loss of generality, we assume that $t_n = -\frac{T}{2} + (n-1)T$ for $n = 1, 2, 3, \ldots$, where T is the period of the cam forcing cycle (note that $T = 2\pi/\omega$). Namely, we have

$$x_2 = P(x_1). \tag{4.1}$$

To construct P we would need to flow forward using the system evolution from x_1 to x_2 for time T taking into account the possible occurrence of impacts and therefore applying Newton's restitution law as required. Alternatively, as shown in [17], it is possible to construct P as the composition of three submappings: (i) an affine transformation $P_{1,T/2}$ from the stroboscopic

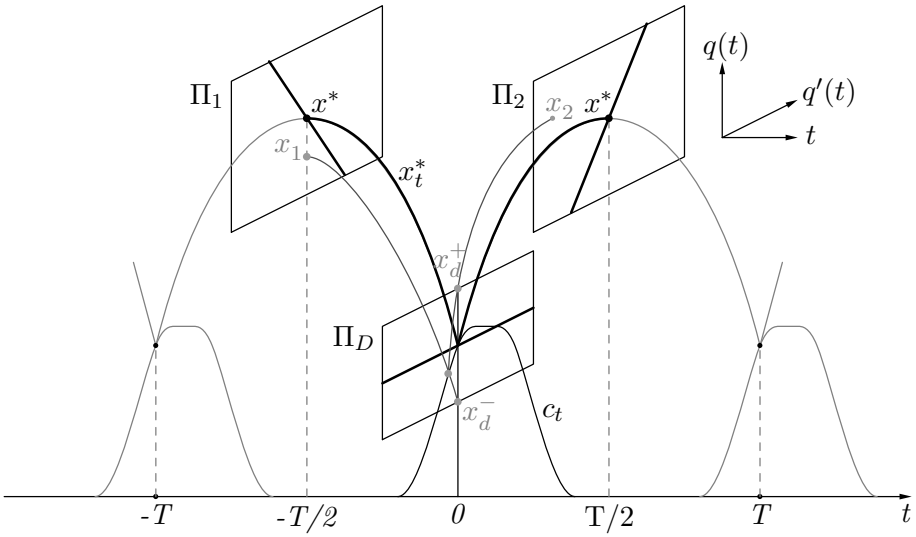


Figure 6. *Global map composition.*

plane Π_1 at $t_1 = -\frac{T}{2}$ to the plane Π_D going through the corner-impact point at $t = 0$; (ii) an appropriate zero-time discontinuity mapping (ZDM) P_D on Π_D accounting for the presence of the discontinuity; and again (iii) an affine transformation $P_{2,T/2}$ from the plane Π_D at $t = 0$ back to the stroboscopic plane Π_2 at $t_2 = \frac{T}{2}$. Specifically, while $P_{1,T/2}$ and $P_{2,T/2}$ are fixed time maps that account for the follower evolution away from the cam as if no impact had occurred, the ZDM represents the correction that needs to be made to the system trajectories because of the presence of impacts. Figure 6 represents the global map composition. This means that we can write

$$P = P_{2,T/2} \circ P_D \circ P_{1,T/2}, \tag{4.2}$$

where $P_{1,T/2} : \Pi_1 \mapsto \Pi_D$ will map the state from the initial condition x_1 on the stroboscopic plane Π_1 to a point x_d^- on the discontinuity plane Π_D as if no impacts had occurred. $P_D : \Pi_D \mapsto \Pi_D$ will then map x_d^- to the point x_d^+ appropriately correcting the evolution for the presence of impacts (see Figure 7). Finally, $P_{2,T/2} : \Pi_D \mapsto \Pi_2$ will map x_d^+ to a point x_2 back onto the stroboscopic plane Π_2. In so doing, as discussed in [17, 12], the effect of the system discontinuities due to impacts is taken into account by the ZDM, P_D, which is therefore often termed as the local normal form map in the context of the theory of discontinuity-induced bifurcations [26].

4.1.1. Derivation of $P_{1,T/2}$ and $P_{2,T/2}$. As explained above, the maps $P_{1,T/2}$ and $P_{2,T/2}$ are defined only in terms of the free body dynamics of the follower and the cam rotating period T (depending upon the cam rotational speed ω). Therefore, we can solve (2.1) to get an analytical expression of the flows generating the mappings of interest.

Specifically, we define

$$x_t = \begin{bmatrix} q(t) + \frac{g}{\omega_0^2} \\ q'(t) \end{bmatrix}, \qquad y_t = \begin{bmatrix} c(t) \\ c'(t) \end{bmatrix}$$

as the state vectors for the follower and the cam, respectively.

Then the generalized solution of (2.1) is

$$x_t = e^{-\zeta t}\left(I\cos(\omega_s t) + A\sin(\omega_s t)\right)x_0 = \phi_t x_0, \tag{4.3}$$

where $\zeta = \frac{b}{2m}$, $\omega_0 = \sqrt{\frac{k}{m}}$, $\omega_s = \sqrt{\omega_0^2 - \zeta^2}$, I is the identity matrix, $\phi_t x_0$ represents the system flow for time t starting from the initial condition x_0, and

$$A = \begin{bmatrix} \frac{\zeta}{\omega_s} & \frac{1}{\omega_s} \\ -\frac{\omega_0^2}{\omega_s} & -\frac{\zeta}{\omega_s} \end{bmatrix}.$$

Note that, in general, the system flow operator can be expressed as

$$\phi_t = \frac{e^{-\zeta t}}{\omega_s}\begin{bmatrix} \omega_s\cos(\omega_s t) + \zeta\sin(\omega_s t) & \sin(\omega_s t) \\ -\omega_0^2\sin(\omega_s t) & \omega_s\cos(\omega_s t) - \zeta\sin(\omega_s t) \end{bmatrix}. \tag{4.4}$$

The submapping $P_{i,T/2}$ can then be easily obtained using (4.3) as

$$\begin{aligned} P_{i,T/2}(x) &= e^{-\zeta T/2}\left(I\cos(\omega_s T/2) + A\sin(\omega_s T/2)\right)x \\ &:= \phi_{\frac{T}{2}}x. \end{aligned} \tag{4.5}$$

4.1.2. Derivation of P_D. As explained in [12], the ZDM can be obtained by an appropriate composition of backward and forward flows so that the overall time spent following backward and forward is zero. As explained earlier, the ZDM is the correction that maps the point $x_d^- \in \Pi_D$ onto the point $x_d^+ \in \Pi_D$, taking into account the presence of impacts in the trajectory of interest. In what follows we assume that only one impact occurs over one cycle of the periodic orbit of interest as we suppose to be sufficiently close to the bifurcating orbit x_t^* shown in Figure 6. Figure 7 shows a schematic diagram that describes the construction of the ZDM, close to the corner-impact bifurcations. Without loss of generality we assume that the origin is placed at the Poincaré section Π_D. To analytically derive the mapping $x_d^+ = P_D(x_d^-)$ we need to perform the following steps:

1. Starting from x_d^-, we find the time t_i at which the impact occurs. Namely, t_i is obtained by looking at the difference, $(q(t) - c(t))$, between the follower position and the cam position close to $t = 0$. Given a vector z, we indicate by $[z]_1$ its first component. Then $q(t) = [x_t]_1 - \frac{g}{\omega_0^2}$, and therefore, close to x_d^-, t_i can be obtained as the nearest solution of the equation

$$H(x_{-t_i}^-, t_i) := \left[x_{-t_i}^- - y_{-t_i}\right]_1 = h\cdot\left[\phi_{-t_i}x_d^- - y_{-t_i}\right] = 0, \tag{4.6}$$

where $h = [\ 1 \quad 0\ ]$.
Hence, t_i is implicitly defined by the equation $H(x_{-t_i}^-, t_i) = 0$. Once t_i is found, the preimpact state of the system, $x_{-t_i}^-$, can also be obtained as

$$x_{-t_i}^- = \phi_{-t_i}x_d^-. \tag{4.7}$$

Note that t_i can be either negative or positive according to whether the impact occurs to the left or to the right of $t = 0$.

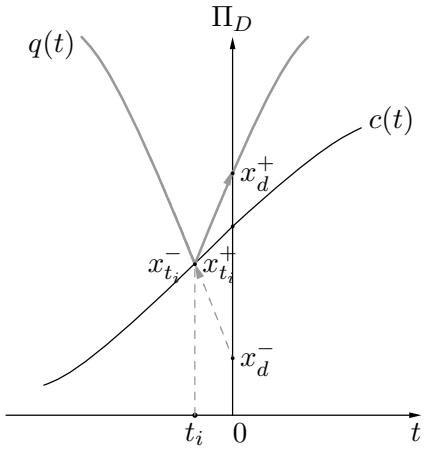


Figure 7. *ZDM construction.*

2. Using the restitution law (2.2), we can then write the postimpact state of the follower $x^+_{-t_i}$ as

$$x^+_{-t_i} = x^-_{-t_i} + R(x^-_{-t_i} - y_{-t_i}) = \rho(x^-_{-t_i}, y_{-t_i}), \tag{4.8}$$

where

$$R = \begin{bmatrix} 0 & 0 \\ 0 & -(1+r) \end{bmatrix}.$$

3. Finally, to obtain x_d^+, we flow forward for time t_i starting from the postimpact state $x^+_{-t_i}$ found at the previous step. In so doing, the state of the follower $x_d^+ \in \Pi_D$ can be computed as

$$x_d^+ = \phi_{t_i} x^+_{-t_i}. \tag{4.9}$$

Using (4.7), (4.8), and (4.9), we can then explicitly write the ZDM as

$$x_d^+ = P_D(x_d^-) = (\mathbf{I} + \phi_{t_i} R \phi_{-t_i})\, x_d^- - \phi_{t_i} R y_{-t_i}, \tag{4.10}$$

with t_i defined implicitly by (4.6).

4.1.3. Constructing the stroboscopic map. Composing the submappings $P_{1,T/2}$, $P_{2,T/2}$, and P_D given by (4.5) and (4.10), we can then construct the stroboscopic Poincaré map, P, of the system close to the corner-impact bifurcation point from a generic $x_n \in \Pi_n$ to $x_{n+1} \in \Pi_{n+1}$ as

$$\begin{aligned} x_{n+1} = P(x_n, T) &= P_{2,T/2}(P_D(P_{1,T/2}(x_n))) \\ &= \phi_{\frac{T}{2}} \left((\mathbf{I} + \phi_{t_i} R \phi_{-t_i})\, \phi_{\frac{T}{2}} x_n - \phi_{t_i} R y_{-t_i} \right), \end{aligned} \tag{4.11}$$

where t_i is implicitly defined by the equation $H(x_n, t_i) = h \cdot (\phi_{\frac{T}{2} - t_i} x_n - y_{-t_i}) = 0$.

Note that the fixed point (x^* associated to the periodic solution existing for a fixed value of the cam period $T = T^*$) can be obtained by solving (4.11) for $x_{n+1} = x_n = x^*$, i.e.,

$$x^* = -\left[\mathbf{I} - \phi_{T^*} + \phi_{\frac{T^*}{2}} R \phi_{\frac{T^*}{2}}\right]^{-1} \phi_{\frac{T}{2}} R y_0, \tag{4.12}$$

with $t_i^* = 0$.

In what follows we are interested in studying such a mapping locally to the corner-impact bifurcation point detected when $\omega = \omega^* = 673.234445$ rpm, corresponding to a period $T^* = 0.08912199969159$ s. The fixed point associated to the bifurcating orbit is $x^* = \begin{bmatrix} 5.09700788184250 & 0 \end{bmatrix}'$. These values were detected first numerically and then obtained analytically by solving (4.12) through an algebraic manipulation software. (For the sake of brevity we leave out the computer algebra here.)

4.2. A locally piecewise-linear continuous map. Let δx_n and δT be sufficiently small variations of the state and parameter from the bifurcation point x^*, T^*. We can then linearize the map $x_{n+1} = P(x_n, T)$ in (4.11) about this point as

$$\delta x_{n+1} = \frac{\partial P(x^*, T^*)}{\partial x_n} \delta x_n + \frac{\partial P(x^*, T^*)}{\partial T} \delta T. \tag{4.13}$$

For the computation of $\frac{\partial P}{\partial x_n}$ it is essential to take into account the implicit dependence of t_i on x_n and T. Hence, using implicit differentiation, we have

$$\frac{\partial P(x_n, T)}{\partial x_n} = \frac{\partial P(x_n)}{\partial x_n} + \frac{\partial P(t_i)}{\partial t_i} \frac{\partial t_i(x_n)}{\partial x_n}. \tag{4.14}$$

Using (4.11), we can then write

$$\frac{\partial P(x_n)}{\partial x_n} = \phi_{\frac{T}{2}} (I + \phi_{-t_i} R \phi_{t_i}) \phi_{\frac{T}{2}}, \tag{4.15}$$

$$\frac{\partial P(t_i)}{\partial t_i} = \phi_{\frac{T}{2}} \left(\phi'_{t_i} R \phi_{-t_i} - \left(\phi_{t_i} R \phi'_{-t_i} \right) \phi_{\frac{T}{2}} x_n - \phi'_{t_i} R y_{-t_i} + \phi_{t_i} R y'_{-t_i} \right). \tag{4.16}$$

Moreover, using the implicit differentiation theorem, from (4.6) we have

$$\frac{\partial H(x_n, t_i(x_n))}{\partial x_n} = \frac{\partial H(x_n)}{\partial x_n} + \frac{\partial H(t_i)}{\partial t_i} \frac{\partial t_i(x_n)}{\partial x_n} = 0.$$

The above expression can be used to compute the remaining term in (4.14) as

$$\frac{\partial t_i(x_n)}{\partial x_n} = -\left(\frac{\partial H(t_i)}{\partial t_i} \right)^{-1} \frac{\partial H(x_n)}{\partial x_n}, \tag{4.17}$$

where

$$\frac{\partial H(t_i)}{\partial t_i} = -h \cdot \left(\phi'_{\frac{T}{2} - t_i} x_n - y'_{-t_i} \right),$$
$$\frac{\partial H(x_n)}{\partial x_n} = h \cdot \phi_{\frac{T}{2} - t_i},$$

and $h = \begin{bmatrix} 1 & 0 \end{bmatrix}$.

After substituting (4.15), (4.16), and (4.17) into (4.14) we obtain

$$\begin{aligned} (4.18)\quad & \left.\frac{\partial P(x_n,T)}{\partial x_n}\right|_{\substack{x_n=x^* \\ T=T^*}} \\ & = \phi_{\frac{T}{2}^*} \left((I+R) + \frac{\left(R\phi_0' - \phi_0' R\right)\phi_{\frac{T}{2}^*} x^* + \phi_0' R y_0 - R y_0'}{h \cdot \left(\phi'_{\frac{T}{2}^*} x^* - y_0'\right)} \right) \phi_{\frac{T}{2}^*}. \end{aligned}$$

In an analogous way, for the computation of $\frac{\partial P}{\partial T}$, it is essential to take into account the implicit dependence of t_i on x_n and T. Hence, by using implicit differentiation, we have

$$(4.19)\quad \frac{\partial P(x_n,T)}{\partial T} = \frac{\partial P(T)}{\partial T} + \frac{\partial P(t_i)}{\partial t_i}\frac{\partial t_i(T)}{\partial T}.$$

Using (4.11), we can then write

$$\begin{aligned} (4.20)\quad & \frac{\partial P(T)}{\partial T} \\ & = \left(\phi_T' + \frac{1}{2}\phi'_{\frac{T}{2}+t_i} R \phi_{\frac{T}{2}-t_i} + \frac{1}{2}\phi_{\frac{T}{2}+t_i} R \phi'_{\frac{T}{2}-t_i}\right) x_n - \frac{1}{2}\phi'_{\frac{T}{2}+t_i} R y_{-t_i} - \phi_{\frac{T}{2}+t_i} R \frac{\partial y_{-t_i,T}}{\partial T}. \end{aligned}$$

Again, from (4.6) we have

$$\frac{\partial H(x_n, t_i(x_n))}{\partial T} = \frac{\partial H(T)}{\partial T} + \frac{\partial H(t_i)}{\partial t_i}\frac{\partial t_i(T)}{\partial T} = 0,$$

which can be used to compute the remaining term in (4.19). Namely, we obtain

$$(4.21)\quad \frac{\partial t_i(T)}{\partial T} = -\left(\frac{\partial H(t_i)}{\partial t_i}\right)^{-1} \frac{\partial H(T)}{\partial T},$$

where

$$\begin{aligned} \frac{\partial H(t_i)}{\partial t_i} &= -h \cdot \left(\phi'_{\frac{T}{2}-t_i} x_n - y'_{-t_i}\right), \\ \frac{\partial H(T)}{\partial T} &= h \cdot \left(\frac{1}{2}\phi'_{\frac{T}{2}-t_i} x_n - \frac{\partial y_{-t_i,T}}{\partial T}\right), \end{aligned}$$

and

$$\frac{\partial y_{t,T}}{\partial T} = \begin{bmatrix} -\frac{t}{T}c'(t) \\ -\frac{1}{T}c'(t) - \frac{t}{T}c''(t) \end{bmatrix}.$$

Finally, substituting (4.16), (4.20), and (4.21) into (4.19) yields

$$\begin{aligned}(4.22)\quad \left.\frac{\partial P(x_n,T)}{\partial T}\right|_{\substack{x_n=x^*\\T=T^*}} &= \left(\phi'_{T^*}+\frac{1}{2}\phi'_{\frac{T}{2}^*}R\phi_{\frac{T}{2}^*}+\frac{1}{2}\phi_{\frac{T}{2}^*}R\phi'_{\frac{T}{2}^*}\right)x^*-\frac{1}{2}\phi'_{\frac{T}{2}^*}Ry_0-\phi_{\frac{T^*}{2}}R\frac{\partial y_{0,T^*}}{\partial T}\\ &\quad+\phi_{\frac{T}{2}^*}\left(\left(R\phi'_0-\phi'_0R\right)\phi_{\frac{T}{2}^*}x^*+\phi'_0Ry_0-Ry'_0\right)\cdot\frac{h\cdot\left(\frac{1}{2}\phi'_{\frac{T}{2}^*}x^*-\frac{\partial y_{0,T^*}}{\partial T}\right)}{h\cdot\left(\phi'_{\frac{T}{2}^*}x^*-y'_0\right)}.\end{aligned}$$

We can then explicitly compute these quantities for the cam-follower system of interest. In particular, after some algebraic manipulation, we have

$$(4.23)\qquad A:=\frac{\partial P}{\partial x_n}(x^*,T^*)=\phi_{\frac{T}{2}^*}\begin{bmatrix}-r & 0\\ -\frac{(1+r)(2\zeta c'_0+c''_0+\omega_0^2q_d^*)}{q_d'^*-c'_0} & -r\end{bmatrix}\phi_{\frac{T}{2}^*}$$

and

$$\begin{aligned}(4.24)\quad B:=\frac{\partial P}{\partial T}(x^*,T^*)&=\frac{1}{2}\phi_{\frac{T}{2}^*}\begin{bmatrix}q_d^*\\ -rq_d'^*+(1+r)c'_0\end{bmatrix}+\frac{1}{2}\phi'_{\frac{T}{2}^*}\begin{bmatrix}q_d^*\\ -rq_d'^*-\frac{2(1+r)}{T^*}c'_0\end{bmatrix}\\ &\quad+\frac{1}{2}\phi_{\frac{T}{2}^*}\frac{(1+r)q_d'^*}{q_d'^*-c'_0}\begin{bmatrix}q_d'^*-c'_0\\ 2\zeta c'_0+c''_0+\omega_0^2q_d^*\end{bmatrix}.\end{aligned}$$

Note that both the matrices A and B as defined by (4.23)–(4.24) depend on the value of the second derivative of the cam acceleration c''_0 at the impact point. Therefore, the map is actually piecewise-linear locally to the bifurcation point where the cam acceleration is discontinuous; i.e.,

$$c''^-_0:=\lim_{t\to0^-}c''(t)\neq\lim_{t\to0^-}c''(t):=c''^+_0.$$

Then, the local map can be expressed as

$$(4.25)\qquad \delta x_{n+1}=\begin{cases}A^-\delta x_n+B^-\delta T & \text{if } C\cdot\delta x_n+D\cdot\delta T<0,\\ A^+\delta x_n+B^+\delta T & \text{if } C\cdot\delta x_n+D\cdot\delta T>0,\end{cases}$$

where

$$A^\pm=\frac{\partial P^\pm}{\partial x},\quad B^\pm=\frac{\partial P^\pm}{\partial T},$$

with the index $\pm$ indicating whether the matrices are evaluated with $c''_0=c''^-_0$ or $c''_0=c''^+_0$.

We have established that close to the corner-impact bifurcation point, the dynamics of the follower can be studied by means of the local mapping (4.25).

Now, from (4.11), the global Poincaré map is known to be a continuous function of the cam position and velocity through the term y_{-t_i}. Moreover, the map is independent from the

cam acceleration. It follows that the map is continuous at the bifurcation point; i.e., we must have that

$$A^-\delta x_n + B^-\delta T = A^+\delta x_n + B^+\delta T$$

when

$$C\delta x_n + D\delta T = 0.$$

Therefore, we have

$$C = h\cdot(A^+ - A^-) \quad \text{and} \quad D = h\cdot(B^+ - B^-).$$

Substituting the numerical values of the map parameters for the cam-follower system of interest, we obtain the following analytical estimates of the map matrices:

$$A^- = \begin{bmatrix} 0.82093496821478 & 0.01346530915655 \\ 2.52012201452530 & 0.82093496821478 \end{bmatrix},\ B^- = \begin{bmatrix} -51.62757990297 \\ -5455.79455977621 \end{bmatrix},$$

$$A^+ = \begin{bmatrix} 0.68571072072040 & -0.07351052377964 \\ 2.30988433707948 & 0.68571072072040 \end{bmatrix},\ B^+ = \begin{bmatrix} 208.11740649865 \\ -5051.96030903248 \end{bmatrix},$$

and

$$C = [-0.13522424749438\ \ -0.08697583293619],\ D = 259.7449864016200.$$

4.2.1. Numerical validation. We will now validate our numerical findings by comparing the map (4.25), which was derived analytically, with the numerical estimates of the mapping obtained by means of simulation and an optimized fitting algorithm close to the bifurcation point.

To derive such an estimate, we use an accurate event-driven numerical algorithm to simulate the cam dynamics over one period starting from a set of M different initial conditions and parameter values—namely, say, $\delta\bar{x}_n$, the vector of M possible perturbations of x^*, and $\delta\bar{T}$, the vector of M possible perturbations of T. We then simulate the cam dynamics from each of the perturbed initial conditions and parameter values to obtain the vector $\delta\bar{x}_{n+1} = x^* - x_{n+1}$ after one period. We repeat the set of simulations twice, once with the cam acceleration set to $c_0''^+$ and once with the acceleration set to $c_0''^-$. In so doing, we numerically obtain the vectors

$$\delta\bar{x}^{\pm}_{n+1} = \begin{bmatrix} \delta\bar{x}^1_{n+1} & \dots & \delta\bar{x}^m_{n+1} & \dots & \delta\bar{x}^M_{n+1} \end{bmatrix}.$$

We then use a least-squares fitting algorithm to estimate the matrices $\hat{A}^{\pm}$ and $\hat{B}^{\pm}$ that minimize the error

$$e = \left\| \delta\bar{x}^{\pm}_{n+1} - \begin{bmatrix} \hat{A}^{\pm} & | & \hat{B}^{\pm} \end{bmatrix} \begin{bmatrix} \delta\bar{x}_n \\ \delta\bar{T} \end{bmatrix} \right\|^2.$$

The estimated map matrices found using this numerical strategy are

$$\hat{A}^- = \begin{bmatrix} 0.82093497830369 & 0.01346530945739 \\ 2.52012201542191 & 0.82093496286678 \end{bmatrix}, \hat{B}^- = \begin{bmatrix} -51.62757113994 \\ -5455.79411324739 \end{bmatrix},$$

$$\hat{A}^+ = \begin{bmatrix} 0.68571065978423 & -0.07351053029558 \\ 2.30988432418263 & 0.68571073479454 \end{bmatrix}, \hat{B}^+ = \begin{bmatrix} 208.11731732063 \\ -5051.95951604729 \end{bmatrix}.$$

We notice that these numerical estimates are almost identical (up to at least five decimal places) to those obtained analytically earlier in the paper. This validates our analysis and shows the reliability of the analytical derivation used to get a leading order estimate of the Poincaré map close to the bifurcation point under investigation.

4.3. Classification of the nonsmooth bifurcation scenario. We can now use the locally derived map (analytical or numerical) to classify and explain the bifurcation scenario due to the corner-impact bifurcation detected in the cam-follower system of interest. In particular, the map derived above is a piecewise-linear continuous map. As the cam rotational speed is increased, the period T of the forcing provided by the cam varies. Correspondingly, at the corner-impact bifurcation point ($\delta T = 0$), the map fixed point undergoes a border-collision. The Feigin strategy for border-collision bifurcations can then be used to classify the corner-impact bifurcation scenario [13].

The idea is to start by recasting the map (4.25) into a canonical form following the procedure presented in [4]. Specifically, we do the following.

1. We eliminate the term depending on δT by considering an appropriate change of coordinates. In particular, if we say that c_1 and c_2 are the coefficients of C, we choose
$$\delta\tilde{x}_n^1 = \delta x_n^1 + D\frac{\mu}{c_1},$$
$$\delta\tilde{x}_n^2 = \delta x_n^2,$$
so that the map becomes
$$\delta\tilde{x}_{n+1} = \begin{cases} A^-\delta\tilde{x}_n + \tilde{B}\delta T & \text{if } C\cdot\delta\tilde{x}_n < 0, \\ \\ A^+\delta\tilde{x}_n + \tilde{B}\delta T & \text{if } C\cdot\delta\tilde{x}_n > 0, \end{cases}$$
where
$$\tilde{B} = \begin{bmatrix} b_1^- - \frac{a_{11}^-}{c_1}d \\ b_2^- - \frac{a_{21}^-}{c_1}d \end{bmatrix} = \begin{bmatrix} b_1^+ - \frac{a_{11}^+}{c_1}d \\ b_2^+ - \frac{a_{21}^+}{c_1}d \end{bmatrix} = \begin{bmatrix} 1525.26226128059 \\ -615.02768162765 \end{bmatrix},$$
with $a_{ij}^{\pm}$ being the coefficients of $A^{\pm}$.
2. Then, using the strategy presented in [4, 11], we consider the change of coordinates $x = W^{-1}\tilde{x}$, where the matrix W is obtained as $W = T^-O^-$ with
$$O^- = \begin{bmatrix} C \\ CA^- \end{bmatrix}, \; T^- = \begin{bmatrix} 1 & 0 \\ d_1^- & 1 \end{bmatrix},$$
where d_1^- is the linear coefficient of the characteristic polynomial of A^- given by $p^-(\lambda) = \lambda^2 + d_1^-\lambda + d_2^-$. Applying such a similarity transformation, the map matrices become
$$\bar{A}^- = \begin{bmatrix} 1.64186993642956 & 1 \\ -0.64 & 0 \end{bmatrix}, \; \bar{A}^+ = \begin{bmatrix} 1.37142144144080 & 1 \\ -0.64 & 0 \end{bmatrix},$$
and
$$\bar{B} = \begin{bmatrix} 152.75990 \\ 207,79599 \end{bmatrix}, \qquad \bar{C} = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

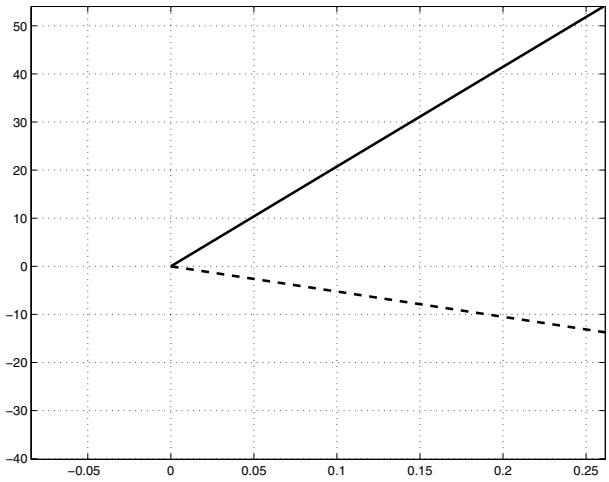


Figure 8. *Numerical bifurcation diagram of the local map* (4.25) *with the analytically estimated matrices. The border-collision when* $\delta T = 0$ *corresponds to the corner-impact bifurcation point at* $\omega \approx 673.2$ *rpm. Note that as predicted a nonsmooth fold scenario is observed with no fixed point existing for* $\delta T < 0$ *and two coexisting fixed points, one stable and the other unstable for* $\delta T > 0$.

As explained in [13, 4], we can now classify the type of bifurcation scenario observed at the bifurcation point under investigation by computing the map eigenvalues on both sides of the boundary. For the case under investigation, we have that (i) the eigenvalues of A^- are $\lambda_1^- = 1.0052$ and $\lambda_2^- = 0.6367$; (ii) the eigenvalues of A^+ are $\lambda_{1,2}^+ = 0.6857 \pm \mathbf{j}0.4120$. Hence, according to Feigin's classification strategy, since the total number of real eigenvalues greater than unity on both sides of the boundary is odd, the bifurcating fixed point will undergo a nonsmooth saddle node bifurcation and will cease to exist [13]. This is in perfect agreement with what is observed numerically in the local bifurcation scenario in the map in Figure 8.

Therefore, we can explain the sudden transition to chaos observed in the cam-follower system under investigation as due to the occurrence of a corner-impact bifurcation. Namely, the corner-impact is associated to a nonsmooth-fold scenario causing the disappearance of the stable impacting solution undergoing the bifurcation. This causes trajectories to leave the local neighborhood where they are confined before the bifurcation and converge toward the stable coexisting chaotic attractor when ω is decreased below the corner-impact bifurcation point.

Hence, we can conclude that corner-impact bifurcations in cam-follower systems can indeed lead to dramatic changes of the system qualitative behavior including sudden transitions from periodic solutions to chaos.

5. Conclusions.

We have studied a novel type of discontinuity-induced bifurcation in a class of mechanical devices widely used in applications: cam-follower systems. Using a representative second order model of the follower, we have shown that its dynamics can undergo several bifurcations including sudden transitions to chaos as the cam rotational speed is varied. We analyzed in detail the corner-impact bifurcation of a one-periodic solution characterized by one impact per period. In particular, we observed that the system's behavior undergoes dra-

matic changes when the impact occurs at a point where the cam profile is discontinuous. Using the concept of discontinuity mappings, we analytically derived the Poincaré map associated to the bifurcating orbit in the case where the cam profile has a discontinuous acceleration. Then, using the classification strategy for border-collision bifurcations, we proved that the corner-impact causes the fixed point associated to the bifurcating orbit to undergo a nonsmooth saddle node bifurcation. Namely, the fixed point ceases to exist, with the trajectories being attracted toward a chaotic invariant set.

We wish to emphasize the following.

- The analysis presented above applies with minor changes to the case of impact oscillators forced by signals with discontinuous second derivative. As shown above, this leads to maps which are locally piecewise-linear continuous close to a corner-impact bifurcation point. This extends the analysis presented in [8] for the case of an impact oscillator forced by a function with discontinuous first derivative. We conjecture that the properties of the local mapping depend on the degree of discontinuity of the forcing signal. This is the subject of ongoing work.
- As shown in [12], discontinuity-induced bifurcations in flows are usually associated to maps which are not piecewise-linear. Grazing bifurcations of limit cycles are known to be associated to maps with square-root singularities in impacting systems and Filippov systems [26] or maps with higher order nonlinear terms in the case of piecewise-smooth continuous (PWSC) flows. The only cases in the literature where the map was indeed found to be piecewise-linear continuous were corner-collisions in PWSC systems and grazing sliding bifurcations in Filippov systems. So far, no evidence was given of a bifurcation event in impacting systems associated to locally piecewise-linear continuous maps. The corner-impact bifurcation scenario presented in this paper fills this gap in the literature.
- We believe cam-follower systems are a particularly useful set-up to show generically the behavior of impacting systems with discontinuous forcing.

Finally, the results presented here can pave the way to future work toward a better understanding of the complex dynamics of cam-follower systems. This can lead to less conservative solutions to detachment avoidance, hopefully without resorting to highly stiff closing springs, and maybe active control strategies.

Appendix A. Cam profile. We report below the analytical description of the representative cam profile considered in this paper. As shown in Figure 9, in this case the cam profile is the result of a geometrically based design.

The lift profile $c(\theta)$ can be defined from the construction as a piecewise-smooth function of the angle θ as

$$c(\theta) = \begin{cases} c_0(\theta) & \text{if } 0 < \theta \le \frac{\pi}{2} - \theta_1, \\ c_1(\theta) & \text{if } \frac{\pi}{2} - \theta_1 < \theta \le \frac{\pi}{2} - \theta_2, \\ c_2(\theta) & \text{if } \frac{\pi}{2} - \theta_2 < \theta \le \frac{\pi}{2} - \theta_3, \\ c_3(\theta) & \text{if } \frac{\pi}{2} - \theta_3 < \theta \le \pi, \end{cases} \tag{A.1}$$

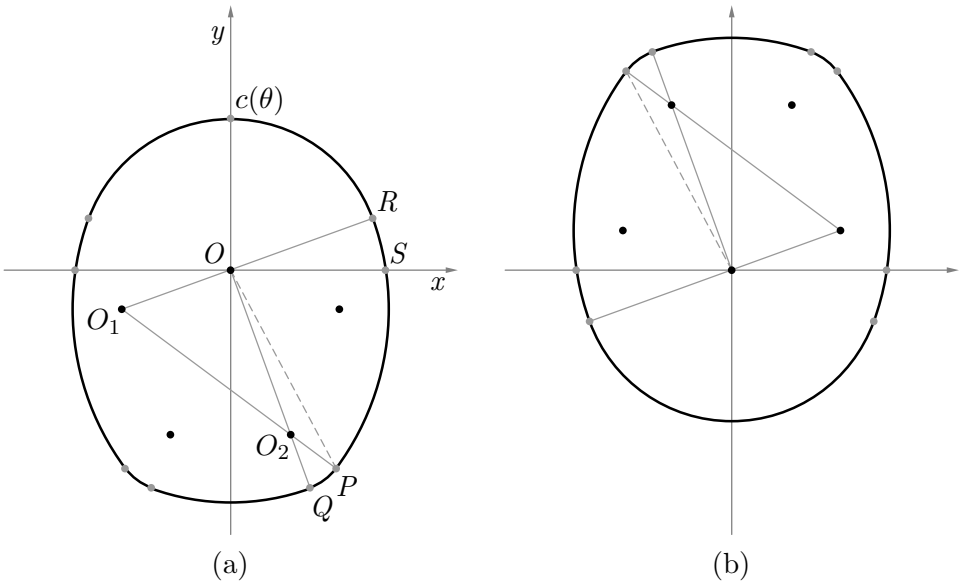


Figure 9. *Cam profile definition.* (a) $\theta = 0$. (b) $\theta = \pi$.

$$
\begin{aligned}
c_0(\theta) &= \rho_0,\\
c_1(\theta) &= -\kappa_1 \sin(\theta + \theta_1) + (\rho_1^2 - \kappa_1^2 \cos(\theta + \theta_1)^2)^{\frac{1}{2}},\\
c_2(\theta) &= \kappa_2 \sin(\theta + \theta_3) + (\rho_2^2 - \kappa_2^2 \cos(\theta + \theta_3)^2)^{\frac{1}{2}},\\
c_3(\theta) &= \rho_3,
\end{aligned}
$$

where $\theta_1 = \angle SOR$, $\theta_2 = \angle SOP$, and $\theta_3 = \angle SOQ$. Additionally, κ_i and ρ_i are constant parameters given by our particular geometrical construction of the cam as (see Figure 9)

$$
\begin{aligned}
\kappa_1 &= \|\overline{OO_1}\|, \quad \rho_0 = \|\overline{OR}\|, \quad \rho_2 = \|\overline{O_2P}\|,\\
\kappa_2 &= \|\overline{OO_2}\|, \quad \rho_1 = \|\overline{O_1R}\|, \quad \rho_3 = \|\overline{OQ}\|.
\end{aligned} \tag{A.2}
$$

Acknowledgment. The authors wish to thank the anonymous reviewers whose comments led to a consistent revision of the original version of this manuscript.

REFERENCES

[1] K. Akiba, A. Shimizu, and H. Sakai, *A Comprehensive Simulation of High Speed Driven Valve Trains*, SAE technical paper 810865, SAE International, Warrendale, PA, 1981.
[2] R. Alzate, M. di Bernardo, U. Montanaro, and S. Santini, *Experimental and numerical verification of bifurcations and chaos in cam-follower impacting systems*, Nonlinear Dynam., to appear.
[3] P. Barkan, *Calculation of high speed valve motion with a flexible overhead linkage*, SAE Trans., 61 (1953), pp. 687–700.
[4] M. Di Bernardo, Ch. Budd, A. Champneys, and P. Kowalczyk, *Bifurcations and Chaos in Piecewise-Smooth Dynamical Systems: Theory and Applications*, Springer-Verlag, New York, 2007.
[5] B. Brogliato, *Nonsmooth Mechanics*, 2nd ed., Springer-Verlag, New York, 1999.
[6] Ch. Budd and F. Dux, *Chattering and related behavior in impact oscillators*, Philos. Trans. R. Soc. London Ser. A, 347 (1994), pp. 385–389.

[7] Ch. Budd, F. Dux, and A. Cliffe, *The effect of frequency and clearance variations on single-degree-of-freedom impact oscillators*, J. Sound Vibration, 3 (1996), pp. 475–502.
[8] Ch. Budd and P. Piiroinen, *Corner bifurcations in non-smoothly forced impact oscillators*, Phys. D, 220 (2006), pp. 127–145.
[9] A. Cardona, E. Lens, and N. Nigro, *Optimal design of cams*, Multibody Syst. Dyn., 7 (2002), pp. 285–305.
[10] T. D. Choi, O. J. Eslinger, C. T. Kelley, J. W. David, and M. Etheridge, *Optimization of automotive valve train components with implicit filtering*, Optimization and Engineering, 1 (2000), pp. 9–27.
[11] M. di Bernardo, *Normal forms of border collisions in high-dimensional maps*, in Proceedings of the IEEE International Symposium on Circuits and Systems, Bangkok, Thailand, 2003, pp. 76–79.
[12] M. di Bernardo, C. J. Budd, and A. R. Champneys, *Corner collision implies border-collision bifurcation*, Phys. D, 154 (2001), pp. 171–194.
[13] M. di Bernardo, M. I. Feigin, S. J. Hogan, and M. E. Homer, *Local analysis of c-bifurcations in n-dimensional piecewise-smooth dynamical systems*, Chaos Solitons Fractals, 10 (1999), pp. 1881–1908.
[14] M. di Bernardo, G. Osorio, and S. Santini, *Chattering and complex behavior of a cam-follower system*, in Proceedings of the Fifth EUROMECH Nonlinear Dynamics Conference, Eindhoven, The Netherlands, 2005, pp. 345–354.
[15] T. L. Dresner and P. Barkan, *New methods for the dynamic analysis of flexible single-input and multi-input cam-follower systems*, J. Mechanical Design, 117 (1995), p. 151.
[16] B. C. Fabien, *The design of dwell-rise-dwell cams with reduced sensibility to parameter variation*, J. Franklin Inst. B, 332 (1995), pp. 195–209.
[17] A. Nordmark and H. Dankowicz, *On the origin and bifurcations of stick-slip oscillations*, Phys. D, 136 (2000), pp. 280–302.
[18] J. Heywood, *Internal Combustion Engine Fundamentals*, McGraw–Hill, New York, 1998.
[19] J. Hogan, L. Higham, and T. C. L. Griffin, *Dynamics of a piecewise linear map with a gap*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 463 (2006), pp. 49–65.
[20] M. P. Koster, *Vibrations of Cam Mechanisms*, Phillips Technical Library Series, Macmillan Press, London, 1974.
[21] M. Kushwaha and H. Rahnejat, *Valve-train dynamics: A simplified tribo-elasto-multi-body analysis*, Proc. Inst. Mechanical Engineers, Part K: Journal of Multi-Body Dynamics, 214 (2001), pp. 1464–4193.
[22] R. I. Leine, D. H. van Campen, and B. L. van de Vrande, *Bifurcations in nonlinear discontinuous systems*, Nonlinear Dynam., 23 (2000), pp. 105–164.
[23] R. I. Leine, B. Brogliato, and H. Nijmeijer, *Periodic motion and bifurcations induced by the Painlevé paradox*, Eur. J. Mech. A Solids, 21 (2002), pp. 869–896.
[24] R. I. Leine and H. Nijmeijer, *Dynamics and Bifurcations in Non-Smooth Mechanical Systems*, Springer-Verlag, New York, 2004.
[25] A. R. Champneys, M. di Bernardo, and C. J. Budd, *Corner collision implies border collision*, Phys. D, 160 (2001), pp. 222–254.
[26] A. R. Champneys, M. di Bernardo, and C. J. Budd, *Normal form maps for grazing bifurcations in n-dimensional piecewise smooth systems*, Phys. D, 154 (2001), pp. 171–194.
[27] I. Merillas, G. Osorio, M. di Bernardo, E. Fossas, and P. Piiroinen, *Complex Dynamics of Cam Follower Systems*, Internal report, SICONOS Project, Dipartimento di Informatica ed Sistemitica, Università degli Studi di Napoli Federico Secondo, Naples, Italy, 2006.
[28] A. Nordmark, *Non-periodic motion caused by grazing incidence in an impact oscillator*, J. Sound Vibration, 145 (1991), pp. 279–297.
[29] M. H. Fredriksson and A. B. Nordmark, *Bifurcations caused by grazing incidence in many degrees of freedom impact oscillators*, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 1261–1276.
[30] R. L. Norton, *Cam Design and Manufacturing Handbook*, 1st ed., Industrial Press, New York, 2002.
[31] R. L. Norton, D. Eovaldi, J. R. Westbrook, and R. L. Stene, *Effect of the Valve-Cam Ramps on Valve Train Dynamics*, SAE Paper 1999-01-0801, SAE International, Warrendale, PA, 1999.

[32] F. Peterka, *Impact oscillator*, in Applied Nonlinear Dynamics and Chaos of Mechanical Systems with Discontinuities A 28, B. de Kraker and M. Wiercigroch, eds., World Scientific, Singapore, 2000, pp. 103–126.

[33] P. T. Piiroinen, L. N. Virgin, and A. R. Champneys, *Chaos and period-adding; experimental and numerical verification of the grazing bifurcation*, J. Nonlinear Sci., 14 (2004), pp. 383–404.

[34] E. Raghavacharyulu and J. S. Rao, *Jump phenomena in cam-follower systems: A continuous-mass-model approach*, in Proceedings of the American Society of Mechanical Engineers Winter Annual Meeting, ASME International, New York, NY, 1976, pp. 1–8.

[35] C. Grebogi and S. Banerjee, *Border collision bifurcations in two-dimensional piecewise smooth maps*, Phys. Rev. E, 59 (1999), pp. 4053–4061.

[36] S. Seidlitz, *Valve train dynamics—a computer study*, SAE technical paper 890620, SAE International, Warrendale, PA, 1989.

[37] M. Teodorescu, V. Votsios, H. Rahnejat, and D. Taraza, *Jounce and impact in cam-tappet conjunction induced by the elastodynamics of valve train system*, Meccanica, 41 (2006), pp. 157–171.

[38] M. D. Todd and L. N. Virgin, *An experimental impact oscillator*, Chaos Solitons Fractals, 8 (1997), pp. 699–714.

[39] T. S. Tumer and Y. Samim Unlusoy, *Nondimensional analysis of jump phenomenon in force-closed cam mechanisms*, Mech. Mach. Theory, 6 (1991), pp. 421–432.

[40] H. S. Yan, M. Tsai, and M. H. Hsu, *An experimental study of the effect of the cam speed on cam-follower systems*, J. Mech. Mach. Theory, 31 (1996), pp. 397–412.

[41] Zh. Zhusubaliyev and E. Mosekilde, *Bifurcations and Chaos in Piecewise-Smooth Dynamical Systems*, World Scientific Series on Nonlinear Science Series A: Monographs and Treatises 44, World Scientific, River Edge, NJ, 2003.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 39–62
© 2008 Society for Industrial and Applied Mathematics

Quantitative Characteristic of Rotating Stall and Surge for Moore–Greitzer PDE Model of an Axial Flow Compressor*

MingQing Xiao†

Abstract. A commonly used mathematical model for axial flow compressors that captures the flow behavior of a compression system, known as the Moore–Greitzer model, consists of a PDE and two ODEs. The PDE describes the dynamical behavior of disturbances in the inlet region of the compression system, and the two ODEs describe the coupling of the disturbances with the mean flow. In this paper, we obtain a quantitative characteristic Δ, depending on the compressor geometry, to identify the type of oscillations of the system. More specifically, the sign of Δ indicates the physical oscillations predominated by rotating stall or by surge. In mathematical terminology, these three types of oscillations are distinctive Hopf bifurcations occurring in the system as the throttle coefficient decreases, and they present quite different dynamical behaviors in axial engine compressors. Estimations of oscillation frequencies corresponding to surge and rotating stall, respectively, are also given in the paper. Numerical simulations are provided to demonstrate different types of flow oscillation of the system.

Key words. Moore–Greitzer PDE model, axial flow compressors, rotating stall and surge, Hopf bifurcations

AMS subject classifications. 34D45, 34K15, 35B40, 35K55

DOI. 10.1137/060658254

1. Introduction.

The problem of compressor instability has been studied for many years because engine efficiency could be reduced significantly by rotating stall and surge (e.g., [14], [12], [9], and references therein). Rotating stall, which corresponds to a traveling wave of gas around the annulus of the compressor, occurs when a nonaxisymmetric flow disturbance develops (around the annulus of the rotor) and causes a drastic reduction in the performance of the compressor. On the other hand, surge is an axisymmetric oscillation across the compression system. It is a low-frequency, large-amplitude oscillation of the mean flow rate in the compressor which induces a high blade, causing stress levels and possibly reversed flow which affects flow conditions throughout the entire compression system. Such instability phenomena can damage engine components during operations.

Moore and Greitzer [14] developed a relatively simple model that captures the dynamical behavior of a compression system. This model consists of a PDE, which describes the behavior of disturbances in the inlet region of compression systems, and two ODEs, which describe the coupling of the disturbances with the mean flow. Considerable research has been carried out on the analysis and control of the stall and the surge, using simplified models obtained by a Galerkin projection of the PDE describing the stall dynamics onto its *first* or *first several Fourier models* (for example, see [12], [10], [2], [11], [19], and references therein), but relatively

*Received by the editors April 26, 2006; accepted for publication (in revised form) by D. Barkley June 27, 2007; published electronically January 16, 2008. This research was supported in part by NSF DMS-0605181.
http://www.siam.org/journals/siads/7-1/65825.html
†Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (mxiao@math.siu.edu).

little research has been conducted heretofore on the analysis of the full PDE model. Although the single harmonic Galerkin method is a useful approximation for general transients and provides a correct picture of the nonlinear effect of rotating stall on performances, it cannot accurately represent a wave of the relaxation type of transition which arises in fully developed rotating stall. Experimental observations of the behavior of disturbance velocity potential in compression systems indicate that its shape is often far from being sinusoidal [14]. Recently studies on the analysis (and control) of the full PDE model of Moore and Greitzer, mainly focusing on low-speed compressors, were conducted by Banaszuk, Hauksson, and Mezić [3], Birnir and Hauksson [4], [5], [6], Chung and Titi [1], Xiao and Başar [18], [20], and Xiao [17].

To the best of our knowledge, currently there is no quantitative criteria available in the literature for identifying which type of flow oscillations, such as rotating stall or surge, will predominate the dynamics in the analysis and control of the full Moore–Greitzer model. As Moore and Greitzer point out in [14], these two types of oscillations differ fundamentally in that the flow dominated by stall is nonaxisymmetric, while the flow dominated by surge is symmetric. Moreover, from the point of view of control, surge is linearly controllable and rotating stall is not linearly controllable (by the throttle coefficient). Thus the control strategies for these two types of oscillations need considerably different methodologies. This motivates us to look for a quantitative criterion which can identify the different types of oscillations occurring inside a compressor.

In this paper we provide a useful indicator Δ which depends on the geometric structure of a compressor and is computable from the parameters of the system for a given axial flow engine compressor. The sign of Δ predicts which type of oscillations will develop and ultimately predominate the dynamical behavior of the system. Let us first recall the stability analysis of nonlinear evolution equations in a Banach space by studying its linearized system. Consider Cauchy problems of form

$$\dot{x} = Ax + f(x), \qquad x(0) = x_0, \tag{1.1}$$

where $x(t)$ takes values in a Banach space $(X, \|\cdot\|)$, A generates a continuous semigroup on X, f is Fréchet differentiable, with Fréchet derivative df, and the mapping $x \to df_x$ is continuous from X to $\mathcal{L}(X)$ ($\mathcal{L}(X)$ is the space of linear maps of X onto itself with the usual norm-topology). Let $x_e \in X$ be an equilibrium solution; that is, x_e is a steady solution of $Ax + f(x) = 0$. x_e is said to be locally exponentially asymptotically stable if there is a neighborhood N of x_e and positive numbers c, α such that if x is a solution of (1.1) with $x(0) \in N$, then $x(t)$ exists for all $t > 0$ and

$$\|x(t) - x_e\| \le ce^{-\alpha t}\|x(0) - x_e\|, \qquad t > 0. \tag{1.2}$$

Now we consider the linearized system of (1.1) at the equilibrium x_e:

$$\dot{x} = Ax + df_{x_e}(x), \qquad x(0) = x_0. \tag{1.3}$$

It is known that if $A + df_{x_e}$ generates a compact C_0 semigroup $S(t)$ for $t > 0$, then the spectrum of $A + df_{x_e}$ consists solely of a point spectrum (e.g., see [15]). In this case x_e is asymptotically stable if and only if there exists an $\alpha < 0$ such that the eigenvalues of $A + df_{x_e}$

lie in the half-space $\operatorname{Re} z \leq \alpha$. The connection of stability between (1.1) and (1.3) in this case is given by Smoller [16] as follows: if x_e is an asymptotically stable solution of (1.3), then x_e is also a locally asymptotically stable solution of (1.1). Thus the stability of the system around x_e will be determined by the eigenvalues of $A + df_{x_e}$ provided that the real parts of all eigenvalues are nonzero. Although when some real parts of the eigenvalues become zero, $A+df_{x_e}$ is not able to provide any stability information, and other methods such as the center manifold approach or the averaging integral method must be employed to study the nonlinear effect in order to determine the stability of the system [20]; however, for the Hopf bifurcation case, $A + df_{x_e}$ still offers us the information of periods of bifurcating closed orbits, which are useful for the study of oscillations of the system.

The axial flow engine compressor PDE model, also called the Moore–Greitzer PDE model, can be put into an abstract evolution form,

$$\dot{x} = Ax + f(\mu, x), \qquad x(0) = x_0, \tag{1.4}$$

in an infinite-dimensional Hilbert space H, where μ is a real parameter varied in $(0, \infty)$. Both equilibrium solutions and the spectrum of $A + df_{x_e}(\mu)$ depend on the parameter μ. As μ decreases, the spectrum of $A + df_{x_e}(\mu)$ moves from the left half plane to the right half plane, which results in an oscillation of the system. In this paper we provide a quantitative criterion which can identify three different types of Hopf bifurcations by checking $\Delta > 0$, or $\Delta < 0$, or $\Delta = 0$. These three types of Hopf bifurcations correspond to the occurrence of three types of physical oscillations predominated by surge, or rotating stall, or a mixture of both. Estimations of oscillation frequencies corresponding to surge and rotating stall, respectively, are also provided in the paper.

2. Axial flow engine compressor model. The basic compression system is shown in Figure 1. It consists of an inlet duct, a compressor, a downstream duct, a plenum, and a throttle. The compressor operates between the inlet duct and the downstream duct. The flow enters the compressor from the inlet duct and exits into the plenum through the downstream duct. The throttle controls the flow through the system at the plenum exit in order to model the turbine. The compressor geometry is shown in Figure 2. Variable ϕ in Figure 1 represents the local, unsteady axial velocity at the compressor face, which depends on both the angle θ around the wheel and (dimensionless) time $\xi = t$. Ψ is the pressure in the plenum, which is defined as

$$\Psi := \frac{\text{exit static pressure} - \text{inlet total pressure}}{\text{density} \times \text{mean rotary speed}^2}.$$

The annulus-averaged axial flow coefficient Φ of ϕ around the wheel is defined as

$$\Phi(t) := \frac{1}{2\pi} \int_0^{2\pi} \phi(\theta, t) d\theta.$$

The model consists of three states, Φ, Ψ, and g, where $\Phi = \Phi(t)$ is the annulus-averaged axial flow coefficient, and $\Psi = \Psi(t)$ is the annulus-averaged total-to-static pressure rise coefficient. We denote $\phi(t, \theta, \eta)$ as the upstream disturbance velocity at any point (θ, η) of the compressor, where η represents the position along the axial direction of the compressor. Let

$$g = \phi_\eta|_{\eta=0},$$

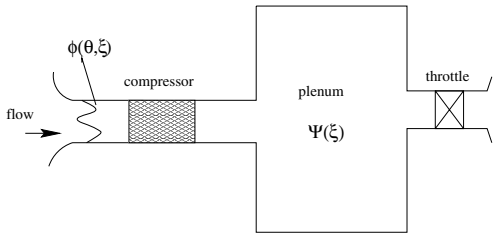


Figure 1. *An outline of a compression system.*

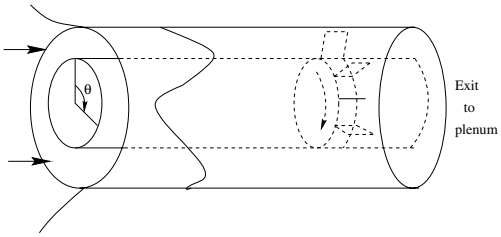


Figure 2. *Compressor geometry.*

which represents the derivative of the upstream disturbance velocity with respect to η at the duct entrance $\eta = 0$. The free variables in the system are t and θ, where t is the dimensionless time variable, and θ is the circumferential angle around the compressor annulus. Before the flow enters the duct, the upstream disturbance velocity ϕ satisfies the Laplace equation,

$$\phi_{\theta\theta} + \phi_{\eta\eta} = 0, \quad (\theta, \eta) \in [0, 2\pi] \times (-\infty, 0), \tag{2.1}$$

and $\phi \equiv 0$ at $\eta = -\infty$. At the duct entrance, we denote $\phi_\eta|_{\eta=0} = g$ and the pressure rise coefficient satisfies

$$\Psi(t) = \psi_c(\Phi + g) - l_c\tfrac{d\Phi}{dt} - \tfrac{\partial}{\partial t}(m\phi + \tfrac{1}{a}g) + \tfrac{1}{2a}g_\theta - \tfrac{\nu}{2a}g_{\theta\theta}, \tag{2.2}$$

where ψ_c is the characteristic function of the compressor, which will be given later; m is the duct parameter; a is the internal compressor lag; l_c is the length of the compressor; and ν is the viscous coefficient. By computing the circumferential mean of boundary condition (2.2), the annulus-averaged axial flow coefficient Φ has a dynamics in the form of

$$l_c\frac{d\Phi}{dt} = -\Psi + \frac{1}{2\pi}\int_0^{2\pi} \psi_c(\Phi + g)d\theta, \tag{2.3}$$

which describes the change of mass flow through the compressor. The annulus-averaged total-to-static pressure rise coefficient Ψ satisfies the pressure-balance equation

$$l_c\frac{d\Psi}{dt} = \frac{1}{4B^2}(\Phi - \mu\sqrt{\Psi}), \tag{2.4}$$

where B is the plenum/compressor volume ratio, which is called the Greitzer B-parameter, and μ is the throttle coefficient, which is used to control the mass flow through the throttle.

It accounts for dynamic pressure changes downstream of the compressor exit, in the plenum and across the throttle.

The compressor characteristic is an inherent feature of the compressor and has a cubic form in terms of Φ (see [14]), as shown in Figure 3,

$$\psi_c(\Phi) = \psi_{c_0} + H\left[1 + \frac{3}{2}\left(\frac{\Phi}{W} - 1\right) - \frac{1}{2}\left(\frac{\Phi}{W} - 1\right)^3\right], \tag{2.5}$$

where $\psi_{c_0} > 0$ is the shut-off value parameter, $H > 0$ is the semiheight parameter, and $W > 0$ is the semiwidth parameter.

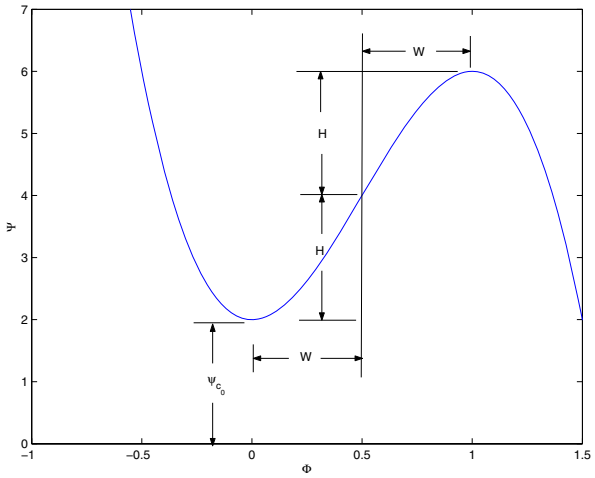


Figure 3. *Notation used in the definition of compressor characteristic with* $w = 0.5$, *and* $H = \psi_{c_0} = 2$.

The throttle characteristic is defined to be

$$\Phi = \mu\sqrt{\Psi}, \tag{2.6}$$

which characterizes the mass flow through the throttle, where $\mu > 0$ is the throttle coefficient which controls the flow through the throttle. When μ is changed from small to large, it implies that the throttle is being opened up and more flow is leaving the plenum (see Figure 1).

Let us denote the intersection of the compressor characteristic $\Psi = \psi_c(\Phi)$ and throttle characteristic $\Phi = \mu\sqrt{\Psi}$ by $(\phi_e(\mu), \psi_e(\mu))$ (see Figure 4). Then $(0, \phi_e(\mu), \psi_e(\mu))$ is called an unstalled or nominal equilibrium point of (2.2)–(2.4).

The viscous coefficient ν usually is small in axial flow engine compression systems; thus throughout this paper we assume[1]

$$\frac{\nu W}{3aH} \leq 1. \tag{2.7}$$

[1]If $\frac{\nu W}{3aH} > 1$, then the system becomes less interesting, since in this case surge will always become predominant.

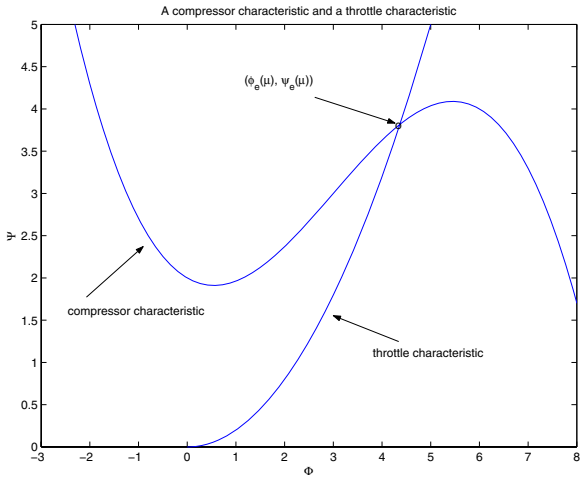


Figure 4. *Notation used in the definition of a nominal equilibrium point.*

This assumption implies that the slope of $\Psi = \psi_c(\Phi)$ at the reflection point $\Phi = W$ is not smaller than $\frac{\nu}{2a}$. This can be seen by observing

$$\frac{\nu}{2a} \leq \frac{3H}{2W} = \psi_c'(W).$$

3. Main results.

Theorem 3.1.

1. *When the throttle coefficient* $\mu \geq \frac{2W}{\sqrt{2H+\psi_{c_0}}}$, *the state* $(g(t,\theta), \Phi(t), \Psi(t))$ *governed by* (2.2)–(2.4) *is locally asymptotically stable and as* $t \to +\infty$

$$\int_0^{2\pi} |g(t,\theta)|^2 \, d\theta \to 0, \qquad \Phi(t) \to \phi_e(\mu), \qquad \Psi(t) \to \psi_e(\mu).$$

2. *Suppose* $\frac{\nu W}{3aH} \leq 1$. *Let*

$$\Delta = \frac{\psi_{c_0} + H\left[1 + \frac{3}{2}\sqrt{1-\frac{\nu W}{3aH}} - \frac{1}{2}\left(\sqrt{1-\frac{\nu W}{3aH}}\right)^3\right]}{W\left(1+\sqrt{1-\frac{\nu W}{3aH}}\right)} - \frac{a}{4B^2\nu}. \tag{3.1}$$

Then, as the throttle coefficient μ *is decreasing, oscillations inside the compressor occur. If* $\Delta > 0$, *the frequency of the flow oscillation is smaller than* $1/(4\sqrt{2}\pi B l_c)$ *(i.e., the flow oscillation is predominated by surge). While* $\Delta < 0$, *the frequency of the flow oscillation is greater than* $1/\left(4\pi\left(\frac{3aH}{\nu W} + am\right)\right)$ *(i.e., the oscillation is predominated by rotating stall). When* $\Delta = 0$ *a motion composed of the above two types of oscillations is expected.*

The above main theorem implies that the sign of

$$\Delta = \frac{\psi_{c_0} + H\left[1 + \frac{3}{2}\sqrt{1 - \frac{\nu W}{3aH}} - \frac{1}{2}\left(\sqrt{1 - \frac{\nu W}{3aH}}\right)^3\right]}{W\left(1 + \sqrt{1 - \frac{\nu W}{3aH}}\right)} - \frac{a}{4B^2\nu}$$

describes how the spectrum moves across the imaginary axis. The quantity Δ depends on the geometry structure (parameters W, H, a, ψ_{c_0}, B) of the axial flow compressor system as well as the viscous coefficient ν. An important observation for Δ is that it can also be written as

$$\Delta = \frac{\psi_e^{(\nu)}}{\phi_e^{(\nu)}} - \frac{a}{4B^2\nu},$$

where

$$\phi_e^{(\nu)} := W\left(1 + \sqrt{1 - \frac{\nu W}{3aH}}\right), \qquad \psi_e^{(\nu)} = \psi_c\left(\phi_e^{(\nu)}\right),$$

and, according to the definition of the compressor characteristic ψ_c, one can see that

$$\psi_e^{(\nu)} := \psi_c\left(\phi_e^{(\nu)}\right) = \psi_{c_0} + H\left[1 + \frac{3}{2}\sqrt{1 - \frac{\nu W}{3aH}} - \frac{1}{2}\left(\sqrt{1 - \frac{\nu W}{3aH}}\right)^3\right].$$

Thus the sign of Δ implies the relationship between two quantities: the slope of the line segment through two points $(\phi_e^{(\nu)}, \psi_e^{(\nu)})$, $(0,0)$ and the fraction $\frac{a}{4B^2\nu}$. For the high speed compressor, parameter B (the plenum/compressor volume ratio) is large, which leads to $\Delta > 0$ in most cases. In this case we shall show that a pair of eigenvalues $\gamma_{\pm 1}$ crosses the imaginary axis with nonzero speed, and a Hopf bifurcation occurs at $\mu = \hat{\mu}$ according to the Hopf bifurcation theorem (see, for example, [7], [8]). Experiments show that for a high-speed compressor there is often a low frequency with large-amplitude flow oscillation through the compressor, which is called *surge*, and it can induce vibrations in other components of the compression system, such as connected piping. On the other hand, for the low-speed compressor, B is small and usually leads to $\Delta < 0$. In this case another pair of eigenvalues $\lambda_{\pm 1}$ crosses the imaginary axis with nonzero speed, and the compressor will enter another type of high-frequency oscillation, called *rotating stall* from the experiments. Thus the sign of Δ is a quantitative criterion for identifying the dynamics predominated by surge or by rotating stall. If $\Delta = 0$, the dynamics of the system becomes even more complicated at $\mu = \mu_\nu$ since two Hopf bifurcations appear simultaneously, as two pairs of eigenvalues $\lambda_{\pm 1}$ and $\gamma_{\pm 1}$ cross the imaginary axis with nonzero speed, and in such a case a mixed type of oscillation takes place. Therefore, Δ predicts the system dynamics as the throttle of the plenum is being closed (μ is reduced). We will provide details in the following sections.

3.1. Abstract formulation. Let $\dot{L}^2(0, 2\pi)$ be the space of all square integrable 2π-periodic complex-valued functions with zero average, that is,

$$\int_0^{2\pi} \phi(\theta) d\theta = 0$$

for any $\phi \in \dot{L}^2(0,2\pi)$. Note that for any $\phi \in \dot{L}^2(0,2\pi)$, its Fourier series is given by

$$\phi(\theta) = \sum_{n\in\mathbf{Z}\setminus\{0\}} \phi_n e^{in\theta}, \quad \text{where} \quad \phi_n = \int_0^{2\pi} \phi(\theta) e^{-in\theta} d\theta,$$

where $\mathbf{Z}$ stands for the set of all integers. Boundary condition $\phi = 0$ at $\eta = -\infty$ implies that Laplace's equation (2.1) has a solution of the form

$$\phi(t,\theta,\eta) = \sum_{n\in\mathbf{Z}\setminus\{0\}} \phi_n(t) e^{|n|\eta + in\theta}.$$

We define the flow disturbance at the compressor face to be $g := \phi_\eta|_{\eta=0}$; it thus can be written as

$$g(t,\theta) = \sum_{n\in\mathbf{Z}\setminus\{0\}} |n| g_n e^{in\theta}.$$

We next introduce a linear operator $\Pi : \dot{L}^2(0,2\pi) \to \dot{L}^2(0,2\pi)$ by

$$\Pi(\phi) = \sum_{n\in\mathbf{Z}\setminus\{0\}} \left\{1 + \frac{am}{|n|}\right\} \phi_n e^{in\theta}$$

for any $\phi = \sum_{n\in\mathbf{Z}\setminus\{0\}} \phi_n e^{in\cdot} \in \dot{L}^2(0,2\pi)$. It is not difficult to see that Π is a positive definite, self-adjoint linear operator on $\dot{L}^2(0,2\pi)$. Thus

$$\langle \phi, \psi \rangle_\Pi := \langle \phi, \Pi\psi \rangle_{L^2(0,2\pi)}$$

defines an equivalent inner product on $\dot{L}^2(0,2\pi)$, and we denote by $\dot{L}^2_\Pi(0,2\pi)$ the space which consists of elements of $\dot{L}^2(0,2\pi)$ with inner product $\langle\cdot,\cdot\rangle_\Pi$. By using the operator Π, (2.2) can equivalently be written as

$$\frac{\partial g}{\partial t} = \Pi^{-1}\left(\frac{\nu}{2}\frac{\partial^2 g(t,\theta)}{\partial\theta^2} - \frac{1}{2}\frac{\partial g(t,\theta)}{\partial\theta}\right) + a\Pi^{-1}\left(\psi_c(\Phi(t)+g(t,\theta)) - \frac{1}{2\pi}\int_0^{2\pi} \psi_c(\Phi(t)+g(t,\theta))d\theta\right).$$

Let $X = L^2_\Pi(0,2\pi)\times\mathbb{R}\times\mathbb{R}$, with inner product

$$\langle x_1, x_2\rangle = a^{-1}\langle g_1, g_2\rangle_{\dot{L}^2_\Pi(0,2\pi)} + l_c\Phi_1\Phi_2 + (4l_cB^2)\Psi_1\Psi_2, \tag{3.2}$$

where $x_i = (g_i, \Phi_i, \Psi_i)^T \in X$, $i = 1,2$, and let the norm on this space be defined by $\|x\| := \sqrt{\langle x, x\rangle}$ for $x \in X$. The system defined by (2.2), (2.3), and (2.4) can thus be written as

$$\frac{\partial}{\partial t}\begin{bmatrix} g \\ \Phi \\ \Psi \end{bmatrix} = \begin{bmatrix} \Pi^{-1}\left(\frac{\nu}{2}\frac{\partial^2}{\partial\theta^2} - \frac{1}{2}\frac{\partial}{\partial\theta}\right) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} g \\ \Phi \\ \Psi \end{bmatrix} + \begin{bmatrix} a\Pi^{-1}(\psi_c(\Phi+g) + \overline{\psi}_c) \\ \frac{1}{l_c}(\overline{\psi}_c - \Psi) \\ \frac{1}{4l_cB^2}(\Phi - \mu\sqrt{\Psi}) \end{bmatrix}, \tag{3.3}$$

where

$$\overline{\psi}_c = \frac{1}{2\pi}\int_0^{2\pi} \psi_c(\Phi+g)d\theta,$$

or in an evolution form of

$$\frac{dx}{dt} = Ax + f(\mu, x), \tag{3.4}$$

where $x = (g, \Phi, \Psi)^T$ and the linear operator A is defined by

$$A = \begin{bmatrix} \Pi^{-1}\left(\frac{\nu}{2}\frac{\partial^2}{\partial\theta^2} - \frac{1}{2}\frac{\partial}{\partial\theta}\right) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.5}$$

with

$$D(A) = \left\{ x \in X \;\middle|\; \frac{\partial g}{\partial\theta}, \frac{\partial^2 g}{\partial\theta^2} \in \dot{L}^2_\Pi(0, 2\pi), \text{ and } g(0) = g(2\pi), \text{ where } x = (g, \Phi, \Psi)^T \in X \right\},$$

and the function $f(\mu, x)$ is defined to be

$$f(\mu, x) = \begin{bmatrix} a\Pi^{-1}\left(\psi_c(\Psi - g) - \overline{\psi}_c\right) \\ \frac{1}{l_c}(\overline{\psi}_c - \Psi) \\ \frac{1}{4l_cB^2}(\Phi - \mu\sqrt{\Psi}) \end{bmatrix},$$

where $x = (g, \Phi, \Psi)^T$.

Remark. When we study (3.4), all parameters such as B, l_c, a, ν, W, H, and ψ_{c_0} are fixed except the throttle coefficient μ, which is varied in $(0, \infty)$.

We will show the main result through a sequence of lemmas.

Lemma 3.1. *The operator A defined in (3.5) is the infinitesimal generator of an analytic compact C_0 semigroup on X.*

Proof of Lemma 3.1 is provided in the appendix. We would like to point out here that Chung and Titi provide a detailed discussion of the analyticity of the solution of (3.4) by establishing Geverey regularity of the system [1]. Their main purpose is to show the existence of a global invariant manifold; thus some subtle technique has to be applied to deal with the term $\sqrt{\Psi}$ since it is not an analytic function near $\Psi = 0$. In this paper we focus on local dynamics of the systems since in reality small perturbations occur more often. Lemma 3.1 can lead to local analyticity of the solution without heavy exposition of other mathematical tools.

A direct calculation gives the following.

Lemma 3.2. *The Fréchet derivative of $f(\mu, x)$ at $x_e(\mu) = (0, \phi_e(\mu), \psi_e(\mu))$ is*

$$df_{x_e}(\mu) = \begin{bmatrix} a\Pi^{-1}\left(\psi_c'(\phi_e(\mu)) - \overline{\psi}_c{}'\right) & 0 & 0 \\ \frac{1}{l_c}\overline{\psi}_c{}' & \frac{1}{l_c}\overline{\psi}_c{}' & -\frac{1}{l_c} \\ 0 & \frac{1}{4l_cB^2} & -\frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e(\mu)}} \end{bmatrix}, \tag{3.6}$$

where ψ_c' and $\overline{\psi}_c'$ are the Fréchet derivatives of ψ_c and $\overline{\psi}_c$, respectively, at x, and for any $x = (g, \Phi, \Psi) \in X$

$$df_{x_e}(\mu)(x) = \begin{bmatrix} a\Pi^{-1}\left(\psi_c'(\phi_e(\mu))g\right) \\ \frac{1}{l_c}\left(\psi_c'(\phi_e(\mu))\Phi - \Psi\right) \\ \frac{1}{4l_cB^2}\left(\Phi - \frac{\mu\Psi}{2\sqrt{\psi_e(\mu)}}\right) \end{bmatrix}. \tag{3.7}$$

Notice that $df_{x_e}(\mu)$ is a bounded operator on X. According to Proposition 3.1.4 and Corollary 3.2.2 of [15], we have the following lemma.

Lemma 3.3. *$A+df_{x_e}(\mu)$ is the infinitesimal generator of an analytic compact C_0 semigroup on X.*

Lemma 3.4 implies that the eigenvalues of $A + df_{x_e}(\mu)$ will determine the stability of the system around the equilibrium point x_e if their real parts are not equal to zero, according to the discussion given in section 1.

Lemma 3.4. *The point spectrum of $A + df_{x_e}(\mu)$ is given by*

$$\sigma\left(A + df_{x_e}(\mu)\right) = \{\gamma_{\pm1}(\mu), \lambda_n(\mu),\ n = \pm1, \pm2, \dots\},$$

where

$$\lambda_n(\mu) = \frac{a|n|}{|n| + am}\left(\psi_c'\left(\phi_e\right) - \frac{\nu}{2a}n^2 - \frac{1}{2a}ni\right), \tag{3.8}$$

and

$$\gamma_{\pm1}(\mu) = \frac{\frac{1}{l_c}\psi_c'(\phi_e) - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}} \pm \sqrt{\left(\frac{1}{l_c}\psi_c'(\phi_e) - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}}\right)^2 + \frac{\psi_c'(\phi_e)}{l_c^2B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2B^2}}}{2}, \tag{3.9}$$

where (ϕ_e, ψ_e) satisfies $\phi_e = \psi_c(\phi_e)$, $\phi_e = \mu\sqrt{\psi_e}$. The eigenvector corresponding to $\lambda_n(\mu)$ is given by $\vec{v}_n = (e^{in\theta}, 0, 0)$.

Proof. Recall that

$$A + df_{x_e}(\mu) = \begin{bmatrix} \Pi^{-1}\left(\frac{\nu}{2}\frac{\partial^2}{\partial\theta^2} - \frac{1}{2}\frac{\partial}{\partial\theta}\right) + a\Pi^{-1}\left(\psi_c' - \overline{\psi}_c{}'\right) & 0 & 0 \\ \frac{1}{l_c}\overline{\psi}_c{}' & \frac{1}{l_c}\overline{\psi}_c{}' & -\frac{1}{l_c} \\ 0 & \frac{1}{4l_cB^2} & -\frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}} \end{bmatrix} \tag{3.10}$$

with

$$D(A+df_{x_e}(\mu)) = \left\{x \in X \mid \frac{\partial g}{\partial\theta}, \frac{\partial^2 g}{\partial\theta^2} \in \dot{L}^2_\Pi(0, 2\pi), \text{ and } g(0) = g(2\pi), \text{ where } x = (g, \Phi, \Psi)^T \in X\right\}.$$

In order to obtain the eigenvalues and eigenvectors, we need to solve the boundary value problem

$$\Pi^{-1}\left(\frac{\nu}{2}\frac{\partial^2 v^{(1)}(\theta)}{\partial\theta^2} - \frac{1}{2}\frac{\partial v^{(1)}(\theta)}{\partial\theta}\right) + a\Pi^{-1}\psi_c' v^{(1)}(\theta) = \lambda v^{(1)}(\theta), \tag{3.11}$$

$$\frac{1}{l_c}\psi_c'(\phi_e)v^{(2)} - \frac{1}{l_c}v^{(3)} = \lambda v^{(2)}, \tag{3.12}$$

$$\frac{1}{4l_cB^2}v^{(2)} - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}}v^{(3)} = \lambda v^{(3)}, \tag{3.13}$$

where $v^{(1)}(0) = v^{(1)}(2\pi)$.

The solutions of (3.11)–(3.13) are given by

$$\lambda_n(\mu) = \frac{a|n|}{|n| + am}\left(\psi_c'(\phi_e(\mu)) - \frac{\nu}{2a}n^2 - \frac{1}{2a}ni\right), \quad \vec{v}_n = (e^{in\theta}, 0, 0), \quad n = \pm 1, \pm 2, \dots, \tag{3.14}$$

and

$$\gamma_{\pm 1}(\mu) = \frac{\frac{1}{l_c}\psi_c'(\phi_e) - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}} \pm \sqrt{\left(\frac{1}{l_c}\psi_c'(\phi_e) - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}}\right)^2 + \frac{\psi_c'(\phi_e)}{l_c^2B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2B^2}}}{2}, \tag{3.15}$$

$\vec{v}_{\gamma_i} = (0, \hat{\phi}_i, \hat{\psi}_i)$, $i = \pm 1$, where $\hat{\phi}_i, \hat{\psi}_i$ satisfy

$$\frac{1}{l_c}\psi_c'(\phi_e)\hat{\phi}_i - \frac{1}{l_c}\hat{\psi}_i = \gamma_i\hat{\phi}_i, \tag{3.16}$$

$$\frac{1}{4l_cB^2}\hat{\phi}_i - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}}\hat{\psi}_i = \gamma_i\hat{\psi}_i. \tag{3.17}$$

We thus have obtained the spectrum of $A + df_{x_e}(\mu)$, which is

$$\sigma\left(A + df_{x_e}(\mu)\right) = \{\gamma_{\pm 1}(\mu), \lambda_n(\mu),\ n = \pm 1, \pm 2, \dots\}, \tag{3.18}$$

and the proof is complete. ■

Let X_1 = the closure of span$\{e^{in\theta};\ n = \pm 1, \pm 2, \dots\}$ and $X_2 = \mathbb{R}^2$. Clearly we have $X = X_1 \oplus X_2$.

Lemma 3.5. *Let the semigroup $T_1^\mu(t)$ and $T_2^\mu(t)$ be the restrictions of $e^{t(A+df_{x_e}(\mu))}$ to X_1 and X_2, respectively. Then X_1 is $T_1^\mu(t)$-invariant and X_2 is $T_2^\mu(t)$-invariant. Thus for any $(g_0, \Phi_0, \Psi_0) \in X$ we have*

$$e^{t(A+df_{x_e}(\mu))}\begin{pmatrix} g_0 \\ \Phi_0 \\ \Psi_0 \end{pmatrix} = \begin{pmatrix} T_1^\mu(t)g_0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ T_2^\mu(t)\begin{pmatrix} \Phi_0 \\ \Psi_0 \end{pmatrix} \end{pmatrix}. \tag{3.19}$$

Proof. Recall that

$$A + df_{x_e}(\mu) = \begin{bmatrix} \Pi^{-1}\left(\frac{\nu}{2}\frac{\partial^2}{\partial\theta^2} - \frac{1}{2}\frac{\partial}{\partial\theta}\right) + a\Pi^{-1}\left(\psi_c' - \overline{\psi}_c'\right) & 0 & 0 \\ \frac{1}{l_c}\overline{\psi}_c' & \frac{1}{l_c}\overline{\psi}_c' & -\frac{1}{l_c} \\ 0 & \frac{1}{4l_cB^2} & -\frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}} \end{bmatrix}. \tag{3.20}$$

For any $(g, \Phi, \Psi) \in D(A)$, we have

$$(A + df_{x_e}(\mu))\begin{pmatrix} g \\ \Phi \\ \Psi \end{pmatrix} = \begin{pmatrix} \Pi^{-1}\left(\frac{\nu}{2}g_{\theta\theta} - \frac{1}{2}g_\theta\right) + a\Pi^{-1}\left(\psi_c'(\phi_e)g\right) \\ \frac{1}{l_c}\left(\psi_c'(\phi_e)\Phi - \Psi\right) \\ \frac{1}{4l_cB^2}\Phi - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}}\Psi \end{pmatrix}, \tag{3.21}$$

since $\overline{\psi}_c' g = 0$ according to the definition of X. Thus the restriction of $A + df_{x_e}(\mu)$ to x_1 is

$$(A + df_{x_e}(\mu))\Big|_{X_1}(g) = \Pi^{-1}\left(\frac{\nu}{2}g_{\theta\theta} - \frac{1}{2}g_\theta\right) + a\Pi^{-1}\left(\psi_c'(\phi_e)g\right) \in X_1, \tag{3.22}$$

and that to x_2 is

$$(A + df_{x_e}(\mu))\Big|_{X_2}\begin{pmatrix}\Phi\\ \Psi\end{pmatrix} = \begin{pmatrix} \frac{1}{l_c}\left(\psi_c'(\phi_e)\Phi - \Psi\right)\\ \frac{1}{4l_cB^2}\Phi - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}}\Psi\end{pmatrix} \in X_2, \tag{3.23}$$

which implies that the restrictions of $A + df_{x_e}(\mu)$ to X_1 and X_2 are invariant, respectively. Therefore, the restrictions of $e^{t(A+df_{x_e}(\mu))}$ to X_1 and X_2 are invariant, respectively. ■

Before we proceed in a further discussion, let us make some important comments based on Lemmas 3.1–3.6.

1. The semigroup $e^{t(A+df_{x_e}(\mu))}$ is asymptotically stable if and only if both $T_1^\mu(t)$ and $T_2^\mu(t)$ are asymptotically stable.
2. The semigroup $T_1^\mu(t)$ is determined by the evolution of the upstream disturbance velocity at the duct entrance, while the semigroup $T_2^\mu(t)$ is decided by dynamics of the flow coefficient and the pressure rise coefficient.
3. As we will see later, as μ is reduced, both $T_1^\mu(t)$ and $T_2^\mu(t)$ will lose asymptotical stability (i.e., their spectra will cross the unit disk, respectively). Resulting dynamical behaviors (rotating stall or surge) of the system depend on which one loses its asymptotic stability first.
4. We will show that the eigenvalues of $T_1^\mu(t)$ and $T_2^\mu(t)$ move across the unit circle with nonzero speeds as μ is reduced. Thus bifurcations to periodic orbits take place.
5. The center manifold theorem indicates that the periodic orbit (rotating stall) due to the spectrum of $T_1^\mu(t)$ is in a two-dimensional subspace of X_1, while the periodic orbit (surge) due to the spectrum of $T_2^\mu(t)$ is in X_2.

In the following, we will show that quantitative Δ given by (3.1) can identify two different dynamical behaviors of the system: rotating stall and surge.

Lemma 3.6. *When $\mu \geq \frac{2W}{\sqrt{2H+\psi_{c_0}}}$, all eigenvalues of $A + df_{x_e}(\mu)$ are in the left half plane.*

Proof. Notice that $\mu \geq \frac{2W}{\sqrt{\psi_{c_0}+2H}}$ implies that $\phi_e(\mu) \geq 2W$, and hence $\psi_c'(\phi_e(\mu)) \leq 0$. Now we consider $\mathrm{Re}(\lambda_n)$ and $\mathrm{Re}(\gamma_{\pm 1})$, respectively. Notice that

$$\begin{aligned}\mathrm{Re}(\lambda_n(\mu)) &= \frac{a|n|}{|n| + am}\left(\psi_c'\left(\phi_e(\mu)\right) - \frac{\nu}{2a}n^2\right) \leq \frac{a|n|}{|n| + am}\left(-\frac{\nu}{2a}n^2\right)\\ &\leq -\frac{\nu}{2(1+am)} \qquad \text{for} \quad n = \pm 1, \pm 2, \ldots.\end{aligned}$$

Next we denote $\chi(\mu)$ as

$$\chi(\mu) = \frac{1}{l_c}\psi_c'(\phi_e(\mu)) - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e(\mu)}}, \tag{3.24}$$

and notice that $\mu = \phi_e/\sqrt{\psi_e}$, and when $\mu \geq \frac{2W}{\sqrt{\psi_{c_0}+2H}}$ (that is, $\phi_e \geq 2W$),

$$\frac{d\chi}{d\mu}(\mu) = \frac{1}{l_c}\frac{d}{d\mu}\psi_c'(\phi_e(\mu)) - \frac{1}{4l_cB^2}\frac{d}{d\mu}\left(\frac{\phi_e(\mu)}{\psi_e(\mu)}\right) < 0, \tag{3.25}$$

since $\psi_c'(\phi_e(\mu))$ is a decreasing function of μ and $\phi_e(\mu)/\psi_e(\mu)$ is an increasing function of μ. Hence when $\mu \geq \mu_0$ one has

$$\chi(\mu) \leq \chi(\mu_0) = \frac{1}{l_c}\psi_c'(\phi_e(\mu_0)) - \frac{1}{4l_cB^2}\frac{\mu_0}{2\sqrt{\psi_e(\mu_0)}} = -\frac{1}{4l_cB^2}\frac{\mu_0}{2\sqrt{\psi_e(\mu_0)}}$$
$$= -\frac{1}{4l_cB^2}\frac{\phi_e(\mu_0)}{2\psi_e(\mu_0)} = -\frac{W}{4l_cB^2(\psi_{c_0}+2H)}$$

since $\psi_c'(\phi_e(\mu_0)) = 0$. Recall that $\gamma_{\pm 1}$ is given by

$$\gamma_{\pm 1}(\mu) = \frac{\frac{1}{l_c}\psi_c'(\phi_e) - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}} \pm \sqrt{\left(\frac{1}{l_c}\psi_c'(\phi_e) - \frac{1}{4l_cB^2}\frac{\mu}{2\sqrt{\psi_e}}\right)^2 + \frac{\psi_c'(\phi_e)}{l_c^2B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2B^2}}}{2}$$
$$= \frac{\chi(\mu) \pm \sqrt{\chi^2(\mu) + \frac{\psi_c'(\phi_e)}{l_c^2B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2B^2}}}{2}.$$

According to the definition of χ given in (3.24), $\gamma_{\pm 1}$ can be further written as

$$\gamma_{\pm 1}(\mu) = \frac{\chi(\mu) \pm \sqrt{\left(\chi(\mu) - \frac{2}{l_c}\psi_c'(\phi_e)\right)^2 - \frac{1}{l_c^2B^2}}}{2}.$$

Notice that $\psi_c'(\phi_e) \leq 0$ and $\chi(\mu) < 0$; thus $\operatorname{Re}\gamma_{\pm 1}(\mu) < 0$, and the proof is complete. ■

Now we are ready to prove part 1 of Theorem 3.1.

Lemma 3.7. *When $\mu > \frac{W}{\sqrt{\psi_{c_0}+H}} := \mu_{00}$, the real parts of eigenvalues $\lambda_n(\mu)$ and $\gamma_{\pm 1}(\mu)$ cross the imaginary axis with nonzero speeds. Specifically, if there are $\mu_1 \geq \mu_{00}$ and $\mu_2 \geq \mu_{00}$ such that $\operatorname{Re}(\lambda_n(\mu_1)) = 0$ and $\operatorname{Re}(\gamma_{\pm 1}(\mu_2)) = 0$, respectively, then*

$$\frac{d}{d\mu}\operatorname{Re}\left(\lambda_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_1} < 0, \qquad \frac{d}{d\mu}\operatorname{Re}\left(\gamma_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_2} < 0.$$

Proof. Recall that (ϕ_e, ψ_e) satisfies $\psi_e = \psi_c(\phi_e)$, $\phi_e = \mu\sqrt{\psi_e}$. Notice that $\mu \geq \frac{W}{\sqrt{\psi_{c_0}+H}}$ implies $\phi_e(\mu) \geq W$. Moreover, $\phi_e(\mu)$ increases as μ increases, which leads to

$$\frac{d\phi_e(\mu)}{d\mu} > 0.$$

Thus for $n = \pm 1, \pm 2, \ldots$, we have

$$\frac{d}{d\mu}\operatorname{Re}\left(\lambda_n(\mu)\right) = \frac{|n|a}{|n| + am}\frac{d^2\psi_c(\phi_e(\mu))}{d\phi_e^2}\frac{d\phi_e(\mu)}{d\mu} < 0$$

since $\Psi = \psi_c(\Phi)$ is concave downward for $\Phi > W$.

For $\operatorname{Re}(\gamma_{\pm 1})$, recall that

$$\gamma_{\pm 1}(\mu) = \frac{\chi(\mu) \pm \sqrt{\chi^2(\mu) + \frac{\psi_c'(\phi_e)}{l_c^2B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2B^2}}}{2}. \tag{3.26}$$

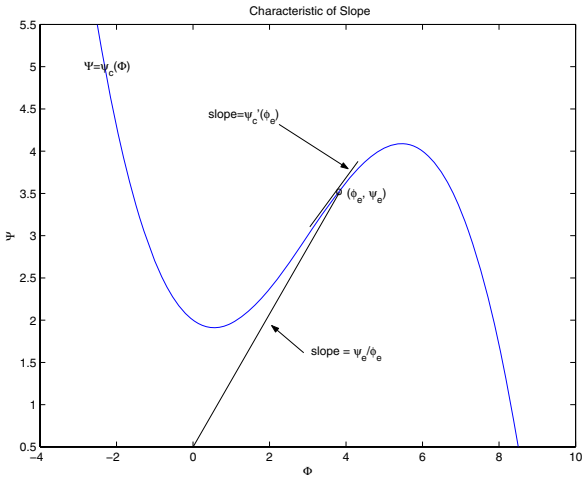


Figure 5. *Slope on the compressor characteristic at* (ϕ_e, ψ_e) *and slope for the line segment between* $(0, 0)$ *and* (ϕ_e, ψ_e).

Notice that

$$\psi_c'(\phi_e) < \frac{\psi_e}{\phi_e}$$

for $\mu > \mu_{00}$ since the slope of the tangent line of $\Psi = \psi_c(\Phi)$ is always smaller than the slope of the line segment between the origin and (ϕ_e, ψ_e), which can be seen in Figure 5. Hence it yields

$$\frac{\psi_c'(\phi_e)}{l_c^2 B^2}\frac{\mu}{2\sqrt{\psi_e}} = \frac{\psi_c'(\phi_e)}{l_c^2 B^2}\frac{\phi_e}{2\psi_e} \le \frac{1}{2l_c^2 B^2},$$

which implies

$$\frac{\psi_c'(\phi_e)}{l_c^2 B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2 B^2} \le -\frac{1}{2l_c^2 B^2} < 0.$$

Thus according to (3.26) $\mathrm{Re}(\gamma_{\pm 1}(\mu)) = 0$ if and only if $\chi(\mu) = 0$ as μ decreases from μ_0. Now we differentiate $\gamma_{\pm 1}(\mu)$ with respect to μ:

$$\frac{d\gamma_{\pm 1}(\mu)}{d\mu} = \frac{1}{2}\left(\frac{d\chi(\mu)}{d\mu} + \frac{2\chi(\mu)\frac{d\chi(\mu)}{d\mu} + \frac{d}{d\mu}\left(\frac{\psi_c'(\phi_e)}{l_c^2 B^2}\frac{\mu}{2\sqrt{\psi_e}}\right)}{2\sqrt{\chi^2(\mu) + \frac{\psi_c'(\phi_e)}{l_c^2 B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2 B^2}}}\right).$$

Applying $\chi(\mu_1) = 0$, we have

$$\frac{d\gamma_{\pm 1}(\mu)}{d\mu}\Big|_{\mu=\mu_1} = \frac{1}{2}\left(\frac{d\chi(\mu)}{d\mu} - i\frac{\frac{d}{d\mu}\left(\frac{\psi_c'(\phi_e)}{l_c^2 B^2}\frac{\mu}{2\sqrt{\psi_e}}\right)}{2\sqrt{-\frac{\psi_c'(\phi_e)}{l_c^2 B^2}\frac{\mu}{2\sqrt{\psi_e}} + \frac{1}{l_c^2 B^2}}}\right)_{\mu=\mu_1}.$$

Therefore,

$$\frac{d}{d\mu}\mathrm{Re}\,\gamma_{\pm 1}(\mu)\Big|_{\mu=\mu_1} = \frac{1}{2}\frac{d\chi(\mu)}{d\mu}\Big|_{\mu=\mu_1} < 0,$$

since $\psi_c'(\phi_e(\mu))$ is a decreasing function of μ and $\phi_e(\mu)/\psi_e(\mu)$ is an increasing function of μ for $\mu \geq \mu_{00}$. ■

Let

$$\phi_e^{(\nu)} := W\left(1+\sqrt{1-\frac{\nu W}{3aH}}\right), \qquad \psi_e^{(\nu)} = \psi_c\left(\phi_e^{(\nu)}\right);$$

then according to the definition of the compressor characteristic ψ_c one can see that

$$\psi_e^{(\nu)} := \psi_c\left(\phi_e^{(\nu)}\right) = \psi_{c_0} + H\left[1+\frac{3}{2}\sqrt{1-\frac{\nu W}{3aH}} - \frac{1}{2}\left(\sqrt{1-\frac{\nu W}{3aH}}\right)^3\right],$$

and $\phi_e^{(\nu)}$ satisfies the equation

$$\psi_c'(\Phi)\,|_{\Phi=\phi_e^{(\nu)}} = \frac{\nu}{2a}.$$

Let us now denote

$$\mu_\nu = \frac{W\left(1+\sqrt{1-\frac{\nu W}{3aH}}\right)}{\sqrt{\psi_{c_0} + H\left[1+\frac{3}{2}\sqrt{1-\frac{\nu W}{3aH}} - \frac{1}{2}\left(\sqrt{1-\frac{\nu W}{3aH}}\right)^3\right]}} = \frac{\phi_e^{(\nu)}}{\sqrt{\psi_e^{(\nu)}}}.$$

One can verify that $\frac{W}{\sqrt{\psi_0+H}} < \mu_\nu < \frac{2W}{\sqrt{\psi_0+2H}}$ (see Figure 3).

Lemma 3.8.

1. *If $\Delta > 0$, there exists a positive number $\hat{\mu}$ with $\mu_\nu < \hat{\mu} < \frac{2W}{\sqrt{\psi_{c_0}+2H}}$ such that when $\mu \in (\hat{\mu}, \infty)$ the spectrum of $A + df_{x_e}(\mu)$ is contained in the left half plane, and when $\mu = \hat{\mu}$, a pair of eigenvalues $\gamma_{\pm 1}$ crosses the imaginary axis with*

$$\frac{d}{d\mu}\operatorname{Re}\left(\gamma_{\pm 1}(\mu)\right)\Big|_{\mu=\hat{\mu}} < 0. \tag{3.27}$$

2. *If $\Delta < 0$, then when $\mu \in (\mu_\nu, \infty)$ the point spectrum of $A + df_{x_e}(\mu)$ is contained in the left half plane, and when $\mu = \mu_\nu$, a pair of eigenvalues $\lambda_{\pm 1} \neq \gamma_{\pm 1}$ crosses the imaginary axis with*

$$\frac{d}{d\mu}\operatorname{Re}\left(\lambda_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_\nu} < 0. \tag{3.28}$$

3. *If $\Delta = 0$, then when $\mu \in (\mu_\nu, \infty)$ the point spectrum of $A + df_{x_e}(\mu)$ is contained in the left half plane, and when $\mu = \mu_\nu$, two pairs of eigenvalues $\lambda_{\pm 1}, \gamma_{\pm 1}$ cross the imaginary axis with*

$$\frac{d}{d\mu}\operatorname{Re}\left(\lambda_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_\nu} < 0,$$
$$\frac{d}{d\mu}\operatorname{Re}\left(\gamma_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_\nu} < 0.$$

Proof. 1. Suppose $\Delta > 0$. According to the definition of the compressor characteristic given in (2.5), one can see when $\phi_e = 2W$ we have $\psi_c'(\phi_e) = 0$. The corresponding throttle coefficient μ_0 can be obtained as

$$\mu_0 = \frac{\phi_e}{\sqrt{\psi_e}} = \frac{2W}{\sqrt{\psi_{c_0} + 2H}}.$$

At $\mu = \mu_0$ we have

$$\begin{aligned} \mathrm{Re}(\lambda_n(\mu_0)) &= \frac{a|n|}{|n| + am}\left(\psi_c'(\phi_e) - \frac{\nu}{2a}n^2\right) < 0, \\ \mathrm{Re}(\gamma_{\pm 0}(\mu_1)) &< 0. \end{aligned}$$

As μ is reduced from μ_0 to μ_ν, the eigenvalues λ_n, $n = \pm 1, \pm 2, \ldots$, become

$$\begin{aligned} \mathrm{Re}(\lambda_{\pm 1}(\mu_\nu)) &= \frac{a}{1 + am}\left(\psi_c'(\phi_e^{(\nu)}) - \frac{\nu}{2a}\right) = 0, \\ \mathrm{Re}(\lambda_n(\mu_\nu)) &= \frac{a|n|}{|n| + am}\left(\psi_c'(\phi_e^{(\nu)}) - \frac{\nu}{2a}n^2\right) < 0, \quad n = \pm 2, \pm 3, \ldots. \end{aligned}$$

Notice that

$$\begin{aligned} \chi(\mu_\nu) &= \frac{1}{l_c}\psi_c'(\phi_e^{(\nu)}) - \frac{1}{4l_cB^2}\frac{\mu_\nu}{2\sqrt{\psi_e^{(\nu)}}} = \frac{1}{l_c}\left(\frac{\nu}{2a} - \frac{1}{4l_cB^2}\frac{\phi_e^{(\nu)}}{2\psi_e^{(\nu)}}\right) \\ &= \frac{\nu\phi_e^{(\nu)}}{2al_c\psi_e^{(\nu)}}\left(\frac{\psi_e^{(\nu)}}{\phi_e^{(\nu)}} - \frac{a}{4B^2\nu}\right) = \frac{\nu\phi_e^{(\nu)}}{2al_c\psi_e^{(\nu)}}\Delta > 0, \end{aligned}$$

which indicates $\mathrm{Re}\,\gamma_1(\mu_\nu) \geq \frac{\chi(\mu_\nu)}{2} > 0$. Since $\mathrm{Re}\,\gamma_{\pm 1}(\mu)$ is a continuous function of μ, the intermediate theorem implies that there exists $\hat{\mu}$ with $\mu_\nu < \hat{\mu} < \mu_0$ such that $\mathrm{Re}\,\gamma_{\pm 1}(\hat{\mu}) = 0$ and $\mathrm{Re}\,\gamma_{\pm 1}(\mu) < 0$ when $\mu > \hat{\mu}$. Since $\hat{\mu} > \frac{W}{\sqrt{\psi_{c_0}+H}} = \mu_{00}$, applying Lemma 3.7 leads to

$$\frac{d}{d\mu}\,\mathrm{Re}\,(\gamma_{\pm 1}(\mu))\Big|_{\mu=\hat{\mu}} < 0.$$

2. Suppose $\Delta < 0$. When $\mu = \mu_\nu$ we have

$$\mathrm{Re}(\lambda_{\pm 1}(\mu_\nu)) = \frac{a}{1 + am}\left(\psi_c'\left(\phi_e^{(\nu)}\right) - \frac{\nu}{2a}\right) = 0$$

and

$$\begin{aligned} \mathrm{Re}(\lambda_n(\mu_\nu)) &= \frac{a|n|}{|n| + am}\left(\psi_c'\left(\phi_e\right) - \frac{\nu}{2a}n^2\right) < 0, \quad n = \pm 2, \pm 3, \ldots, \\ \chi(\mu_\nu) &= \frac{\nu\phi_e^{(\nu)}}{2al_c\psi_e^{(\nu)}}\Delta < 0. \end{aligned}$$

Lemma 3.7 implies that both $\mathrm{Re}(\lambda_n(\mu))$ and $\chi(\mu)$ are decreasing functions of μ when $\mu > \frac{W}{\sqrt{\psi_{c_0}+H}}$. Hence when $\mu > \mu_\nu$ we have

$$\mathrm{Re}(\lambda_n(\mu_\nu)) < 0, \quad n = \pm 2, \ldots, \quad \text{and} \quad \mathrm{Re}(\gamma_{\pm 1}(\mu)) < 0.$$

Thus in this case $\lambda_{\pm 1}$ crosses the imaginary axis and the rest of the spectrum of $A + df_{x_e}(\mu)$ is contained in the left half plane when $\mu > \mu_\nu$. Applying Lemma 3.7 again we know that

$$\frac{d}{d\mu} \mathrm{Re}\left(\lambda_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_\nu} < 0.$$

3. Suppose $\Delta = 0$. Then when $\mu = \mu_\nu$ we have

$$\mathrm{Re}(\lambda_{\pm 1}(\mu_\nu)) = \frac{a}{1+am}\left(\psi_c'\left(\phi_e^{(\nu)}\right) - \frac{\nu}{2a}\right) = 0,$$
$$\chi(\mu_\nu) = \frac{\nu\phi_e^{(\nu)}}{2al_c\psi_e^{(\nu)}}\Delta = 0,$$

and

$$\mathrm{Re}(\lambda_n(\mu_\nu)) = \frac{a|n|}{|n|+am}\left(\psi_c'\left(\phi_e^{(\nu)}\right) - \frac{\nu}{2a}n^2\right) = \frac{a|n|}{|n|+am}\left(\frac{\nu}{2a} - \frac{\nu}{2a}n^2\right) < 0, \quad n = \pm 2, \pm 3 \ldots.$$

Lemma 3.7 leads to

$$\frac{d}{d\mu} \mathrm{Re}\left(\lambda_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_\nu} < 0, \qquad \frac{d}{d\mu} \mathrm{Re}\left(\gamma_{\pm 1}(\mu)\right)\Big|_{\mu=\mu_\nu} < 0,$$

and the proof of the lemma is complete. ■

Lemma 3.9. *If $\Delta > 0$, the frequency of the flow oscillation is smaller than $1/(4\sqrt{2}\pi Bl_c)$ (i.e., the flow oscillation is predominated by surge). While $\Delta < 0$, the frequency of the flow oscillation is greater than $1/\left(4\pi\left(\frac{3aH}{\nu W} + am\right)\right)$ (i.e., the oscillation is predominated by rotating stall).*

Proof. When $\Delta > 0$, as μ is decreasing from μ_0, according to Lemma 3.8 $\gamma_{\pm 1}(\mu)$ crosses the imaginary axis with nonzero speed at $\mu = \hat{\mu}$. In this case the bifurcating periodic orbit lies in X_2, that is, in the $\Phi\Psi$-plane. According to the Hopf bifurcation theorem [13], the period of the closed orbit is about $\frac{2\pi}{|\gamma_{\pm 1}(\hat{\mu})|}$. In the proof of Lemma 3.7 we know that

$$\gamma_{\pm 1}(\hat{\mu}) = \frac{\pm\sqrt{\frac{\psi_c'(\phi_e)}{l_c^2B^2}\frac{\hat{\mu}}{2\sqrt{\psi_e}} - \frac{1}{l_c^2B^2}}}{2}$$

and

$$\frac{\psi_c'(\phi_e)}{l_c^2B^2}\frac{\mu}{2\sqrt{\psi_e}} - \frac{1}{l_c^2B^2} < -\frac{1}{2l_c^2B^2}.$$

Hence we have

$$\text{the period} \approx \frac{2\pi}{|\gamma_{\pm 1}(\hat{\mu})|} > \frac{2\pi}{\frac{1}{2}\sqrt{\frac{1}{2l_c^2B^2}}} = 4\sqrt{2}\pi Bl_c.$$

Therefore, the frequency of the flow oscillation is smaller than $1/(4\sqrt{2}\pi B l_c)$.

When $\Delta < 0$, as μ is decreasing from μ_0, $\lambda_{\pm 1}(\mu)$ crosses the imaginary axis with nonzero speed at $\mu = \mu_\nu$ according to Lemma 3.8 again. Now the bifurcating periodic orbit lies in a two-dimensional subspace of X_1, and $\lambda_{\pm 1}(\mu_\nu) = \pm \frac{1}{2(1+am)} i$. Thus according to the Hopf bifurcation theorem the period of the flow oscillation is about $\frac{2\pi}{\frac{1}{2(1+am)}} = 4\pi(1+am)$. Notice that $\frac{\nu W}{3aH} \leq 1$; we have

$$4\pi(1+am) \leq 4\pi \left(\frac{3aH}{\nu W} + am \right),$$

which implies that the frequency is greater than $1/\left(4\pi\left(\frac{3aH}{\nu W} + am\right)\right)$. Therefore, we finish the proof of Theorem 3.1. ■

4. Numerical simulations. Following Moore and Greitzer [14], in the simulations the parameter values for the compressor characteristic (2.5) are chosen as

$$H = 0.18, \qquad W = 0.25, \qquad \psi_{c_0} = 1.67H.$$

In this case the peak of the compressor characteristic is at $(\Phi, \Psi) = (0.5, 0.6606)$. Without loss of generality we also set operator Π, defined in section 3.1, to be the identity of X, since Π does not affect the stability of the system (see Lemma 3.4). Other parameters for system (2.2)–(2.4) are set to be

$$l_c = 8, \qquad a = 1/3.5, \qquad \nu = 0.1.$$

The initial disturbance is set to be $g(0, \theta) = 0.005 \sin\theta$, and the initial flow and initial pressure rise are set near the peak of the compressor characteristic, that is, $(\Phi(0), \Psi(0)) = (0.51, 0.66)$. The simulations for the full model (3.3) are conducted by using the *Mathematica* PDE package.

Asymptotically stable case. In this case we choose $\mu = 0.66$ which is greater than $\frac{2W}{\sqrt{2H+\psi_{c_0}}} = 0.615$. The simulation is given by Figure 6, which is consistent with the first part of the main results.

Surge case. We pick $B = 2$. Based on parameters we set, we have $\Delta = 1.19706 > 0$, $\mu_\nu = 0.589996$. We choose $\mu = 0.6$ (so that μ is the value between μ_ν and $\frac{2W}{\sqrt{2H+\psi_{c_0}}}$), and simulation is provided in Figure 7. Simulation shows that the frequency of the flow is about $1/500$, which is smaller than $1/(4\sqrt{2}\pi B l_c) \approx 1/284$. Moreover, one can see that some inverse flow appears, which is one of the features of surge behavior.

Rotating stall case. In this case we set $B = 0.5$. Then we have $\Delta = -1.48151 < 0$. We set $\mu = 0.565$, which is smaller than $\mu_\nu = 0.589996$. Simulation is presented in Figure 8. From the simulation, one can see that the frequency of the flow oscillation is about $1/50$, which is greater than $1/\left(4\pi\left(\frac{3aH}{\nu W}\right)\right) \approx 1/77$ ($m = 0$ due to $\Pi =$ identity). Moreover, the amplitudes of both Φ and Ψ are relatively smaller than those in previous (surge) case.

A mixture case. If we set $B = 0.72058$, then $\Delta = -0.000017316$. We pick $\mu = 0.572$, and simulation is given in Figure 9. The frequency of the flow oscillation is about $1/100$, which lies between the frequency of surge and the frequency of rotating stall. The amplitudes of Φ and Ψ are similar to those in the surge case. Simulation shows that the dynamics of the system evolves as a combination of rotating stall and surge.

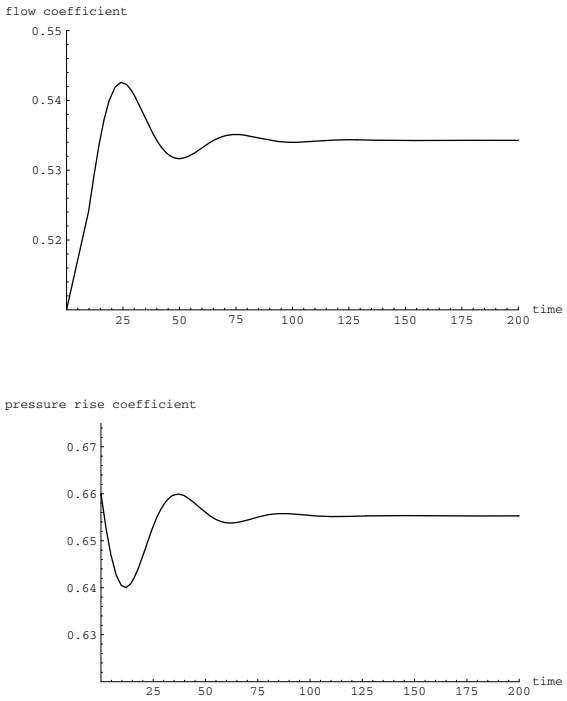


Figure 6. *Asymptotically stable case with* $B = 0.5$ *and* $\mu = 0.66 \geq 0.615 = \frac{2W}{\sqrt{2H+\psi_{c_0}}}$.

5. Concluding remarks. In this paper we provide a quantity Δ for predicting the type of oscillations of the Moore–Greitzer PDE model of an axial flow compressor. $\Delta > 0$ indicates that the dynamics will be predominated by surge, an oscillation which is induced by the flow coefficient Φ and the pressure coefficient Ψ. $\Delta < 0$ implies that the dynamics will be predominated by rotating stall g, an oscillation that results from the flow velocity disturbance at the duct entrance of a compressor. If $\Delta = 0$, then the dynamics will be governed by both. These instability phenomena can be further quantified by applying the infinite-dimensional version of the Hopf bifurcation theorem (see, e.g., [21], [7], and [8]). To avoid an overwhelming presentation in this paper, such results will be reported elsewhere.

6. Appendix.

Lemma 3.1. *The operator* A *defined in* (3.5) *is the infinitesimal generator of an analytic compact* C_0 *semigroup on* X.

Proof. We first claim that A generates an analytic C_0 semigroup on X. Notice that $D(A) = \left(H^2(0,2\pi) \cap X\right) \times \mathbb{R} \times \mathbb{R}$ is dense in X. Let $y \in X$ and $\lambda = \rho e^{i\vartheta}$ with $\rho > 0$, $-\frac{\pi}{2} < \vartheta < \frac{\pi}{2}$. Consider the eigenvalue problem

$$\lambda x - Ax = y.$$

By taking the inner product of both sides of the above identity with $e^{-i\vartheta}x$, and then taking the real part, we arrive at

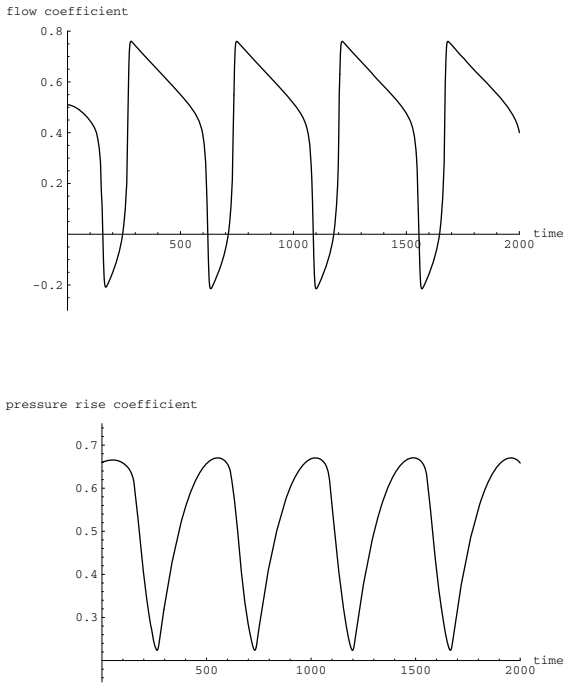


Figure 7. *Surge oscillation* $\Delta \approx 1.197 > 0$ *with* $B = 2$ *and* $\mu = 0.6$.

$$\rho \cos\vartheta \|x\|^2 - \cos\vartheta \operatorname{Re}\langle Ax, x\rangle + \sin\vartheta \operatorname{Im}\langle Ax, x\rangle = \operatorname{Re}[e^{-i\vartheta}\langle x, y\rangle], \tag{6.1}$$

where $\operatorname{Re}\langle Ax, x\rangle$ and $\operatorname{Im}\langle Ax, x\rangle$ stand for the real part and the imaginary part of $\langle Ax, x\rangle$, respectively. Next we show that there exists ϑ_0 with $0 < \vartheta_0 < \frac{\pi}{2}$ such that when $|\vartheta| < \frac{\pi}{2} + \vartheta_0$ we have

$$-\cos\vartheta \operatorname{Re}\langle Ax, x\rangle + \sin\vartheta \operatorname{Im}\langle Ax, x\rangle \geq 0 \tag{6.2}$$

for any $x \in D(A)$. Let $x = (g, \Phi, \Psi)$ with $g = g_1 + ig_2$. The conjugate function of g is denoted by $g^* = g_1 - ig_2$. A direct calculation shows that

$$\begin{aligned}
\operatorname{Re}\langle Ax, x\rangle &= \operatorname{Re}\int_0^{2\pi} \left(\frac{\nu}{2a}\frac{\partial^2 g}{\partial\theta^2} - \frac{1}{2a}\frac{\partial g}{\partial\theta}\right) g^* d\theta \\
&= \int_0^{2\pi} \left(\frac{\nu}{2a}\frac{\partial^2 g_1}{\partial\theta^2} - \frac{1}{2a}\frac{\partial g_1}{\partial\theta}\right) g_1 d\theta + \int_0^{2\pi} \left(\frac{\nu}{2a}\frac{\partial^2 g_2}{\partial\theta^2} - \frac{1}{2a}\frac{\partial g_2}{\partial\theta}\right) g_2 d\theta \\
&= -\frac{\nu}{2a}\int_0^{2\pi} \left(\left|\frac{\partial g_1}{\partial\theta}\right|^2 + \left|\frac{\partial g_2}{\partial\theta}\right|^2\right) d\theta
\end{aligned}$$

and

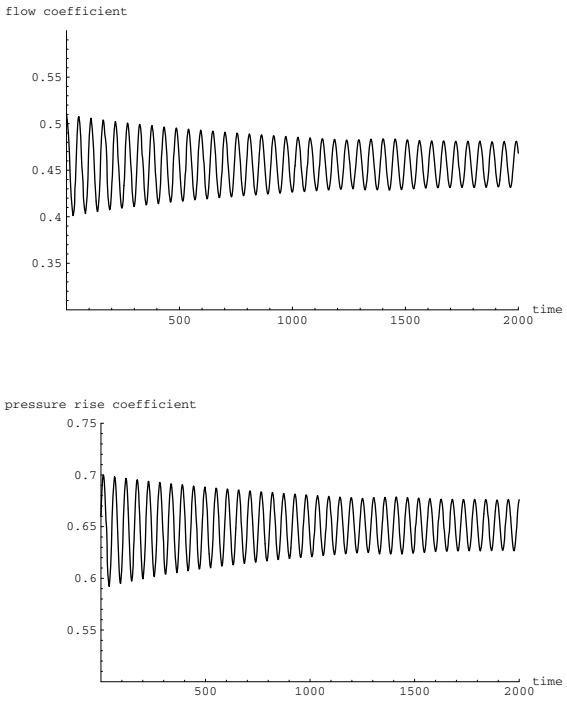


Figure 8. *Rotating stall oscillation* $\Delta \approx -1.482 < 0$ *with* $B = 0.5$ *and* $\mu = 0.572$.

$$\begin{aligned}\operatorname{Im}\langle Ax, x\rangle &= \operatorname{Im}\int_0^{2\pi}\left(\frac{\nu}{2a}\frac{\partial^2 g}{\partial\theta^2} - \frac{1}{2a}\frac{\partial g}{\partial\theta}\right) g^* d\theta \\ &= -\int_0^{2\pi}\left(\frac{\nu}{2a}\frac{\partial^2 g_1}{\partial\theta^2} - \frac{1}{2a}\frac{\partial g_1}{\partial\theta}\right) g_2 d\theta + \int_0^{2\pi}\left(\frac{\nu}{2a}\frac{\partial^2 g_2}{\partial\theta^2} - \frac{1}{2a}\frac{\partial g_2}{\partial\theta}\right) g_1 d\theta \\ &= \frac{1}{2a}\int_0^{2\pi} 2\frac{\partial g_1}{\partial\theta} g_2 d\theta.\end{aligned}$$

According to Poincaré's inequality we know that for any $g \in \dot{L}(0, 2\pi)$ we have

$$\int_0^{2\pi} |g|^2 d\theta \le 4\pi^2 \int_0^{2\pi} \left|\frac{\partial g}{\partial\theta}\right|^2 d\theta. \tag{6.3}$$

According to Cauchy's inequality and (6.3) we obtain

$$\left|\frac{\operatorname{Im}\langle Ax, x\rangle}{\operatorname{Re}\langle Ax, x\rangle}\right| = \left|\frac{\displaystyle\int_0^{2\pi} 2\left|\frac{\partial g_1}{\partial\theta} g_2\right| d\theta}{\displaystyle\nu\int_0^{2\pi}\left(\left|\frac{\partial g_1}{\partial\theta}\right|^2 + \left|\frac{\partial g_2}{\partial\theta}\right|^2\right) d\theta}\right|$$

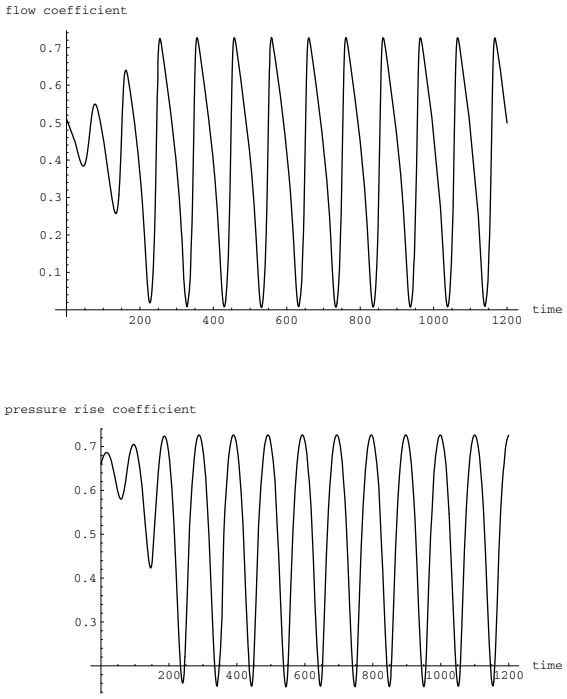


Figure 9. *A mixed oscillation* $\Delta \approx 0$ *with* $B = 0.72061$ *and* $\mu = 0.572$.

$$\leq \frac{2\left(\int_0^{2\pi}\left|\frac{\partial g_1}{\partial\theta}\right|^2 d\theta\right)^{\frac{1}{2}}\left(\int_0^{2\pi}|g_2|^2 d\theta\right)^{\frac{1}{2}}}{\nu\int_0^{2\pi}\left(\left|\frac{\partial g_1}{\partial\theta}\right|^2+\left|\frac{\partial g_2}{\partial\theta}\right|^2\right)d\theta}$$

$$\leq \frac{4\pi\left(\int_0^{2\pi}\left|\frac{\partial g_1}{\partial\theta}\right|^2 d\theta\right)^{\frac{1}{2}}\left(\int_0^{2\pi}\left|\frac{\partial g_2}{\partial\theta}\right|^2 d\theta\right)^{\frac{1}{2}}}{\nu\int_0^{2\pi}\left(\left|\frac{\partial g_1}{\partial\theta}\right|^2+\left|\frac{\partial g_2}{\partial\theta}\right|^2\right)d\theta} \leq \frac{2\pi}{\nu}.$$

On the other hand, one can see that

$$-\cos\vartheta\,\mathrm{Re}\,\langle Ax,x\rangle+\sin\vartheta\,\mathrm{Im}\,\langle Ax,x\rangle=-|\langle Ax,x\rangle|\cos(\vartheta+\varphi), \tag{6.4}$$

where

$$\tan\varphi=\frac{\mathrm{Im}\,\langle Ax,x\rangle}{\mathrm{Re}\,\langle Ax,x\rangle}. \tag{6.5}$$

Thus, if we let $\vartheta_0 := \tan^{-1}\frac{2\pi}{\nu}$, then for $|\vartheta| < \frac{\pi}{2}+\vartheta_0$ we have $\cos(\vartheta+\varphi) > 0$, and thus $-\cos\vartheta\,\mathrm{Re}\,\langle Ax,x\rangle+\sin\vartheta\,\mathrm{Im}\,\langle Ax,x\rangle \geq 0$.

From (6.1) it follows that when $|\vartheta| < \frac{\pi}{2} + \vartheta_0$ we have $\|x\| \le (\rho \cos \frac{\vartheta_0}{2})^{-1} \|y\|$, so that

$$\|(\lambda I - A)^{-1}\| \le \frac{1}{|\lambda| \cos \frac{\vartheta_0}{2}}, \qquad |\arg \lambda| < \frac{\pi}{2} + \vartheta_0,$$

and hence

$$\rho(A) \supset \Sigma(\vartheta) = \left\{ \lambda \in \mathbf{C} : \ |\arg \lambda| < \frac{\pi}{2} + \vartheta_0 \right\}.$$

Therefore, it follows that A is the infinitesimal generator of an analytic semigroup on X.

Next we claim that A also generates a compact C_0 semigroup on X. Since the semigroup $T(t)$ generated by A is analytic, it is continuous in the uniform operator topology for $t > 0$. Furthermore, the embedding $H^2_\pi(0, 2\pi) \hookrightarrow L^2_\pi(0, 2\pi)$ is compact and $R(\lambda, A)X \subseteq D(A) \subseteq H^2_\pi(0, 2\pi)$. Thus $R(\lambda, A)$ is a compact operator. Therefore, $T(t)$, $t > 0$, is a compact semigroup, and this completes the proof. ■

Acknowledgment. The author would like to thank the anonymous reviewers for their suggestions which have been very helpful in the improvement of the paper.

REFERENCES

[1] Y. Chung and E. S. Titi, *Inertial manifolds and Gevrey regularity for the Moore-Greitzer model of an axial-flow compressor*, J. Nonlinear Sci., 13 (2003), pp. 1–25.

[2] D. C. Liaw and E. H. Abed, *Active control of compressor stall inception: A bifurcation theoretic approach*, Automatica J. IFAC, 32 (1996), pp. 109–115.

[3] A. Banaszuk, H. A. Hauksson, and I. Mezić, *A backstepping controller for a nonlinear partial differential equation model of compression system instabilities*, SIAM J. Control Optim., 37 (1999), pp. 1503–1537.

[4] B. Birnir and H. A. Hauksson, *A finite dimensional attractor of the Moore–Greitzer PDE model*, SIAM J. Appl. Math., 59 (1998), pp. 636–650.

[5] B. Birnir and H. A. Hauksson, *Basic control for the viscous Moore–Greitzer partial differential equation*, SIAM J. Control Optim., 38 (2000), pp. 1554–1580.

[6] B. Birnir and H. A. Hauksson, *The basic attractor of the viscous Moore-Greitzer equation*, J. Nonlinear Sci., 11 (2001), pp. 169–192.

[7] M. Crandall and P. Rabinowitz, *The Hopf bifurcation theorem in infinite dimensions*, Arch. Rational Mech. Anal., 67 (1977), pp. 53–72.

[8] M. Crandall and P. Rabinowitz, *Mathematical theory of bifurcation*, in Bifurcation Phenomena in Mathematical Physics and Related Topics, C. Bardos and D. Bessis, eds., Reidel, Dodrecht, 1980.

[9] I. J. Day, E. M. Greitzer, and N. A. Cumpsty, *Prediction of compressor performance in rotating stall*, J. Engineering for Power, 100 (1978), pp. 1–14.

[10] J. S. Humber and A. J. Krener, *Dynamics and control of entrained solutions in multi-mode Moore-Greitzer compressor models*, Internat. J. Control, 71 (1998), pp. 807–821.

[11] W. Kang, G. Gu, A. Sparks, and S. Banda, *Bifurcation test functions and surge control for axial flow compressors*, Automatica J. IFAC, 35 (1999), pp. 229–239.

[12] F. E. McCaughan, *Bifurcation analysis of axial flow compressor stability*, SIAM J. Appl. Math., 50 (1990), pp. 1232–1253.

[13] J. E. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications*, Springer-Verlag, New York, 1976.

[14] F. K. Moore and E. M. Greitzer, *A theory of post-stall transients in axial compression systems: Part I–development of equations*, ASME J. Engineering for Gas Turbines and Power, 108 (1986), pp. 68–76.

[15] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[16] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
[17] M. Xiao, *Stabilization of the full model compression system*, in Proceedings of the 37th IEEE Confernece on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1998, pp. 2575–2580.
[18] M. Xiao and T. Başar, *Rotating stall control of MG3 compressor models governed by partial differential equations*, in Proceedings of the IFAC Congress, Beijing, China, 1999, pp. 183–188.
[19] M. Xiao and T. Başar, *Analysis and control of multi-mode axial flow compression system models*, ASME J. Dynamic Systems, Measurement and Control, 122 (2000), pp. 393–401.
[20] M. Xiao and T. Başar, *Center manifold of the viscous Moore–Greitzer PDE model*, SIAM J. Appl. Math., 61 (2000), pp. 855–869.
[21] M. Xiao and W. Kang, *Control of Hopf bifurcations for infinite-dimensional nonlinear systems*, in New Trends in Nonlinear Dynamics and Control and Their Application, W. Kang, M. Xiao, and C. Borges, eds., Springer-Verlag, New York, 2003, pp. 101–116.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 63–78
© 2008 Society for Industrial and Applied Mathematics

Pattern Formation in a Model of a Vibrated Granular Layer*

D. M. Winterbottom†, S. M. Cox†, and P. C. Matthews†

Abstract. A phenomenological model for pattern formation in a vertically vibrated granular layer is examined in order to investigate its nonlinear dynamics. The model comprises two coupled partial differential equations: one describes the evolution of the short-scale pattern, while the other enforces conservation of granular material. In a layer of moderate horizontal extent, the model predicts that a variety of exotic regular patterns may be stable, according to the system parameters. The usual cubic-order amplitude equations are unable to determine the stable solution over a significant parameter range; we compute the corresponding fifth-order terms necessary to resolve this degeneracy. When spatial modulation of the pattern is taken into account, in a sufficiently wide layer, a stability analysis of regular one-dimensional roll and two-dimensional square patterns demonstrates that each may suffer a modulational instability, which tends to localize the pattern. The corresponding modulational stability boundaries, for both rolls and squares, coincide with the transition between stable rolls and squares in the unmodulated problem. As a consequence, in a suitably large container, squares are always unstable, and corresponding numerical simulations indicate highly localized worm- or chain-like patterns. The numerical simulations and stability results are compared with appropriate experimental results.

Key words. pattern formation, phenomenological model, vibrated granular layer

AMS subject classifications. 76T25, 74J30

DOI. 10.1137/06067540X

1. Introduction.

The near-threshold behavior of a vertically vibrated granular layer has been the subject of considerable attention in recent years. Experimentally, a diverse collection of extended cellular and localized patterns has been observed, including rolls, squares, hexagons, interfaces, and oscillons [1, 25, 26, 33, 34]. Theoretically, however, there is no widely accepted continuum description for the highly complicated rheology of granular media [2, 3, 6, 19]; granular materials are known to exhibit some of the dynamical properties of fluids and solids, as well as their own unique properties [15, 16, 19]. Aranson and Tsimring [3] have provided a comprehensive recent review of the behavior of granular media.

A thin horizontal layer of granular material on a flat plate that oscillates with vertical displacement $z = \mathcal{A}\sin(2\pi ft)$ can form a pattern, which is governed by the driving frequency f and the nondimensional acceleration amplitude $\Gamma = 4\pi^2\mathcal{A}f^2/g$, where g is the acceleration due to gravity [25, 26, 33]. Almost irrespective of the driving frequency, patterned states appear at $\Gamma \approx 2.4$, where rolls, squares, or oscillons may be found (all oscillating at half the driving frequency). At low frequencies, a hysteretic transition to a regular square pattern is observed, while above some critical frequency, rolls are selected [25]; localized structures such

*Received by the editors November 11, 2006; accepted for publication (in revised form) by M. Silber July 3, 2007; published electronically January 16, 2008.
http://www.siam.org/journals/siads/7-1/67540.html
†School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK (david.winterbottom@gmail.com, stephen.cox@nottingham.ac.uk, paul.matthews@nottingham.ac.uk).

as oscillons are observed in the slightly subcritical region. At higher accelerations ($\Gamma > 3.72$), these patterns become unstable and are replaced by hexagons. Further increases in Γ lead to the appearance of interfaces (also known as kinks) and, subsequently, quarter-harmonic patterns [26].

Several theoretical approaches to this phenomenon have been developed, including simulations of molecular dynamics [6, 12, 31], hydrodynamic-type and phenomenological models [11, 14, 27], and semicontinuum models [8, 17, 18, 30, 35]. In particular, granular hydrodynamic models have had considerable success in describing granular flow in appropriate parameter regimes [3]. However, while many features of the experiments have been reproduced, no model offers a complete description of the rich variety of observed phenomena [19].

In this paper, we consider the phenomenological model proposed by Tsimring and Aranson [2, 32], which comprises a Ginzburg–Landau-type amplitude equation for an order parameter, coupled to an equation expressing the net conservation of granular material. A weakly nonlinear analysis of the emerging pattern reveals how the constraint that the granular material be conserved leads to a mechanism for localization of the pattern. The resulting weakly nonlinear equations allow the computation of the stability boundaries for regular patterns; beyond these boundaries in parameter space localized patterns are indeed found in numerical simulations. We emphasize that although the underlying model [2, 32] is purely phenomenological, the mechanism by which the localization of the pattern takes place relies essentially only on the physical constraint of mass conservation, and so we expect our results to be robust, and to extend in principle to more sophisticated models of granular media.

We also expect our results to be applicable to vibrated fluid layers where, experimentally, a wide range of patterns are observed [5, 23] but, as with granular layers, a number of theoretical models are employed [37, 38]. Indeed, the Ginzburg–Landau-type amplitude equation (2.1), with no coupling to a mean-field, has been studied previously as a model for parametric surface waves [20, 37].

The structure of the paper is as follows. In section 2 we recall the phenomenological order-parameter model of Tsimring and Aranson [2, 32] and review the corresponding linear stability results. In section 3 we develop two weakly nonlinear models: one describes a spatially regular pattern comprising modes with up to four distinct wavevectors, and the other describes the spatial modulations of the pattern that may arise in a layer of greater horizontal extent. The stability of various patterns is then determined and the results confirmed and extended through numerical simulation. Finally, in section 4 we discuss the implications of these results and present our conclusions.

2. Order-parameter model. Since there is no generally accepted continuum model, our starting point is the phenomenological model [2, 32]

$$\psi_t = \gamma\psi^* - (1 - \mathrm{i}\omega)\psi + (1 + \mathrm{i}b)\nabla^2\psi - |\psi|^2\psi - \rho\psi, \tag{2.1}$$

$$\rho_t = \beta\nabla^2\rho + \alpha\vec{\nabla}\cdot(\rho\vec{\nabla}|\psi|^2) \tag{2.2}$$

that incorporates the essential physical ingredients of a subharmonic pattern-forming instability and mass conservation. The order parameter $\psi(x, y, t)$ characterizes the complex amplitude of the subharmonic pattern, with the disturbance to the planar layer surface being

$$h = \psi \mathrm{e}^{\mathrm{i}\pi f t} + \text{c.c.}, \tag{2.3}$$

while $\rho(x, y, t)$ represents the local mass of the granular layer per unit area.

As described by Blair et al. [7], the linear terms in (2.1) follow from an expansion of the linear dispersion relation for surface perturbations in the form $h \propto \exp(\Lambda(k)t + \mathrm{i}kx)$ for small k. If we retain only the terms $\Lambda(k) = -\Lambda_0 - \Lambda_2 k^2$ in this expansion then we obtain the linearized evolution equation $h_t = -\Lambda_{0r}(1 + \mathrm{i}\Lambda_{0i}/\Lambda_{0r})h + \Lambda_{2r}(1 + \mathrm{i}\Lambda_{2i}/\Lambda_{2r})\nabla^2 h$, where subscripts r and i denote real and imaginary parts, respectively. If we now remove the factors Λ_{0r} and Λ_{2r} by rescaling time and space variables, we see that ψ satisfies

$$\psi_t + \mathrm{i}\pi f\psi = -(1 + \mathrm{i}\Lambda_{0i}/\Lambda_{0r})\psi + (1 + \mathrm{i}\Lambda_{2i}/\Lambda_{2r})\nabla^2\psi,$$

and hence we obtain the linear terms in (2.1), with $\omega = -\Lambda_{0i}/\Lambda_{0r} - \pi f$ and $b = \Lambda_{2i}/\Lambda_{2r}$. Since the subharmonic oscillation in the physical variable h is already accounted for in (2.3), the simplest subharmonic pattern is realized with ψ of the time-independent form $\psi = \Psi(x, y)$.

Equation (2.1), with no coupling to the mean-field $\rho(x, y, t)$, has been studied previously as a model for a vibrated fluid layer [20, 37] as well as for the large-Γ regime of a vibrated granular layer [4, 7], where ρ is assumed constant. The term $\gamma\psi^*$ provides the parametric driving required for standing waves to become excited, where the control parameter γ is the normalized amplitude of the parametric forcing. The coupling term $-\rho\psi$ indicates that increasing the depth of the layer makes the system more stable, as observed in experiments [25].

Equation (2.2) describes the conservation of mass in the layer; the distributive mechanisms are a diffusive flux (proportional to $-\vec{\nabla}\rho$) and a flux proportional to $-\rho\vec{\nabla}|\psi|^2$ corresponding to particles escaping from regions of large fluctuation. The mass diffusion constant β is expected to be proportional to the energy of the plate vibrations and should increase with the driving frequency f [32]. It is the rapid diffusive smoothing which allows the role of ρ to be discounted in investigations of the high-acceleration regime of a vibrated granular layer [4, 7]. Since ρ is conserved, its mean value ρ_0 is effectively another parameter in the model. Since the model comprising (2.1) and (2.2) involves pattern formation coupled to a conservation law, it is likely to exhibit localized patterns [24].

Perturbations to the uniform solution $\psi = 0$, $\rho = \rho_0$ (corresponding to a flat homogeneous layer) with wavenumber k give rise to a neutral curve of the form

$$\gamma^2 = (1 + k^2 + \rho_0)^2 + (\omega - bk^2)^2. \tag{2.4}$$

The nature of the primary bifurcation as the control parameter γ is increased then depends on the sign of $\omega b - 1 - \rho_0$. For $\omega b - 1 - \rho_0 > 0$, stability is lost at $\gamma = \gamma_c$, where

$$\gamma_c^2 = \frac{[\omega + b(1 + \rho_0)]^2}{1 + b^2}; \tag{2.5}$$

corresponding perturbations have finite critical wavenumber k_c, determined by

$$k_c^2 = \frac{\omega b - 1 - \rho_0}{1 + b^2}. \tag{2.6}$$

By contrast, for $\omega b - 1 - \rho_0 < 0$, the instability is first manifest through disturbances on the largest spatial scales (i.e., as the wavenumber $k \to 0$) with corresponding threshold

$$\gamma_c^2 = (1 + \rho_0)^2 + \omega^2. \tag{2.7}$$

We restrict our attention in this paper to the case of instability to finite-wavelength perturbations; henceforth, we assume $\omega b - 1 - \rho_0 > 0$.

3. Amplitude equations. We suppose the system to be close to the primary bifurcation of the uniform solution, near the onset of a pattern-forming instability at finite wavelength. We proceed with a weakly nonlinear framework, setting $\gamma = \gamma_c + \epsilon^2\gamma_2$ and expanding ψ and ρ about the uniform state $\psi = 0$, $\rho = \rho_0$ as series in the small parameter ϵ:

$$\psi = \epsilon\psi_1 + \epsilon^2\psi_2 + \epsilon^3\psi_3 + \cdots, \tag{3.1}$$

$$\rho = \rho_0 + \epsilon^2\rho_2 + \cdots. \tag{3.2}$$

Note that only even powers of ϵ appear in the expansion for ρ because the forcing in (2.2) is quadratic in ψ. Making the appropriate substitutions in (2.1) and (2.2), and collecting linear terms, we find

$$\partial_t\psi_1 = \gamma_c\psi_1^* - (1 - \mathrm{i}\omega)\psi_1 + (1 + \mathrm{i}b)\nabla^2\psi_1 - \rho_0\psi_1. \tag{3.3}$$

This equation inherits from (2.1) a rotational symmetry: all modes with a given wavenumber k have equal rate of growth or decay, regardless of their orientation. The modes that can be realized in practice are dictated by the container; for analytical simplicity, we assume that periodic boundary conditions are applied in x and y. Here we examine two pattern formation problems: the first involves the competition between two sets of modes, aligned at some angle to one another, and the second involves the competition between four modes.

3.1. No spatial modulation. Consider a two-mode rhombic standing wave ansatz

$$\psi_1 = (b + s + \mathrm{i})\left\{\left[\tilde{A}\mathrm{e}^{\mathrm{i}kx} + \tilde{B}\mathrm{e}^{\mathrm{i}k(x\cos\theta + y\sin\theta)}\right] + \text{c.c.}\right\}, \tag{3.4}$$

where $\tilde{A}$ and $\tilde{B}$ are complex amplitudes, evolving on the slow time scale $\tilde{T} = \epsilon^2 t$, θ parameterizes the angle between the two modes, and we have introduced

$$s = \sqrt{1 + b^2}.$$

Note that the argument of ψ_1, indicated by the factor $b + s + \mathrm{i}$, follows from (3.3).

At $O(\epsilon^2)$ we are permitted to choose $\psi_2 = 0$, while ρ_2 is found to be slaved to the quadratic self-interactions of ψ_1. Finally at $O(\epsilon^3)$, applying the appropriate solvability condition yields the evolution equations for A and B (compare with equation (4) of [2, 32]):

$$\tilde{A}_{\tilde{T}} = r\tilde{A} - 2s(s + b)(3 - \phi)|\tilde{A}|^2\tilde{A} - 4s(s + b)(3 - 2\phi)|\tilde{B}|^2\tilde{A}, \tag{3.5}$$

$$\tilde{B}_{\tilde{T}} = r\tilde{B} - 2s(s + b)(3 - \phi)|\tilde{B}|^2\tilde{B} - 4s(s + b)(3 - 2\phi)|\tilde{A}|^2\tilde{B}, \tag{3.6}$$

where

$$r = \frac{\gamma_2}{\gamma_c}\left(\frac{\omega}{b} + 1 + \rho_0\right), \qquad \phi = \frac{\alpha\rho_0}{\beta}.$$

Note that these equations are independent of the angle θ between the two modes. We assume henceforth that ϕ, which measures the strength of coupling between ψ and ρ, satisfies $0 < \phi < 3$, so that the primary bifurcation to a single mode is supercritical. The interesting dynamical

behavior of (3.5) and (3.6) then arises for $r > 0$, which we assume henceforth. To reduce (3.5) and (3.6) to a canonical form, we introduce

$$(A, B) = \sqrt{\frac{2s(s+b)(3-\phi)}{r}}(\tilde{A}, \tilde{B}), \qquad T = r\tilde{T}; \tag{3.7}$$

then the governing equations (3.5) and (3.6) become

$$A_T = A - |A|^2 A - \lambda |B|^2 A, \tag{3.8}$$

$$B_T = B - |B|^2 B - \lambda |A|^2 B, \tag{3.9}$$

where the nonlinear coupling coefficient is

$$\lambda = 2\left(\frac{3-2\phi}{3-\phi}\right). \tag{3.10}$$

The equations (3.8) and (3.9) reveal two nontrivial solutions [21]: rolls (e.g., $|A| = 1$, $B = 0$), which are stable when $\lambda > 1$, and rhombs (i.e., $|A|^2 = |B|^2 = 1/(1+\lambda)$), which are stable when $-1 < \lambda < 1$. Thus rolls are predicted to be stable for $\phi < 1$, while rhombs (which include squares as the special case $\theta = \pi/2$) are stable when $1 < \phi < 9/5$. This picture is consistent with experimental observations, since β is expected to increase (and hence ϕ is expected to decrease) with increasing f [32].

When there is no coupling between the order-parameter equation (2.1) and ρ (see [4, 7, 20, 37]), $\alpha = 0$, and so $\phi = 0$, and the nonlinear coupling coefficient is simply $\lambda = 2$. Thus, in this case the predictions of the amplitude equations are insensitive to the model parameters, and only rolls can be stable.

An extended version of the amplitude equations (3.8) and (3.9) was derived in [2, 32] including terms up to $O(\epsilon^5)$, by making the simplifying approximation that ρ is slaved to $|\psi|^2$ in (2.2). These extended equations allow a phase diagram to be constructed, illustrating the stability regions of rolls and squares in Γ-ϕ parameter space. However, the region of validity of the inherent approximations is limited to large α and β. Furthermore, we shall see below, when we describe our numerical simulations of (2.1) and (2.2), that the *global* conservation of ρ provides a mechanism for the generation of highly localized patterns, and this mechanism is lost by imposing a *local* enslavement of ρ to $|\psi|^2$.

We turn now to a lattice-periodic ansatz comprising four modes, such as may arise in a periodic square box of side $L = n\sqrt{5}\Lambda_c$ for $n \in \mathbb{N}$, where $\Lambda_c = 2\pi/k_c$ is the critical wavelength. Thus [13]

$$\begin{aligned}\psi_1 = (b+s+\mathrm{i})\Bigg\{\Bigg[&\tilde{A}\exp\left\{\frac{\mathrm{i}k}{\sqrt{5}}(2x+y)\right\} + \tilde{B}\exp\left\{\frac{\mathrm{i}k}{\sqrt{5}}(x+2y)\right\} \\ &+ \tilde{C}\exp\left\{\frac{\mathrm{i}k}{\sqrt{5}}(-x+2y)\right\} + \tilde{D}\exp\left\{\frac{\mathrm{i}k}{\sqrt{5}}(-2x+y)\right\}\Bigg] + \text{c.c.}\Bigg\}.\end{aligned} \tag{3.11}$$

Performing a weakly nonlinear expansion as for the roll–rhomb ansatz (3.4), we obtain, at

$O(\epsilon^3)$, after a rescaling as in (3.7), the amplitude equations

$$A_T = A - |A|^2 A - \lambda A\left(|B|^2 + |C|^2 + |D|^2\right), \tag{3.12}$$

$$B_T = B - |B|^2 B - \lambda B\left(|A|^2 + |C|^2 + |D|^2\right), \tag{3.13}$$

$$C_T = C - |C|^2 C - \lambda C\left(|A|^2 + |B|^2 + |D|^2\right), \tag{3.14}$$

$$D_T = D - |D|^2 D - \lambda D\left(|A|^2 + |B|^2 + |C|^2\right), \tag{3.15}$$

where the nonlinear coupling coefficient λ is just that for the rhombic case, given by (3.10). We find that the only steady solutions to these four amplitude equations that may be stable in some region of parameter space are rolls (e.g., $|A| = 1$, $B = C = D = 0$), which are stable for $\phi < 1$, and four-mode solutions with $|A|^2 = |B|^2 = |C|^2 = |D|^2 = 1/(1+3\lambda)$, which are stable when $1 < \phi < 21/13$. The amplitude equations also allow stationary solutions in the form of squares (e.g., $|A| = |C| = 1/(1+\lambda)$, $B = D = 0$) and rhombs (e.g., $|A| = |B| = 1/(1+\lambda)$, $C = D = 0$), but these solutions are always unstable.

Since (3.12)–(3.15) are invariant under phase translations ($A \mapsto Ae^{i\theta}$), higher-order terms are required to resolve the relationships between the phases of the four modes in the latter case. However, it has been shown using the equivariant branching lemma that such solutions exist in two forms—"supersquares" and "antisquares" [13], with stability being determined by the coefficients of some quintic-order terms in the amplitude equations. To determine the necessary coefficients, we write for each amplitude an evolution equation of the form $\tilde{A}_t = \epsilon^2 f_2 + \epsilon^4 f_4 + \cdots$, where, as indicated above, prior to any rescaling,

$$f_2 = r\tilde{A} - 2s(s+b)(3-\phi)|\tilde{A}|^2\tilde{A} - 4s(s+b)(3-2\phi)\tilde{A}(|\tilde{B}|^2 + |\tilde{C}|^2 + |\tilde{D}|^2).$$

The relevant terms for distinguishing between the stability of supersquares and antisquares arise in f_4, which may be written in the form

$$f_4 = \tilde{A}F_4(|\tilde{A}|^2, |\tilde{B}|^2, |\tilde{C}|^2, |\tilde{D}|^2) + b_{41}\tilde{B}^2\tilde{C}^{*2}\tilde{D} + b_{42}\tilde{A}^*\tilde{B}\tilde{C}\tilde{D}^{*2}$$

for some real-valued coefficients b_{41} and b_{42}. Fortunately, for the purpose of determining the stability of supersquares and antisquares it is not necessary to compute F_4; it proves necessary only to calculate the quantity $b_{41} + 2b_{42}$. It turns out that antisquares are stable if $1 < \phi < 21/13$ and $b_{41} + 2b_{42} < 0$; supersquares are stable if the latter inequality is reversed. We find

$$b_{41} = 12s^2(s+b)^2\left\{\frac{(453b\gamma_c - 32sk_c^2)(\phi-1)^2}{40s^3k_c^4} - \frac{\phi^2}{\rho_0}\right\}, \tag{3.16}$$

$$b_{42} = 2b_{41}. \tag{3.17}$$

Thus we conclude that (provided $1 < \phi < 21/13$) antisquares are stable when

$$\frac{\phi^2}{(\phi-1)^2} > \delta, \tag{3.18}$$

where

$$\delta = \frac{(453b\gamma_c - 32sk_c^2)\rho_0}{40s^3k_c^4}$$

(and supersquares are stable when the inequality in (3.18) is reversed). It follows that, as we increase ϕ from $\phi = 1$, antisquares are stable, at least at first. If $\delta < 441/64$, then (3.18) is satisfied for all ϕ in the range $1 < \phi < 21/13$, so supersquares are never stable. Alternatively, if $\delta > 441/64$, then antisquares are stable for $1 < \phi < \phi_1$, while supersquares are stable for $\phi_1 < \phi < 21/13$, where $\phi_1 = 2\sqrt{\delta}/(\sqrt{\delta} - 1)$.

3.1.1. Rotational degeneracy of the amplitude equation coefficients. We now briefly comment on the fact that the cross-coupling coefficients in (3.5) and (3.6) are independent of the angle θ between the two sets of modes. Because of this degeneracy, there is no selection of θ in the cubic-order amplitude equations. One might ask whether one could augment the phenomenological model for ψ and ρ to overcome this degeneracy. The obvious candidate respecting the symmetry $\psi \mapsto -\psi$ is the addition of a term $\psi_t = \cdots - p\psi|\vec{\nabla}\psi|^2$. However, it turns out that the corresponding contribution to the cross-terms (e.g., the term proportional to $A|B|^2$ in the A_T equation) then vanishes identically. Other terms, such as quadratic terms of the form $\psi_t = \cdots + p\psi^2 + q\psi^{2*}$, succeed in resolving the degeneracy, in that the cross-coupling term depends on θ, but require more substantial justification: such terms might represent, for example, an additional component of the forcing of the layer, proportional to $\sin(3\pi ft)$ or $\cos(3\pi ft)$.

3.2. Small-container numerical simulations. The analytical results of the preceding section can be used to predict interesting parameter regimes in which to numerically simulate the model equations (2.1) and (2.2) in a container of moderate horizontal extent (so that long-wavelength modulational effects are not relevant). For all our simulations, we set the parameter values

$$\beta = \rho_0 = 0.3, \quad b = 1, \quad \omega = 2.5,$$

and use the mobility coefficient α (which equals ϕ with this choice of parameters) as a control parameter. For these parameter values, we find $b_{41} + 2b_{42} = P(\alpha - \alpha_1)(\alpha - \alpha_2)$, where $P = 12425(3 + 2\sqrt{2})/4 > 0$, $\alpha_1 \approx 0.7475$, and $\alpha_2 \approx 1.5100$. Thus, since $21/13 \approx 1.6154$, antisquares are predicted to be stable for $1 < \alpha < \alpha_2$, and supersquares are predicted to be stable for $\alpha_2 < \alpha < 21/13$.

Our simulations use periodic boundary conditions, implemented with a Fourier spectral method. The initial conditions are the uniform state plus small-amplitude noise, and the time-stepping is carried out with the exponential time differencing method [9]. If we choose a square container with side $L = n\Lambda_c$, $n \in \mathbb{N}$, then the dominant modes are perpendicular to and aligned with the periodic boundaries. In this situation, the leading-order pattern is of the form (3.4) with $\theta = \pi/2$. Near-threshold simulations in such containers confirm the predictions of (3.8)–(3.9), displaying stable rolls for $\phi < 1$ and stable squares for $1 < \phi < 3/2$.

Alternatively, if we choose $L = n\sqrt{5}\Lambda_c$, $n \in \mathbb{N}$, then the dominant modes take the form of those in (3.11). Again simulations support the theory: near onset, rolls are stable for $\phi < 1$ and a four-mode solution is stable for $1 < \phi < 3/2$. Figure 1 shows simulations for $\phi = 1.1$: Figure 1(a) shows a stable four-mode solution, which takes the form of antisquares, as predicted above. An accompanying video (67540_01.mov [334KB]) shows a random initial condition leading to stable antisquares for parameters as in Figure 1(a). As far as we are aware this is the first example of a pattern-forming system exhibiting stable antisquares. However,

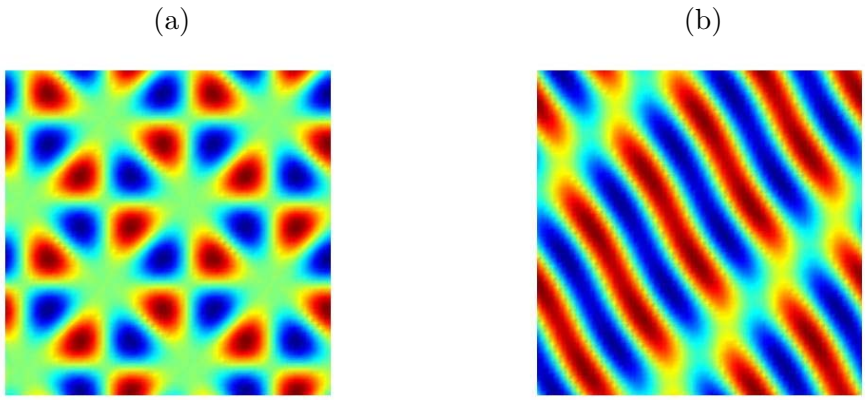


Figure 1. *Plots of $Re(\psi)$ from spectral simulations of (2.1) and (2.2) with a container size $L = 2\sqrt{5}\Lambda_c$ and periodic boundary conditions:* (a) *antisquares found at* 0.5% *above threshold;* (b) *split rolls at* 1% *above threshold. Further above threshold, roll patterns are found. The parameter values for these simulations are $\alpha = 1.1$, $\beta = \rho_0 = 0.3$, $b = 1$, $\omega = 2.5$; these values of β, ρ_0, b, and ω are used for all simulations reported here.*

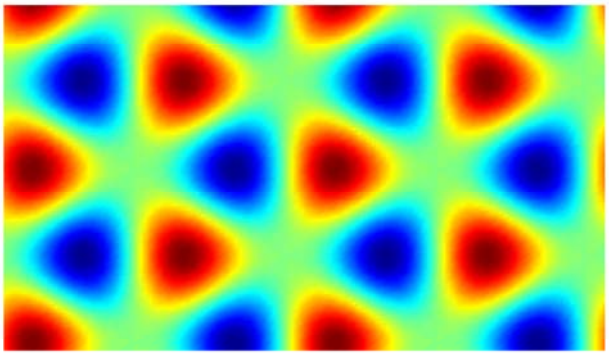


Figure 2. *Plots of $Re(\psi)$ from spectral simulations of (2.1) and (2.2) with a container size $4\Lambda_c \times 4\Lambda_c/\sqrt{3}$ and periodic boundary conditions, showing a pattern of triangles. The parameter values are as for Figure 1*(a).

the range of validity of the weakly nonlinear approximation is found to be fairly small; at 1% above threshold, antisquares are no longer realized, and instead we find rolls or roll-like patterns (Figure 1(b)).

The theoretical results predict that supersquares are stable only in a very small region of parameter space near the point at which the bifurcation becomes subcritical, but we were unable to find stable supersquares in this region in our numerical simulations. However, if we change the parameters to $\beta = \rho_0 = 2$, $b = 1$, $\omega = 4$, the theory predicts a larger region of stable supersquares, $1.0867 < \alpha < 1.6154$, and stable supersquares were found in numerical simulations in this range.

Since the model equations do not select a preferred angle between modes when $\phi > 1$, this angle is in fact determined by the imposed periodic lattice. As an illustration of this, Figure 2

shows a simulation in a periodic domain of size $4\Lambda_c \times 4\Lambda_c/\sqrt{3}$, designed to permit patterns that reside on a hexagonal lattice. In this case a stable pattern of regular triangles is found.

3.3. Spatial modulation. The weakly nonlinear analysis presented above holds when the pattern spectrum consists of discrete modes whose wavenumbers are close to the critical circle $k^2 = k_c^2$. However, in a spatially extended domain, the fact of mass conservation leads to the existence of slowly evolving modes of large wavelength (i.e., with wavenumber close to zero), representing adjustments of large-scale inhomogeneities in the thickness of the layer. These slowly varying modes, even though linearly damped through the diffusive term in the equation of mass conservation, must be included in any amplitude-equation description of the near-threshold dynamics of the pattern [10, 22, 24]. Previous theoretical work has indicated the appropriate amplitude equations for a system undergoing a stationary bifurcation to a pattern [10, 22, 24], and we sketch below how this analysis may straightforwardly be extended to the present problem of parametric forcing.

To this end, we restrict our attention to mutually perpendicular modes ($\theta = \pi/2$) and consider the planforms

$$\psi_1 = (b + s + \mathrm{i})\left(\tilde{A}(\tilde{X}, \tilde{Y}, \tilde{T}) \exp(\mathrm{i}kx) + \tilde{B}(\tilde{X}, \tilde{Y}, \tilde{T}) \exp(\mathrm{i}ky) + \text{c.c.}\right), \tag{3.19}$$

$$\rho_2 = \tilde{C}(\tilde{X}, \tilde{Y}, \tilde{T}) + \text{q.t.}, \tag{3.20}$$

where "q.t." denotes quadratic interaction terms involving $\tilde{A}$ and $\tilde{B}$ (proportional to $\mathrm{e}^{2\mathrm{i}kx}$, $\mathrm{e}^{\mathrm{i}k(x+y)}$, etc.). The amplitudes $\tilde{A}, \tilde{B}$ are complex, but $\tilde{C}$, the large-scale mode arising from mass conservation, is real.

As before, the evolution of the amplitudes takes place on the slow time scale $\tilde{T} = \epsilon^2 t$, while spatial modulations occur over the scales given by $\tilde{X} = \epsilon x$ and $\tilde{Y} = \epsilon y$. Examining the problems that arise at successive orders in ϵ, we find at $O(\epsilon^3)$, after a rescaling of amplitudes and time as in (3.7), and a spatial scaling

$$(X, Y) = \sqrt{\frac{b(\omega + b(1 + \rho_0))}{2(1 + b^2)(\omega b - 1 - \rho_0)}}\,(\tilde{X}, \tilde{Y}),$$

the amplitude equations

$$A_T = A + A_{XX} - |A|^2 A - \lambda |B|^2 A - AC, \tag{3.21}$$

$$B_T = B + B_{YY} - |B|^2 B - \lambda |A|^2 B - BC, \tag{3.22}$$

where λ is given in (3.10). In addition, at $O(\epsilon^4)$ the evolution equation for C is found to be (after the same rescaling of variables that leads to (3.21) and (3.22))

$$C_T = \sigma \nabla^2 C + \mu \nabla^2 \left(|A|^2 + |B|^2\right), \tag{3.23}$$

where

$$\sigma = \frac{\beta}{2}\left[\frac{\omega b + b^2(1 + \rho_0)}{(\omega b - 1 - \rho_0)(1 + b^2)}\right], \tag{3.24}$$

$$\mu = \frac{\alpha \rho_0}{3 - \phi}\left[\frac{\omega b + b^2(1 + \rho_0)}{(\omega b - 1 - \rho_0)(1 + b^2)}\right]. \tag{3.25}$$

The average of C over X and Y is conserved, according to (3.23), corresponding to conservation of mass in (2.2). We must set this average to zero, so that the average of ρ is ρ_0. The amplitude equations (3.21)–(3.23) are precisely those of Cox and Matthews [10] (in the special case that their parameter $\mu_- = 0$), and so we may directly apply their modulational stability results for rolls and squares.

We first consider rolls: a family of roll solutions to (3.21)–(3.23) exists in the form $A = Q\mathrm{e}^{\mathrm{i}qX}$, $B = C = 0$, where $Q^2 = 1 - q^2$, corresponding to rolls of wavenumber $k_c + \epsilon q$. Of course, we recover the results of section 3.1 regarding the stability of rolls to disturbances with exactly critical wavenumber: rolls are unstable to squares if $\lambda < 1$ (i.e., if $\phi > 1$). An analysis of the modulational stability of rolls [10] indicates that a new instability, leading to amplitude modulation, replaces the usual Eckhaus instability and, when $\mu > 0$, is more widespread than the case when there is no conserved mode. Indeed, all rolls may be modulationally unstable if $\mu > \sigma$, with the last rolls to succumb to the instability being those at band-center ($q = 0$). We thus focus on the band-center rolls. For these rolls, in terms of the original variables, the region of modulational instability corresponds to $1 < \phi < 3$, which is identical to that for instability to perturbations in a perpendicular mode. Since we expect the time-scale for the growth of perpendicular modes to be less than that for modulational modes, we do not expect to see long-wavelength instability of rolls in simulations, except under contrived circumstances, and we expect instead to see only the transition to squares. This coincidence of stability conditions is also found in the modified Swift–Hohenberg model considered in [10] and is a consequence of the fact that in these simple models the cross-coupling coefficient λ does not depend on the angle between the modes.

Squares take the form $A = Q\mathrm{e}^{\mathrm{i}qX}$, $B = Q\mathrm{e}^{\mathrm{i}qY}$, $C = 0$ with $Q^2 = (1 - q^2)/(1 + \lambda)$ and are stable to roll-type perturbations only while $|\lambda| < 1$, i.e., $1 < \phi < 9/5$. All square patterns are found to be unstable to modulational perturbations when $2\mu > (1 + \lambda)\sigma$ [10]. This implies that squares are modulationally unstable while $\phi > 1$, again matching the roll-perturbation stability boundary. This leads to the remarkable result that all square patterns are unstable, provided the container size is suitably large, and so a regular square pattern can never be seen in corresponding simulations. A more detailed analysis indicates that the initial deformation of a modulationally unstable square pattern will be primarily to its amplitude if the wavenumber $k + \epsilon q$ lies near the band-center, but to its phase otherwise [10]. Nevertheless, in a finite computational box, one would generally expect some modulation of both the amplitude and phase; in very long boxes, different modulational scalings are appropriate [28].

3.4. Large-container numerical simulations. For a numerical extension of the preceding analytical results, we simulate (2.1) and (2.2) in a large square container (so that long-wavelength modulations to the pattern can occur), again with periodic boundary conditions. Extensive simulations support the predictions regarding the modulational stability of both rolls and squares. Figure 3 illustrates the predicted long-wavelength instability of squares. We see that the initial manifestation of the instability is an amplitude modulation in the form of a large-scale square "superstructure" (a limited analysis of such a superstructure [10] is consistent with the square shape observed here). The final state reached in this example is a highly localized oscillon chain. Similar patterns have been found through a model based on the assumption of nearest-pattern interaction [17, 18], which indicates that the observed state is not merely an artifact of the present model. We find such localized states, composed of straight or

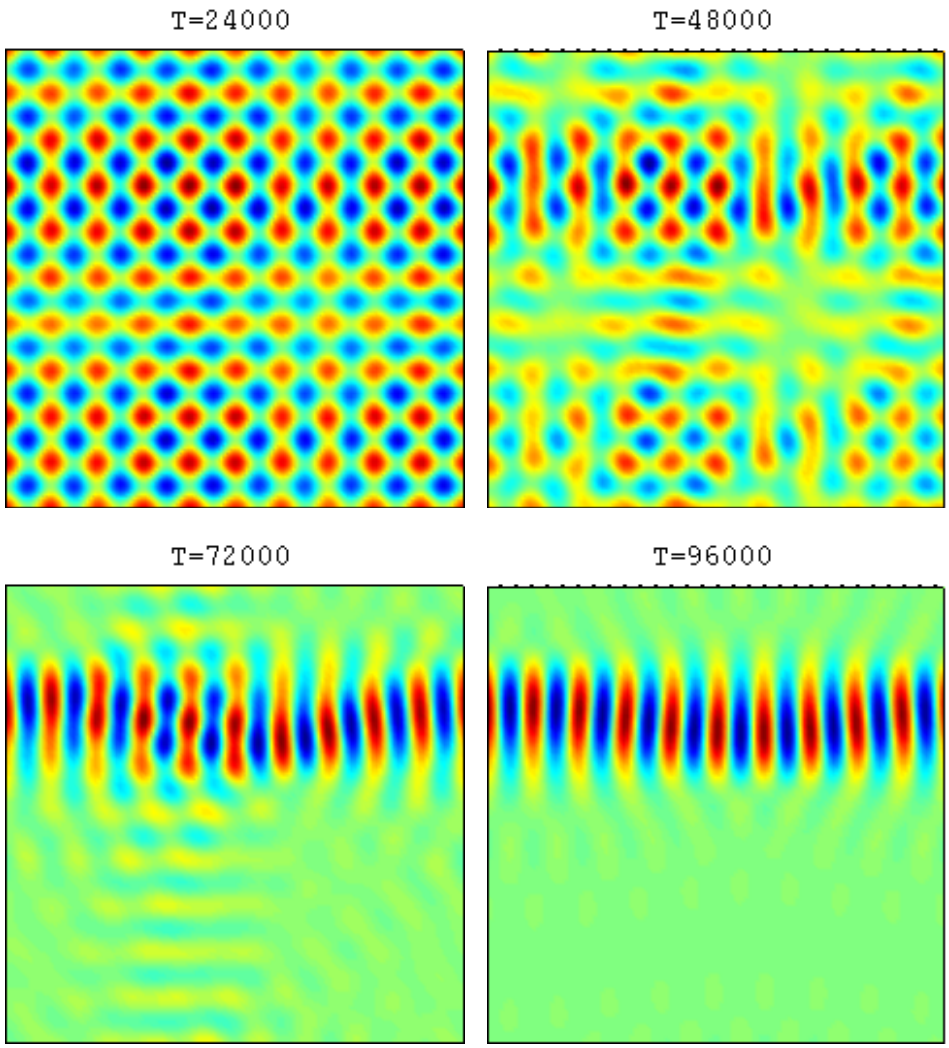


Figure 3. *Plots of $Re(\psi)$ depicting four snapshots of the long-wavelength instability of square patterns. The initial manifestation of the instability takes the form of a square modulation (cf. Figure* 5(a) *of* [10]*), while the final state is an oscillon chain (cf. Figure* 10(f) *of* [17]*), spreading across the entire width of the domain. For this simulation, parameter values are the same as for Figure* 1 *except the container size is $L = 10\Lambda_c$, γ is* 0.1% *above critical, and $\alpha = \phi = 1.1$. The initial condition is periodic squares perturbed by small amplitude noise.*

wavy oscillon chains, widely in our simulations in intermediate to large containers with $\phi > 1$ [36]. In large containers we also commonly observe worm-like patterns [29], several of which can coexist in any given simulation, in various orientations in different parts of the domain.

Although no corresponding analysis is currently available, an analogous modulational instability for antisquares can be seen in the simulations illustrated in snapshot in Figure 4. After an initial condition of small-amplitude noise, we see regular antisquares initially being selected before a long-wavelength amplitude modulation takes place in the form of a one-dimensional superstructure. Subsequently, the system evolves to a single oscillon chain.

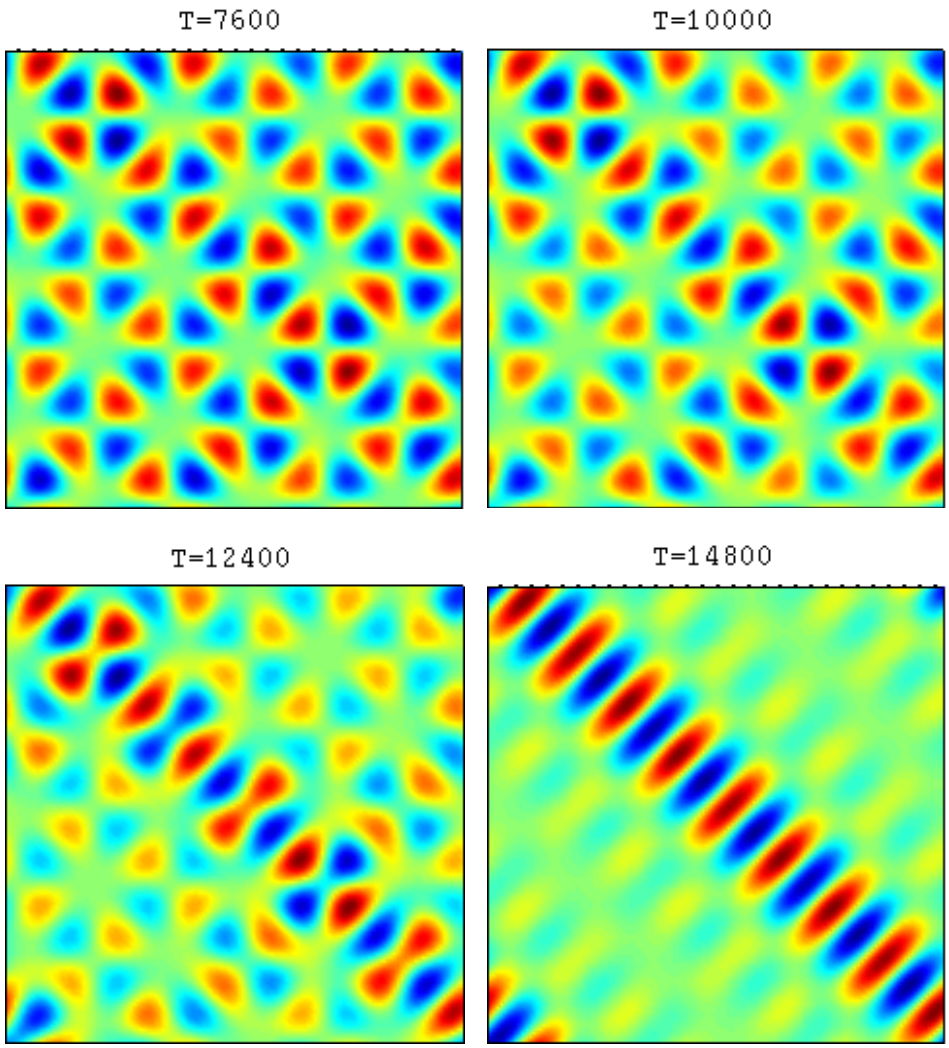


Figure 4. *Plots of $Re(\psi)$ depicting four snapshots of the long-wavelength instability of antisquare patterns. The initial manifestation of the instability takes the form of a roll modulation, while the final state is a single oscillon chain. For this simulation, the container size is $L = 3\sqrt{5}\Lambda_c$, γ is 0.1% above critical, and α is 1.25. The initial condition is small-amplitude noise.*

Simulating the model equations at increasing values of ϕ shows the degree of localization to increase as ϕ moves away from 1. This phenomenon is illustrated in Figure 5, where the final states are shown for four values of ϕ. We see roll solutions for $\phi = 0.9$, while rhombs are found at $\phi = 1.1$. Subsequently, at $\phi = 1.2$, we find a disordered cellular pattern, where a slight localization can be observed. Finally, at $\phi = 1.3$, a similar cellular pattern is apparent but with a far stronger degree of localization. Indeed, the final state can be thought of as comprising several oscillon chains. This behavior is characteristic of much of parameter space. An accompanying video (67540_02.mov [604KB]) shows a typical example of increasing localization leading to oscillon chains for $\phi = 1.2$, with γ chosen to be 0.2% above critical.

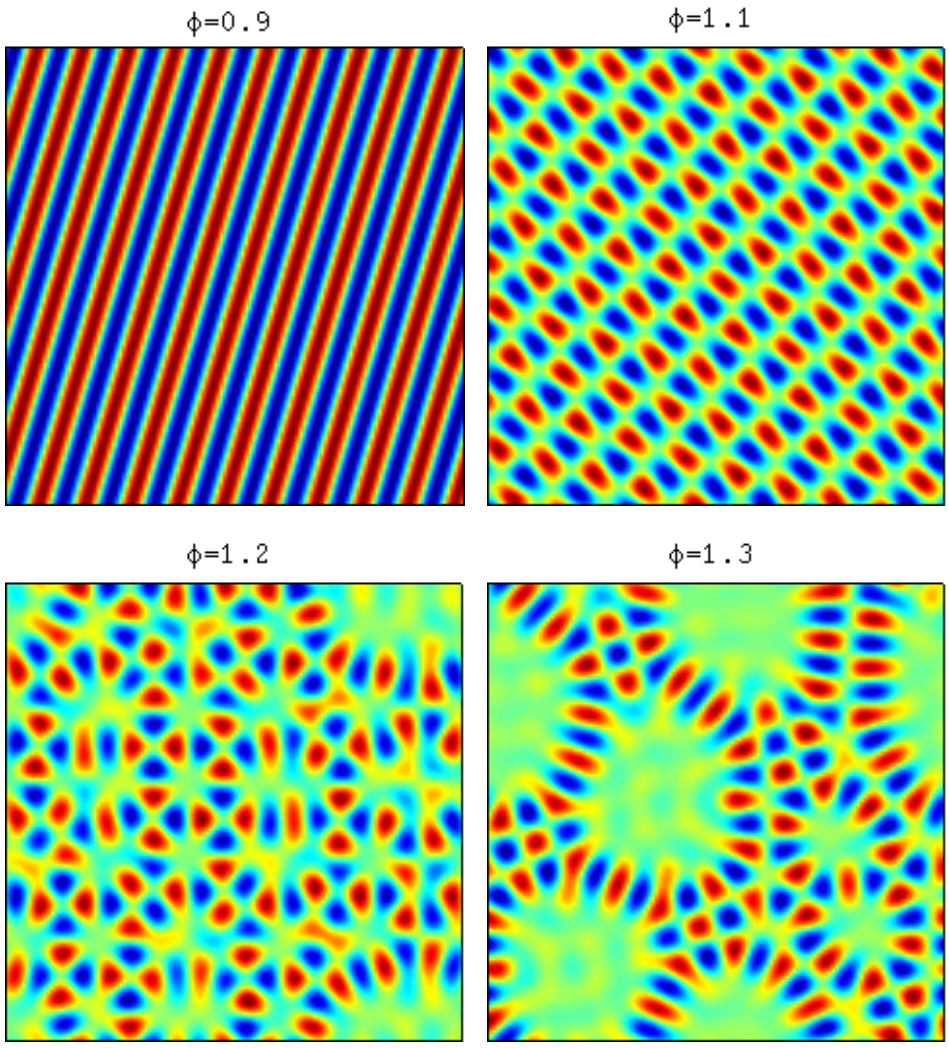


Figure 5. *Plots of $Re(\psi)$ illustrating the steady states realized as ϕ is increased. The parameter values are the same as in Figure 1, the container size is $L = 10.5\Lambda_c$, γ is 0.5% above critical, and the initial condition is noise of amplitude 10^{-3}.*

4. Discussion. The analysis presented here extends that of Tsimring and Aranson [2, 32] in providing a more complete description of the stability of patterns in their phenomenological model for pattern formation in a vertically vibrated granular layer. Extensive numerical experiments on (2.1) and (2.2) have supported all the analysis.

In small domains, a variety of unusual patterns can be stable, including antisquares and triangles. In larger domains, our results show that rolls and squares may suffer a modulational instability, which is strongly influenced by the constraint of mass conservation. It turns out that rolls are effectively immune to this modulational instability because it arises only where a regular roll pattern is already unstable to a regular pattern of squares. By contrast, squares are always susceptible; both these modulational results have been confirmed numerically. The

fact that squares are always unstable would appear to be in conflict with the experimental results, where stable square patterns are commonly observed [25]; it may thus indicate a shortcoming in the phenomenological model of (2.1) and (2.2). However, other explanations for the discrepancy are not ruled out. For example, since the modulational instability leads to long-wavelength perturbations becoming unstable first, the instability can take place only if perturbations on a sufficiently large spatial scale can be accommodated in the container. So for finite containers squares will generally be stable close to the roll–square transition, but will lose stability as the driving frequency is decreased (effectively increasing ϕ). A further possible explanation for the absence of observations of the instability of squares is that the long-wavelength nature of the instability corresponds to a very slow temporal growth of the disturbance. Thus, square patterns may be observed for a significant period of time before the long-wavelength modulation manifests itself; this has certainly been apparent in our numerical simulations. However, the exact coincidence of the roll–square transition and the modulational instability is a consequence of the simplicity of the model.

Since we have considered only a supercritical bifurcation from the uniform state, bistability arguments are not necessary in explaining the existence of strongly localized solutions. Rather, it is the influence of the conserved quantity that provides the localization mechanism: particles tend to flee regions in which ψ is large, leading to an increased flux of particles and a reduction in the local density ρ. This, in turn, decreases the damping in the ψ equation (2.1). Such mass redistribution, and corresponding localization, is observed in numerical simulations of models such as that of Eggers and Riecke [14]. The feedback loop just described enables localized structures to remain stable. In effect, a large-scale redistribution of the granular medium causes some regions to be locally "supercritical" and others locally "subcritical," while the system as a whole is supercritical. Many other models of a vertically vibrated granular layer include in them the principle of conservation of mass, and one might expect corresponding long-wavelength effects to be applicable in a wide range of models. However, few continuous order-parameter models exist; most are of a semicontinuum or stroboscopic nature and the weakly nonlinear analysis carried out here is not immediately transferable.

Our interest in (2.1) and (2.2) is not exclusively in the context of vibrated granular layers; such a phenomenological model could also account for a vibrated fluid layer. In this situation, the complex order-parameter ψ corresponds to the velocity potential at the free surface [37], while ρ represents the displacement from the undisturbed fluid height (the volume of fluid being conserved). In addition, aside from its application to granular and fluid layers, the phenomenological model (2.1) and (2.2) is an interesting and unusual model of pattern formation in its own right. It is unusual in the sense that the coupling to a mean field enables the selection of cellular patterns to be determined by the parameter values rather than being solely dependent on symmetries as in other common models such as the Swift–Hohenberg equation. What makes (2.1) and (2.2) particularly interesting is how the coupling with a large-scale mode facilitates the near-threshold exhibition of a plethora of cellular, but predominantly localized, patterns.

REFERENCES

[1] I. S. Aranson, D. Blair, W. K. Kwok, G. Karapetrov, U. Welp, G. W. Crabtree, and V. M. Vinokur, *Controlled dynamics of interfaces in a vibrated granular layer*, Phys. Rev. Lett., 82 (1998), pp. 731–734.

[2] I. S. Aranson and L. S. Tsimring, *Formation of periodic and localized patterns in an oscillating granular layer*, Phys. A, 249 (1998), pp. 103–110.

[3] I. S. Aranson and L. S. Tsimring, *Patterns and collective behavior in granular media: Theoretical concepts*, Rev. Modern Phys., 78 (2006), pp. 641–692.

[4] I. S. Aranson, L. S. Tsimring, and V. M. Vinokur, *Hexagons and interfaces in a vibrated granular layer*, Phys. Rev. E, 59 (1999), pp. R1327–R1330.

[5] H. Arbell and J. Fineberg, *Spatial and temporal dynamics of two interacting modes in parametrically driven surface waves*, Phys. Rev. Lett., 81 (1998), pp. 4384–4387.

[6] C. Bizon, M. D. Shattuck, J. B. Swift, W. D. McCormick, and H. L. Swinney, *Patterns in 3D vertically oscillated granular layers: Simulation and experiment*, Phys. Rev. Lett., 80 (1998), pp. 57–60.

[7] D. Blair, I. S. Aranson, G. W. Crabtree, V. Vinokur, L. S. Tsimring, and C. Josserand, *Patterns in thin vibrated granular layers: Interfaces, hexagons and superoscillons*, Phys. Rev. E, 61 (2000), pp. 5600–5610.

[8] E. Cerda, F. Melo, and S. Rica, *Model for subharmomic waves in granular materials*, Phys. Rev. Lett., 79 (1997), pp. 4570–4573.

[9] S. M. Cox and P. C. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.

[10] S. M. Cox and P. C. Matthews, *Instability and localisation of patterns due to a conserved quantity*, Phys. D, 175 (2003), pp. 196–219.

[11] C. Crawford and H. Riecke, *Oscillon-type structures and their interaction in a Swift–Hohenberg model*, Phys. D, 129 (1999), pp. 83–92.

[12] J. R. de Bruyn, C. Bizon, M. D. Shattuck, D. Goldman, J. B. Swift, and H. L. Swinney, *Continuum-type stability balloon in oscillated granular layers*, Phys. Rev. Lett., 81 (1998), pp. 1421–1424.

[13] B. Dionne, M. Silber, and A. C. Skeldon, *Stability results for steady, spatially periodic planforms*, Nonlinearity, 10 (1997), pp. 321–353.

[14] J. Eggers and H. Riecke, *Continuum description of vibrated sand*, Phys. Rev. E, 59 (1999), pp. 4476–4483.

[15] I. Goldhirsch, *Rapid granular flows*, in Annual Review of Fluid Mechanics, Annu. Rev. Fluid Mech. 35, Annual Reviews, Palo Alto, CA, 2003, pp. 267–93.

[16] H. M. Jaeger, S. R. Nagel, and R. P. Behringer, *Granular solids, liquids and gases*, Rev. Modern Phys., 68 (1996), pp. 1259–1273.

[17] S.-O. Jeong, T.-W. Ko, and H.-T. Moon, *Oscillons, kinks and patterns in a model for a periodically forced medium*, Phys. D, 164 (2002), pp. 71–84.

[18] S.-O. Jeong, H.-T. Moon, and T.-W. Ko, *Nearest pattern interaction and global pattern formation*, Phys. Rev. E, 62 (2000), pp. 7778–7780.

[19] L. P. Kadanoff, *Built upon sand: Theoretical ideas inspired by granular flows*, Rev. Modern Phys., 71 (1999), pp. 435–444.

[20] S. V. Kiyahsko, L. N. Korzinov, M. I. Rabinovich, and L. S. Tsimring, *Rotating spirals in a Faraday experiment*, Phys. Rev. Lett., 54 (1995), pp. 5037–5040.

[21] E. Knobloch, *Doubly diffusive waves*, in Double Diffusive Motions, The American Society of Mechanical Engineers, New York, 1985, pp. 17–22.

[22] N. L. Komarova and A. C. Newell, *Nonlinear dynamics of sand banks and sand waves*, J. Fluid Mech., 415 (2000), pp. 285–321.

[23] A. Kudrolli, B. Pier, and J. P. Gollub, *Superlattice patterns in surface waves*, Phys. D, 123 (1998), pp. 99–111.

[24] P. C. Matthews and S. M. Cox, *Pattern formation with a conservation law*, Nonlinearity, 13 (2000), pp. 1293–1320.

[25] F. Melo, P. Umbanhowar, and H. L. Swinney, *Transition to parametric wave patterns in a vertically oscillated granular layer*, Phys. Rev. Lett., 72 (1994), pp. 172–175.
[26] F. Melo, P. B. Umbanhowar, and H. L. Swinney, *Hexagons, kinks and disorder in oscillated granular layers*, Phys. Rev. Lett., 75 (1995), pp. 3838–3841.
[27] H.-K. Park and H.-T. Moon, *Square to stripe transition and superlattice patterns in vertically oscillated granular layers*, Phys. Rev. E, 65 (2002), 051310.
[28] M. R. E. Proctor, *Finite amplitude behaviour of the Matthews–Cox instability*, Phys. Lett. A, 292 (2001), pp. 181–187.
[29] H. Riecke and G. D. Granzow, *Localization of waves without bistability: Worms in nematic electrocon vection*, Phys. Rev. Lett., 81 (1998), pp. 333–336.
[30] D. H. Rothman, *Oscillons, spiral waves, and stripes in a model of vibrated sand*, Phys. Rev. E, 57 (1998), pp. R1239–R1242.
[31] T. Shinbrot, *Competition between randomizing impacts and inelastic collisions in granular pattern formation*, Nature, 389 (1997), pp. 574–576.
[32] L. S. Tsimring and I. G. Aranson, *Localized and cellular patterns in a vibrated granular layer*, Phys. Rev. Lett., 79 (1997), pp. 213–216.
[33] P. Umbanhowar, F. Melo, and H. Swinney, *Localized excitations in a vertically vibrated granular layer*, Nature, 382 (1996), pp. 793–796.
[34] P. B. Umbanhowar and H. L. Swinney, *Wavelength scaling and square/stripe and grain mobility transition in vertically oscillated granular layers*, Phys. A, 288 (2000), pp. 344–362.
[35] S. C. Venkataramani and E. Ott, *Spatiotemporal bifurcation phenomena with temporal period doubling*, Phys. Rev. Lett., 80 (1998), pp. 3495–3498.
[36] D. M. Winterbottom, *Pattern Formation with a Conservation Law*, Ph.D. thesis, University of Nottingham, Nottingham, UK, 2005, http://etheses.nottingham.ac.uk/archive/00000180/.
[37] W. Zhang and J. Viñals, *Secondary instabilities and spatiotemporal chaos in parametric surface waves*, Phys. Rev. Lett., 74 (1995), pp. 690–693.
[38] W. Zhang and J. Viñals, *Square patterns and quasipatterns in weakly damped Faraday waves*, Phys. Rev. E, 53 (1996), pp. R4283–R4286.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 79–107
© 2008 Society for Industrial and Applied Mathematics

Near Invariance for Markov Diffusion Systems*

Fritz Colonius†, Tobias Gayer†, and Wolfgang Kliemann‡

Abstract. A concept of "near invariance" is developed starting from sets that are actually invariant under smaller perturbations. This is based on a theory for system dynamics of Markov diffusion processes illuminating the idea of "large" noise perturbations turning invariant sets for smaller noise ranges into transient sets. The controllability behavior of associated deterministic systems plays a crucial role. This setup also allows for numerical computation of nearly invariant sets, the exit times from these sets, and the exit locations under varying perturbation ranges. Three examples with additive perturbations are included: a one degree of freedom system with double well potential and the escape equation without and with periodic excitation.

Key words. almost invariance, Markov diffusions, control sets

AMS subject classifications. 37H20, 60J60, 03B05

DOI. 10.1137/040618539

1. Introduction.

Almost invariance is an often used concept for stochastic dynamical systems that intends to describe sets such that the system

- stays within a set in the state space for a "long" time,
- exits from the set only under "large" noise perturbations,
- and may return to this set at a later, much "longer" time.

Hence almost invariance tries to describe a transient phenomenon of stochastic systems, but on "large" time intervals. The interpretation of "large" time intervals and "large" perturbations usually depends on the application one has in mind.

Recently, these phenomena have found renewed interest. This includes approaches based on transfer operator theory combined with set oriented numerics (Dellnitz and Junge [11], [12], Froyland [20], [21], and Froyland and Dellnitz [22]) and graph theoretic methods (Dellnitz et al. [14]) as well as extensions of metastability in the classical Freidlin–Wentzell theory [19] in Huisinga, Meyn, and Schütte [27] and the analysis of dominant eigenvalues of transfer operators (Schütte, Huisinga, and Deuflhard [37]; Deuflhard et al. [16]). An important approach is also developed in the work of Bovier [2] and Bovier et al. [3], [4].

Applications of almost invariant sets include, e.g., the analysis of molecular dynamics, where they can symbolize conformations of a molecule that are essential for its chemical properties (Deuflhard and Schütte [15]); Mezic [34] proposes a different dynamical systems explanation of conformation dynamics based on an interplay between local and global interconnections for coupled oscillator networks. Similar problems of almost invariance occur,

*Received by the editors November 8, 2004; accepted for publication (in revised form) by K. Mischaikow July 10, 2007; published electronically January 16, 2008.
http://www.siam.org/journals/siads/7-1/61853.html
†Institut für Mathematik, Universität Augsburg, 86135 Augsburg, Germany (fritz.colonius@math.uni-augsburg.de, tobias.gayer@math.uni-augsburg.de).
‡Department of Mathematics, Iowa State University, Ames, IA 50011 (kliemann@iastate.edu).

e.g., in dynamical astronomy (Dellnitz et al. [13]), in the analysis of dynamic reliability when one tries to estimate rare occurrences of system failure due to large perturbations (see, e.g., Colonius, Häckl, and Kliemann [5]), and in other models in engineering and science.

The goal of this paper is to develop a theory that

- defines a plausible concept of "nearly invariant sets" based on the actual system dynamics of Markov diffusion processes,
- illuminates the idea of "large" noise perturbations turning invariant sets for smaller noise ranges into transient sets,
- explores the idea of invariance over "large" time intervals,
- and allows for numerical computation of nearly invariant sets, the exit times from these sets, and the exit locations under varying perturbation ranges.

Thus the concept of near invariance captures essential features of "almost invariance." Our approach is, roughly, as follows:

- We consider Markov diffusion models (i.e., the system does not anticipate future behavior of the noise) with perturbations entering as parameter or additive noise into the system dynamics, which are modeled as a set of ordinary differential equations

$$\dot{x} = X_0(x) + \sum_{i=1}^{m} \xi_i(t,\omega) X_i(x) \tag{1}$$

  on a finite dimensional C^∞ manifold M, where the C^∞ vector field X_0 describes the unperturbed dynamics and $\xi(t,\omega) = (\xi_i(t,\omega),\ i = 1,\dots,m)$ is the vector of random perturbation processes with C^∞ dynamics $X_1,\dots,X_m$. We model ξ as a function $\xi = f(\eta)$ of a background noise η, $f : N \to U$, where N is the state space of the background noise and $U \subset \mathbb{R}^m$ is the set of perturbation values. We assume η to be a stationary, ergodic Markov process.
- The noise range is treated as a parameter $\rho \geq 0$ of the system by introducing a family $f^\rho : N \to U^\rho$, $\rho \geq 0$, of functions such that the sets U^ρ of perturbation values increase with ρ. Setting $U^0 = \{0\}$, we recover the unperturbed dynamics of the system (1).
- We identify the invariant sets of the stochastic system (1), depending on the noise range. Under mild conditions, the invariant control sets of an associated control system are the supports of the invariant measures of (1) and they form the cores of the invariant sets for the system.
- Analyzing the change of the invariant sets as the noise range $\rho \geq 0$ increases leads to the study of the loss of invariance, specifically to the analysis of bifurcation points ρ_0 where an invariant set loses its invariance and becomes transient or "nearly invariant."
- Finally, we study the exit time distributions from invariant sets as they become transient under the influence of larger perturbations.

This approach develops a concept for near invariance starting from sets that are actually invariant under smaller perturbations. In other approaches the term "almost invariance" is used to describe the behavior in certain regions, usually in relation to an invariant probability measure with support on the whole state space; see, e.g., Huisinga, Meyn, and Schütte in [27]. In the approach outlined above, such a reference measure need not exist, and we suggest the term "near invariance" for the concept developed here.

Though our analytical approach applies to systems in arbitrary finite dimension, numerical evaluations appear possible for low dimensional systems only thus restricting the range of applicability.

It is worth noting that recently, in the context of random diffeomorphisms, problems similar to near invariance have been analyzed by Zmarrou and Homburg [41]. They analyze average escape times from sets as functions of a bifurcation parameter.

In section 2 we describe the setup used in this paper and recall some background material on Markov diffusion systems and their qualitative behavior, based on the analysis of associated control systems with varying control range. Section 3 presents the definition of near invariance together with the main result on the existence of nearly invariant sets. Theorem 3.3 and Corollary 3.4 describe the bifurcation points where an invariant and a variant set merge to generate a nearly invariant set. The rest of this section is devoted to the study of the exit sets from variant sets. Section 4 discusses the numerical computation of exit times for nearly invariants sets and the corresponding exit locations. Section 5 analyzes three examples in some detail: a one degree of freedom system with double well potential and additive perturbation and two perturbed versions of the escape equation, without and with an extra periodic excitation; see, e.g., [40], [35], [17], [26], and the references therein. The latter example is three dimensional and at the present limit of our computational possibilities. The appendix, section 6, contains some background information on parameter dependent deterministic control systems that is used throughout the paper.

2. Markov diffusion systems and associated control systems. In this section we recall some facts about Markov diffusion systems, their relations to associated control systems, and the support theorem of Stroock and Varadhan. We start from the system

$$\dot{x} = X_0(x) + \sum_{i=1}^{m} f_i(\eta_t) X_i(x) \tag{2}$$

on a finite dimensional, C^∞ manifold M with C^∞ vector fields $X_0, \dots, X_m$ as in section 1. First we specify our assumptions on the background noise η. Let N be a compact connected finite dimensional C^∞ manifold on which the stochastic differential equation

$$d\eta = Y_0(\eta)dt + \sum_{j=1}^{l} Y_j(\eta) \circ dW_j \tag{3}$$

is defined. Here $W = (W_j)$ is an l dimensional Wiener process, $Y_0, \dots, Y_l$ are C^∞ vector fields on N, and "$\circ$" denotes the Stratonovich stochastic differential. The compactness of the noise space N rules out excitation processes with Gaussian statistics, and thus (3) can be regarded as a realistic model of physical systems with bounded noise. We assume that (3) admits at least one stationary Markov solution. Imposing the Lie algebra rank condition

$$\dim \mathcal{LA}\{Y_1, \dots, Y_l\}(q) = \dim N \text{ for all } q \in N \tag{4}$$

as a nondegeneracy condition on N guarantees that this stationary solution is unique (see Kunita [31]) and can be extended to a stationary Markov solution η_t^*, $t \in \mathbb{R}$.

The noise process $\xi_t := f_i(\eta_t)$ in (2) is defined in the following way: Let $U \subset \mathbb{R}^m$ be a compact convex set with $0 \in \operatorname{int} U$ and $U = \operatorname{cl} \operatorname{int} U$. Let

$$f : N \to U$$

be a continuous surjective function such that there exists a closed connected subset $L \subset N$. $f|_L$ is C^1 and $Df(\eta)$ has full rank for all $\eta \in L$ with $f(\eta) \in U$; see [7]. Then $\xi_t := f(\eta_t^*)$ is a stationary process with values in U.

We model variations in the size of the noise by introducing a parameter $\rho \geq 0$ and the noise ranges U^ρ, satisfying the same assumption as U above. We consider the process η_t^* as a background noise, which for every ρ is mapped into the stochastic perturbation space $\mathcal{U}^\rho = \{u : \mathbb{R} \to U^\rho, \textit{measurable}\}$ by a continuous surjective function

$$f^\rho : N \to U^\rho,$$

which satisfies the assumptions on f above. Combining this perturbation model with system (1), we arrive at the Markov diffusion system

$$\begin{aligned} d\eta &= Y_0(\eta)dt + \textstyle\sum_{j=1}^{\ell} Y_i(\eta) \circ dW_j, \quad \eta_0 = \eta_0^*, \\ \dot{x} &= X_0(x) + \textstyle\sum_{i=1}^{m} f_i^\rho(\eta_t) X_i(x) \end{aligned} \tag{5}$$

on the state space $M \times N$, for which we assume the existence and uniqueness of a strong solution for all $t \geq 0$. This system is degenerate since the Wiener process acts only on the second component. Note that, in general, the component $x(t)$ by itself is not Markovian. The pair process $(x(t), \eta_t)$ is, however, a Markov diffusion process for all ρ, if the initial random variable x_0 in M is independent of the increments of the Wiener process. Compare, in particular, especially [29] for results on degenerate diffusions along these lines, and [7] and [8] for more details on our setting in general.

The system (5) can be analyzed using control theory via the support theorem presented by Stroock and Varadhan in [38]. To make this more precise, we set up the control system associated with (5) to be

$$\begin{aligned} \dot{\eta} &= Y_0(\eta) + \textstyle\sum_{j=1}^{\ell} w_j(t) Y_i(\eta), \\ \dot{x} &= X_0(x) + \textstyle\sum_{i=1}^{m} f_i^\rho(\eta_t) X_i(x), \end{aligned} \tag{6}$$

where $w \in \mathcal{W} := \{w : [0, \infty) \to \mathbb{R}^l, \text{piecewise constant}\}$, and we assume the Lie algebra rank condition (4) for the η-component. Furthermore, we want the pair system (5) to be regular; i.e., we want the topological support of its transition probabilities from each point $(x, p) \in M \times N$ to have nonvoid interior in $M \times N$. This is guaranteed by

$$\dim \mathcal{LA} \left\{ \begin{pmatrix} X_0 + \sum \eta_i X_i(x) \\ Y_0 + \sum w_j Y_j \end{pmatrix}, w \in \mathbb{R}^l \right\} \begin{pmatrix} x \\ \eta \end{pmatrix} = \dim M + \dim N \tag{7}$$

for all $(x, \eta) \in M \times N$ (see Meyn and Tweedie [33] for a relaxation of this condition). Instead of (6) it will be sufficient to consider the system

$$\dot{x}(t) = X_0(x(t)) + \sum_{i=1}^{m} u_i(t)\, X_i(x(t)), \quad u \in \mathcal{U}^\rho; \tag{8}$$

see the appendix, section 6, for definitions and notation of control systems. Note that the condition (7) implies local accessibility for the x-component (8).

We fix $\rho \geq 0$ for the remainder of this section and drop it in the notation. For all $(x, \eta) \in M \times N$ the orbits $\mathcal{O}^+(x, \eta)$ of system (6) are of the form $\operatorname{cl} \mathcal{O}^+(x, \eta) = \operatorname{cl} \mathcal{O}^+(x) \times N$, where $\mathcal{O}^+(x)$ is the forward orbit of the system (8) from $x \in M$. In particular, the invariant control sets $\hat{C} \subset M \times N$ of (6) correspond one-to-one to the invariant control sets $C \subset M$ of (8) via $\hat{C} = C \times N$. This follows from Lemma 3.17 in [7]. (We remark that in the statement of that lemma one has to add the surjectivity assumption for f which is used in the proof.) Therefore, the global control structure of the x-component (8) determines the control structure of the pair process (6).

The natural probability space to work in is $\hat{\Omega} := \mathcal{C}(\mathbb{R}_0^+, M \times N) = \{\omega : R_0^+ \to M \times N, \textit{continuous}\}$ and for fixed initial conditions $(x, q) \in M \times N$ the pair process (5) induces a probability measure $\hat{P}_{(x,q)}$ on $\hat{\Omega}$. By $\hat{P}_{(x,\eta^*)}$ we denote the measure corresponding to the stationary Markov solution $\{\eta_t^*, t \geq 0\}$ in the η-component. Its marginal distribution on $\Omega := \mathcal{C}(\mathbb{R}_0^+, M)$ will be denoted by P_x, $x \in M$. The trajectories of the pair process are $(\varphi(t, (x, q), \omega), \eta(t, q, \omega))$ for $(x, q) \in M \times N$, and we will write the x-component under $\{\eta_t^*, t \geq 0\}$ as $\varphi(t, x, \omega)$, $x \in M$. Then the "transition probability" from $x \in M$ to a set $A \subset M$ in time $t \geq 0$ is

$$P(t, x, A) = P_x(\varphi(t, x, \omega) \in A). \tag{9}$$

Using the tube method introduced by Arnold and Kliemann in [1], it follows (compare with [28]) from the support theorem that

$$\operatorname{supp} P(t, x, \cdot) = \operatorname{cl} \left\{ \begin{matrix} y \in M \mid \text{there is a piecewise continuous} \\ u \in \mathcal{U} \text{ such that } \varphi(t, x, u) = y \end{matrix} \right\}. \tag{10}$$

It now follows from [29] and [7] that the invariant Markov probability measures μ of (5) have support given by $\operatorname{supp} \mu = C \times N$, where C is an invariant control set of (8), and these measures are unique on each set of this form. We call *ergodic sets* those invariant control sets C of (8) such that $C \times N$ is the support of some invariant Markov measure, which includes, in particular, all bounded invariant control sets. All points in $M \times N$ outside of invariant control sets are transient.

To describe the consequences of the support theorem for the relationship between the Markov diffusion process (5) and the control system (8) in more detail, we define the *first entrance time* of (5) to a set $A \subset M$ from a point $x \in M$ as the random variable

$$\tau_x(A) := \inf\{t \geq 0, \varphi(t, x, \omega) \in A\},$$

and the *first exit time* of (5) from a set $A \subset M$ starting at a point $x \in M$ as the random variable

$$\sigma_x(A) := \inf\{t \geq 0, \varphi(t, x, \omega) \notin A\}.$$

The corresponding *exit location* is given as

$$h_x(A)(\omega) := \begin{cases} y \in M, \ y = \varphi(\sigma_x(A), x, \omega) & \text{for } \sigma_x(A)(\omega) < \infty, \\ \emptyset & \text{for } \sigma_x(A)(\omega) = \infty. \end{cases}$$

Due to Theorem 3.19 in [7], for invariant control sets $C \subset M$ of system (8) the equation $P_x(\sigma_x(C) < \infty) = 0$ holds for all $x \in C$. For bounded variant control sets $D \subset M$, on the other hand, it holds that $P_x(\sigma_x(D) < \infty) = 1$ for all $x \in D$. Under the measure P_x we even have that the expectation of the sojourn time $E_x[\sigma_x(D)]$ is finite (see [5, Theorem 11]).

3. Near invariance and mergers of control sets. If a bounded invariant control set C^ρ for $\rho \leq \rho_0$ becomes variant for $\rho > \rho_0$, then the corresponding ergodic set of the Markov process disappears and becomes transient. Nevertheless, although the disappearance of an ergodic set changes the global behavior of a stochastic system considerably, we expect the system to experience large exit times from the resulting variant control set as long as ρ is close to ρ_0 (see [25] for an example that can serve as a prototype of this phenomenon). This behavior is captured more generally in the following definition.

Definition 3.1. *Consider the family of Markov diffusion systems* $(5)^\rho$. *A closed set* $A \subset M$ *with* $\operatorname{int} A \neq \emptyset$ *is* nearly invariant *in* $x_0 \in \operatorname{int} A$ *for* $\rho > \rho_0$ *if*

(i) $\sigma^\rho_{x_0}(A) < \infty$ *with positive probability for* $\rho > \rho_0$, *and*

(ii) *for all* $x \in A$ *one has* $\sigma^\rho_x(A) \nearrow \infty$ *almost surely for* $\rho \searrow \rho_0$ *and* $\sigma^{\rho_0}_x(A) = \infty$ *almost surely.*

If A *is nearly invariant in every* $x_0 \in \operatorname{int} A$, *the set* A *is called* nearly invariant.

The following theorem reduces the search for nearly invariant sets to the search for closed sets A which are invariant for the control range U^{ρ_0} and lose their invariance under increased control ranges.

Theorem 3.2. *Suppose the Markov diffusion systems* $(5)^\rho$ *satisfy the Lie algebra rank conditions* (7) *and* (4) *and that* U^ρ *increases upper semicontinuously with respect to* $\rho \in (\rho_*, \rho^*)$. *Let* $x_0 \in \operatorname{int} A$ *for some closed set* $A \subset M$, $\operatorname{int} A \neq \varnothing$, *and consider* $\rho_0 \in (\rho_*, \rho^*)$. *Then the set* A *is nearly invariant in* x_0 *if and only if the set* A *is positively invariant for* ρ_0, *and for each* $\rho > \rho_0$

$$\operatorname{int}(\mathcal{O}^{\rho,+}(x_0) \setminus A) \neq \emptyset. \tag{11}$$

Proof. First we show that from positive invariance of A and upper semicontinuity of U^ρ at $\rho = \rho_0$ property (ii) of Definition 3.1 follows. By Lemma 6.1, $\operatorname{int} A$ is also positively invariant and hence $\sigma^{\rho_0}_x(A) = \infty$ almost surely. Now assume, contrary to the other assertion, that there are $x \in A$, a positive time $T > 0$, and $\rho_n \searrow \rho_0$ such that $P_x(\sigma^{\rho_n}_x(A) < T) > 0$. Then from (10) it follows that for all ρ_n there is a control $u_n \in \mathcal{U}^{\rho_n}$ with $\varphi(T, x, u_n) \notin A$, and, due to continuity, there are positive times $t_n < T$ such that $\varphi(t_n, x, u_n) \in \partial A$. Since U^ρ is increasing, we can look upon the sequence u_n as a sequence in the compact set $\mathcal{U}^{\rho_1}$ endowed with the weak*-topology. Then there are subsequences, called (t_n) and (u_n) again, such that $t_n \to t_*$ and $u_n \to u_*$. By (20) it follows that $\varphi(t_n, x, u_n) \to \varphi(t_*, x, u_*)$. Now observe that on a bounded interval weak*-convergence in L_∞ implies weak convergence in L_2; and here a subsequence of a weakly convergent sequence converges pointwise. Hence upper semicontinuity of the closed sets U^ρ implies that $u_* \in \mathcal{U}^{\rho_0}$, because if $u_*(t)$ was not in U^{ρ_0} for some t, this would contradict $u_n(t) \in U^{\rho_n}$ for all n. Then by continuity it follows that $\varphi(t_*, x, u_*) \in \partial A$, contradicting the positive invariance of $\operatorname{int} A$.

Next we prove that assumption (11) implies property (i) of near invariance by showing that $P_{x_0}(\sigma^\rho_{x_0}(A) < \infty) > 0$ for all $\rho > \rho_0$. Pick $\rho > \rho_0$; then there are some open set

$V \subset \operatorname{int}(\mathcal{O}^{\rho,+}(x_0) \setminus A)$, a positive time $t_0 < \infty$, and a piecewise constant control $u_0 \in \mathcal{U}^\rho$ such that $\varphi(t_0, x_0, u_0) \in V$. By continuous dependence of the solutions of $(8)^\rho$ on u, there is an open neighborhood $\mathcal{V}(u_0) \subset \mathcal{U}^\rho$ such that $\varphi(t_0, x_0, u) \in V$ for all $u \in \mathcal{V}(u_0)$. The support theorem implies that $P(\eta \in \mathcal{C}(\mathbb{R}_0^+, N),\ f^\rho(\eta) \in \mathcal{V}(u_0)) > 0$. Since the trajectories of (5) are continuous, we obtain

$$\begin{aligned} P_{x_0}(\sigma^\rho_{x_0}(A) < \infty) &\geq P_{x_0}(\sigma^\rho_{x_0}(A) < t_0) \\ &\geq P(\eta \in \mathcal{C}(\mathbb{R}_0^+, N),\ f^\rho(\eta) \in \mathcal{V}(u_0)) > 0. \end{aligned} \tag{12}$$

For the converse implication assume that A is nearly invariant in $x_0 \in \operatorname{int} A$ for $\rho > \rho_0$. Then $\sigma^\rho_{x_0}(A) < \infty$ with positive probability for $\rho > \rho_0$. Thus for every $\rho > \rho_0$ there is a realization of η and a time T such that with $u^\rho := f^\rho(\eta) \in \mathcal{U}^\rho$

$$\varphi(T, x_0, u^\rho) \notin A.$$

Thus $\varphi(T, x_0, u^\rho) \in \mathcal{O}^{\rho,+}(x_0) \setminus A$. Local accessibility of (8) implies that

$$\mathcal{O}^{\rho,+}(x_0) \subset \operatorname{cl} \operatorname{int} \mathcal{O}^{\rho,+}(x_0).$$

Since A is closed, we see that for every $\rho > \rho_0$ condition (11) holds.

It remains to show that the set A is positively invariant for ρ_0. This follows from $\sigma^{\rho_0}_x(A) = \infty$ almost surely. In fact, if A is not positively invariant, we obtain a contradiction using the same reasoning as above in the proof that (11) implies property (i) of near invariance. ■

This result shows that we have to look for closed sets which are positively invariant for ρ_0 and lose their invariance for $\rho > \rho_0$. Naturally, the sets A that are nearly invariant for all $x_0 \in \operatorname{int} A$ are of particular interest. These sets are specified in the following theorem. Recall from section 6 that $\mathbf{A}^{inv}(I)$ denotes the largest invariant set in the domain of attraction of a set I.

Theorem 3.3. (i) *Let the assumptions of Theorem 3.2 be satisfied and let C^{ρ_0} be a compact invariant control set for ρ_0. For each $\rho > \rho_0$ denote by C^ρ the unique control set of $(8)^\rho$ for which $C^{\rho_0} \subset C^\rho$. Suppose that there is $x \in \operatorname{int} C^{\rho_0}$ with*

$$\operatorname{int}(\mathcal{O}^{\rho,+}(x) \setminus C^{\rho_0}) \neq \emptyset \text{ for all } \rho > \rho_0. \tag{13}$$

Then the invariant control set C^{ρ_0} is nearly invariant for $\rho > \rho_0$.

(ii) *For every compact set $K \subset M$ the intersection $\mathbf{A}^{inv}(C^{\rho_0}) \cap K$ is nearly invariant for ρ_0 if the intersection is positively invariant for ρ_0.*

(iii) *If the invariant control set C^{ρ_0} is nearly invariant for $\rho > \rho_0$ and bounded, then $P_{x_o}\{\sigma^\rho_{x_0}(C^{\rho_0}) < \infty\} = 1$ for all $x_0 \in C^{\rho_0}$ and all $\rho > \rho_0$.*

(iv) *Condition (13) is satisfied, in particular, if C^{ρ_0} merges with a variant control set D^{ρ_0} with nonvoid interior, i.e., $D^{\rho_0} \subset C^\rho$ for all $\rho > \rho_0$, or if all $(u, x) \in \mathcal{U}^{\rho_0} \times C^{\rho_0}$ are inner pairs of system $(8)^\rho$ for every $\rho > \rho_0$; compare with the appendix, section 6.*

Proof. (i) We show that C^{ρ_0} is nearly invariant for $\rho > \rho_0$. Since $\operatorname{int}(C^\rho \setminus C^{\rho_0}) \neq \emptyset$ and C^ρ is a control set, there are $y \in \operatorname{int}(C^\rho \setminus C^{\rho_0})$ and $x \in \operatorname{int} C^{\rho_0}$ such that $y \in \mathcal{O}^{\rho,+}(x)$. Due

to continuity, it follows that there is an open neighborhood $V(y) \subset \operatorname{int}(C^\rho \setminus C^{\rho_0})$ of y such that $V(y) \subset \mathcal{O}^{\rho,+}(C^{\rho_0})$, and therefore condition (11) holds.

(ii) Condition (13) implies that (11) is satisfied for every $x_0 \in A := \mathbf{A}^{inv}(C^{\rho_0})$, since

$$\mathcal{O}^{\rho,+}(x) \subset \mathcal{O}^{\rho,+}(x_0) \text{ for all } x_0 \in \mathbf{A}^{inv}(C^{\rho_0}).$$

(iii) According to [29], all points $x \in M$ are either recurrent or transient, and points in variant control sets are transient. Furthermore, the first exit time from bounded sets of transient points is almost surely finite.

(iv) If C^{ρ_0} merges with a variant control set D^{ρ_0} with nonvoid interior, one has $D^{\rho_0} \cap C^{\rho_0} = \emptyset$ and $D^{\rho_0} \subset C^\rho$ for $\rho > \rho_0$, and therefore condition (13) is satisfied. Finally, from the assumption that all $(u,x) \in \mathcal{U}^{\rho_0} \times C^{\rho_0}$ are inner pairs of system $(8)^\rho$ for $\rho > \rho_0$, it ensues that $C^{\rho_0} \subset \operatorname{int} C^\rho$ according to Theorem 6.4. Therefore there is some open set $V \subset C^\rho \setminus C^{\rho_0}$, and condition (13) holds. ■

This theorem shows that control sets C^{ρ_0} that are invariant for the perturbation range ρ_0, but variant for $\rho > \rho_0$, are the key nearly invariant sets of a stochastic system. They are contained in the variant control sets $D^\rho \supset C^{\rho_0}$ as "nearly invariant" sets. If these nearly invariant sets are also bounded, then property (i) of Definition 3.1 holds with probability 1. In this situation, we also have the following consequence.

Corollary 3.4. *Let the assumptions of Theorem 3.2 be satisfied and let C^{ρ_0} be a compact invariant control set for ρ_0. For each $\rho > \rho_0$ denote by C^ρ the unique control set of $(8)^\rho$ for which $C^{\rho_0} \subset C^\rho$. Assume that C^{ρ_0} merges with a variant control set D^{ρ_0} with nonvoid interior, i.e., $D^{\rho_0} \subset C^\rho$ for all $\rho > \rho_0$. If C^ρ is bounded, then $P_x\{\sigma_x^\rho(C^\rho) < \infty\} = 1$ for all $x \in C^\rho$, $\rho > \rho_0$, and $\sigma_x(C^\rho)$ has finite expectation. This holds, in particular, for $x \in C^{\rho_0}$.*

The proof of this lemma is a direct consequence of Theorem 11 in [5].

We now analyze how the stochastic system can exit from variant control sets. The following propositions show how the continuity results for exit boundaries of control sets (see section 6) can be translated to the stochastic situation.

Proposition 3.5. *Suppose the family of Markov diffusion systems $(5)^\rho$ fulfills the Lie algebra rank conditions (4) and (7) for all $\rho \in [\rho_*, \rho^*]$.*

Let $D^\rho \subset M$ be a bounded variant control set of $(8)^\rho$ with nonvoid interior such that $D^{\rho_*} \subset D^\rho$, and let $x \in D^\rho$. For each ρ we define a probability measure on M via

$$Q_x(D^\rho)(A) := P_x(\omega \in \Omega,\, h_x(D^\rho)(\omega) \in A) \quad \textit{for all Borel sets } A \subset M,$$

with support $\operatorname{cl} \partial_{ex} D^\rho$. If the mapping $\rho \to \operatorname{cl} D^\rho$ is continuous in the Hausdorff distance at ρ_0 and if the perturbation range U^ρ increases lower semicontinuously at ρ_0, then the support of $Q_x(D^\rho)$ changes continuously.

Proof. Recall that $P_x(\sigma_x(D) < \infty) = 1$ for a bounded variant control set D with $x \in D$, and since all trajectories $\varphi(t,x,\omega)$ are continuous, $Q_x(D^\rho)$ is a probability measure. Equation (10) implies that $\operatorname{supp} Q_x(D^\rho) = \operatorname{cl} \partial_{ex} D^\rho$ by definition of $\partial_{ex} D^\rho$. The desired continuity follows from the deterministic situation in Theorem 6.5. ■

Finally, we study the exit locations when an invariant control set merges with a variant control set. The deterministic situation is described in Theorem 6.5.

Proposition 3.6. *Suppose the family of Markov diffusion systems* (5)$^\rho$ *fulfills the Lie algebra rank conditions* (4) *and* (7) *for all* $\rho \in [\rho_*, \rho^*]$. *For* $\rho_o \in (\rho_*, \rho^*)$ *let* C^{ρ_0} *and* D^{ρ_0} *be an invariant and a variant control set, respectively, satisfying the conditions of Theorem* 6.6.

Then for the stochastic system (5)$^{\rho_0}$ *we have for the first entrance time* $\tau_x(C^{\rho_0})$ *to the set* C^{ρ_0} *that the probability* $p_x := P_x(\tau_x(C^{\rho_0}) < \infty) < 1$ *for* $x \in D^{\rho_0}$. *By*

$$Q_{x \not\to C^{\rho_0}}(D^\rho)(A) := \tfrac{1}{1-p_x} P_x\big(\omega \in \Omega, \quad h_x(D^\rho) \in A \text{ and } \tau_x(C^{\rho_0}) = \infty\big) \\ \text{for all Borel sets } A \subset M$$

a probability measure is defined on M *with support* $\operatorname{cl} \partial^{ex \not\to C^{\rho_0}} D^\rho$. *Furthermore, for the variant control set* $F^\rho \supset C^{\rho_0} \cup D^{\rho_0}$ *we have that*

$$\operatorname{supp} Q_x(F^\rho) \to \operatorname{supp} Q_{x \not\to C^{\rho_0}}(D^{\rho_0}) \text{ for } \rho \searrow \rho_0$$

in the Hausdorff metric.

Proof. We first show that $p_x < 1$ for $x \in D^{\rho_0}$. Since it is assumed that the exit boundary of D^{ρ_0} can be nontrivially decomposed into $\partial^{ex \to C^{\rho_0}} D^{\rho_0}$ and $\partial^{ex \not\to C^{\rho_0}} D^{\rho_0}$, it follows that $\operatorname{cl} \partial^{ex \not\to C^{\rho_0}} D^{\rho_0} \neq \emptyset$. Then (10) implies $p_x < 1$.

Thus $Q_{x \not\to C^{\rho_0}}(D^{\rho_0})$ is well defined and $Q_{x \not\to C^{\rho_0}}(D^{\rho_0})(M) = 1$. As before, due to (10) and the continuity of the trajectories, $\operatorname{supp} Q_{x \not\to C^{\rho_0}}(D^{\rho_0}) = \operatorname{cl} \partial^{ex \not\to C^{\rho_0}} D^{\rho_0}$. Now the asserted right continuity follows from Theorem 6.6. ∎

4. Computation of exit times and exit locations for nearly invariant sets. In this section we present an algorithm to compute exit times of stochastic systems from sets, based on set oriented methods as they were developed for dynamical systems by Dellnitz, Hohmann, and Junge (see [10], [11]) and for control systems by Szolnoki (cf. [39]). We start from the setup in Theorem 3.3 and Corollary 3.4: For the parameter interval $[\rho_*, \rho^*]$ we assume that there is a "bifurcation point" ρ_0 such that C^{ρ_0} is an invariant control set that is contained in a variant control set C^ρ for $\rho > \rho_0$. According to Theorem 3.3, points x in the set C^{ρ_0} and in $\mathbf{A}^{inv}(C^{\rho_0}) \cap K$ of the stochastic system (5)$^{\rho_0}$ can be expected to be identified in the analysis of system (5)$^\rho$ for $\rho > \rho_0$, with $\rho - \rho_0$ small, by significantly large first exit times. However, it is impossible to analytically compute $\sigma_x(C^\rho)$ in general. We know, however, that for bounded variant C^ρ we have $P_x(\sigma_x(C^\rho) < \infty) = 1$ for all $x \in C^\rho$. For more detailed information on exit time distributions, one has to use numerical methods.

The following algorithm produces a numerical approximation to the distribution of exit times from sets in the state space. We will concentrate here on the distribution $P_x\{\sigma_x(C^\rho) \leq t\}$, $t \geq 0$, for bounded variant control sets C^ρ of the system (5)$^\rho$.

Algorithm.

Step 1. Compute the bounded variant control set $C^\rho \subset M$ of the control system (8)$^\rho$.

Step 2. Choose a compact set $K \subset M$ with $\operatorname{cl} C^\rho \subset \operatorname{int} K$ and define a partition $\mathcal{P}$ of K into finitely many boxes B_i. Define the collection $\mathcal{C} = \{B_1, B_2, \ldots, B_N\}$ of all boxes in $\mathcal{P}$ that have nonvoid intersection with C^ρ, and denote by B_{N+1} the "sink box" which symbolizes the area outside of $\bigcup_{i=1}^N B_i$. Since $C^\rho \subset \bigcup_{i=1}^N B_i$, and we are interested in the first exit time, one box suffices to cover the area of "no return."

Step 3. Choose a discretization time $T > 0$, and compute the "transition probabilities" $p_{ij} := \frac{1}{m(B_i)} \int_{B_i} P(T, y, B_j)\, dy$ for the ensuing discretized system, with $P(T, y, B_j)$ as defined in (9) for $i = 1, \ldots, N$. Here $m(\cdot)$ denotes the Lebesgue measure. We set $p_{N+1,j} = 1$ for $j = 1, \ldots, N+1$. The resulting matrix $P := (p_{ij}) \in \mathbb{R}^{(N+1)\times(N+1)}$ is row stochastic and hence the transition matrix of a certain Markov chain on the box space.

Step 4. Compute the cumulative distribution function (cdf) of the first exit time $\sigma_x(C^\rho)$ for $x \in B_i$: $P\{\sigma_x(C^\rho) \le nT\}$ is approximated by the ith entry in the last column $(p^{(n)}_{i,N+1})$ of P^n. Specifically, for a given time T_{exit} we find n_e with $(n_e - 1)T \le T_{exit} \le n_e T$, and the last column of P^{n_e} approximates the probability to exit C^ρ from B_i until time T_{exit}.

For the approximation of the control sets, numerical methods have been developed in [39] relying on subdivision techniques for the numerical analysis of dynamical systems from [10], [11]. These references also describe the generation of a partition $\mathcal{P}$ and of the boxes.

For the approximation of the dynamics of $(5)^\rho$ we have created a Markov chain on a finite box partition. After choosing a discretization time T in Step 3, the transition probabilities between the states are computed by Monte Carlo simulation. This idea is rather old and goes back to Metropolis, Ulam, and von Neumann (see [32]). Although in the meantime many sophisticated variants for different disciplines have been developed, there are no general error estimates available; hence one can never be sure that the Monte Carlo simulation recognizes all relevant behavior of the stochastic system. This is especially problematic if one wants to compute stationary measures or long time simulations of stochastic processes that visit certain areas of the state space only infrequently. There have been some developments to overcome these problems for specific systems. For instance, for systems with purely additive noise, the deterministic part and the noise influence can be decoupled, as has been done by Fischer in [17] and Fischer and Kreuzer in [18] following some work by Froyland [20]. Subsequent application of the so-called exhaustion algorithm produces some error bounds for such systems. In the algorithm described above we start from a given partition $\mathcal{P}$, a fixed discretization time T, and several starting points within each box B_i. Hence this algorithm does not follow a simulated trajectory of one initial point over a long time period, and it has proven to be quite reliable.

To approximate the dynamics of $(5)^\rho$, in Step 3 we first simulate a large number of trajectories $\hat{\eta}^l$, $l = 1, \ldots, s_1$, of the background noise process η. For this we choose initial values in the compact space N according to the stationary solution η^* (provided this is known) and approximate solutions of the stochastic differential equation (3) until time T. Strong schemes are the methods of choice for the approximation because information about the whole solution path of $(3)^\rho$ is needed for solving the x-component of $(5)^\rho$ (see Kloeden and Platen [30] for an introduction to numerical methods for stochastic differential equations).

Subsequently, s_2 starting points x^k are picked in each box B_i. From each starting point, the solution of the x-component of $(5)^\rho$ is approximated for all samples $\hat{\eta}^l$ generating $s_1 s_2$ target points, denoted by $\hat{\varphi}(T, x^k, \hat{\eta}^l)$. The transition probability from box B_i to B_j is then approximated by

$$p_{ij} = \frac{1}{m(B_i)} \int_{B_i} P(T, x, B_j)\, dx \approx \frac{1}{s_1 s_2} \sum_{k=1}^{s_2} \sum_{l=1}^{s_1} \chi_{B_j}\left(\hat{\varphi}(T, x^k, \hat{\eta}^l)\right),$$

where χ_{B_j} denotes the characteristic function of the set B_j. The question as to how many

boxes, starting points, and sample paths of the background process should be used depends on the properties of the system, the time length T, and the box size—and, of course, on the availability of computing resources. While the number of boxes $N+1$ is mainly limited by available memory (note that it is necessary to multiply full matrices with $(N+1)^2$ entries in Step 4), we have observed that the algorithm is more sensitive to a change of the noise realization than to a change of the initial values within a box. It seems that the solution trajectories $\eta_\cdot^*(\omega)$ of (3) are less smooth than the solutions of the system (2). Therefore, it is reasonable to increase the number of realizations of the background noise at the expense of initial values in each box when computing resources become an issue.

Repeated multiplication of the matrix P with itself in Step 4 may pose a problem for fine partitions, particularly in higher dimensions. When computing the cdf of the first exit time, this problem cannot be avoided. If one is interested mainly in the probability of exit until some large time T_{exit}, one can save certain iterations: Instead of performing $n_e = \frac{T_{exit}}{T}$ multiplications with P, we find $\widehat{n} = \max\{n \in N,\, 2^{\widehat{n}} \le n_e\}$ and compute $P^{2^{\widehat{n}}}$ in $\widehat{n}$ steps. If $2^{\widehat{n}} < n_e$, we continue the same process with $n_e - 2^{\widehat{n}}$, etc., until P^{n_e} is computed. (Of course, bases other than 2 can be used and sometimes lead to fewer factors in the decomposition of n_e.) For $T_{exit} = 10^4$ and $T = 10^{-2}$, this process results in 25 matrix multiplications instead of 10^6. If the cdf of the first exit time is not required in a resolution corresponding to n_e time intervals, one can proceed similarly by expressing the size of the desired resolution in powers of a prime, e.g., of 2. In our example, choosing a resolution of $10^3 T$, we compute P^{1000} with 14 multiplications, and then P^{1000k}, $k = 2, \dots, 1000$, resulting in 1013 steps.

Recall that for bounded variant control sets C^ρ the expected exit time from a point $x \in C^\rho$ is finite and given by

$$E[\sigma_x(C^\rho)] = \int_0^\infty t \, dP_\sigma,$$

where P_σ is the distribution of σ_x. This expected value can be approximated by

$$\hat{E}[\sigma_x(C^\rho)] = T \sum_{n=1}^{\infty} n \, (p_{i,N+1}^{(n)} - p_{i,N+1}^{(n-1)}) \qquad \text{for } x \in B_i.$$

For the actual computation, naturally an upper limit n_{max} on n has to be chosen, which results in an approximation of the expected exit time before $n_{max}T$.

To compute the exit locations for the system (2), we again approximate its dynamics by the Markov chain defined in Step 3. For an initial value $x \in C^\rho$ we identify the box B_i with $x \in B_i$. As before, $p_{i,j}^{(n)}$ is the probability to reach the state B_j from B_i in n steps. If $B_j \neq B_{N+1}$, and if $p_{j,N+1}^{(n+1)} > 0$, then the Markov chain exits from $\mathcal{C}$ in step $n+1$. In this case the state B_j is an exit state for the chain, starting from B_i. Let h_i denote the corresponding random exit location. We then have

$$P\{h_i = B_j\} = \sum_{n=0}^{\infty} p_{ij}^{(n)} p(j, N+1),$$

and this distribution approximates that of $h_x(C^\rho)$ as defined in section 2. In practice, again one will have to choose a maximal time $T_{exit} \in \mathbb{N}$, and the finite sum with $T_{exit}+1$ terms is computed.

5. Examples.

5.1. A perturbed escape equation. As a first example we will present some results for the perturbed escape equation. It describes the movement of a particle with unit mass in the potential $V(x) = \frac{1}{2}x^2 - \frac{1}{3}x^3$ with inertia and linear viscous damping under the influence of some perturbation. This equation has attracted great interest and has been analyzed thoroughly (see, e.g., [40], [35], or [17] and the references therein). We consider the perturbed escape equation

$$\ddot{x} + \gamma\dot{x} + x - x^2 = \rho \sin \eta_t$$

with a background noise process η_t on the one dimensional sphere $\mathbb{S}^1$. The Wiener process on this sphere is considered as the one dimensional Wiener process on $\mathbb{R}$ modulo 2π. For $t \geq 0$ and $\bar{x}, \bar{y} \in \mathbb{S}^1$ and $x, y \in \mathbb{R}$ such that $\bar{x} \equiv x \bmod 2\pi$ and $\bar{y} \equiv y \bmod 2\pi$, the transition densities of this process, resulting from the corresponding normally distributed process on $\mathbb{R}$, are given by

$$p(t, \bar{x}, \bar{y}) = \frac{1}{\sqrt{2\pi t}} \sum_{n=-\infty}^{\infty} \exp\left(-\frac{(y - x + 2n\pi)^2}{2t}\right).$$

The sum on the right-hand side converges uniformly and absolutely. Then, for an integrable nonnegative function $f : \mathbb{S}^1 \to \mathbb{R}$, it holds that

$$\begin{aligned} U_t f(\bar{x}) &:= \textstyle\int_{S^1} p(t, \bar{x}, \bar{y})\, f(\bar{y})\, d\bar{y} \\ &= \tfrac{1}{\sqrt{2\pi t}} \textstyle\int_0^{2\pi} \left(\sum_{n=-\infty}^{\infty} \exp\left(-\frac{(y-x+2n\pi)^2}{2t}\right)\right) f(y)\, dy \\ &= \tfrac{1}{\sqrt{2\pi t}} \textstyle\int_{-\infty}^{\infty} \exp\left(-\frac{(y-x)^2}{2t}\right) f(y \bmod 2\pi)\, dy. \end{aligned}$$

The function $f(\bar{x}) \equiv \frac{1}{2\pi}$ fulfills $U_t f(\bar{x}) = f(\bar{x})$. Thus $f(\bar{x})$ is the unique stationary density of the noise process because (4) obviously holds.

The perturbed escape equation driven by this background process is given by

$$\begin{aligned} \dot{x}(t) &= y(t), \\ \dot{y}(t) &= -\gamma\, y(t) - x(t) + x(t)^2 + \rho \sin(\eta_t), \\ d\eta_t &= dW_t \bmod 2\pi. \end{aligned} \tag{14}$$

As we saw, the stationary process η_t^* has the uniform distribution on $\mathbb{S}^1$ as its one dimensional distribution.

The associated controlled version of this equation on $\mathbb{R}^2$ reads

$$\begin{pmatrix} \dot{x}(t) \\ \dot{y}(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ -\gamma\, y(t) - x(t) + x(t)^2 \end{pmatrix} + \begin{pmatrix} 0 \\ u(t) \end{pmatrix}, \tag{15}$$

where $u(t) \in U^\rho := [-\rho, \rho]$. For our computations we set the damping coefficient γ to 0.1. Computation of the control sets using the method described in section 4 yields for $\rho = 0.04$ the existence of one invariant control set $C^{0.04}$ that contains the stable fixed point $(0, 0)$ of the uncontrolled equation and one variant control set $D^{0.04}$ containing the hyperbolic fixed point $(1, 0)$ of the uncontrolled equation (cf. Figure 1). Increasing the control range, one finds that

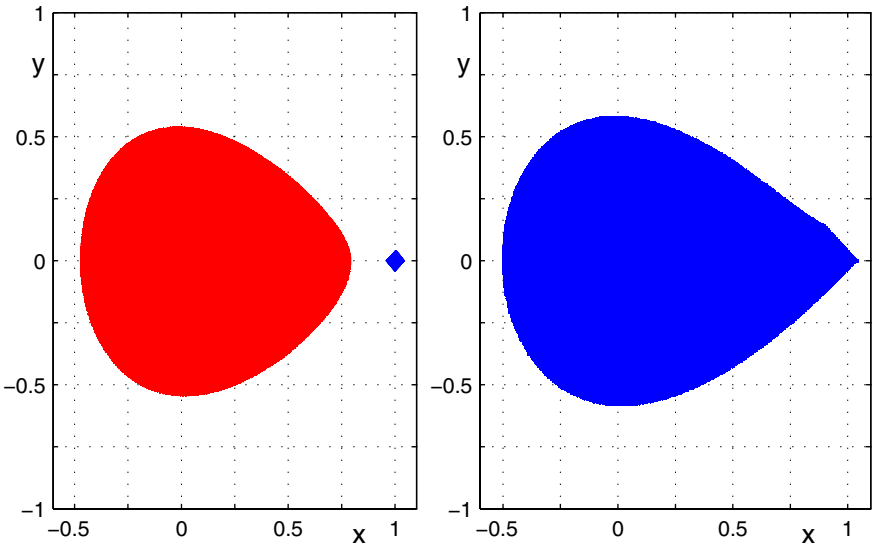


Figure 1. *Control sets for the controlled escape equation for $\rho = 0.04$ (left) and $\rho = 0.045$ (right).*

the two control sets merge for some ρ_0 close to 0.0411 (see [23]) to form one variant control set. The assumptions of Theorem 6.6 are satisfied for this example.

For the computation of the exit times from the merged control set we set $\rho = 0.15$ and distinguish two different scenarios. The first one explores the exit time distribution for a very short time, i.e., $T_{exit} \leq 1.0$. In this case we choose a fine partition of the compact set K containing $D^{0.15}$. The second one aims at long times, and we choose a coarser partition to accelerate the computation time. In both cases we pick only the center of each box as the initial value because the system (14) proves to be more sensitive to a variation in the noise sample than to a small change of the initial value.

In order to approximate the background noise process in the short time case ($T_{exit} \leq 1.0$), we choose $\hat{\eta}_0^l = l \cdot 2\pi/100$ for $l = \{0, 1, \ldots, 99\}$ as initial values to represent the uniform distribution of η_t^*. Then the background noise part of (14) is solved for each of these initial values with step size 0.1 until time 1.0, generating 100 sample paths $\hat{\eta}^l$ of the Wiener process on $\mathbb{S}^1$. For this integration, a simple Euler scheme can be used efficiently because drift and diffusion coefficients are both constant. The exit probability from a box B_i is then approximated directly by solving the (x, y)-component for each sample $\hat{\eta}^l$ starting at the center of B_i. This way, the upper left graph in Figure 2 was produced, where different colors represent different exit probabilities until time $T_{exit} = 1.0$. The other three graphs in Figure 2 follow the same procedure for $T_{exit} = 5$, 30, and 220.

To compute the distribution of the exit times $\sigma_x(D^{0.15})$, which requires large time intervals, we follow the same scheme to integrate the Wiener process, but compute more samples by starting from $\hat{\eta}_0^l = l \cdot 2\pi/10000$ for $l = \{0, 1, \ldots, 9999\}$ to compensate for the increased box sizes. Once again, the approximation of the (x, y)-component for each sample $\hat{\eta}^l$ starts at the center of B_i. Here the limiting factor for the number of boxes is the multiples of the transition matrix P that are to be computed. Multiples P^n of P are computed for $n = 2, \ldots, 1500$. The

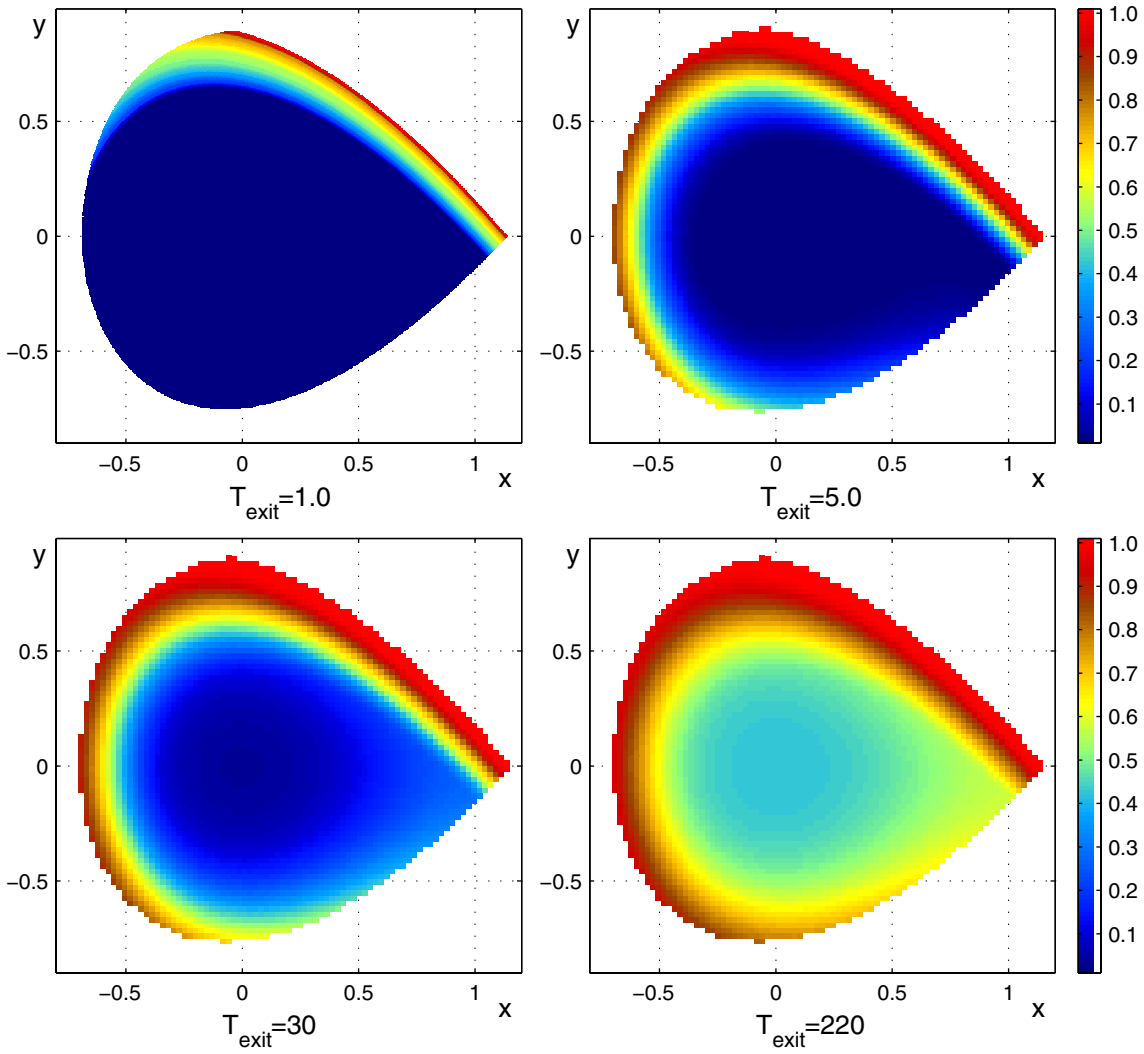


Figure 2. *Exit probabilities from* $D^{0.15}$ *for* $\rho = 0.15$ *until* T_{exit}.

minimum over all boxes of the exit probabilities $\min_i p_{i,N+1}^{(1500)}$ until $T_{exit} = 1500$ is then 0.98, and the computation is terminated. The left-hand graph in Figure 3 shows the distribution of the exit probability until time $n = 1500$ for the initial value $(0, 0)$, and Figure 4 shows the distribution for the initial values $(0.0, -0.5)$ and $(0.9, -0.1)$, now on a logarithmic scale. Both graphs show an exponential tail for the exit time distribution. Indeed, these numerically computed distributions (after some oscillations during the initial settling-in period) closely resemble a three-parameter Weibull distribution, which is the standard model for lifetime distributions in reliability theory. The oscillations stem from the deterministic dynamics of system (14). Computing an unperturbed solution that starts not too far away from $(0, 0)$ on the positive x-axis, one obtains a time of roughly 6.5 before the trajectory intersects the positive x-axis again. This is exactly the average distance between two maxima in the

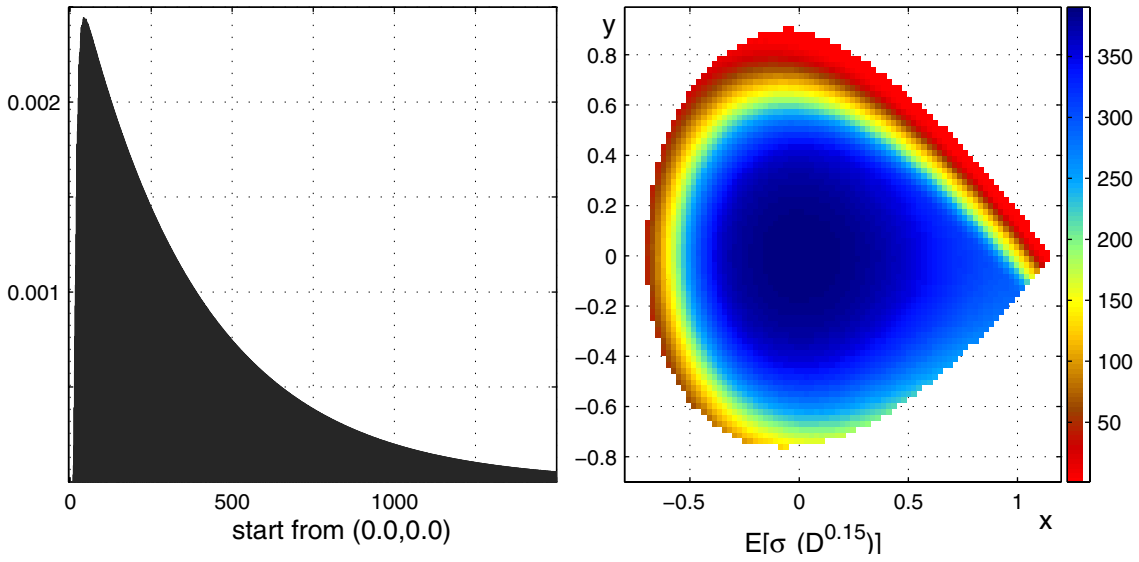


Figure 3. *Exit time distribution and expected exit times until time* 1500.

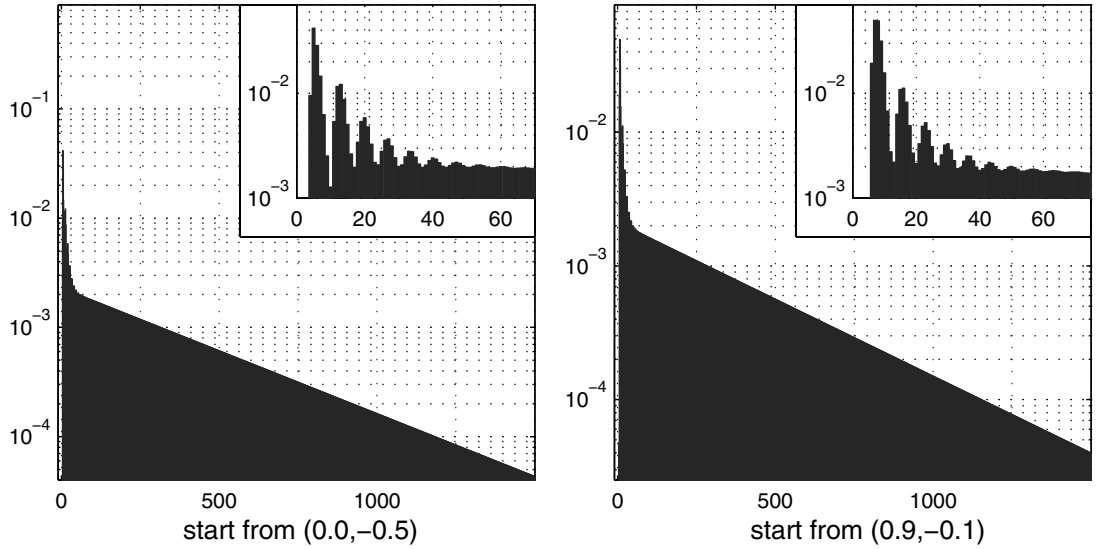


Figure 4. *Exit time distribution starting from* $(0.0, -0.5)$ *and* $(0.9, -0.1)$.

histograms of the distributions. The right-hand graph in Figure 3 shows the expected value of the exit time from all boxes in $D^{0.15}$. These expected times reflect the separation between long sojourn times in the formerly invariant region and short ones outside this area; compare with Figure 1.

5.2. A system with perturbed double well potential. Next we investigate a particle in a two-well potential and consider the following equation:

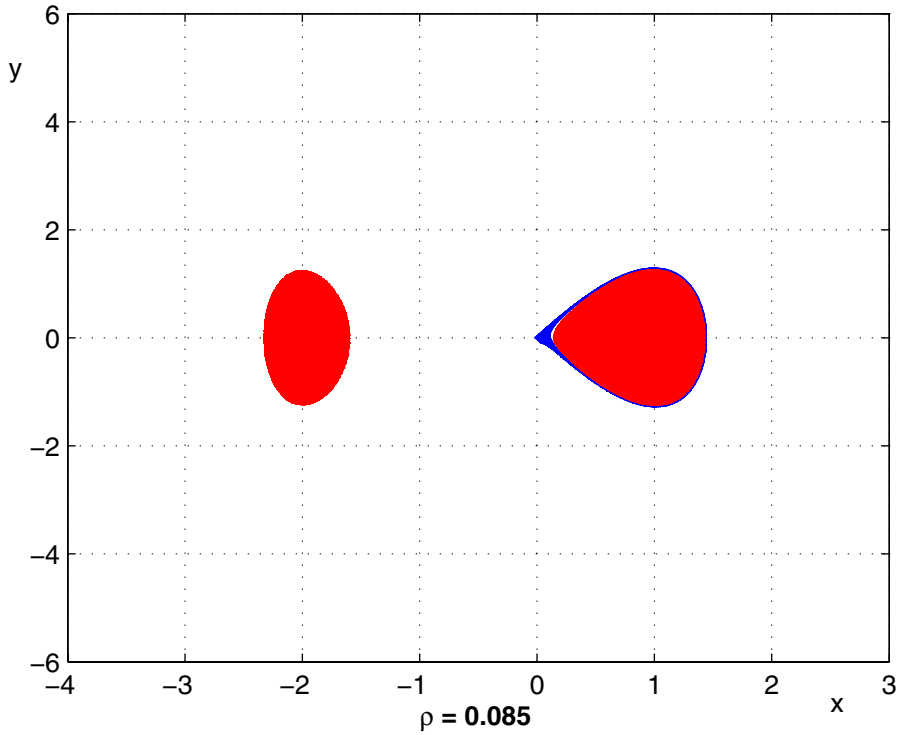


Figure 5. *Control sets for the double well potential at* $\rho = 0.085$.

$$
\begin{aligned}
\dot{x}(t) &= y(t),\\
\dot{y}(t) &= -\gamma\, y(t) - x(t)\,(2x^2(t) + 2\,x(t) - 4) + \rho \sin(\eta_t),\\
d\eta_t &= dW_t \bmod 2\pi,
\end{aligned}
\tag{16}
$$

with associated control system

$$
\begin{pmatrix} \dot{x}(t) \\ \dot{y}(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ -\gamma\, y(t) - x(t)\,(2x^2(t) + 2\,x(t) - 4) \end{pmatrix} + \begin{pmatrix} 0 \\ u(t) \end{pmatrix}, \tag{17}
$$

where again $u(t) \in U^\rho := [-\rho, \rho]$ and the damping coefficient γ is set to 0.1. For $\rho = 0.07$ there are two invariant control sets $C_1^{0.07}$ and $C_2^{0.07}$ that contain the stable fixed points $(1, 0)$ and $(-2, 0)$, respectively, of the uncontrolled equation and one variant control set $D^{0.07}$ containing the hyperbolic fixed point $(0, 0)$ of the uncontrolled equation. Increasing the control range, one finds that the control sets $C_1^{\rho_0}$ and D^{ρ_0} merge for some ρ_0 close to 0.085 and form one variant control set (see Figure 5). Note that before the merger of the control sets, the variant control set increases discontinuously and forms a ring around the invariant control set.

At some ρ_1 close to $\rho = 0.2$ the remaining control sets $C_2^{\rho_1}$ and D^{ρ_1} merge in a similar way (see Figures 6, 7, and 8).

Thus the corresponding stochastic system (16) possesses one nearly invariant region $C_1^{\rho_0}$ and one nearly invariant region $C_2^{\rho_1}$. Figure 9 shows the exit probabilities until the given exit

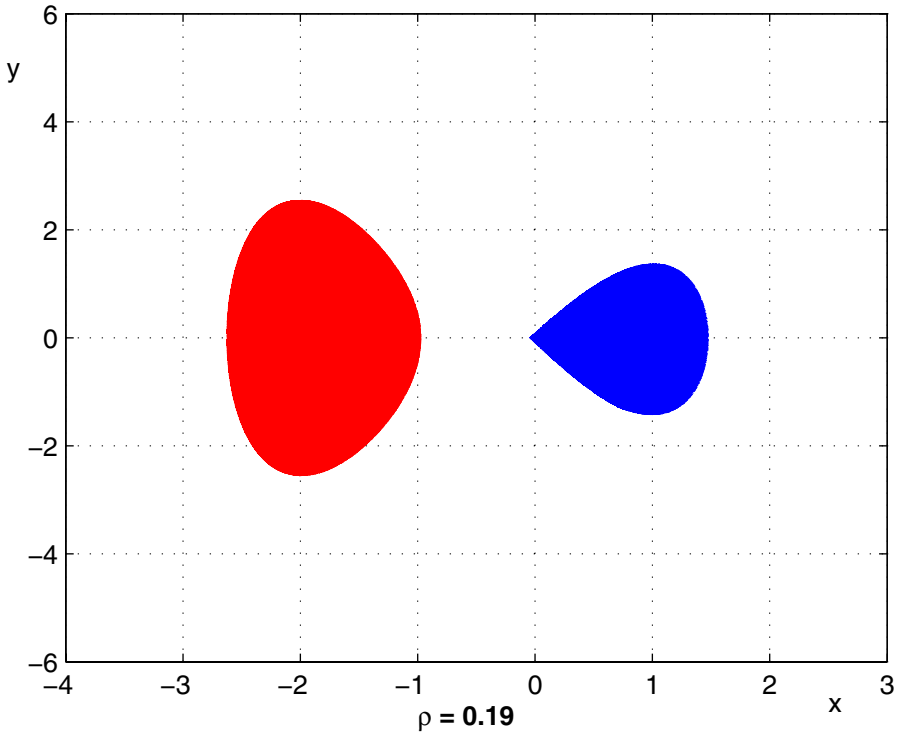


Figure 6. *Control sets for the double well potential at* $\rho = 0.19$.

times from the colored subsets for $\rho = 0.4$. Again, a comparison of the regions of large exit time in Figure 9 with the invariant control sets $C_1^{\rho_0}$ in Figure 5 and $C_2^{\rho_1}$ in Figure 7 show remarkable agreement. Also the invariant domains of attraction of the control sets become visible in Figure 9 as regions, whose exit times are rather large.

5.3. The escape equation with periodic excitation. Our third example is a perturbed escape equation with a periodic excitation and the same noise process as above. The standard way of removing the periodic time dependence leads to a three dimensional system which we analyze via its Poincaré sections. Specifically, we consider

$$\begin{aligned}
\dot{x}(t) &= y(t),\\
\dot{y}(t) &= -\gamma y(t) - x(t) - x(t)^2 + F\sin z(t) + \rho\sin\eta_t,\\
\dot{z}(t) &= \omega \bmod 2\pi,\\
d\eta_t &= dW_t \bmod 2\pi
\end{aligned}$$

with parameters

$$F = 0.06,\ \omega = 0.85,\ \gamma = 0.1, \text{ and } \rho = 0.02. \tag{18}$$

The somewhat involved control set structure of the associated control system has been studied in detail in [26]. For $\rho = 0.0$ there are two orbitally stable periodic solutions and two hyperbolic

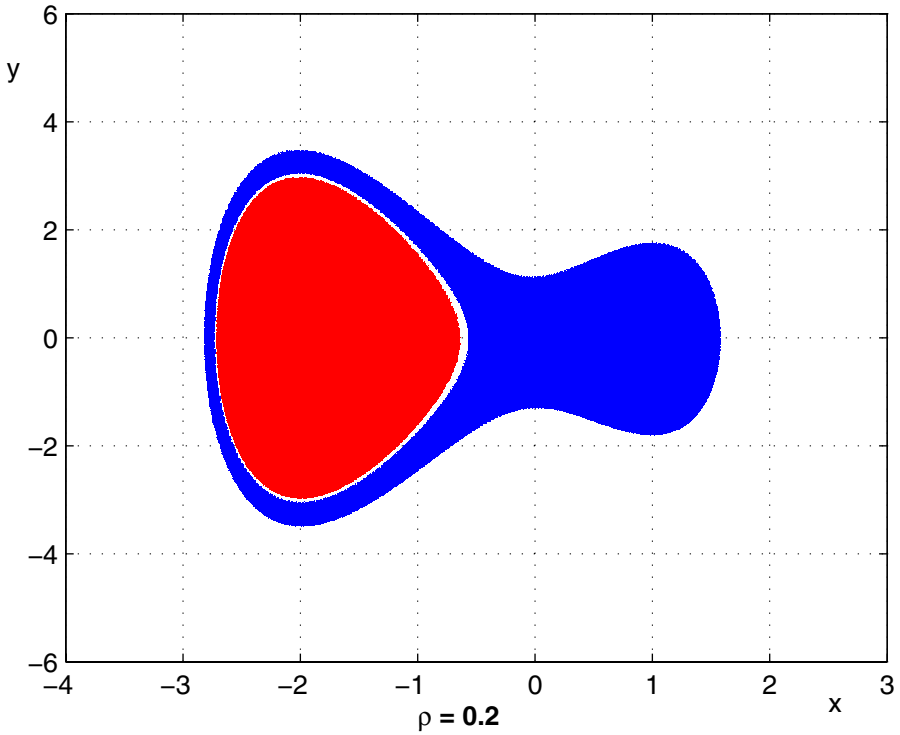


Figure 7. *Control sets for the double well potential at* $\rho = 0.2$.

periodic solutions. For small amplitude, e.g., for $\rho = 0.005$, they are included in the interior of control sets $D_1^{0.005}, D_2^{0.005}, D_3^{0.005}, D_4^{0.005}$. Figure 10 shows a slice at the phase π/ω; the potential hill top is to the right. Here $D_1^{0.005}$ and $D_3^{0.005}$ (in red) are invariant control sets, while $D_2^{0.005}$ (in the potential well) and $D_4^{0.005}$ (on the potential hill top) are variant control sets. The white regions around $D_1^{0.005}$ and $D_3^{0.005}$ show their domains of attraction $\mathbf{A}(D_1^{0.005})$ and $\mathbf{A}(D_3^{0.005})$, respectively.

For $\rho = 0.0085$, the two control sets $D_1^{0.005}$ and $D_2^{0.005}$ have merged into a variant control set $D_{12}^{0.0085}$, while $D_3^{0.0085}$ and $D_4^{0.0085}$ remain distinct. For $\rho = 0.01$ also, the control sets $D_{12}^{0.0085}$ and the invariant control set $D_3^{0.0085}$ have merged into an invariant control set $D_{123}^{0.01}$, and, finally, for $\rho \geq 0.013$ also, the control set $D_4^{0.01}$ has merged with $D_{123}^{0.01}$ forming a variant control set D_{1234}^{ρ}. In this latter situation, no invariance properties prevail.

We remark that the results presented in [26] have to be slightly modified: For the periodic control $u(t) = 0.0064 \sin \omega t, t \in \mathbb{R}$, there is a hill top periodic solution which, for $\rho > 0.0064$, is contained in the interior of D_4^{ρ}. Numerical results show that its stable and unstable manifolds have transversal intersections. Hence, for these ρ-values, there exists a homoclinic orbit which is also contained in the control set D_4^{ρ} (compare with [9]).

Figures 11–13 show, for Poincaré sections at the phase π/ω, the exit probabilities from initial points in the control set $D_{1234}^{0.2}$ for different exit times (note that the color coding differs). One sees, as expected, that exit is highly probable from a first area above the hill top. It is also probable from an area below the hill top. Here, in fact, an intersection point of the

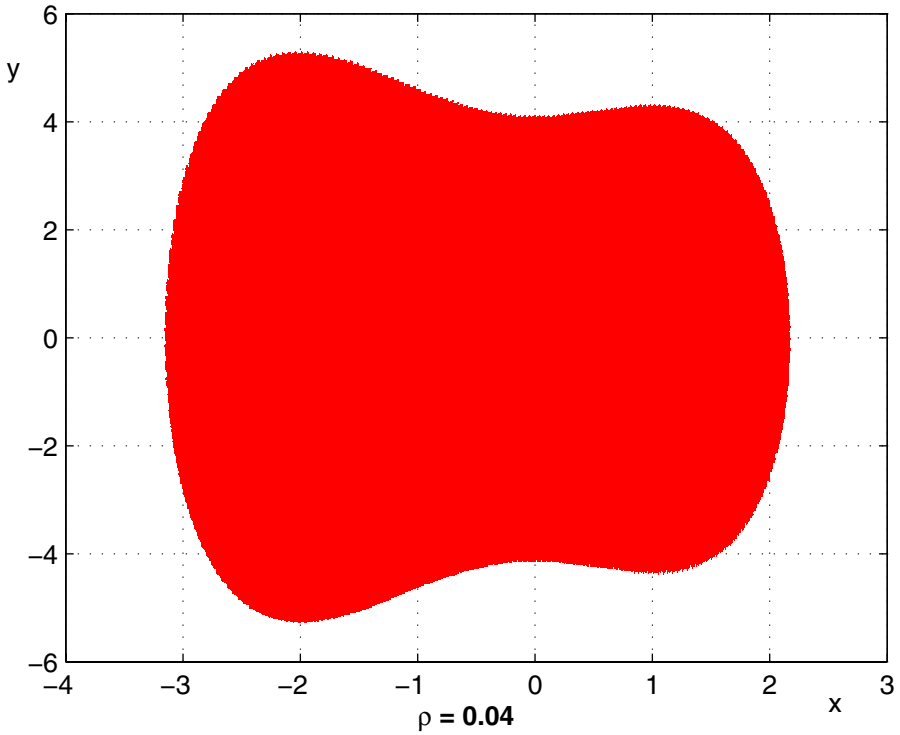


Figure 8. *Control sets for the double well potential at* $\rho = 0.4$.

stable and the unstable manifolds of the hill top periodic solution lies, and one iteration of the Poincaré map leads into the first area.

One also notes remarkable differences of exit probabilities from other areas of the control set. This is explored in more detail in Figures 14–16, which show slices through the domain of attraction $\mathbf{A}(D_{1234}^{0.2})$ for different exit times.

In Figure 16 one can discern two areas color coded by blue and brown. A comparison to Figure 10 reveals that they correspond to the domains of attraction of the two invariant control sets D_1^ρ and D_3^ρ (before their merging). They are separated by an area which corresponds to the (variant) control set D_2^ρ in the potential well. These results illustrate that the near invariance property is still present for control range $\rho = 0.02$, which is well above the control ranges where the two invariant control sets D_1^ρ and D_3^ρ lose their invariance.

We remark that, particularly due to memory requirements as discussed above, a direct numerical analysis of the three dimensional problem would be much harder. Furthermore, Poincaré sections are convenient for visualization of the results.

6. Appendix: Some background on nonlinear control systems.

In this appendix, we recall some facts on nonlinear control systems. See, for example, [6] for more information.

6.1. Accessibility and control sets.

Consider the control-affine system (8) given by

$$\dot{x}(t) = X_0(x(t)) + \sum_{i=1}^{m} u_i(t)\, X_i(x(t)) \tag{19}$$

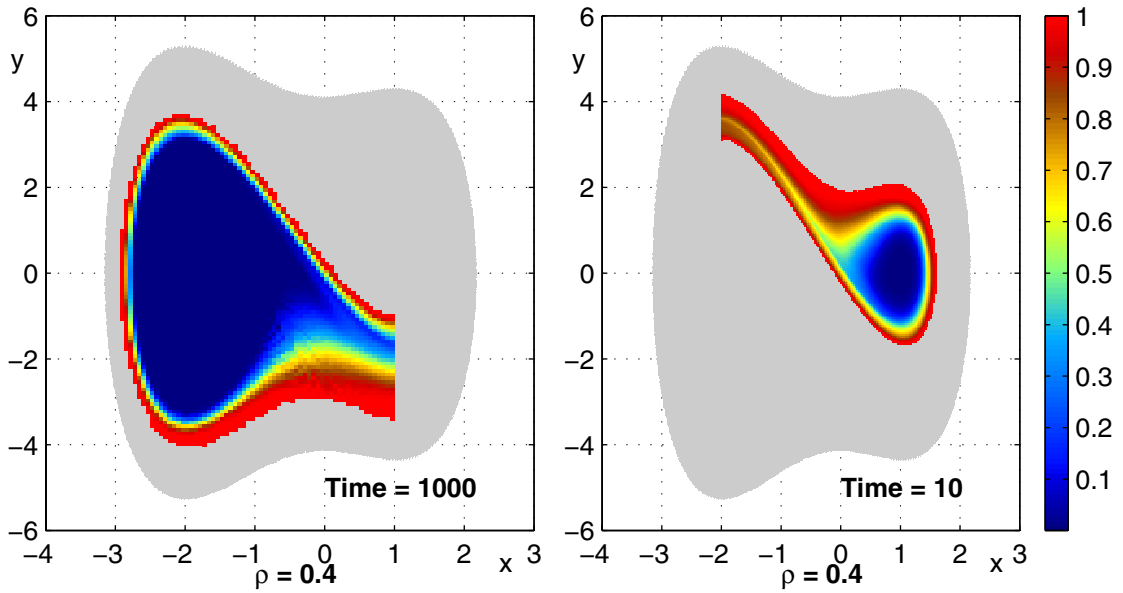


Figure 9. *Exit probabilities from the colored region around $C_1^{\rho_0}$ until time $T = 10$ (right) and from the colored region around $C_2^{\rho_0}$ until time $T = 1000$ (left) for $\rho = 0.4$. Parts of the invariant domains of attraction $A^{inv}(C_1^{\rho_0})$ and $A^{inv}(C_2^{\rho_0})$ become visible.*

with C^∞ vector fields $X_0, \dots, X_m$ on a C^∞ manifold M of dimension $d < \infty$. We obtain a family of systems by specifying an increasing family of compact convex control ranges $0 \in \operatorname{int} U^\rho \subset \mathbb{R}^m$ with $U^\rho = \operatorname{cl} \operatorname{int} U^\rho$ for all $\rho \in [\rho_*, \rho^*]$ and define corresponding sets of control functions $\mathcal{U}^\rho = \{u : \mathbb{R} \to U^\rho, \text{ measurable}\}$. Setting $u \equiv 0$ models the uncontrolled system. We assume that there exists a unique solution $\varphi(t, x, u)$ of (19) for each ρ, for every $u \in \mathcal{U}^\rho$, for every initial state $x \in M$, and for all $t \in (-\infty, \infty)$. If the dependence on ρ is not important, we will simply omit the notation of ρ in the following.

The positive and negative orbits at time $t > 0$ are

$$\mathcal{O}_t^+(x) = \{\varphi(t, x, u),\, u \in \mathcal{U}\}, \qquad \mathcal{O}_t^-(x) = \{\varphi(-t, x, u),\, u \in \mathcal{U}\},$$

and we set

$$\mathcal{O}_{\le T}^+(x) = \bigcup_{t \in [0,T]} \mathcal{O}_t^+(x), \qquad \mathcal{O}_{\le T}^-(x) = \bigcup_{t \in [0,T]} \mathcal{O}_t^-(x),$$

$$\mathcal{O}^+(x) = \bigcup_{t \in [0,\infty)} \mathcal{O}_t^+(x), \qquad \mathcal{O}^-(x) = \bigcup_{t \in [0,\infty)} \mathcal{O}_t^-(x),$$

respectively. A set $D \subset M$ with nonvoid interior is a control set if it is a maximal set with the property $D \subset \operatorname{cl} \mathcal{O}^+(x)$ for every $x \in D$. A control set C with $C = \operatorname{cl} \mathcal{O}^+(x)$ for every $x \in C$ is an invariant control set; the others are called variant. Throughout we assume that system (8) is locally accessible, i.e.,

$$\operatorname{int} \mathcal{O}_{\le T}^+(x) \neq \varnothing \text{ and } \operatorname{int} \mathcal{O}_{\le T}^-(x) \neq \varnothing \text{ for all } T > 0.$$

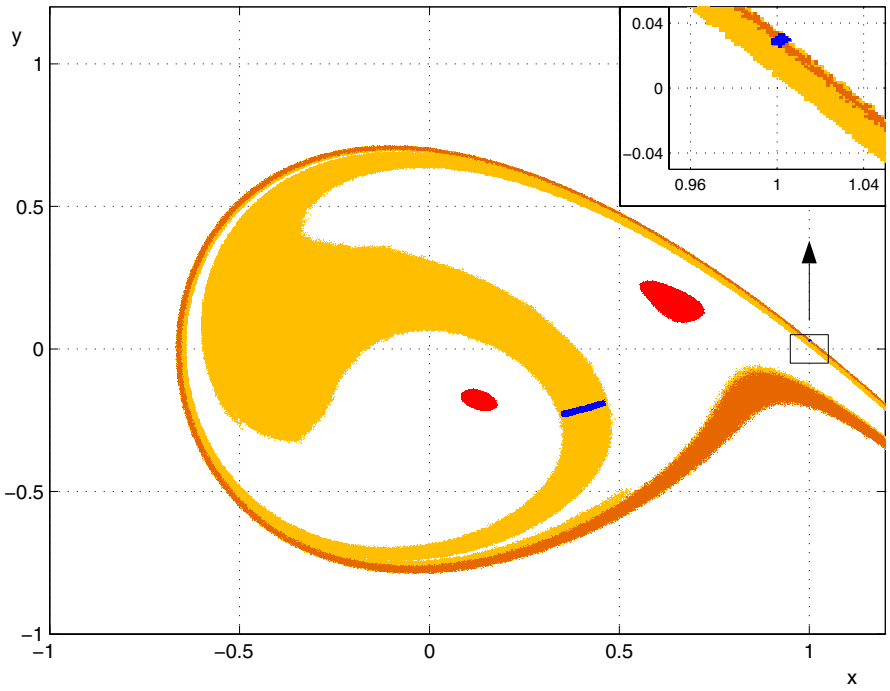


Figure 10. *Control sets and domains of attraction for* $\rho = 0.005$.

This is guaranteed by the Lie algebra rank condition $\dim \mathcal{LA}\{X_0, \ldots, X_m\}(x) = d$ for all $x \in M$. We endow the set of control functions $\mathcal{U} \subset L_\infty(\mathbb{R}, \mathbb{R}^m)$ with the weak*- (or L_1-) topology, which makes $\mathcal{U}$ a compact metric space. Then for $t_n \to t$, $x_n \to x$, and $u_n \to u$ in $\mathcal{U}$, it follows that

$$\varphi(t_n, x_n, u_n) \to \varphi(t, x, u). \tag{20}$$

We note the following lemma which states that the interior of a positively invariant set is positively invariant.

Lemma 6.1. *Suppose that* $I \subset M$ *is closed and satisfies* $\varphi(t, x, u) \in I$ *for all* $t \geq 0$, $x \in I$, *and* $u \in \mathcal{U}$. *Then* $\varphi(t, x, u) \in \operatorname{int} I$ *for all* $x \in \operatorname{int} I$, $u \in \mathcal{U}$, *and* $t \geq 0$.

Proof. Suppose that there are $x \in \operatorname{int} I$, $t > 0$, and $u \in \mathcal{U}$ with $\varphi(t, x, u) \notin \operatorname{int} I$. Then $\tau := \sup\{s \in (0, t],\, \varphi(t, x, u) \in \operatorname{int} I\}$ satisfies $\varphi(\tau, x, u) \in \partial I$. Hence there is a neighborhood V of $\varphi(\tau, x, u)$ with $V \cap (M \setminus I) \neq \varnothing$. Continuous dependence on initial conditions implies that there are $y \in \operatorname{int} I$ with $\varphi(\tau, y, u) \notin I$ contradicting the positive invariance of I. ∎

Invariant control sets and hence their interiors are positively invariant. For a set $I \subset M$ with nonvoid interior the domain of attraction is

$$\mathbf{A}(I) = \left\{x \in M,\, \operatorname{cl} \mathcal{O}^+(x) \cap \operatorname{int} I \neq \varnothing\right\}.$$

Domains of attraction are open, since by local accessibility $\operatorname{cl} \mathcal{O}^+(x) = \operatorname{cl} \operatorname{int} \mathcal{O}^+(x)$. We define the invariant domain of attraction as the largest invariant set contained in $\mathbf{A}(I)$ (sometimes called its invariance kernel).

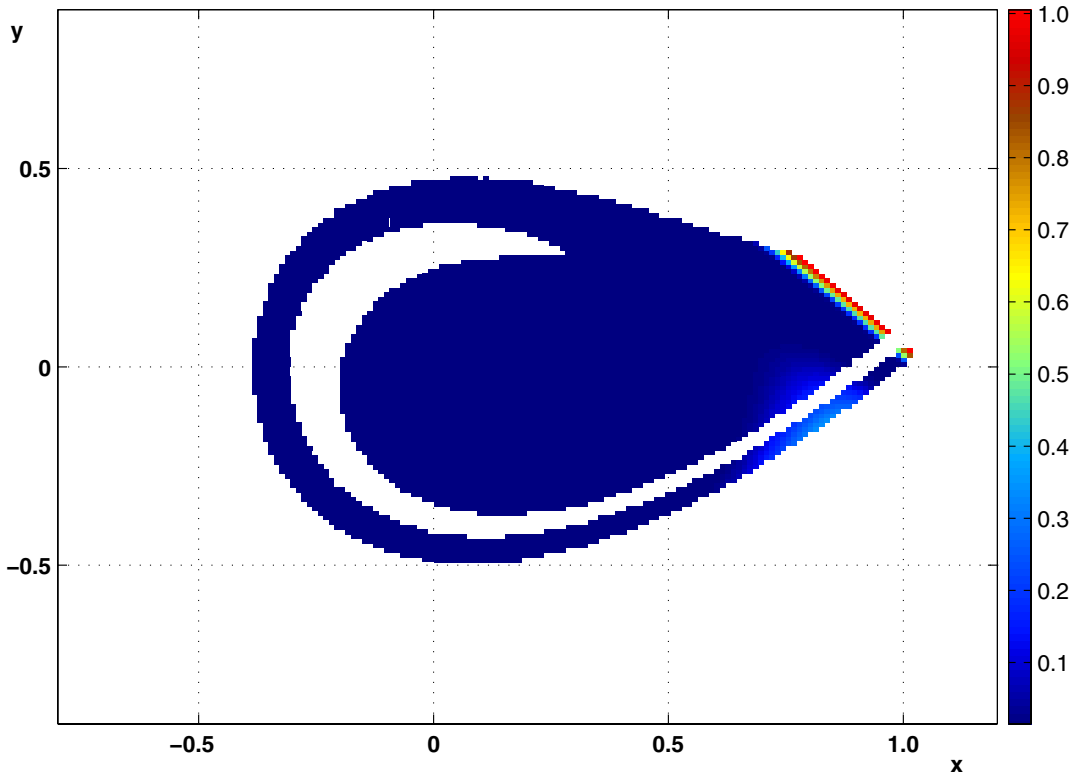


Figure 11. *Exit probabilities from the control set $D_{1234}^{0.2}$ until time $T_{exit} = 2\pi/\omega$.*

Definition 6.2. *For $I \subset M$ the invariant domain of attraction is*

$$\mathbf{A}^{inv}(I) = \{x \in \mathbf{A}(I),\, \varphi(t,x,u) \in \mathbf{A}(I) \text{ for all } u \in \mathcal{U} \text{ and } t \in \mathbb{R}_+\}.$$

This set is related to invariant control sets by the following observation.

Proposition 6.3. *Assume that $\mathbf{A}(I) \cap K$ is positively invariant for a compact set K. Then*

$$\mathbf{A}^{inv}(I) \cap K = \left\{ x \in \mathbf{A}(I) \cap K, \begin{array}{c} \text{if } C \subset \operatorname{cl} \mathcal{O}^+(x) \text{ is an invariant} \\ \text{control set, then } C \cap \operatorname{int} I \neq \varnothing \end{array} \right\}, \tag{21}$$

and this set is compact. Furthermore, $\operatorname{int}[\mathbf{A}^{inv}(I) \cap K]$ *is positively invariant.*

Proof. Let $x \in \mathbf{A}^{inv}(I) \cap K$ and suppose that $C \subset \operatorname{cl} \mathcal{O}^+(x)$ is an invariant control set. Then $\operatorname{int} C \subset \mathcal{O}^+(x)$. If $C \cap \operatorname{int} I = \varnothing$, invariance of $\operatorname{int} C$ implies that we can find $y \in C \cap \mathcal{O}^+(x)$, which is not in $\mathbf{A}(I)$, contradicting $x \in \mathbf{A}^{inv}(I)$. For the converse, let $x \in \mathbf{A}(I) \cap K$ be in the set on the right-hand side of (21). Consider $\varphi(t,x,u)$ with $u \in \mathcal{U}$ and $t \in \mathbb{R}_+$. Then by [6, Theorem 3.2.8] there is an invariant control set $C \subset \operatorname{cl} \mathcal{O}^+(x) \cap K$. Then $C \cap \operatorname{int} I \neq \varnothing$ and it follows that $\varphi(t,x,u) \in \mathbf{A}(I)$, and hence $x \in \mathbf{A}^{inv}(I) \cap K$. This proves the other inclusion. In order to see closedness, let $x_n \in \mathbf{A}^{inv}(I) \cap K$ with $x_n \to x$. Then $x \in K$ and, again by [6, Theorem 3.2.8], there is an invariant control set $C \subset \operatorname{cl} \mathcal{O}^+(x) \cap K$. We find $T > 0$ and

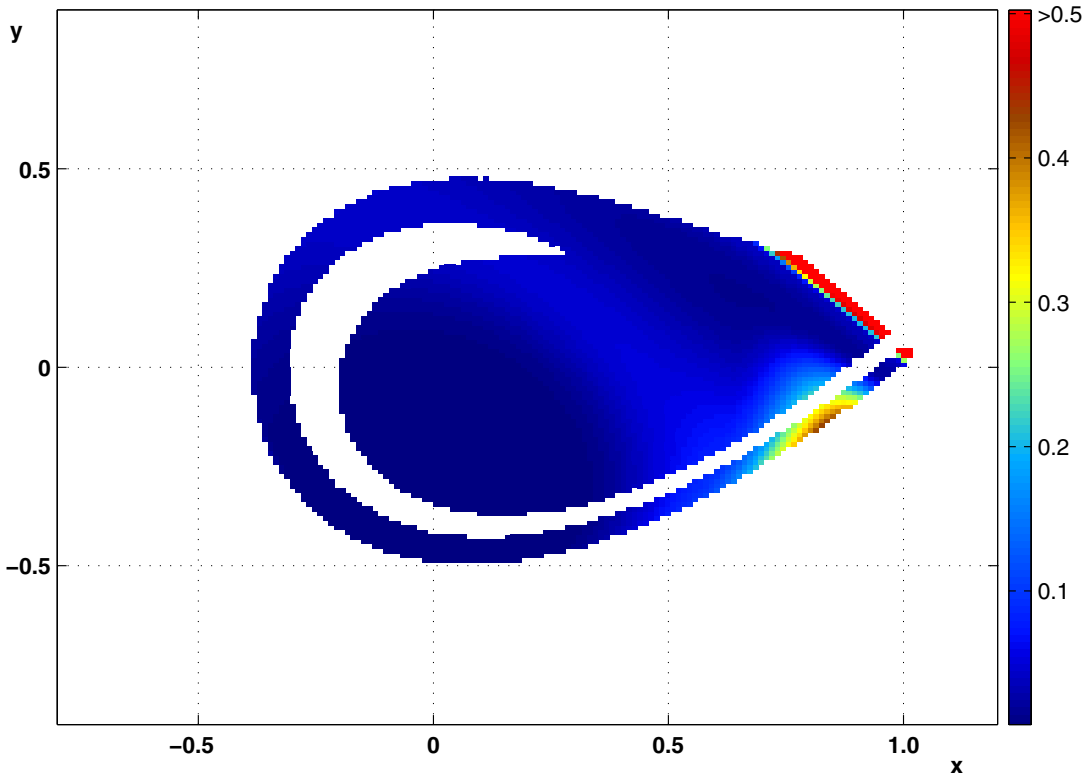


Figure 12. *Exit probabilities from the control set* $D^{0.2}_{1234}$ *until time* $T_{exit} = 10 * 2\pi/\omega$.

$u \in \mathcal{U}$ with $\varphi(T, x, u) \in \operatorname{int} C$. Then for n large enough, also $\varphi(T, x_n, u) \in \operatorname{int} C$ and hence $C \subset \operatorname{cl} \mathcal{O}^+(x_n)$. Now (21) implies $C \cap \operatorname{int} I \neq \varnothing$ and $x \in \mathbf{A}^{inv}(I) \cap K$ follows. Invariance of the interior follows by Lemma 6.1. ■

Note also that every invariant control set C satisfies $C \subset \mathbf{A}^{inv}(C)$, but not necessarily $C \subset \operatorname{int} \mathbf{A}^{inv}(C)$.

6.2. Parameter dependent control systems. In this section we describe the behavior of control sets under perturbations of the control range. Here, in addition to control sets, also chain control sets are needed. A nonvoid set $E \subset M$ is a chain control set for (19) if it is a maximal set such that for all $x \in E$ there is a control $u \in \mathcal{U}$ with $\varphi(t, x, u) \in E$ for all $t \in \mathbb{R}$, and for every $\varepsilon > 0, T > 0$ any two points $x, y \in E$ can be connected by controlled (ε, T)-chains; i.e., there are

$$n \in \mathbb{N},\ x_0 = x, \ldots, x_n = y,\ u_0, \ldots, u_{n-1} \in \mathcal{U}, \text{ and } T_0, \ldots, T_{n-1} > T$$

with

$$\mathrm{d}(\varphi(T_i, x_i, u_i), x_{i+1}) < \varepsilon \text{ for all } i = 0, \ldots, n-1.$$

For a given interval $[\rho_*, \rho^*]$ of parameters, we denote by $(19)^\rho$ the corresponding control system with control range U^ρ, $\rho \in [\rho_*, \rho^*]$. For every control set D^{ρ_*} and every chain control set E^{ρ_*}

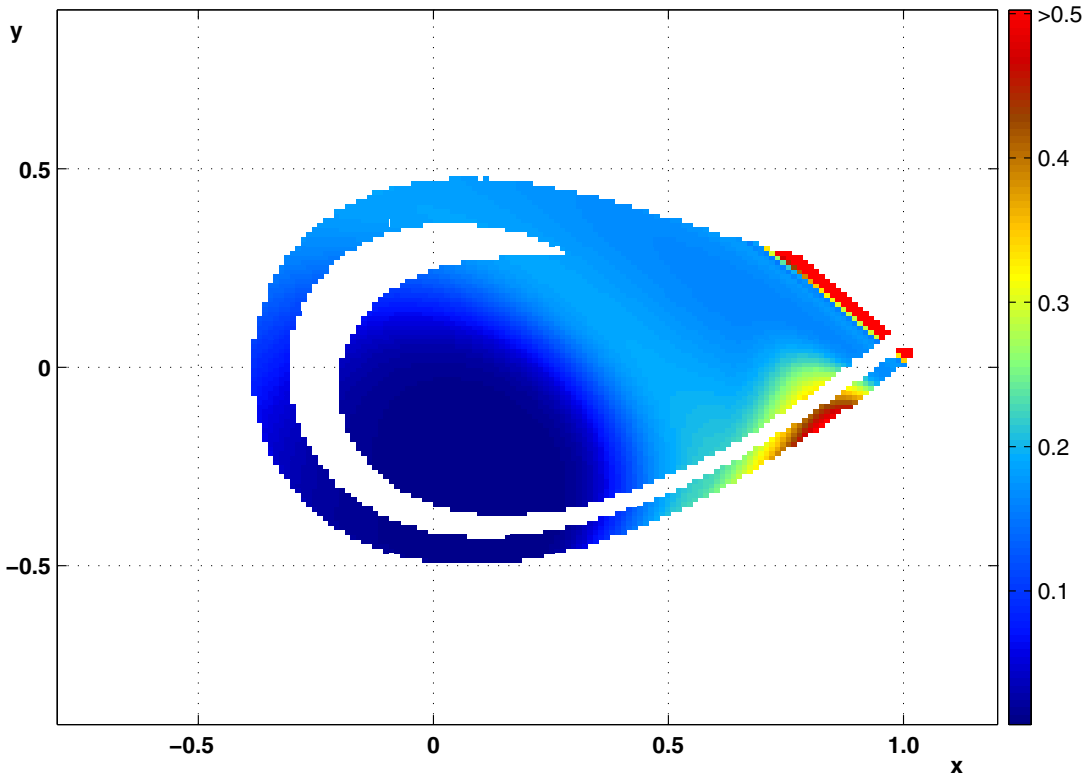


Figure 13. *Exit probabilities from the control set* $D_{1234}^{0.2}$ *until time* $T_{exit} = 100 * 2\pi/\omega$.

of the system $(19)^{\rho_*}$ there are unique control sets D^ρ and unique chain control sets E^ρ for each $\rho \in [\rho_*, \rho^*]$ such that $D^{\rho_*} \subset D^\rho$ and $E^{\rho_*} \subset E^\rho$. If all involved sets are bounded, it is well known that the increasing compact-valued mappings $\rho \mapsto \operatorname{cl} D^\rho$ and $\rho \mapsto \operatorname{cl} E^\rho$ are continuous with respect to the Hausdorff metric at all but countably many ρ-values (Scherbina's lemma [36]).

In order to obtain stronger results on the behavior of control sets and chain control sets, the following inner-pair condition is needed. A pair $(x, u) \in M \times \mathcal{U}$ is called an *inner pair* of the control system (19) if there exists $T > 0$ such that $\phi(T, x, u) \in \operatorname{int} \mathcal{O}^+(x)$. The family of systems $(19)^\rho$ is said to satisfy the *inner-pair condition* if for all $\rho_1 < \rho_2$ each pair $(x, u) \in M \times \mathcal{U}^{\rho_1}$ is an inner pair of the ρ_2-system $(19)^{\rho_2}$. We say that a set $K \subset M$ fulfills the *no-return condition* if $x \in \mathcal{O}^+(K) \cap K^c$ implies that $\mathcal{O}^+(x) \cap K = \emptyset$, where K^c denotes the complement of K in M.

The following theorem (see [6, Lemma 4.7.3, Lemma 4.7.4, and Theorem 4.7.5]) describes the close relation between control sets and chain control sets if the inner-pair condition holds.

Theorem 6.4. *Consider the family of control-affine systems* $(19)^\rho$ *for* $\rho \in [\rho_*, \rho^*]$, *where* $\rho \mapsto U^\rho$ *is continuous with respect to the Hausdorff metric. Let* D^{ρ_*} *be a control set and* E^{ρ_*} *be a chain control set of* $(19)^{\rho_*}$ *such that* $D^{\rho_*} \subset E^{\rho_*}$. *Then for all* ρ *it holds that* $D^\rho \subset E^\rho$, *where the sets* D^ρ *and* E^ρ *are defined as above. Suppose* $E^{\rho^*} \subset K$ *for a compact set* $K \subset M$

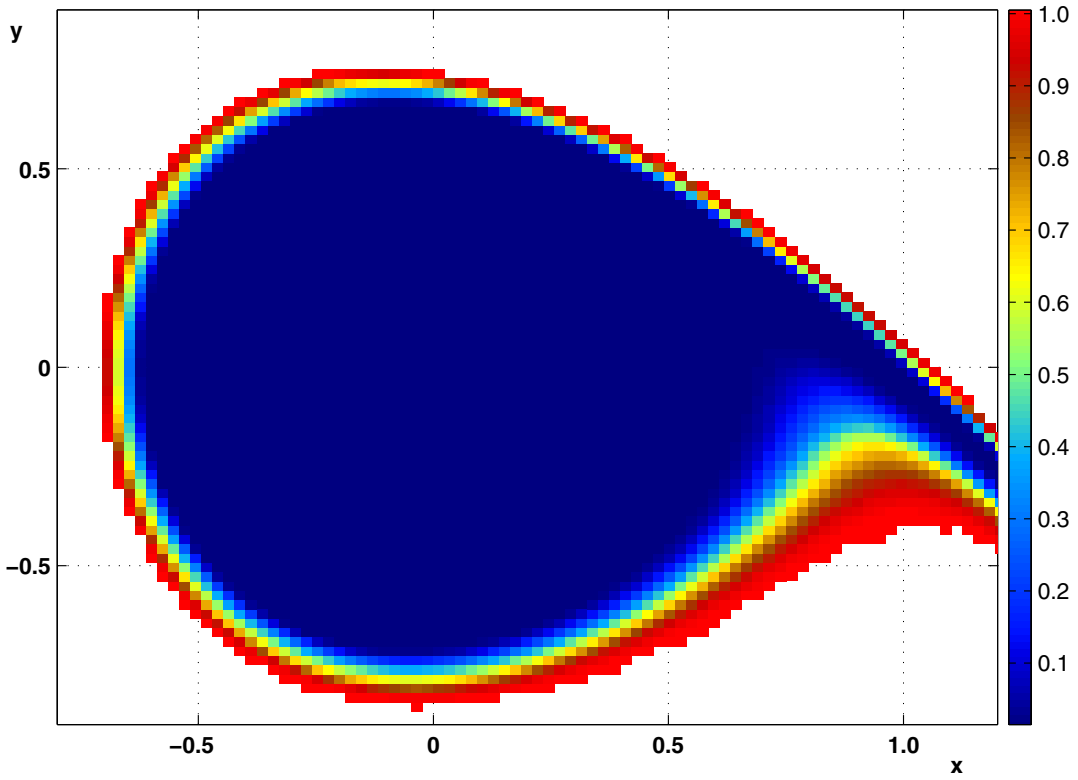


Figure 14. *Exit probabilities from the domain of attraction* $\mathbf{A}(D_{1234}^{0.2})$ *until time* $T_{exit} = 2\pi/\omega$.

that fulfills the no-return condition for the ρ^**-system, and assume that the family* (19)$^\rho$ *satisfies the inner-pair condition in* $[\rho_*, \rho^*]$.

Then for $\rho_1 < \rho_2$ *in* $(\rho_*, \rho^*]$ *it holds that* $\operatorname{cl} D^{\rho_1} \subset E^{\rho_1} \subset \operatorname{int} D^{\rho_2}$, *and for all up to at most countably many* ρ*-values, the equation* $\operatorname{cl} D^\rho = E^\rho$ *is satisfied. The map* $(\rho_*, \rho^*) \to \mathcal{C}(K) : \rho \mapsto \operatorname{cl} D^\rho$ *is continuous at* ρ *if and only if* $\operatorname{cl} D^\rho = E^\rho$*; the same is true for the map* $\rho \mapsto E^\rho$. *Here* $\mathcal{C}(K)$ *denotes the space of compact subsets of* K.

In [24] it is shown that the inner-pair condition holds for an important class of systems that includes, in particular, the escape equation (15) and the double well equation (17).

We also need some results on the boundaries of control sets D. Define the entrance and exit boundaries by

$$\begin{aligned} \partial^{ex} D &:= \{x \in \partial D \mid \text{there is } y \in \operatorname{int} D \text{ such that } x \in \mathcal{O}^+(y)\}, \\ \partial^{en} D &:= \{x \in \partial D \mid \text{there is } y \in \operatorname{int} D \text{ such that } y \in \mathcal{O}^+(x)\}, \end{aligned} \tag{22}$$

and the tangential boundary $\partial^{tg} D := \partial D \setminus (\partial^{ex} D \cup \partial^{en} D)$. The sets $\partial^{ex} D$ and $\partial^{en} D$ are disjoint and open in ∂D, and $\partial^{tg} D$ is closed in ∂D. Furthermore, $\partial^{tg} D = \operatorname{cl} \partial^{ex} D \cap \operatorname{cl} \partial^{en} D$ and $\operatorname{int}_{\partial D} \partial^{tg} D = \emptyset$. The following theorem from [24] shows that exit and entrance boundaries change continuously if the control range U^ρ increases lower semicontinuously and if the control

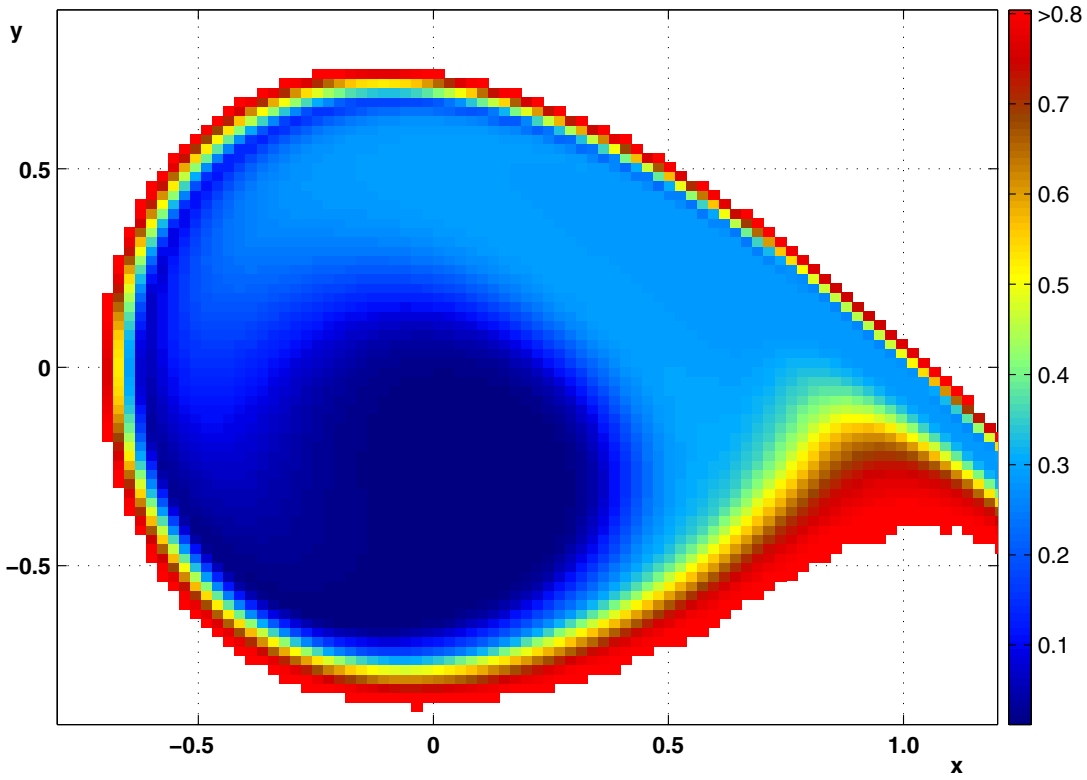


Figure 15. *Exit probabilities from the domain of attraction* $\mathbf{A}(D_{1234}^{0.2})$ *until time* $T_{exit} = 254 * 2\pi/\omega$.

sets themselves change continuously.

Theorem 6.5. *Consider the set-valued mapping* $[\rho_*, \rho^*] \to \mathcal{C}(M)$, $\rho \mapsto \operatorname{cl} D^\rho$, *as in the previous theorem, where now* D^{ρ_*} *is a control set of* $(19)^{\rho_*}$ *and* D^ρ *denotes the unique control set of* $(19)^\rho$ *with* $D^{\rho_*} \subset D^\rho$. *If this map is continuous in the Hausdorff distance at* $\rho_0 \in (\rho_*, \rho^*)$, D^{ρ^*} *is bounded, and if the control range* U^ρ *increases lower semicontinuously at* ρ_0, *then the mappings* $\rho \mapsto \partial D^\rho$, $\rho \mapsto \operatorname{cl} \partial^{ex} D^\rho$, *and* $\rho \mapsto \operatorname{cl} \partial^{en} D^\rho$ *are continuous in the Hausdorff distance at* ρ_0.

Next we will examine more closely how an invariant control set C loses its invariance when merging with a variant control set D while the control range U^ρ is increased. For this we introduce two further specifications of exit boundaries: the part from where under all admissible controls exactly one invariant control set C can be reached, and the part from where C cannot be reached at all. We denote the first set by

$$\partial^{ex \to C} D := \left\{ \begin{array}{c} x \in \partial^{ex} D \mid \mathcal{O}^+(x) \text{ bounded, and if for some invariant} \\ \text{control set } C' \subset M \text{ we have } C' \cap \mathcal{O}^+(x) \neq \emptyset, \text{ then } C = C' \end{array} \right\}$$

and the second one by

$$\partial^{ex \not\to C} D := \{x \in \partial^{ex} D \mid \mathcal{O}^+(x) \cap C = \emptyset\}.$$

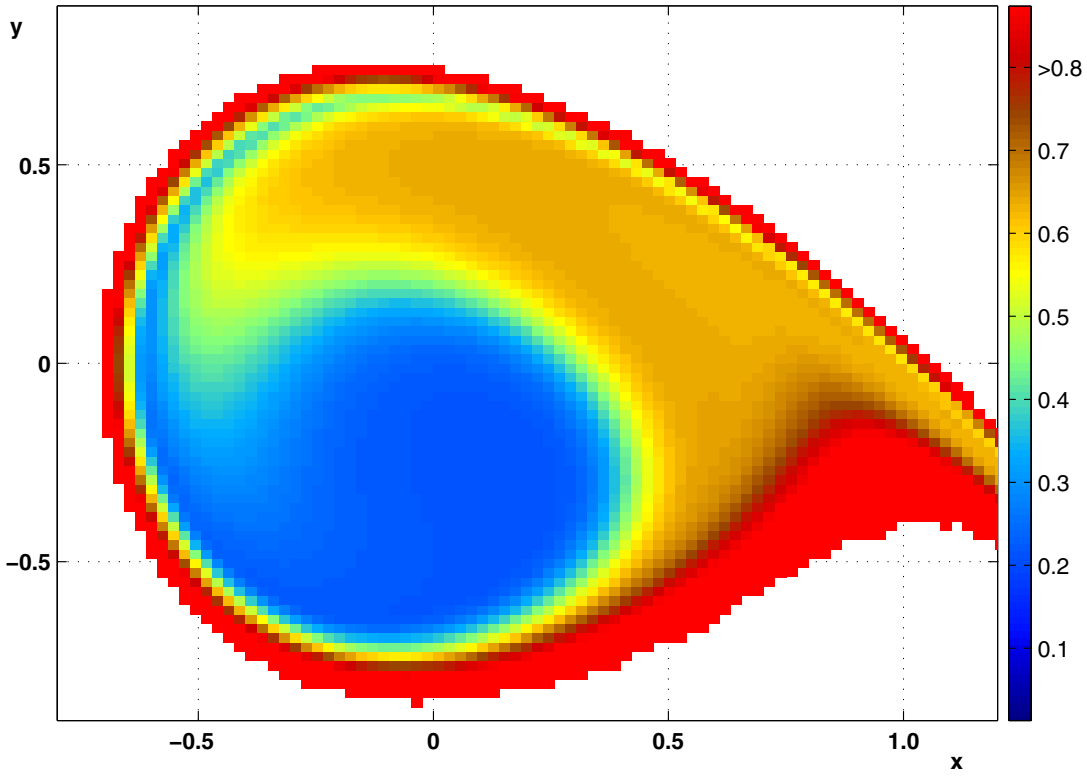


Figure 16. *Exit probabilities from the domain of attraction* $\mathbf{A}(D_{1234}^{0.2})$ *until time* $T_{exit} = 1000 * 2\pi/\omega$.

Note that from [6, Theorem 3.2.8] it follows that $\mathcal{O}^+(x) \subset \mathcal{O}^-(C) \cap \mathcal{O}^+(D)$ for all $x \in \partial^{ex \to C} D$.

If the exit boundary of D^{ρ_0} can be decomposed into $\partial^{ex \to C^{\rho_0}} D^{\rho_0}$ and $\partial^{ex \not\to C^{\rho_0}} D^{\rho_0}$, then the exit boundary of the merged set is continuous in the following sense [24].

Theorem 6.6. *Let* $K \subset M$ *be a compact set such that all control sets of the control systems* $(19)^\rho$ *have void intersection with the boundary of* K. *Assume that system* $(19)^{\rho_0}$ *has precisely one invariant control set* $C^{\rho_0} \subset K$ *and one variant control set* $D^{\rho_0} \subset K$ *such that* $C^{\rho_0} \cap \operatorname{cl} D^{\rho_0} \neq \emptyset$. *For each* $\rho > \rho_0$ *let there be precisely one variant control set* $F^\rho \subset K$ *of* $(19)^\rho$ *and* $C^{\rho_0} \cup D^{\rho_0} \subset F^\rho$. *Suppose that* $\operatorname{cl} F^\rho$ *are chain control sets of* $(19)^\rho$ *for each* $\rho > \rho_0$ *and* $\operatorname{cl}(\mathcal{O}^{\rho_0,-}(C^{\rho_0}) \cap \mathcal{O}^{\rho,+}(D^{\rho_0}))$ *is a chain control set of* $(19)^{\rho_0}$. *Finally, assume that* U^ρ *depends continuously on* ρ *with respect to the Hausdorff metric at* ρ_0 *and let* $\delta^{ex \to C^{\rho_0}} D^{\rho_0}$ *and* $\delta^{ex \not\to C^{\rho_0}} D^{\rho_0}$ *be a nontrivial decomposition of* $\delta^{ex} D^{\rho_0}$.

Then $\operatorname{cl} \partial^{ex} F^\rho \to \operatorname{cl} \partial^{ex \not\to C^{\rho_0}} D^{\rho_0}$ *in the Hausdorff metric for* $\rho \searrow \rho_0$.

Acknowledgments. The algorithms used have been implemented into the MATLAB version of the program package GAIO by Junge. Thus the box handling algorithms from Junge could be used. The control sets are found using methods based on Szolnoki [39]. The necessary solvers for stochastic differential equations and the routines for the computation of the transition matrix were added into the GAIO structure.

REFERENCES

[1] L. Arnold and W. Kliemann, *Qualitative theory of stochastic systems*, in Probabilistic Analysis and Related Topics, Vol. 3, A. T. Bharucha-Reid, ed., Academic Press, New York, 1983, pp. 1–79.

[2] A. Bovier, *Metastability and ageing in stochastic dynamics,* in Dynamics and Randomness II, A. Maas, S. Martinez, and J. San Martin, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004, pp. 17–79.

[3] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, *Metastability in reversible diffusion processes.* I. *Sharp asymptotics for capacities and exit times*, J. Eur. Math. Soc. (JEMS), 6 (2004), pp. 399–424.

[4] A. Bovier, V. Gayrard, and M. Klein, *Metastability in reversible diffusion processes.* II. *Precise asymptotics for small eigenvalues*, J. Eur. Math. Soc. (JEMS), 7 (2005), pp. 69–99.

[5] F. Colonius, G. Häckl, and W. Kliemann, *Dynamic reliability of nonlinear systems under random excitation*, in Vibration and Control of Stochastic Dynamical Systems, L. A. Bergman and B. F. Spencer, eds., ASME DE, 84 (1) (1995), pp. 1007–1024.

[6] F. Colonius and W. Kliemann, *The Dynamics of Control*, Birkhäuser Boston, Boston, 2000.

[7] F. Colonius and W. Kliemann, *Topological, smooth, and control techniques for perturbed systems*, in Stochastic Dynamics, H. Crauel and M. Gundlach, eds., Springer-Verlag, New York, 1999, pp. 181–208.

[8] F. Colonius, F. J. de la Rubia, and W. Kliemann, *Stochastic models with multistability and extinction levels*, SIAM J. Appl. Math., 56 (1996), pp. 919–945.

[9] F. Colonius, A. Marquardt, E. Kreuzer, and W. Sichermann, *A numerical study of capsizing: Comparing control set analysis and Melnikov's method*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., submitted, 2006.

[10] M. Dellnitz and A. Hohmann, *A subdivision algorithm for the computation of unstable manifolds and global attractors*, Numer. Math., 75 (1997), pp. 293–317.

[11] M. Dellnitz and O. Junge, *Set oriented numerical methods for dynamical systems*, in Handbook of Dynamical Systems III: Towards Applications, B. Fiedler, G. Iooss, and N. Kopell, eds., World Scientific, Singapore, 2002, pp. 221–264.

[12] M. Dellnitz and O. Junge, *Almost invariant sets in Chua's circuit*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 7 (1997), pp. 2475–2485.

[13] M. Dellnitz, O. Junge, W. S. Koon, F. Lekien, M. W. Lo, J. E. Marsden, K. Padberg, R. Preis, S. D. Ross, and B. Thiere, *Transport in dynamical astronomy and multibody problems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 699–727.

[14] M. Dellnitz, M. Hessel-von Molo, Ph. Metzner, R. Preis, and Ch. Schütte, *Graph algorithms for dynamical systems*, in Analysis, Modeling and Simulation of Multiscale Problems, A. Mielke, ed., Springer-Verlag, New York, 2006, pp. 619–645.

[15] P. Deuflhard and Ch. Schütte, *Molecular conformation dynamics and computational drug design*, in Applied Mathematics Entering the 21st Century, Proc. Appl. Math. 116, J. M. Hill and R. Moore, eds., SIAM, Philadelphia, 2004, pp. 91–119.

[16] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte, *Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains*, Linear Algebra Appl., 315 (2000), pp. 39–59.

[17] J. Fischer, *Cell Mapping for Randomly Perturbed Systems. Robustness Analysis of Dynamical Systems*, dissertation, Technische Universität Hamburg-Harburg, Hamburg, Germany, VDI Verlag, Düsseldorf, Germany, 2002.

[18] J. Fischer and E. Kreuzer, *Generalized cell mapping for randomly perturbed dynamical systems*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 769–777.

[19] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

[20] G. Froyland, *Ulam's method for random interval maps*, Nonlinearity, 12 (1999), pp. 1029–1052.

[21] G. Froyland, *Statistically optimal almost-invariant sets*, Phys. D, 200 (2005), pp. 205–219.

[22] G. Froyland and M. Dellnitz, *Detecting and locating near-optimal almost-invariant sets and cycles*, SIAM J. Sci. Comput., 24 (2003), pp. 1839–1863.

[23] T. Gayer, *Controlled and Perturbed Systems under Parameter Variation*, dissertation, Universität Augsburg, Augsburg, Germany, Shaker Verlag, Aachen, Germany, 2003.

[24] T. Gayer, *Control sets and their boundaries under parameter variation*, J. Differential Equations, 201 (2004), pp. 177–200.
[25] T. Gayer, *On Markov chains and the spectra of the corresponding Frobenius–Perron operator*, Stoch. Dyn., 1 (2001), pp. 477–491.
[26] T. Gayer, *Controllability and invariance properties of time-periodic systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 1361–1375.
[27] W. Huisinga, S. Meyn, and Ch. Schütte, *Phase transitions & metastability in Markovian and molecular systems*, Ann. Appl. Probab., 14 (2004), pp. 419–458.
[28] W. Kliemann, *Analysis of nonlinear stochastic systems*, in Analysis and Estimation of Stochastic Mechanical Systems, W. Schiehlen and W. Wedig, eds., Springer-Verlag, New York, 1988, pp. 43–102.
[29] W. Kliemann, *Recurrence and invariant measures for degenerate diffusions*, Ann. Probab., 15 (1987), pp. 690–707.
[30] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, 3rd ed., Springer-Verlag, New York, 1999.
[31] H. Kunita, *Supports of diffusion processes and controllability problems*, in Proceedings of the International Symposium on Stochastic Differential Equations, K. Ito, ed., Wiley, New York, 1978, pp. 163–185.
[32] N. Metropolis and S. Ulam, *The Monte Carlo method*, J. Amer. Statist. Assoc., 44 (1949), pp. 335–341.
[33] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes* II: *Continuous-time processes and sampled chains*, Adv. in Appl. Probab., 25 (1993), pp. 487–517.
[34] I. Mezic, *On the dynamics of molecular conformation*, Proc. Natl. Acad. Sci. USA, 103 (2006), pp. 7542–7547.
[35] H. E. Nusse, E. Ott, and J. A. Yorke, *Saddle-node bifurcations on fractal boundaries*, Phys. Rev. Lett., 75 (1995), pp. 2482–2485.
[36] S. Y. Pilyugin, *The Space of Dynamical Systems with C^0-Topology*, Springer-Verlag, New York, 1994.
[37] Ch. Schütte, W. Huisinga, and P. Deuflhard, *Transfer operator approach to conformational dynamics in biomolecular systems*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, B. Fiedler, ed., Springer-Verlag, New York, 2001, pp. 191–223.
[38] D. Stroock and S. Varadhan, *On the support of diffusion processes with applications to the strong maximum principle*, in Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 3, University of California Press, Berkeley, CA, 1972, pp. 333–359.
[39] D. Szolnoki, *Set oriented methods for computing reachable sets and control sets*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 361–382.
[40] J. M. T. Thompson, *Chaotic phenomena triggering the escape from a potential well*, Proc. Roy. Soc. London Ser. A, 421 (1989), pp. 195–225.
[41] H. Zmarrou and A. J. Homburg, *Bifurcations of stationary measures of random diffeomorphisms*, Ergodic Theory Dynam. Systems, to appear.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 108–133
© 2008 Society for Industrial and Applied Mathematics

Global Synchronization of Linearly Hybrid Coupled Networks with Time-Varying Delay*

Wenwu Yu†, Jinde Cao‡, and Jinhu Lü§

Abstract. Many real-world large-scale complex networks demonstrate a surprising degree of synchronization. To unravel the underlying mechanics of synchronization in these complex networks, a generally linearly hybrid coupled network with time-varying delay is proposed, and its global synchronization is then further investigated. Several effective sufficient conditions of global synchronization are attained based on the Lyapunov function and a linear matrix inequality (LMI). Both delay-independent and delay-dependent conditions are deduced. In particular, the coupling matrix may be nonsymmetric or nondiagonal. Moreover, the derivative of the time-varying delay is extended to any given value. Finally, a small-world network, a regular network, and scale-free networks with network size are constructed to show the effectiveness of the proposed synchronous criteria.

Key words. Lyapunov function, linear matrix inequality (LMI), global synchronization, time-varying delay, complex networks

AMS subject classifications. 94B50, 34C15, 34D20, 92B20

DOI. 10.1137/070679090

1. Introduction.

Over the last few years, complex networks have received increasing attention from all fields of sciences and humanities [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. Networks are everywhere in the real world, such as food-webs, ecosystems, metabolic pathways, the Internet, the World Wide Web, social networks, and global economic markets [1, 2]. The ubiquity of networks in the biological, physical, engineering, and social sciences leads naturally to two important common problems: How does network structure affect network function? How do individual dynamics affect global dynamics?

Despite advances in understanding network structure and dynamical behaviors in idealized cases, relatively little is known about large-scale, real-world complex networks and their

*Received by the editors January 3, 2007; accepted for publication (in revised form) by A. Hagberg July 12, 2007; published electronically January 16, 2008. This work was jointly supported by the National Natural Science Foundation of China under grants 60574043, 60772158, and 60221301, the 973 Program of China under grants 2003CB317004 and 2007CB310800, the Important Direction Project of Knowledge Innovation Program of the Chinese Academy of the Sciences under grant KJCX3-SYW-S01, the Natural Science Foundation of Jiangsu Province of China under grant BK2006093, and the International Joint Project funded by NSFC and the Royal Society of the United Kingdom.

http://www.siam.org/journals/siads/7-1/67909.html

†Department of Mathematics, Southeast University, Nanjing 210096, China, Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China, and Department of Electrical Engineering, Columbia University, New York, NY 10027 (wenwuyu@gmail.com, wy2137@columbia.edu).

‡Department of Mathematics, Southeast University, Nanjing 210096, China (jdcao@seu.edu.cn).

§Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China, and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544 (jhlu@iss.ac.cn, jinhulu@princeton.edu).

dynamical characteristics, especially for the evolving networks [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34]. Historically, many models were proposed to describe various complex networks, including regular graph, random graph, small-world network, scale-free network, evolving networks, etc. [1, 2]. Undoubtedly, these models well describe many real networks in nature, such as social, biological, and engineering networks.

On the other hand, one can also extend the existing network models by introducing dynamical elements into the network nodes [3, 4, 14, 32, 33, 34]. Over the last few years, nonlinear dynamics of complex networks have been intensively investigated. Synchronization is a kind of typical collective behavior and a basic motion in nature [14]. Our intuition is that loosely coupled dynamical systems tend to synchronize with respect to periodic behavior [18]. This synchronization is essentially a form of self-organization. Moreover, it has been demonstrated that many real-world problems have a close relationship with network synchronization. For example, theoretical and experimental results reveal that a mammalian brain not only displays its storage of associative memories but also modulates oscillatory neuronal synchronization by selective perceived attention [6].

Recently, network synchronization has been intensively investigated in various different fields [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34]. For example, some researchers studied the synchronization of coupled connected neural networks [6, 8, 9, 10]; Yu, Cao, and their colleagues explored the synchronization of an array of linearly coupled networks with time-delay [7, 16]; Lü and his colleagues introduced several synchronization criteria for the time-varying complex dynamical networks [16, 32, 33, 34]; and Li and Chen studied the robust adaptive synchronization of some uncertain dynamical networks [15].

In this paper, we introduce a linearly hybrid coupled network with time-varying delay. Based on this network model, several simply sufficient conditions of global network synchronization are then deduced by using the Lyapunov function and a linear matrix inequality (LMI). Both delay-independent and delay-dependent sufficient conditions are also attained. It should be especially emphasized that we do not assume that the coupling matrix is symmetric or diagonal. However, most of the former works on network synchronization are based on this assumption. Furthermore, we extend the derivative of the time-varying delay to any given value. Last but not least, one constructs a small-world network, a regular network, and scale-free networks with network size to verify the effectiveness of the proposed synchronous criteria.

The remainder of this paper is organized as follows: In section 2, the main background of complex networks is briefly outlined, and a generally linearly hybrid coupled network with time-varying delay is proposed. The main theorems and corollaries for global network synchronization are then given in section 3. In section 4, a small-world network, a regular network, and scale-free networks with network size are constructed to show the effectiveness of the proposed global network synchronous criteria. The conclusions are finally drawn in section 5.

2. Preliminaries. Consider a complex dynamical network consisting of N identical nodes with linearly diffusive couplings [3, 4, 5, 14, 32, 33, 34], which is described by

$$\dot{x}_i(t) = f(x_i(t)) + c\sum_{j=1, j\neq i}^{N} G_{ij}\Gamma(x_j(t) - x_i(t)), \quad i = 1, 2, \ldots, N, \tag{1}$$

where $i = 1, 2, \ldots, N$, $x_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{in}(t))^T \in R^n$ is the state vector of the ith node, $f : R^n \longrightarrow R^n$ is continuously differentiable, c is the coupling strength, $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_n) \in R^{n\times n}$ is a 0-1 diagonal matrix with specific $\gamma_i = 1$ and 0 for others, and $G = (G_{ij})_{N\times N}$ is a coupling configuration matrix representing the topological structure of the network, where G_{ij} is defined as follows: if there exists a connection from node i to another node j, then the coupling strength $G_{ij} = G_{ji} = 1$; otherwise, $G_{ij} = G_{ji} = 0$ $(j \neq i)$, and the diagonal elements of matrix G are defined by

$$G_{ii} = -\sum_{j=1, j\neq i}^{N} G_{ij} = -\sum_{j=1, j\neq i}^{N} G_{ji}. \tag{2}$$

Thus network (1) can be rewritten as follows:

$$\dot{x}_i(t) = f(x_i(t)) + c\sum_{j=1}^{N} G_{ij}\Gamma x_j(t), \quad i = 1, 2, \ldots, N. \tag{3}$$

Hereafter, suppose network (3) is connected in the sense that there are no isolate clusters. That is, the coupling configuration matrix G is an irreducible matrix.

However, most real-world complex networks are time-varying. To characterize the real-world evolving networks, Lü and Chen introduced a time-varying network [14, 32, 33, 34] which represents many real biological and engineering networks. Also, there inevitably exists time-delay in many practical complex networks because of the finite information exchanging speed. Considering the time-delay, we propose a simple complex network model as follows [13]. Recently, a linearly coupled complex network was presented and further studied [6, 7, 8]. Considering the time delay, Cao and his colleagues further introduced the following constant delayed complex dynamical network [16].

In this paper, we will consider the following linearly hybrid coupled network with time-varying delay:

$$\dot{x}_i(t) = -Cx_i(t) + Af(x_i(t)) + Bf(x_i(t-\tau)) + I(t) + \sum_{j=1}^{N} G_{ij}Dx_j(t) + \sum_{j=1}^{N} G_{ij}D_\tau x_j(t-\tau(t)), \tag{4}$$

where $i = 1, 2, \ldots, N$, $C = \mathrm{diag}(c_1, c_2, \ldots, c_n) \in R^{n\times n}$ is a diagonal matrix with positive diagonal entries $c_i > 0$, $i = 1, 2, \ldots, n$, $A = (a_{ij})_{n\times n}$ and $B = (b_{ij})_{n\times n}$ are weight and delayed weight matrices, respectively, $I(t) = (I_1(t), I_2(t), \ldots, I_n(t))^T \in R^n$ is an external input vector, $D = (d_{ij}) \in R^{n\times n}$ and $D_\tau = (d_{\tau ij}) \in R^{n\times n}$ are constant and delayed inner coupling matrices of complex networks, respectively, $f(x_i(t)) = (f_1(x_{i1}(t)), f_2(x_{i2}(t)), \ldots, f_n(x_{in}(t)))^T \in R^n$

corresponds to the activation functions of neurons, and $G = (G_{ij})_{N\times N}$ satisfies the diffusive condition (2). In particular, one does not assume that the constant and delayed inner coupling matrices $D = (d_{ij}) \in R^{n\times n}$ and $D_\tau = (d_{\tau ij}) \in R^{n\times n}$ are diagonal matrices.

Denote $x_i(t) = \phi_i(t) \in \mathcal{C}([-r, 0], R)$ $(i = 1, 2, \ldots, N)$, where $r = \sup_{t\in R}\{\tau(t)\}$ and $\mathcal{C}([-r, 0], R)$ is the set of continuous functions from $[-r, 0]$ to R. To simplify, one has the following fundamental assumptions.

A_1: $f_i(x_i)$ $(i = 1, 2, \ldots, n)$ are monotonically nondecreasing on R.

A_2: $f_i(x_i)$ $(i = 1, 2, \ldots, n)$ are Lipschitz continuous; i.e., there exist constants $F_i > 0$ such that

$$|f_i(\alpha_1) - f_i(\alpha_2)| \le F_i|\alpha_1 - \alpha_2| \quad \forall \alpha_1, \alpha_2 \in R. \tag{5}$$

A_3: $\tau(t)$ is a bounded differential function of time t satisfying

$$0 \le \dot{\tau}(t) \le h < 1,$$

where h is a positive real constant.

A_4: The coupling matrix G satisfies the conditions

$$G_{ij} \ge 0, \quad i \ne j, \qquad G_{ii} = -\sum_{j=1,\, j\ne i}^{N} G_{ij}, \quad i = 1, 2, \ldots, N. \tag{6}$$

Before stating the main results, some similar definitions and lemmas are given [6, 7, 8, 9, 10].

Definition 1. *Let* $r = \max_{t\in R}\{\tau(t)\}$. *Set* $\mathbf{S} = \{x = (x_1(s), x_2(s), \ldots, x_N(s)) : x_i(s) \in \mathcal{C}([-r, 0], R),\ x_i(s) = x_j(s),\ i, j = 1, 2, \ldots, N\}$, *which is called the synchronization manifold of network* (4).

Definition 2. *Let* $\widehat{R}$ *be a ring and* $T(\widehat{R}, K) = \{$*the set of matrices with entries* $\widehat{R}$ *such that the sum of the entries in each row is equal to* K *for some* $K \in \widehat{R}\}$.

Definition 3. *The set of* $M_1^N(1)$: $M_1^N(1)$ *is composed of matrices with* N *columns; each row (such as the* i*th row) of* $\widetilde{M} \in M_1^N(1)$ *has exactly one entry* α_i *and one entry* $-\alpha_i$, *where* $\alpha_i \ne 0$, *and all other entries are zeros.*

Definition 4. *The set of* $M_1^N(n)$: $M_1^N(n) = \{\mathbf{M} = M \otimes I_n : M \in M_1^N(1)$, I_n *is the* n*-dimensional identity matrix*$\}$, *where* $\otimes$ *is Kronecker product.*

Definition 5. $M_2^N(n) \subset M_1^N(n)$: *If* $M \in M_2^N(n)$, *for any pair of indices* i *and* j, *there exist indices* $j_1, j_2, \ldots, j_l$ *and* $p_1, p_2, \ldots, p_{l-1}$ *such that* $M_{p_q,i_q} \ne 0$ *and* $M_{p_q,i_{q+1}} \ne 0$ *for all* $1 \le q < l$, *where* $j_1 = i$ *and* $j_l = j$.

Definition 6. *Synchronization manifold* $\mathbf{S}$ *is said to be globally exponentially stable (or network* (4) *is globally exponentially synchronized) if there exist* $\epsilon > 0$, $T > t_0$, *and* $M > 0$ *such that*

$$\|x_i(t) - x_j(t)\| \le Me^{-\epsilon t}, \tag{7}$$

where $\phi_i \in \mathcal{C}([-r, 0], R)$, $t > T$, $i, j = 1, 2, \ldots, N$.

Definition 7. *Synchronization manifold* $\mathbf{S}$ *is said to be globally asymptotically stable (or network* (4) *is globally asymptotically synchronized) if for any* $\varepsilon > 0$, *there exists* $T > t_0$ *such that*

$$\|x_i(t) - x_j(t)\| \le \varepsilon, \tag{8}$$

where $\phi_i, \phi_j \in \mathcal{C}([-r, 0], R)$, $t > T$, $i, j = 1, 2, \ldots, N$.

Lemma 1 (see [9]). *Let* G *be an* $N \times N$ *matrix in the set* $T(\widehat{R}, K)$. *Then the* $(N-1)\times(N-1)$ *matrix* H *satisfies* $MG = HM$, *where* $H = MGJ$,

(9)

$$M = \begin{pmatrix} \mathbf{1} & \mathbf{-1} & & & \\ & \mathbf{1} & \mathbf{-1} & & \\ & & & \ddots & \\ & & & \mathbf{1} & \mathbf{-1} \end{pmatrix}_{((N-1)\times N)}, \quad J = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \cdots & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & \cdots & \mathbf{1} \\ & & \ddots & & \mathbf{1} \\ & & \cdots & \mathbf{1} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}_{(N\times(N-1))},$$

in which $\mathbf{1}$ *is the multiplicative identity of* $\widehat{R}$. *Moreover, the matrix* H *can be rewritten explicitly as follows:* $H_{(i,j)} = \sum_{k=1}^{j} G_{(i,k)} - G_{(i+1,k)}$ *for* $i, j \in \{1, 2, \ldots, N-1\}$.

Lemma 2 (Schur complement [17]). *The LMI*

$$\begin{pmatrix} \widetilde{Q}(x) & \widetilde{S}(x) \\ \widetilde{S}(x)^T & \widetilde{R}(x) \end{pmatrix} > 0$$

is equivalent to one of the following conditions:

(i) $\widetilde{Q}(x) > 0$, $\widetilde{R}(x) - \widetilde{S}(x)^T \widetilde{Q}(x)^{-1} \widetilde{S}(x) > 0$,

(ii) $\widetilde{R}(x) > 0$, $\widetilde{Q}(x) - \widetilde{S}(x) \widetilde{R}(x)^{-1} \widetilde{S}(x)^T > 0$,

where $\widetilde{Q}(x) = \widetilde{Q}(x)^T$, $\widetilde{R}(x) = \widetilde{R}(x)^T$.

Lemma 3 (see [9]). *If matrix* G *is symmetric and also satisfies condition* A_4, *then* G *is irreducible iff there exists a* $p \times N$ *matrix* $M \in M_2^N(1)$, *such that* $G = -M^T M$.

Lemma 4 (see [9]). *Let* $x = (x_1, x_2, \ldots, x_N)^T$, *where* $x_i \in R^n$, $i = 1, 2, \ldots, N$. *Then* $x \in \mathbf{S}$ *iff there exists* $\mathbf{M} \in M_2^N(n)$ *satisfying* $\|\mathbf{M}x\| = 0$.

Denote

$$d(x) = \|\mathbf{M}x\|^2 = x^T \mathbf{M}^T \mathbf{M} x, \quad \mathbf{M} \in M_2^N(n). \tag{10}$$

Then $d(x)$ *is a nonnegative distance function. From the assumptions of* $\mathbf{M}$, *one has* $d(x) \longrightarrow 0$ *iff* $\|x_i(t) - x_j(t)\| \longrightarrow 0$ *for all* $i, j = 1, 2, \ldots, N$.

Lemma 5 (see [19]). *The Kronecker product has the following properties:*

(1) $(\alpha A) \otimes B = A \otimes (\alpha B)$;

(2) $(A + B) \otimes C = A \otimes C + B \otimes C$;

(3) $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.

Lemma 6 (Jensen inequality [29]). *Assume that the vector function* $\omega : [0, r] \in R^{m\times m}$ *is well defined for the following integrations. For any symmetric matrix* $W \in R^{m\times m}$ *and scalar*

$r > 0$, one has

$$r\int_0^r \omega(s)W\omega(s)\,ds \geq \left(\int_0^r \omega(s)\,ds\right)^T W\left(\int_0^r \omega(s)\,ds\right).$$

3. Main results. In this section, several novel criteria are proposed for the global synchronization of complex network (4) based on the Lyapunov function and an LMI.

To simplify the presentation, some notation is given in the following. Let T be a symmetric and irreducible matrix satisfying assumption A_4. From Lemma 3, there exists a $p \times N$ matrix $\widetilde{M} \in M_2^N(1)$ such that $T = -\widetilde{M}^T\widetilde{M}$. Denote $\mathbf{M} = \widetilde{M} \otimes I_n \in M_2^N(n)$, and let $\mathbf{M}_i = (M_{i1}, M_{i2}, \ldots, M_{iN})$ be the ith row of $\mathbf{M}$, where $M_{ii_1} = \alpha_i I_n$, $M_{ii_2} = -\alpha_i I_n$, and $M_{ij} = 0$ for $j \neq i_1, i_2$. Let $A \otimes B$ be the Kroneker product of matrices A and B.

$$\mathbf{C} = I_N \otimes C, \quad \mathbf{C^1} = I_p \otimes C, \quad \mathbf{A} = I_N \otimes A, \quad \mathbf{A^1} = I_p \otimes A,$$

$$\mathbf{B} = I_N \otimes B, \quad \mathbf{B^1} = I_p \otimes B, \quad \mathbf{G} = G \otimes D, \quad \mathbf{G}_\tau = G \otimes D_\tau,$$

$$x_i(t) = (x_{i1}(t), x_{i2}(t), \ldots, x_{in}(t))^T \ (\forall i = 1, 2, \ldots, N), \quad x(t) = (x_1^T(t), x_2^T(t), \ldots, x_N^T(t))^T,$$

$$\mathbf{f}(x(t)) = (f^T(x_1(t)), f^T(x_2(t)), \ldots, f^T(x_N(t)))^T, \quad \mathbf{I}(t) = (I^T(t), I^T(t), \ldots, I^T(t))^T, \tag{11}$$

where I_N is the N dimensional identity matrix.

Then the complex network (4) can be recast as follows:

$$\dot{x}(t) = -\mathbf{C}x(t) + \mathbf{A}\mathbf{f}(x(t)) + \mathbf{B}\mathbf{f}(x(t-\tau(t))) + \mathbf{I}(t) + \mathbf{G}x(t) + \mathbf{G}_\tau x(t-\tau(t)), \quad i = 1, 2, \ldots, N. \tag{12}$$

Theorem 1. *Suppose A_2–A_4 hold. Network (12) is globally exponentially synchronized if there exist positive definite matrices $P = (p_{ij})_{n\times n} \in R^{n\times n}$, $Q = (q_{ij})_{n\times n} \in R^{n\times n}$, and $\Omega = (\Omega_{ij})_{n\times n} \in R^{n\times n}$, a positive definite diagonal matrix $\Sigma = \mathrm{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_n) \in R^{n\times n}$, a symmetric matrix $\Delta = (\Delta_{ij})_{n\times n} \in R^{n\times n}$, and an irreducible symmetric matrix $T = (t_{ij}) \in R^{N\times N}$ satisfying A_4, such that*

$$\Lambda_0 = \begin{pmatrix} -2PC - \Delta + F\Sigma F & PA & PB \\ A^TP & -\Sigma + Q & 0 \\ B^TP & 0 & -(1-h)Q \end{pmatrix} < 0, \tag{13}$$

where $F = \mathrm{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$ and one of the following conditions holds:

$$\text{(i) } \Lambda_{ij} = \begin{pmatrix} \sum_{k=1}^N T_{ik}G_{kj}(PD + D^TP) + T_{ij}(\Omega + \Delta) & \sum_{k=1}^N T_{ik}G_{kj}PD_\tau \\ \sum_{k=1}^N T_{ik}G_{kj}D_\tau^TP & -(1-h)T_{ij}\Omega \end{pmatrix} < 0 \tag{14}$$

$$\forall 1 \leq i < j \leq N.$$

$$\text{(ii) } \widetilde{\Lambda}_{ij} = \begin{pmatrix} 2\sum_{k=1}^n p_{ik}d_{kj}TG + (\Omega_{ij} + \Delta_{ij})T & \sum_{k=1}^n p_{ik}d_{\tau kj}TG \\ \sum_{k=1}^n p_{ik}d_{\tau kj}G^TT & -(1-h)\Omega_{ij}T \end{pmatrix} > 0 \tag{15}$$

$$\forall 1 \leq i, j \leq n.$$

Proof. See Appendix A.

Instead of using inequality (A.4), from assumption A_1, one has

$$\begin{aligned}\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{\Sigma}\mathbf{M}\mathbf{f}(x(t)) &= \sum_{j=1}^{p}\alpha_j^2[f(x_{j_1}(t))-f(x_{j_2}(t))]^T\Sigma[f(x_{j_1}(t))-f(x_{j_2}(t))]\\ &\le \sum_{j=1}^{p} y_j^T(t)F\Sigma[f(x_{j_1}(t))-f(x_{j_2}(t))].\end{aligned} \tag{16}$$

Similarly, following the same steps in part (i) of Theorem 1, one can easily attain the following corollary.

Corollary 1. *Suppose that assumptions A_1–A_4 hold. Then network (12) is globally exponentially synchronized if there exist positive definite matrices $P=(p_{ij})_{n\times n}\in R^{n\times n}$, $Q=(q_{ij})_{n\times n}\in R^{n\times n}$, and $\Omega=(\Omega_{ij})_{n\times n}\in R^{n\times n}$, a positive definite diagonal matrix $\Sigma=\operatorname{diag}(\Sigma_1,\Sigma_2,\ldots,\Sigma_n)\in R^{n\times n}$, a symmetric matrix $\Delta=(\Delta_{ij})_{n\times n}\in R^{n\times n}$, and an irreducible symmetric matrix $T=(t_{ij})\in R^{N\times N}$ satisfying A_4, such that*

$$\begin{pmatrix} -2PC-\Delta & PA+F\Sigma & PB\\ A^TP+\Sigma F & -2\Sigma+Q & 0\\ B^TP & 0 & -(1-h)Q\end{pmatrix}<0, \tag{17}$$

where $F=\operatorname{diag}(F_1,F_2,\ldots,F_n)\in R^{n\times n}$ and

$$\begin{pmatrix} \sum_{k=1}^{N}T_{ik}G_{kj}(PD+D^TP)+T_{ij}(\Omega+\Delta) & \sum_{k=1}^{N}T_{ik}G_{kj}PD_\tau\\ \sum_{k=1}^{N}T_{ik}G_{kj}D_\tau^TP & -(1-h)T_{ij}\Omega\end{pmatrix}<0 \quad \forall 1\le i<j\le N. \tag{18}$$

Denote $e=(1,1,\ldots,1)^T\in R^N$, $J=ee^T$, $U=NI_N-J$. Let $T=-U=J-NI_N$; then $T_{ij}=1$ $(i\ne j)$, $T_{ij}=-(N-1)$ $(i=j)$, $i,j=1,2,\ldots,N$. It is easy to verify that T satisfies assumption A_4. According to Lemma 3, there exists a $p\times N$ matrix $M\in M_2^N(1)$, such that $T=-M^TM$. Since G satisfies assumption A_4, then one has

$$\begin{aligned}\sum_{k=1}^{N}T_{ik}G_{kj} &= (T_{ii}-1)G_{ij}+\sum_{k=1,\,k\ne i}^{N}T_{ik}G_{kj}+G_{ij}\\ &= -NG_{ij}+\sum_{k=1}^{N}G_{kj}=-NG_{ij}.\end{aligned} \tag{19}$$

Therefore, from Theorem 1, one gets the following corollary.

Corollary 2. *Suppose assumptions A_2–A_4 hold. Then network (12) is globally exponentially synchronized if there exist positive definite matrices $P=(p_{ij})_{n\times n}\in R^{n\times n}$, $Q=(q_{ij})_{n\times n}\in R^{n\times n}$, and $\Omega=(\Omega_{ij})_{n\times n}\in R^{n\times n}$, a positive definite diagonal matrix $\Sigma=\operatorname{diag}(\Sigma_1,\Sigma_2,\ldots,\Sigma_n)\in R^{n\times n}$, and a symmetric matrix $\Delta=(\Delta_{ij})_{n\times n}\in R^{n\times n}$, such that*

$$\begin{pmatrix} -2PC-\Delta+F\Sigma F & PA & PB\\ A^TP & -\Sigma+Q & 0\\ B^TP & 0 & -(1-h)Q\end{pmatrix}<0, \tag{20}$$

where $F=\operatorname{diag}(F_1,F_2,\ldots,F_n)\in R^{n\times n}$ and

$$\begin{pmatrix} -NG_{ij}(PD+D^TP)+(\Omega+\Delta) & -NG_{ij}PD_\tau\\ -NG_{ij}D_\tau^TP & -(1-h)\Omega\end{pmatrix}<0 \quad \forall 1\le i<j\le N. \tag{21}$$

Remark 1. In [16], Cao and his colleagues investigated the global synchronization of a coupled complex network with constant time-delay. The main theorem in [16] is Corollary 2 with $h = 0$, where the time-delay is a constant. Therefore, the main result in [16] is a special case of Theorem 1.

Let $G_\tau = 0$; i.e., there is no linearly delayed coupling in network (12) as that in [6, 7, 8]. Let $\Omega = \zeta I_n$, where ζ is a sufficient small positive number. Then one has the following corollary.

Corollary 3. *Suppose assumptions A_2–A_4 hold. Network (12) with $D_\tau = 0$ is globally exponentially synchronized if there exist positive definite matrices $P = (p_{ij})_{n\times n} \in R^{n\times n}$ and $Q = (q_{ij})_{n\times n} \in R^{n\times n}$, a positive definite diagonal matrix $\Sigma = \operatorname{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_n) \in R^{n\times n}$, a symmetric matrix $\Delta = (\Delta_{ij})_{n\times n} \in R^{n\times n}$, and an irreducible symmetric matrix $T = (t_{ij}) \in R^{N\times N}$ satisfying A_4, such that*

$$\begin{pmatrix} -2PC - \Delta + F\Sigma F & PA & PB \\ A^T P & -\Sigma + Q & 0 \\ B^T P & 0 & -(1-h)Q \end{pmatrix} < 0, \tag{22}$$

where $F = \operatorname{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$ and

$$\left(2\sum_{k=1}^{n} p_{ik} d_{kj} TG + \Delta_{ij} T\right) > 0 \quad \forall 1 \le i, j \le n. \tag{23}$$

It should be especially emphasized that the inner coupling matrix D is not necessarily a diagonal matrix in Theorem 1. If D is a diagonal matrix, then one gets the following corollary.

Corollary 4. *Suppose that assumptions A_2–A_4 hold. Assume also that $D_\tau = 0$ and D is diagonal. Network (12) is globally exponentially synchronized if there exist a positive definite matrix $Q = (q_{ij})_{n\times n} \in R^{n\times n}$, positive definite diagonal matrices $P = \operatorname{diag}(p_1, p_2, \ldots, p_n) \in R^{n\times n}$ and $\Sigma = \operatorname{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_n) \in R^{n\times n}$, a symmetric matrix $\Delta = \operatorname{diag}(\delta_1, \delta_2, \ldots, \delta_n) \in R^{n\times n}$, and an irreducible symmetric matrix $T = (t_{ij}) \in R^{N\times N}$ satisfying A_4, such that*

$$\begin{pmatrix} -2PC - \Delta + F\Sigma F & PA & PB \\ A^T P & -\Sigma + Q & 0 \\ B^T P & 0 & -(1-h)Q \end{pmatrix} < 0, \tag{24}$$

where $F = \operatorname{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$ and

$$T\left(2p_i d_i G + \delta_i\right) > 0, \quad i = 1, 2, \ldots, n. \tag{25}$$

Let P, Ω, and Δ be diagonal matrices. According to part (ii) of Theorem 1, one has the following corollary.

Corollary 5. *Suppose assumptions A_2–A_4 hold. Suppose also that D and D_τ are diagonal matrices. Then network (12) is globally exponentially synchronized if there exist a positive definite matrix $Q = (q_{ij})_{n\times n} \in R^{n\times n}$, positive definite diagonal matrices $P = \operatorname{diag}(p_1, p_2, \ldots, p_n) \in R^{n\times n}$, $\Sigma = \operatorname{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_n) \in R^{n\times n}$, and $\Sigma = \operatorname{diag}(\Omega_1, \Omega_2, \ldots, \Omega_n) \in R^{n\times n}$, a*

symmetric matrix $\Delta = \operatorname{diag}(\delta_1, \delta_2, \ldots, \delta_n) \in R^{n\times n}$, *and an irreducible symmetric matrix* $T = (t_{ij}) \in R^{N\times N}$ *satisfying* A_4, *such that*

$$\begin{pmatrix} -2PC - \Delta + F\Sigma F & PA & PB \\ A^T P & -\Sigma + Q & 0 \\ B^T P & 0 & -(1-h)Q \end{pmatrix} < 0, \tag{26}$$

where $F = \operatorname{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$ *and*

$$\begin{pmatrix} 2p_i d_i TG + \Omega_i + \delta_i T & p_i d_{\tau i} TG \\ p_i d_{\tau i} G^T T & -(1-h)\Omega_i T \end{pmatrix} > 0, \quad i = 1, 2, \ldots, n. \tag{27}$$

Remark 2. In [6], Lu and Chen further investigated the synchronization of a coupled connected neural network with constant time-delay. Theorem 3 in [6] is Corollary 4 with $h = 0$. Therefore, the main result in [6] is a special case of Theorem 1. Moreover, in [6], the inner coupling matrix D and the inner delayed coupling matrix D_τ are both diagonal matrices. However, one removes these limit conditions in this paper.

Remark 3. To minimize the number of LMIs in the conditions of Theorem 1, one can apply the following rule: if $N < n$, one can use condition (i) of Theorem 1; otherwise, one can use condition (ii) of Theorem 1.

Since the conditions of Theorem 1 are relatively complex, one will simplify these LMIs by introducing some special $M \in M_2^N(1)$.

Theorem 2. *Suppose that assumptions* A_2–A_4 *hold. Network* (12) *is globally asymptotically synchronized if there exist positive definite matrices* $\mathbf{P} = (p_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, $\mathbf{Q} = (q_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, *and* $\mathbf{R} = (r_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$ *and a positive definite diagonal matrix* $\mathbf{\Sigma} = \operatorname{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_{(N-1)n}) \in R^{(N-1)n\times(N-1)n}$, *such that*

$$\mathbf{\Omega} = \begin{pmatrix} -2\mathbf{PC^1} + \mathbf{PH} + \mathbf{H}^T\mathbf{P} + \mathbf{F\Sigma F} + \mathbf{R} & \mathbf{PH}_\tau & \mathbf{PA^1} & \mathbf{PB^1} \\ \mathbf{H}_\tau^T\mathbf{P} & -(1-h)\mathbf{R} & 0 & 0 \\ \mathbf{A^1}^T\mathbf{P} & 0 & -\mathbf{\Sigma} + \mathbf{Q} & 0 \\ \mathbf{B^1}^T\mathbf{P} & 0 & 0 & -(1-h)\mathbf{Q} \end{pmatrix} < 0, \tag{28}$$

where $F = \operatorname{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$, $\mathbf{F} = I_{N-1} \otimes F$, $H = MGJ$, $\mathbf{H} = (MGJ) \otimes D$, $\mathbf{H}_\tau = (MGJ) \otimes D_\tau$, *and* M *and* J *are defined in* (9).

Proof. See Appendix B.

Instead of using inequality (B.4), from assumption A_1, one has

$$\begin{aligned} \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{\Sigma}\mathbf{M}\mathbf{f}(x(t)) &= \textstyle\sum_{j=1}^{N-1} [f(x_j(t)) - f(x_{j+1}(t))]^T \mathbf{\Sigma}_j [f(x_j(t)) - f(x_{j+1}(t))] \\ &\le \textstyle\sum_{j=1}^{N-1} [x_j(t) - x_{j+1}(t)]^T F \mathbf{\Sigma}_j [x_j(t) - x_{j+1}(t)] \\ &= x^T(t)\mathbf{M}^T\mathbf{F\Sigma M}x(t), \end{aligned} \tag{29}$$

where $\mathbf{\Sigma}_j = \operatorname{diag}(\Sigma_{(j-1)n+1}, \ldots, \Sigma_{jn})$. Following the same steps in Theorem 2, other conditions can be similarly verified. Then the following corollary is obtained.

Corollary 6. *Suppose assumptions A_1–A_4 hold. Network (12) is globally asymptotically synchronized if there exist positive definite matrices $\mathbf{P} = (p_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, $\mathbf{Q} = (q_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, and $\mathbf{R} = (r_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$ and a positive definite diagonal matrix $\mathbf{\Sigma} = \mathrm{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_{(N-1)n}) \in R^{(N-1)n\times(N-1)n}$, such that*

$$\begin{pmatrix} -2\mathbf{PC^1} + \mathbf{PH} + \mathbf{H}^T\mathbf{P} + \mathbf{R} & \mathbf{PH}_\tau & \mathbf{PA^1} + \mathbf{F\Sigma} & \mathbf{PB^1} \\ \mathbf{H}_\tau^T\mathbf{P} & -(1-h)\mathbf{R} & 0 & 0 \\ \mathbf{A^1}^T\mathbf{P} + \mathbf{\Sigma F} & 0 & -2\mathbf{\Sigma} + \mathbf{Q} & 0 \\ \mathbf{B^1}^T\mathbf{P} & 0 & 0 & -(1-h)\mathbf{Q} \end{pmatrix} < 0, \tag{30}$$

where $F = \mathrm{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$, $\mathbf{F} = I_{N-1} \otimes F$, $H = MGJ$, $\mathbf{H} = (MGJ) \otimes D$, $\mathbf{H}_\tau = (MGJ) \otimes D_\tau$, and M and J are defined in (9).

Corollary 7. *Suppose that assumptions A_2–A_4 hold. Network (12) is globally asymptotically synchronized if there exist positive definite matrices $P = (p_{ij})_{n\times n} \in R^{n\times n}$, $Q = (q_{ij})_{n\times n} \in R^{n\times n}$, and $R = (r_{ij})_{n\times n} \in R^{n\times n}$, a positive definite diagonal matrix $\Sigma = \mathrm{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_n) \in R^{n\times n}$, and a symmetric matrix $\Delta = (\Delta_{ij})_{n\times n} \in R^{n\times n}$, such that*

$$\Lambda_0 = \begin{pmatrix} -2PC - \Delta + F\Sigma F & PA & PB \\ A^TP & -\Sigma + Q & 0 \\ B^TP & 0 & -(1-h)Q \end{pmatrix} < 0 \tag{31}$$

and

$$\Xi = \begin{pmatrix} \mathbf{PH} + \mathbf{H}^T\mathbf{P} + \mathbf{R} + \mathbf{\Delta} & \mathbf{PH}_\tau \\ \mathbf{H}_\tau^T\mathbf{P} & -(1-h)\mathbf{R} \end{pmatrix} < 0, \tag{32}$$

where $F = \mathrm{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$, $\mathbf{P} = I_{N-1} \otimes P$, $\mathbf{R} = I_{N-1} \otimes R$, $\mathbf{\Delta} = I_{N-1} \otimes \Delta$, $\mathbf{H} = H \otimes D$, $\mathbf{H}_\tau = H \otimes D_\tau$, $H = MGJ$, and M and J are defined in (9).

Proof. See Appendix C.

When $D_\tau = 0$, there is no linearly delayed coupling in network (12). Let $R = \zeta I_n$ in Corollary 7, where ζ is a sufficiently small positive number. Then Corollary 7 can be simplified as follows.

Corollary 8. *Suppose that assumptions A_2–A_4 hold. Network (12) with $D_\tau = 0$ is globally exponentially synchronized if there exist positive definite matrices $P = (p_{ij})_{n\times n} \in R^{n\times n}$ and $Q = (q_{ij})_{n\times n} \in R^{n\times n}$, a positive definite diagonal matrix $\Sigma = \mathrm{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_n) \in R^{n\times n}$, and a symmetric matrix $\Delta = (\Delta_{ij})_{n\times n} \in R^{n\times n}$, such that*

$$\Lambda_0 = \begin{pmatrix} -2PC - \Delta + F\Sigma F & PA & PB \\ A^TP & -\Sigma + Q & 0 \\ B^TP & 0 & -(1-h)Q \end{pmatrix} < 0$$

and

$$\mathbf{PH} + \mathbf{H}^T\mathbf{P} + \mathbf{\Delta} < 0,$$

where $F = \mathrm{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$, $\mathbf{P} = I_{N-1} \otimes P$, $\mathbf{\Delta} = I_{N-1} \otimes \Delta$, $\mathbf{H} = H \otimes D$, $H = MGJ$, and M and J are defined in (9).

Remark 4. In [7], Wang and Cao further studied the synchronization of an array of linearly coupled networks with time-varying delay. The main theorem, Theorem 2, in [7] is Corollary 8. Therefore, the main result in [7] is a special case of Theorem 2. Moreover, the inner coupling matrix D and inner delayed coupling matrix D_τ are both diagonal matrices in [7]. However, one removes these limit conditions. Furthermore, one also introduces the time-delay in the linear coupling in this paper.

If assumption A_3 is not satisfied, i.e., $\dot{\tau}(t) \geq 1$ for some t, one attains the following synchronous theorem.

Theorem 3. *Suppose assumptions A_2 and A_4 hold. Then network (12) is globally asymptotically synchronized if there exist positive definite matrices $\mathbf{P} = (p_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, $\mathbf{Q} = (q_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, $\mathbf{R} = (r_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, and $\mathbf{T} = (t_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, positive definite diagonal matrices $\mathbf{\Sigma} = \mathrm{diag}(\Sigma_1, \Sigma_2, \ldots, \Sigma_{(N-1)n}) \in R^{(N-1)n\times(N-1)n}$ and $\mathbf{\Lambda} = \mathrm{diag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_{(N-1)n}) \in R^{(N-1)n\times(N-1)n}$, and a matrix $\mathbf{U} = (u_{ij})_{(N-1)n\times(N-1)n} \in R^{(N-1)n\times(N-1)n}$, such that*

(33)
$$\mathbf{\Omega_1} = \begin{pmatrix} \Psi & \mathbf{PH}_\tau + \mathbf{U}^T & \mathbf{PA^1} & \mathbf{PB^1} & 0 & (-\mathbf{C^1} + \mathbf{H})^T\mathbf{T} \\ \mathbf{H}_\tau^T\mathbf{P} + \mathbf{U} & \Psi_1 & 0 & 0 & -\mathbf{U} & \mathbf{H}_\tau^T\mathbf{T} \\ \mathbf{A^1}^T\mathbf{P} & 0 & -\mathbf{\Sigma} + \mathbf{Q} & 0 & 0 & \mathbf{A^1}^T\mathbf{T} \\ \mathbf{B^1}^T\mathbf{P} & 0 & 0 & -(1-h)\mathbf{Q} - \mathbf{\Lambda} & 0 & \mathbf{B^1}^T\mathbf{T} \\ 0 & -\mathbf{U}^T & 0 & 0 & -\frac{1}{r}\mathbf{T} & 0 \\ \mathbf{T}(-\mathbf{C^1} + \mathbf{H}) & \mathbf{TH}_\tau & \mathbf{TA^1} & \mathbf{TB^1} & 0 & -\frac{1}{r}\mathbf{T} \end{pmatrix} < 0,$$

where $\Psi = -2\mathbf{PC^1} + \mathbf{PH} + \mathbf{H}^T\mathbf{P} + \mathbf{R} + \mathbf{F\Sigma F}$, $\Psi_1 = -(1-h)\mathbf{R} - 2\mathbf{U} + \mathbf{F\Lambda F}$, $F = \mathrm{diag}(F_1, F_2, \ldots, F_n) \in R^{n\times n}$, $\mathbf{F} = I_{N-1} \otimes F$, $H = MGJ$, $\mathbf{H} = (MGJ) \otimes D$, $\mathbf{H}_\tau = (MGJ) \otimes D_\tau$, and M and J are defined in (9).

Proof. See Appendix D.

Remark 5. Although Theorems 1–3 and Corollaries 1–8 give some rigorously theoretical conditions for the synchronization of network (12), it is also difficult to fix the suitable parameters of matrixes in these conditions. However, in real-world control systems, one can easily use MATLAB LMI Toolbox to numerically solve these system parameters. For example, in Theorem 1, fixing matrix T as in Corollary 2, one can use MATLAB LMI Toolbox to solve (20) and (21); in Theorems 2–3, one can use MATLAB LMI Toolbox to solve (28) and (33), respectively.

In this paper, the delay-independent and delay-dependent conditions are both further investigated. It is well known that the delay-independent conditions tend to be conservative for small time-delay. In addition, for the coupled networks with time-varying delay, the state estimation criteria proposed in [7] are not applicable to the case in which the derivative of the time-varying delay is larger than 1. In this case, assumption A_3 is not satisfied. In this paper, one overcomes this difficulty in Theorem 3. Moreover, in [6, 7, 8], the coupling matrix G is a diagonal matrix. However, we do not need this assumption in all theorems.

4. Numerical simulations.

To verify the effectiveness of the proposed theorems and corollaries, three simple examples are given in the following.

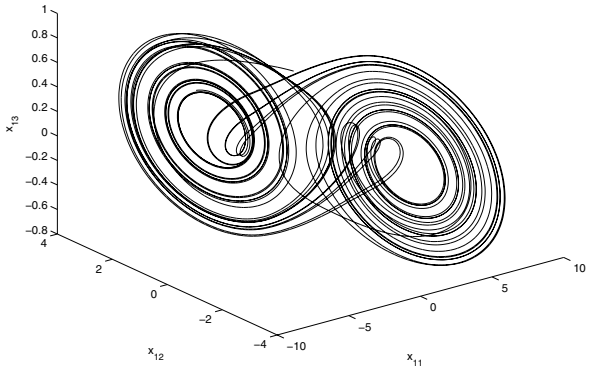


Figure 1. *Phase portrait of a single node.*

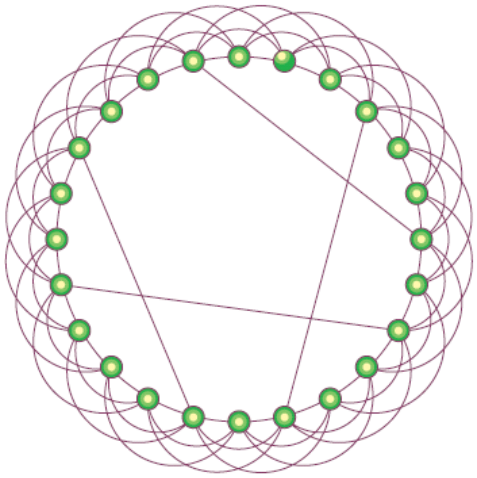


Figure 2. *Small-world network.*

4.1. Synchronization of small-world network. Consider the following Chua circuit described by [37]:

$$\begin{cases} \dot{x}_1 = \theta(-x_1 + x_2 - l(x_1)), \\ \dot{x}_2 = x_1 - x_2 + x_3, \\ \dot{x}_3 = -\beta x_2, \end{cases} \tag{34}$$

where $l(x_1) = bx_1 + 0.5(a-b)(|x_1+1| - |x_1-1|)$. The system (34) is chaotic as shown in Figure 1 when $\theta = 10$, $\beta = 18$, $a = -4/3$, and $b = -3/4$.

Now one considers network (4) with small-world connection as shown in Figure 2 [1], where the single node is given as above, and

$$D = \begin{pmatrix} 4 & 0.4 & 0.4 \\ 0.8 & 4 & 1.2 \\ -0.4 & -0.8 & 4 \end{pmatrix}.$$

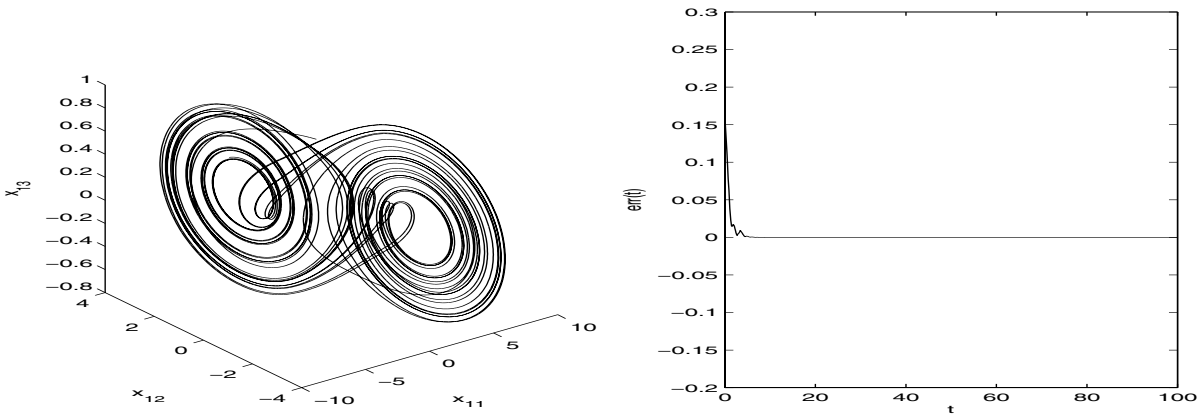


Figure 3. *Phase portrait of single node in network* (4) *and total synchronous error of the small-world network* (4).

According to Theorem 2 and MATLAB LMI Toolbox, one can easily attain the feasible solutions. Then network (4) is globally asymptotically synchronized. The total synchronous error of the small-world network is defined as follows:

$$err(t) = \frac{1}{25}\sum_{i=1}^{2}\sqrt{\sum_{j=1}^{25}[x_{1i}(t) - x_{ji}(t)]^2}.$$

Figure 3 shows the phase portrait of single node in network (4) and the total synchronous error of the small-world network (4). Similarly, one can verify the synchronization of network (4) with other topological structures, such as random graph and scale-free distribution.

4.2. Synchronization of a regular network. Consider the following 2-dimensional delayed system as a node, which is described by

$$\dot{x}(t) = -Cx(t) + Af(x(t)) + Bf(x(t-\tau(t))) + I(t), \tag{35}$$

where $x(t) = (x_1(t), x_2(t))^T$, $f(x(t)) = (\tanh(x_1(t)), \tanh(x_2(t)))^T$, $I(t) = (0,0)^T$,

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 2.0 & -0.1 \\ -5.0 & 3.0 \end{pmatrix}, \quad B = \begin{pmatrix} -1.5 & -0.1 \\ -0.2 & -2.5 \end{pmatrix},$$

and $\tau(t) = 0.03[1 + \sin(40t)]$.

It is easy to verify that assumptions A_1 and A_2 hold for $F = I_2$ and assumption A_3 does not hold for $h = 1.2$, $r = 0.06$. The initial values are given as follows:

$$x_1(s) = 0.4, \quad x_2(s) = 0.6 \qquad \forall s \in [-1, 0].$$

Then system (35) has a chaotic attractor as shown in Figure 4.

Consider a regular network (4), where $A, B, C, I(t), f, \tau(t)$ are given above, and

$$G = \begin{pmatrix} -3 & 1 & 2 \\ 1 & -2 & 1 \\ 2 & 1 & -3 \end{pmatrix}, \quad D = \begin{pmatrix} 4 & 0.4 \\ 0.8 & 4 \end{pmatrix}, \quad D_\tau = \begin{pmatrix} 1 & 0.2 \\ 0.1 & 1 \end{pmatrix}.$$

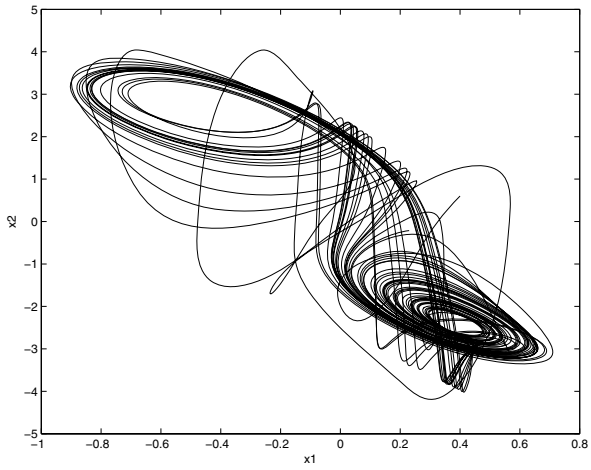


Figure 4. *Phase portrait of a single node.*

From Theorem 3, one gets the feasible solutions as follows:

$$\mathbf{P} = \begin{pmatrix} 9.4243 & -0.9377 & -1.6670 & 0.2372 \\ -0.9377 & 7.7092 & 0.2372 & -1.2747 \\ -1.6670 & 0.2372 & 9.4243 & -0.9377 \\ 0.2372 & -1.2747 & -0.9377 & 7.7092 \end{pmatrix},$$

$$\mathbf{Q} = \begin{pmatrix} 31.5252 & 8.0596 & 2.1015 & -0.8115 \\ 8.0596 & 21.9258 & -0.8115 & 0.8043 \\ 2.1015 & -0.8115 & 31.5252 & 8.0596 \\ -0.8115 & 0.8043 & 8.0596 & 21.9258 \end{pmatrix},$$

$$\mathbf{R} = \begin{pmatrix} 38.3723 & 2.7562 & -3.4028 & -3.1834 \\ 2.7562 & 20.8588 & -3.1834 & 1.0167 \\ -3.4028 & -3.1834 & 38.3723 & 2.7562 \\ -3.1834 & 1.0167 & 2.7562 & 20.8588 \end{pmatrix},$$

$$\mathbf{T} = \begin{pmatrix} 4.5225 & 0.1097 & -0.4335 & -0.1062 \\ 0.1097 & 3.1546 & -0.1062 & -0.0711 \\ -0.4335 & -0.1062 & 4.5225 & 0.1097 \\ -0.1062 & -0.0711 & 0.1097 & 3.1546 \end{pmatrix},$$

$$\mathbf{U} = \begin{pmatrix} 52.2980 & 3.7695 & 4.8829 & 1.0437 \\ 3.7695 & 44.0932 & 1.0437 & 5.5130 \\ 4.8829 & 1.0437 & 52.2980 & 3.7695 \\ 1.0437 & 5.5130 & 3.7695 & 44.0932 \end{pmatrix},$$

$$\mathbf{\Sigma} = \begin{pmatrix} 118.7802 & 0 & 0 & 0 \\ 0 & 66.5769 & 0 & 0 \\ 0 & 0 & 118.7802 & 0 \\ 0 & 0 & 0 & 66.5769 \end{pmatrix},$$

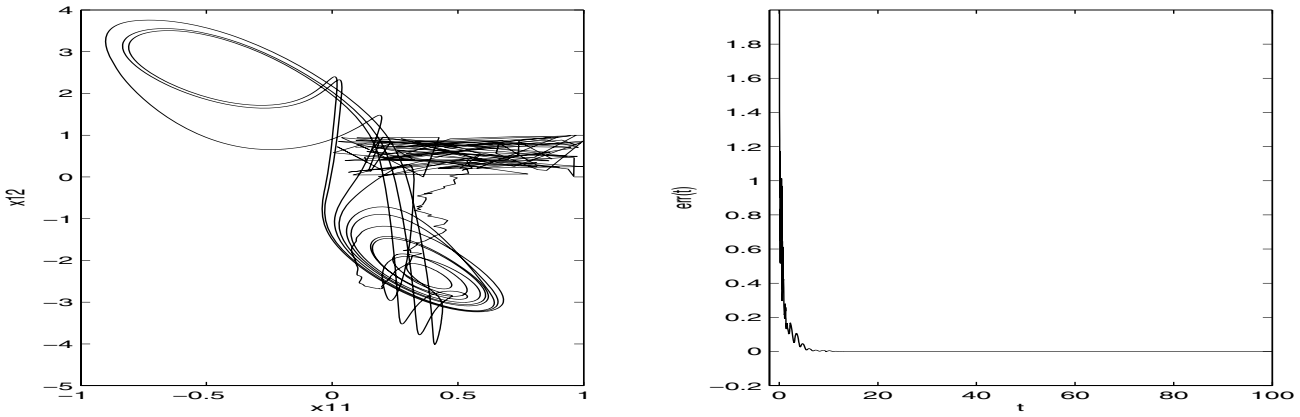


Figure 5. *Total synchronous error of network* (4) *with a regular structure.*

$$\mathbf{\Lambda} = \begin{pmatrix} 35.4724 & 0 & 0 & 0 \\ 0 & 30.6460 & 0 & 0 \\ 0 & 0 & 35.4724 & 0 \\ 0 & 0 & 0 & 30.6460 \end{pmatrix}.$$

According to Theorem 3, network (4) is globally asymptotically synchronized. The total error of network (4) is defined by

$$err(t) = \frac{1}{3}\sum_{i=1}^{2}\sqrt{[x_{1i}(t) - x_{2i}(t)]^2 + [x_{1i}(t) - x_{3i}(t)]^2}.$$

Figure 5 shows the total synchronous error of network (5), where the initial values are given by

$$x_1(s) = \begin{pmatrix} 0.1 \\ -0.3 \end{pmatrix}, \quad x_2(s) = \begin{pmatrix} 0.5 \\ -1 \end{pmatrix}, \quad x_3(s) = \begin{pmatrix} 1 \\ -0.5 \end{pmatrix}.$$

4.3. Synchronization of scale-free networks with network size. The scale-free network model was proposed by Barabási and Albert [40]; they generated the network as follows:

(1) *Growth*: Starting with a small number (m_0) of nodes, at every time step a new node is introduced and connected to m ($\leq m_0$) existing nodes by undirected links.

(2) *Preferential attachment*: The probability that the new node is connected to node i is based on the degree k_i of node i:

$$p_i = \frac{k_i}{\sum_{j=1}^{N} k_j}.$$

After t time steps, this complex network has $N = t+m_0$ nodes and mt links. In the simulation, we take $m_0 = m = 5$, with each node being the same system as (35) except $\tau(t) = \frac{e^t}{1+e^t}$. It is obvious that $0 < \tau(t) < 1$, $\dot{\tau}(t) = \frac{e^t}{(1+e^t)^2} \leq \frac{1}{2} < 1$. G is connected in the scale-free network sense: if there exists a connection from node i to another node j in the scale-free

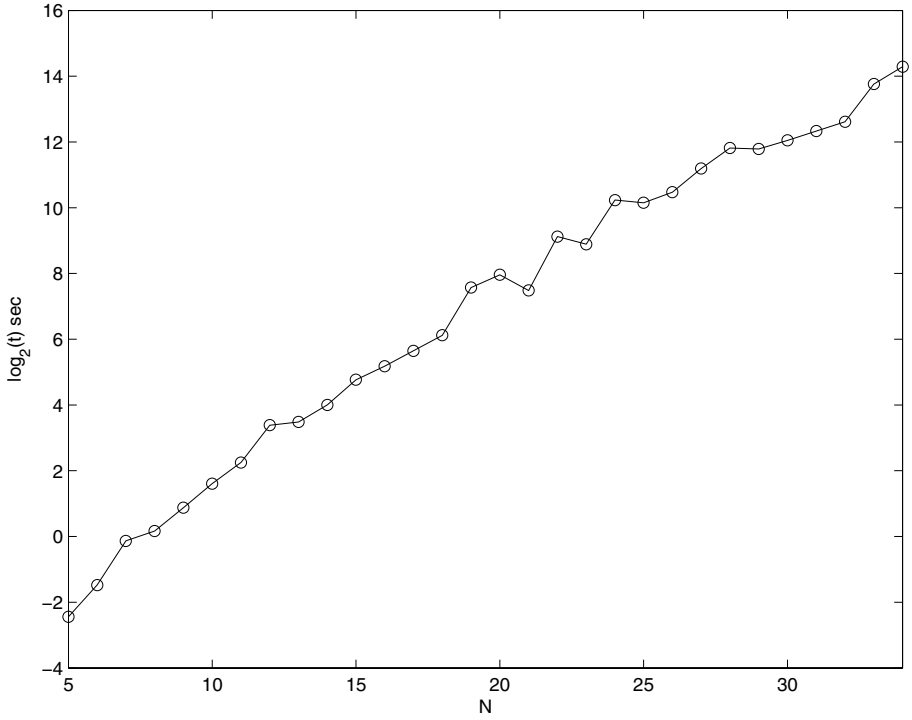


Figure 6. *The time for solving the LMI with network size in scale-free networks.*

network, then the coupling strength $G_{ij} = G_{ji} = 1$; otherwise, $G_{ij} = G_{ji} = 0$ $(j \neq i)$, and $G_{ii} = -\sum_{j=1, j\neq i}^{N} G_{ij}$.

$$D = \begin{pmatrix} 10 & 1 \\ 2 & 10 \end{pmatrix}, \quad D_\tau = \begin{pmatrix} 1 & 0.2 \\ 0.1 & 1 \end{pmatrix}.$$

Theorem 1 is available only when the scale-free network size $N = 5$ or $N = 6$. Therefore, the previous works [6, 7, 16] cannot be used to solve the network with size $N > 6$. However, Theorem 2 in this paper can be used to solve this problem. The sufficient condition in Theorem 2 is satisfied with more larger network size N. The time for solving the LMI in MATLAB Toolbox by using a normal computer and the corresponding network size are given in Figure 6.

Though the size of matrices $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{\Sigma}$ in Theorem 2 are of order $N \times N$, it can provide general and good results which is applicable for ensuring the synchronization of coupled networks with large size N. If the network size N is very large, the computation of LMI conditions is very difficult as in Figure 6 which should be further considered in the near future.

5. Conclusions. We have developed a generally linearly hybrid coupled network with time-varying delay and also further investigated its global synchronization. Based on this model, several effective sufficient conditions of global network synchronization are then deduced by using the Lyapunov function and an LMI. Both delay-independent and delay-

dependent sufficient conditions are attained. It should be especially pointed out that we do not assume that the coupling matrix is symmetric or diagonal. However, most of the former works on network synchronization are based on this assumption. Moreover, we also generalize the derivative of the time-varying delay to any given value in this paper, although most of the former results are based on assumption A_3. To verify the effectiveness of the proposed synchronous criteria, a small-world network, a regular network, and scale-free networks with increasing network size are finally constructed to show the global network synchronization.

The proposed network model builds a platform for the study of network synchronization and other network dynamical behaviors. These network synchronous criteria also provide some new insight for the underlying mechanics of network synchronization. Furthermore, there are some abundant dynamical behaviors in the network which deserve to be also further investigated in the near future, such as the relation between network structure and function, the individual and global dynamics, etc. We will also explore the possible applications for these criteria in the real-world biological and engineering networks.

Appendix A. Proof of Theorem 1.

Let

$$\mathbf{\Sigma} = I_N \otimes \Sigma, \quad \mathbf{\Sigma^1} = I_p \otimes \Sigma, \quad \mathbf{\Delta} = I_p \otimes \Delta, \quad \mathbf{P} = I_p \otimes P,$$

$$\mathbf{Q} = I_p \otimes Q, \quad \mathbf{\Omega} = I_p \otimes \Omega, \quad \mathbf{I_n} = I_N \otimes I_n, \quad \mathbf{I_n^1} = I_p \otimes I_n.$$

Let $y(t) = \mathbf{M}x(t) = (y_1^T(t), y_2^T(t), \ldots, y_p^T(t))^T$, $y_i(t) = (y_{i1}(t), y_{i2}(t), \ldots, y_{in}(t))^T$, $i = 1, 2, \ldots, p$.

According to (13), there exists a sufficiently small $\varepsilon > 0$, such that

$$\widetilde{\Lambda_0} = \begin{pmatrix} -2P(C - \varepsilon I_n) - \Delta + F\Sigma F & PA & PB \\ A^T P & -\Sigma + Q & 0 \\ B^T P & 0 & -(1-h)Q \end{pmatrix} < 0.$$

(i) Consider the Lyapunov candidate

$$\begin{aligned} V(t) = {} & e^{2\varepsilon t} x^T(t) \mathbf{M}^T \mathbf{P} \mathbf{M} x(t) + \int_{t-\tau(t)}^{t} e^{2\varepsilon s} \mathbf{f}^T(x(s)) \mathbf{M}^T \mathbf{Q} \mathbf{M} \mathbf{f}(x(s))\, ds \\ & + \int_{t-\tau(t)}^{t} e^{2\varepsilon s} x^T(s) \mathbf{M}^T \mathbf{\Omega} \mathbf{M} x(s)\, ds. \end{aligned} \tag{A.1}$$

Differentiating $V(t)$ along the trajectories of (12) yields

(A.2)
$$\begin{aligned}\dot{V}(t)|_{(12)} &= 2\varepsilon e^{2\varepsilon t}x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}x(t) + 2e^{2\varepsilon t}x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}\dot{x}(t) + e^{2\varepsilon t}\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t))\\ &\quad - (1-\dot{\tau}(t))e^{2\varepsilon(t-\tau(t))}\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t)))\\ &\quad + e^{2\varepsilon t}x^T(t)\mathbf{M}^T\mathbf{\Omega}\mathbf{M}x(t)\\ &\quad - (1-\dot{\tau}(t))e^{2\varepsilon(t-\tau(t))}x^T(t-\tau(t))\mathbf{M}^T\mathbf{\Omega}\mathbf{M}x(t-\tau(t))\\ &\le 2e^{2\varepsilon t}x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}[-(\mathbf{C}-\varepsilon\mathbf{I_n})x(t) + \mathbf{A}\mathbf{f}(x(t)) + \mathbf{B}\mathbf{f}(x(t-\tau(t))) + \mathbf{I}(t)\\ &\quad + \mathbf{G}x(t) + \mathbf{G}_\tau x(t-\tau(t))] + e^{2\varepsilon t}\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t))\\ &\quad - (1-h)e^{2\varepsilon t}\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t)))\\ &\quad + e^{2\varepsilon t}x^T(t)\mathbf{M}^T(\mathbf{\Omega}+\mathbf{\Delta}-\mathbf{\Delta})\mathbf{M}x(t)\\ &\quad - (1-h)e^{2\varepsilon t}x^T(t-\tau(t))\mathbf{M}^T\mathbf{\Omega}\mathbf{M}x(t-\tau(t)).\end{aligned}$$

From the definition of $\mathbf{M}$, one gets

$$\mathbf{MC} = \mathbf{C^1M}, \quad \mathbf{MA} = \mathbf{A^1M}, \quad \mathbf{MB} = \mathbf{B^1M}, \quad \mathbf{MI_n} = \mathbf{I_n^1M}, \quad \mathbf{MI}(t) = 0.$$

Therefore, one has

(A.3)
$$\begin{aligned}\dot{V}(t)|_{(12)} &\le 2e^{2\varepsilon t}x^T(t)\mathbf{M}^T\mathbf{P}[-(\mathbf{C^1}-\varepsilon\mathbf{I_n^1})\mathbf{M}x(t) + \mathbf{A^1}\mathbf{M}\mathbf{f}(x(t)) + \mathbf{B^1}\mathbf{M}\mathbf{f}(x(t-\tau(t)))\\ &\quad + \mathbf{M}\mathbf{G}x(t) + \mathbf{M}\mathbf{G}_\tau x(t-\tau(t))] + e^{2\varepsilon t}\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t))\\ &\quad - (1-h)e^{2\varepsilon t}\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t)))\\ &\quad + e^{2\varepsilon t}x^T(t)\mathbf{M}^T(\mathbf{\Omega}+\mathbf{\Delta}-\mathbf{\Delta})\mathbf{M}x(t)\\ &\quad - (1-h)e^{2\varepsilon t}x^T(t-\tau(t))\mathbf{M}^T\mathbf{\Omega}\mathbf{M}x(t-\tau(t)).\end{aligned}$$

According to assumption A_2, one gets

(A.4)
$$\begin{aligned}\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{\Sigma}\mathbf{M}\mathbf{f}(x(t)) &= \textstyle\sum_{j=1}^p \alpha_j^2[f(x_{j_1}(t)) - f(x_{j_2}(t))]^T\Sigma[f(x_{j_1}(t)) - f(x_{j_2}(t))]\\ &\le \textstyle\sum_{j=1}^p y_j^T(t)F\Sigma F y_j(t).\end{aligned}$$

Let

$$\xi_j = \begin{pmatrix} y_j^T(t) & \alpha_j(f(x_{j_1}(t)) - f(x_{j_2}(t)))^T & \alpha_j(f(x_{j_1}(t-\tau(t))) - f(x_{j_2}(t-\tau(t))))^T \end{pmatrix}^T.$$

It follows from (A.3)–(A.4) that

(A.5)
$$\begin{aligned}
\dot{V}(t)|_{(12)} &\le e^{2\varepsilon t}\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda}_0\xi_j + 2e^{2\varepsilon t}x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}[\mathbf{G}x(t)+\mathbf{G}_\tau x(t-\tau(t))] + e^{2\varepsilon t}x^T(t)\mathbf{M}^T\\
&\quad\times(\mathbf{\Omega}+\mathbf{\Delta})\mathbf{M}x(t) - (1-h)e^{2\varepsilon t}x^T(t-\tau(t))\mathbf{M}^T\mathbf{\Omega}\mathbf{M}x(t-\tau(t))\\
&= e^{2\varepsilon t}\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda}_0\xi_j + 2e^{2\varepsilon t}x^T(t)(\widetilde{M}^T\otimes I_n)(I_p\otimes P)(\widetilde{M}\otimes I_n)(G\otimes D)x(t)\\
&\quad + 2e^{2\varepsilon t}x^T(t)(\widetilde{M}^T\otimes I_n)(I_p\otimes P)(\widetilde{M}\otimes I_n)(G\otimes D_\tau)x(t-\tau(t))\\
&\quad + e^{2\varepsilon t}x^T(t)(\widetilde{M}^T\otimes I_n)(I_p\otimes(\Omega+\Delta))(\widetilde{M}\otimes I_n)x(t)\\
&\quad - (1-h)e^{2\varepsilon t}x^T(t-\tau(t))(\widetilde{M}^T\otimes I_n)(I_p\otimes\Omega)(\widetilde{M}\otimes I_n)x(t-\tau(t))\\
&= e^{2\varepsilon t}\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda}_0\xi_j + 2e^{2\varepsilon t}x^T(t)(\widetilde{M}^T\widetilde{M}G\otimes PD)x(t) + 2e^{2\varepsilon t}x^T(t)(\widetilde{M}^T\widetilde{M}G\otimes PD_\tau)\\
&\quad\times x(t-\tau(t)) + e^{2\varepsilon t}x^T(t)(\widetilde{M}^T\widetilde{M}\otimes(\Omega+\Delta))x(t) - (1-h)e^{2\varepsilon t}x^T(t-\tau(t))\\
&\quad\times(\widetilde{M}^T\widetilde{M}\otimes\Omega)x(t-\tau(t))\\
&= e^{2\varepsilon t}\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda}_0\xi_j - 2e^{2\varepsilon t}x^T(t)(TG\otimes PD)x(t) - 2e^{2\varepsilon t}x^T(t)(TG\otimes PD_\tau)x(t-\tau(t))\\
&\quad - e^{2\varepsilon t}x^T(t)(T\otimes(\Omega+\Delta))x(t) + (1-h)e^{2\varepsilon t}x^T(t-\tau(t))(T\otimes\Omega)x(t-\tau(t))\\
&= e^{2\varepsilon t}\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda}_0\xi_j - 2e^{2\varepsilon t}\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t)(\sum_{k=1}^{N}T_{ik}G_{kj}PD)x_j(t)\\
&\quad - 2e^{2\varepsilon t}\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t)(\sum_{k=1}^{N}T_{ik}G_{kj}PD_\tau)x_j(t-\tau(t)) - e^{2\varepsilon t}\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t)T_{ij}\\
&\quad\times(\Omega+\Delta)x_j(t) + (1-h)e^{2\varepsilon t}\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t-\tau(t))T_{ij}\Omega x_j(t-\tau(t)).
\end{aligned}$$

Denote $L_{ij}=\sum_{k=1}^{N}T_{ik}G_{kj}$. Then one obtains

(A.6)
$$\begin{aligned}
\sum_{j=1}^{N}L_{ij} &= \sum_{j=1,\,j\neq i}^{N}L_{ij} + L_{ii}\\
&= \sum_{j=1,\,j\neq i}^{N}\sum_{k=1}^{N}T_{ik}G_{kj} + \sum_{k=1}^{N}T_{ik}G_{ki}\\
&= \sum_{k=1}^{N}(\sum_{j=1,\,j\neq i}^{N}T_{ik}G_{kj} + T_{ik}G_{ki})\\
&= \sum_{k=1}^{N}T_{ik}(\sum_{j=1}G_{kj}) = 0.
\end{aligned}$$

Thus one has

(A.7)
$$L_{ii} = -\sum_{j=1,\,j\neq i}^{N}L_{ij},\quad i=1,2,\ldots,N.$$

Let $\eta_{ij} = (x_i^T(t) - x_j^T(t),\, x_i^T(t-\tau(t)) - x_j^T(t-\tau(t)))^T$. From (A.5), one gets

(A.8)
$$\begin{aligned}
\dot{V}(t)|_{(12)} \le{}& e^{2\varepsilon t}\textstyle\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda_0}\xi_j - 2e^{2\varepsilon t}\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t)(L_{ij}PD)x_j(t)\\
&- 2e^{2\varepsilon t}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t)(L_{ij}PD_\tau)x_j(t-\tau(t))\\
&- e^{2\varepsilon t}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t)T_{ij}(\Omega+\Delta)x_j(t)\\
&+ (1-h)e^{2\varepsilon t}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{N}x_i^T(t-\tau(t))T_{ij}\Omega x_j(t-\tau(t))\\
={}& e^{2\varepsilon t}\textstyle\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda_0}\xi_j - 2e^{2\varepsilon t}\sum_{i=1}^{N}\Big(\sum_{j=1,\,j\neq i}^{N}x_i^T(t)(L_{ij}PD)x_j(t) + x_i^T(t)L_{ii}PDx_i(t)\Big)\\
&- 2e^{2\varepsilon t}\textstyle\sum_{i=1}^{N}\Big(\sum_{j=1,\,j\neq i}^{N}x_i^T(t)(L_{ij}PD_\tau)x_j(t-\tau(t)) + x_i^T(t)L_{ii}PD_\tau x_i(t-\tau(t))\Big)\\
&- e^{2\varepsilon t}\textstyle\sum_{i=1}^{N}\Big(\sum_{j=1,\,j\neq i}^{N}x_i^T(t)T_{ij}(\Omega+\Delta)x_j(t) + x_i^T(t)T_{ii}(\Omega+\Delta)x_i(t)\Big) + (1-h)\\
&\times e^{2\varepsilon t}\textstyle\sum_{i=1}^{N}\Big(\sum_{j=1,\,j\neq i}^{N}x_i^T(t-\tau(t))T_{ij}\Omega x_j(t-\tau(t)) + x_i^T(t-\tau(t))T_{ii}\Omega x_i(t-\tau(t))\Big)\\
={}& e^{2\varepsilon t}\textstyle\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda_0}\xi_j + 2e^{2\varepsilon t}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(x_i(t)-x_j(t))^T(L_{ij}PD)(x_i(t)-x_j(t))\\
&+ 2e^{2\varepsilon t}\textstyle\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(x_i(t)-x_j(t))^T(L_{ij}PD_\tau)(x_i(t-\tau(t))-x_j(t-\tau(t)))\\
&+ e^{2\varepsilon t}\textstyle\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(x_i(t)-x_j(t))^T T_{ij}(\Omega+\Delta)(x_i(t)-x_j(t)) - (1-h)\\
&\times e^{2\varepsilon t}\textstyle\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(x_i(t-\tau(t))-x_j(t-\tau(t)))^T T_{ij}\Omega(x_i(t-\tau(t))-x_j(t-\tau(t)))\\
={}& e^{2\varepsilon t}\textstyle\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda_0}\xi_j + e^{2\varepsilon t}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\eta_{ij}^T\Lambda_{ij}\eta_{ij}.
\end{aligned}$$

According to Lemma 4 and (A.8), under conditions (13)–(14), $\dot{V}(y(t)) \le 0$ and $V(t) \le V(0)$. That is, $V(t)$ is a bounded function and $\|y(t)\| = O(e^{-\varepsilon t})$. This completes the proof of part (i).

(ii) Let $y(t) = \mathbf{M}x(t) = (y_1^T(t), y_2^T(t), \ldots, y_p^T(t))^T$, $y_i(t) = (y_{i1}(t), y_{i2}(t), \ldots, y_{in}(t))^T$, $i = 1, 2, \ldots, p$, $\widetilde{x}_j(t) = (x_{1j}(t), x_{2j}(t), \ldots, x_{Nj}(t))^T$, and $\widetilde{y}_j(t) = (y_{1j}(t), y_{2j}(t), \ldots, y_{pj}(t))^T$. Then $\widetilde{y}_j(t) = \widetilde{M}\widetilde{x}_j(t)$ for $j = 1, 2, \ldots, n$.

Following the same method in part (i), from (A.5), one has

(A.9)
$$\begin{aligned}
\dot{V}(t)|_{(12)} \le{}& e^{2\varepsilon t}\textstyle\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda_0}\xi_j - 2e^{2\varepsilon t}x^T(t)(TG\otimes PD)x(t) - 2e^{2\varepsilon t}x^T(t)(TG\otimes PD_\tau)x(t-\tau(t))\\
&- e^{2\varepsilon t}x^T(t)(T\otimes(\Omega+\Delta))x(t) + (1-h)e^{2\varepsilon t}x^T(t-\tau(t))(T\otimes\Omega)x(t-\tau(t))\\
={}& e^{2\varepsilon t}\textstyle\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda_0}\xi_j - 2e^{2\varepsilon t}\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{x}_i^T(t)(\sum_{k=1}^{n}p_{ik}d_{kj}TG)\widetilde{x}_j(t)\\
&- 2e^{2\varepsilon t}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{x}_i^T(t)(\sum_{k=1}^{n}p_{ik}d_{\tau kj}TG)\widetilde{x}_j(t-\tau(t)) - e^{2\varepsilon t}\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{x}_i^T(t)\\
&\times(\Omega_{ij}+\Delta_{ij})T\widetilde{x}_j(t) + (1-h)e^{2\varepsilon t}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{x}_i^T(t-\tau(t))\Omega_{ij}T\widetilde{x}_j(t-\tau(t)).
\end{aligned}$$

Let $\widetilde{\eta}_{ij} = (\widetilde{x}_i^T(t) - \widetilde{x}_j^T(t),\, \widetilde{x}_i^T(t-\tau(t)) - \widetilde{x}_j^T(t-\tau(t)))^T$. According to (A.9), one obtains

$$\dot{V}(t)|_{(12)} \le e^{2\varepsilon t}\sum_{j=1}^{p}\xi_j^T\widetilde{\Lambda_0}\xi_j + e^{2\varepsilon t}\sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{\eta}_{ij}^T\widetilde{\Lambda}_{ij}\widetilde{\eta}_{ij}. \tag{A.10}$$

From Lemma 4 and (A.10), under conditions (13) and (15), $\dot{V}(y(t)) \le 0$ and $V(t) \le V(0)$. That is, $V(t)$ is a bounded function and $\|y(t)\| = O(e^{-\varepsilon t})$. This completes the proof of part (ii).

Appendix B. Proof of Theorem 2. Consider the following Lyapunov candidate:

(B.1)
$$V(t) = x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}x(t) + \int_{t-\tau(t)}^{t} \mathbf{f}^T(x(s))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(s))\,ds + \int_{t-\tau(t)}^{t} x^T(s)\mathbf{M}^T\mathbf{R}\mathbf{M}x(s)\,ds.$$

Differentiating $V(t)$ along the trajectories of (12) yields

(B.2)
$$\begin{aligned}\dot{V}(t)|_{(12)} &= 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}\dot{x}(t) + \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t)) - (1-\dot{\tau}(t))\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\\ &\quad\times \mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t))) + x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t)\\ &\quad - (1-\dot{\tau}(t))x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t))\\ &\le 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}[-\mathbf{C}x(t) + \mathbf{A}\mathbf{f}(x(t)) + \mathbf{B}\mathbf{f}(x(t-\tau(t))) + \mathbf{I}(t) + \mathbf{G}x(t)\\ &\qquad + \mathbf{G}_\tau x(t-\tau(t))]\\ &\quad + \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t)) - (1-h)\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t)))\\ &\quad + x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t) - (1-\dot{\tau}(t))x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t)).\end{aligned}$$

From the definition of $\mathbf{M}$, one has

$$\mathbf{M}\mathbf{C} = \mathbf{C}^{\mathbf{1}}\mathbf{M}, \quad \mathbf{M}\mathbf{A} = \mathbf{A}^{\mathbf{1}}\mathbf{M}, \quad \mathbf{M}\mathbf{B} = \mathbf{B}^{\mathbf{1}}\mathbf{M}, \quad \mathbf{M}\mathbf{I}(t) = 0.$$

Therefore, one obtains

(B.3)
$$\begin{aligned}\dot{V}(t)|_{(12)} &\le 2x^T(t)\mathbf{M}^T\mathbf{P}[-\mathbf{C}^{\mathbf{1}}\mathbf{M}x(t) + \mathbf{A}^{\mathbf{1}}\mathbf{M}\mathbf{f}(x(t)) + \mathbf{B}^{\mathbf{1}}\mathbf{M}\mathbf{f}(x(t-\tau(t))) + \mathbf{M}\mathbf{G}x(t)\\ &\qquad + \mathbf{G}_\tau x(t-\tau(t))]\\ &\quad + \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t)) - (1-h)\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t)))\\ &\quad + x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t) - (1-h)x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t)).\end{aligned}$$

According to assumption A_2, one gets

(B.4)
$$\begin{aligned}\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{\Sigma}\mathbf{M}\mathbf{f}(x(t)) &= \textstyle\sum_{j=1}^{N-1}[f(x_j(t)) - f(x_{j+1}(t))]^T\mathbf{\Sigma}_j[f(x_j(t)) - f(x_{j+1}(t))]\\ &\le \textstyle\sum_{j=1}^{N-1}[x_j(t) - x_{j+1}(t)]^T F\mathbf{\Sigma}_j F[x_j(t) - x_{j+1}(t)]\\ &= x^T(t)\mathbf{M}^T\mathbf{F}\mathbf{\Sigma}\mathbf{F}\mathbf{M}x(t),\end{aligned}$$

where $\mathbf{\Sigma}_j = \operatorname{diag}(\Sigma_{(j-1)n+1}, \ldots, \Sigma_{jn})$. From Lemmas 1 and 5, one has

(B.5)
$$\begin{aligned}2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}\mathbf{G}x(t) &= 2x^T(t)\mathbf{M}^T\mathbf{P}[(M\otimes I_n)(G\otimes D)]x(t)\\ &= 2x^T(t)\mathbf{M}^T\mathbf{P}[MG\otimes D]x(t)\\ &= 2x^T(t)\mathbf{M}^T\mathbf{P}[HM\otimes D]x(t)\\ &= 2x^T(t)\mathbf{M}^T\mathbf{P}[(H\otimes D)(M\otimes I_n)]x(t)\\ &= 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{H}\mathbf{M}x(t)\end{aligned}$$

and

$$
\begin{aligned}
2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}\mathbf{G}_\tau x(t-\tau(t)) &= 2x^T(t)\mathbf{M}^T\mathbf{P}[(M\otimes I_n)(G\otimes D_\tau)]x(t-\tau(t)) \\
&= 2x^T(t)\mathbf{M}^T\mathbf{P}[MG\otimes D_\tau]x(t-\tau(t)) \\
&= 2x^T(t)\mathbf{M}^T\mathbf{P}[HM\otimes D_\tau]x(t-\tau(t)) \\
&= 2x^T(t)\mathbf{M}^T\mathbf{P}[(H\otimes D_\tau)(M\otimes I_n)]x(t-\tau(t)) \\
&= 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{H}_\tau\mathbf{M}x(t-\tau(t)),
\end{aligned} \tag{B.6}
$$

where $H = MGJ$, $\mathbf{H} = (MGJ)\otimes D$, $\mathbf{H}_\tau = (MGJ)\otimes D_\tau$, and M and J are defined in (9).

According to (B.3)–(B.6), one obtains

$$
\begin{aligned}
\dot{V}(t)|_{(12)} \le\ & -2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{C}^{\mathbf{1}}\mathbf{M}x(t) + 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{A}^{\mathbf{1}}\mathbf{M}\mathbf{f}(x(t)) \\
& + 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{B}^{\mathbf{1}}\mathbf{M}\mathbf{f}(x(t-\tau(t))) \\
& + 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{H}\mathbf{M}x(t) + 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{H}_\tau\mathbf{M}x(t-\tau(t)) + x^T(t)\mathbf{M}^T\mathbf{F}\mathbf{\Sigma}\mathbf{F}\mathbf{M}x(t) \\
& - \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{\Sigma}\mathbf{M}\mathbf{f}(x(t)) + \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t)) - (1-h)\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T \\
& \times \mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t))) + x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t) - (1-h)x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t)) \\
=\ & \eta^T(t)\mathbf{\Omega}\eta(t),
\end{aligned} \tag{B.7}
$$

where

$$
\eta(t) = \begin{pmatrix} x^T(t)\mathbf{M}^T & x^T(t-\tau(t))\mathbf{M}^T & \mathbf{f}^T(x(t))\mathbf{M}^T & \mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T \end{pmatrix}^T.
$$

From Lemma 4 and (B.7), under the condition (28), $\dot{V}(t) \le 0$ and $V(t) \le V(0)$. That is, $V(t)$ is a bounded function and $\|\mathbf{M}x(t)\| \longrightarrow 0$. This proof is thus completed.

Appendix C. Proof of Corollary 7. Select the Lyapunov candidate (B.1), where M and J are also defined in (9), $\mathbf{P} = I_{N-1}\otimes P$, $\mathbf{Q} = I_{N-1}\otimes Q$, $\mathbf{R} = I_{N-1}\otimes R$, and $\mathbf{\Sigma} = I_{N-1}\otimes\Sigma$.

From (B.7), one obtains

$$
\begin{aligned}
\dot{V}(t)|_{(12)} &\le \eta^T(t)\mathbf{\Omega}\eta(t) \\
&= \textstyle\sum_{j=1}^{N-1}\xi_j^T\Lambda_0\xi_j + \left(x^T(t)\mathbf{M}^T, x^T(t-\tau(t))\mathbf{M}^T\right)\Xi\left(\mathbf{M}x^T(t), \mathbf{M}x^T(t-\tau(t))\right)^T,
\end{aligned} \tag{C.1}
$$

where

$$
\xi_j = \begin{pmatrix} (x_j(t)-x_{j+1}(t))^T & (f(x_j(t))-f(x_{j+1}(t)))^T & (f(x_j(t-\tau(t)))-f(x_{j+1}(t-\tau(t))))^T \end{pmatrix}^T
$$

and $\mathbf{\Omega}$ is defined in (28). According to Lemma 4 and (C.1), under the conditions (31)–(32), $\dot{V}(t) \le 0$ and $V(t) \le V(0)$. That is, $V(t)$ is a bounded function, and $\|\mathbf{M}x(t)\| \longrightarrow 0$. This

completes the proof.

Appendix D. Proof of Theorem 3. Construct the Lyapunov function as follows:

(D.1)
$$V(t) = x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}x(t) + \int_{t-\tau(t)}^{t} \mathbf{f}^T(x(s))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(s))\,ds + \int_{t-\tau(t)}^{t} x^T(s)\mathbf{M}^T\mathbf{R}\mathbf{M}x(s)\,ds + \int_{-r}^{0} d\theta \int_{t+\theta}^{t} \dot{x}^T(s)\mathbf{M}^T\mathbf{T}\mathbf{M}\dot{x}(s)\,ds.$$

From Lemma 6, differentiating $V(t)$ along the trajectories of (12) results in

(D.2)
$$\begin{aligned}\dot{V}(t)|_{(12)} &= 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}\dot{x}(t) + \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t)) - (1-\dot{\tau}(t))\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T \\ &\quad \times \mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t))) + x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t) \\ &\quad - (1-\dot{\tau}(t))x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t)) \\ &\quad + r\dot{x}^T(t)\mathbf{M}^T\mathbf{T}\mathbf{M}\dot{x}(t) - \int_{t-r}^{t} \dot{x}^T(\theta)\mathbf{M}^T\mathbf{T}\mathbf{M}\dot{x}(\theta)\,d\theta \\ &\le 2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{M}[-\mathbf{C}x(t) + \mathbf{A}\mathbf{f}(x(t)) + \mathbf{B}\mathbf{f}(x(t-\tau(t))) + \mathbf{I}(t) + \mathbf{G}x(t) \\ &\qquad + \mathbf{G}_\tau x(t-\tau(t))] \\ &\quad + \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t)) - (1-h)\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t))) \\ &\quad + x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t) - (1-h)x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t)) + r\dot{x}^T(t)\mathbf{M}^T\mathbf{T}\mathbf{M}\dot{x}(t) \\ &\quad - \frac{1}{r}\left(\int_{t-\tau(t)}^{t} \mathbf{M}\dot{x}(\theta)\,d\theta\right)^T \mathbf{T}\left(\int_{t-\tau(t)}^{t} \mathbf{M}\dot{x}(\theta)\,d\theta\right).\end{aligned}$$

From the definition of $\mathbf{M}$, one has

$$\mathbf{M}\mathbf{C} = \mathbf{C}^{\mathbf{1}}\mathbf{M}, \quad \mathbf{M}\mathbf{A} = \mathbf{A}^{\mathbf{1}}\mathbf{M}, \quad \mathbf{M}\mathbf{B} = \mathbf{B}^{\mathbf{1}}\mathbf{M}, \quad \mathbf{M}\mathbf{I}(t) = 0.$$

According to (D.2), one obtains

(D.3)
$$\begin{aligned}\dot{V}(t)|_{(12)} &\le 2x^T(t)\mathbf{M}^T\mathbf{P}[-\mathbf{C}^{\mathbf{1}}\mathbf{M}x(t) + \mathbf{A}^{\mathbf{1}}\mathbf{M}\mathbf{f}(x(t)) + \mathbf{B}^{\mathbf{1}}\mathbf{M}\mathbf{f}(x(t-\tau(t))) + \mathbf{H}\mathbf{M}x(t) \\ &\qquad + \mathbf{H}_\tau\mathbf{M}x(t-\tau(t))] \\ &\quad + \mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t)) - (1-h)\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t-\tau(t))) \\ &\quad + x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t) - (1-h)x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t)) + r\dot{x}^T(t)\mathbf{M}^T\mathbf{T}\mathbf{M}\dot{x}(t) \\ &\quad - \frac{1}{r}\left(\int_{t-\tau(t)}^{t} \mathbf{M}\dot{x}(\theta)\,d\theta\right)^T \mathbf{T}\left(\int_{t-\tau(t)}^{t} \mathbf{M}\dot{x}(\theta)\,d\theta\right).\end{aligned}$$

Similar to (B.4), one has

$$\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{\Sigma}\mathbf{M}\mathbf{f}(x(t)) \le x^T(t)\mathbf{M}^T\mathbf{F}\mathbf{\Sigma}\mathbf{F}\mathbf{M}x(t) \tag{D.4}$$

and

$$\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{\Lambda}\mathbf{M}\mathbf{f}(x(t-\tau(t))) \le x^T(t-\tau(t))\mathbf{M}^T\mathbf{F}\mathbf{\Lambda}\mathbf{F}\mathbf{M}x(t-\tau(t)). \tag{D.5}$$

From the Leibniz–Newton formula, for any matrix $\mathbf{U}$ with appropriate dimensions, one gets

$$x^T(t-\tau(t))\mathbf{M}^T\mathbf{U}\mathbf{M}\left(x(t)-x(t-\tau(t))-\int_{t-\tau(t)}^{t}\dot{x}(s)\,ds\right)=0. \tag{D.6}$$

Let

$$\xi(t)=\begin{pmatrix} x^T(t)\mathbf{M}^T & x^T(t-\tau(t))\mathbf{M}^T & \mathbf{f}^T(x(t))\mathbf{M}^T & \mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T & \int_{t-\tau(t)}^{t}\dot{x}^T(s)\,ds\,\mathbf{M}^T \end{pmatrix}^T,$$
$$\Pi=\begin{pmatrix} -\mathbf{C^1}+\mathbf{H} & \mathbf{H}_\tau & \mathbf{A^1} & \mathbf{B^1} & 0 \end{pmatrix};$$

then one has

$$\dot{x}^T(t)\mathbf{M}^T\mathbf{T}\mathbf{M}\dot{x}(t)=\xi^T(t)\Pi^T\mathbf{T}\Pi\xi(t). \tag{D.7}$$

Combining this with (D.3)–(D.7), one obtains

(D.8)
$$\begin{aligned}\dot{V}(t)|_{(12)} \le{}& -2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{C^1}\mathbf{M}x(t)+2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{A^1}\mathbf{M}\mathbf{f}(x(t))+2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{B^1}\mathbf{M}\mathbf{f}(x(t-\tau(t)))\\
&+2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{H}\mathbf{M}x(t)+2x^T(t)\mathbf{M}^T\mathbf{P}\mathbf{H}_\tau\mathbf{M}x(t-\tau(t))+x^T(t)\mathbf{M}^T\mathbf{F}\mathbf{\Sigma}\mathbf{F}\mathbf{M}x(t)\\
&-\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{\Sigma}\mathbf{M}\mathbf{f}(x(t))+x^T(t-\tau(t))\mathbf{M}^T\mathbf{F}\mathbf{\Lambda}\mathbf{F}\mathbf{M}x(t-\tau(t))-\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\\
&\times\mathbf{\Lambda}\mathbf{M}\mathbf{f}(x(t-\tau(t)))+\mathbf{f}^T(x(t))\mathbf{M}^T\mathbf{Q}\mathbf{M}\mathbf{f}(x(t))-(1-h)\mathbf{f}^T(x(t-\tau(t)))\mathbf{M}^T\mathbf{Q}\mathbf{M}\\
&\times\mathbf{f}(x(t-\tau(t)))+x^T(t)\mathbf{M}^T\mathbf{R}\mathbf{M}x(t)-(1-h)x^T(t-\tau(t))\mathbf{M}^T\mathbf{R}\mathbf{M}x(t-\tau(t))\\
&+r\xi^T(t)\Pi^T\mathbf{T}\Pi\xi(t)-\frac{1}{r}\left(\int_{t-\tau(t)}^{t}\mathbf{M}\dot{x}(\theta)\,d\theta\right)^T\mathbf{T}\left(\int_{t-\tau(t)}^{t}\mathbf{M}\dot{x}(\theta)\,d\theta\right)\\
&+2x^T(t-\tau(t))\mathbf{M}^T\mathbf{U}\mathbf{M}\left(x(t)-x(t-\tau(t))-\int_{t-\tau(t)}^{t}\dot{x}(s)\,ds\right)\\
={}&\xi^T(t)(\mathbf{\Xi}+r\Pi^T\mathbf{T}\Pi)\xi(t),\end{aligned}$$

where

$$\mathbf{\Xi}=\begin{pmatrix} \Psi & \mathbf{P}\mathbf{H}_\tau+\mathbf{U}^T & \mathbf{P}\mathbf{A^1} & \mathbf{P}\mathbf{B^1} & 0\\ \mathbf{H}_\tau^T\mathbf{P}+\mathbf{U} & -(1-h)\mathbf{R}-2\mathbf{U}+\mathbf{F}\mathbf{\Lambda}\mathbf{F} & 0 & 0 & -\mathbf{U}\\ \mathbf{A^1}^T\mathbf{P} & 0 & -\mathbf{\Sigma}+\mathbf{Q} & 0 & 0\\ \mathbf{B^1}^T\mathbf{P} & 0 & 0 & -(1-h)\mathbf{Q}-\mathbf{\Lambda} & 0\\ 0 & -\mathbf{U}^T & 0 & 0 & -\frac{1}{r}\mathbf{T} \end{pmatrix}$$

and $\Psi=-2\mathbf{P}\mathbf{C^1}+\mathbf{P}\mathbf{H}+\mathbf{H}^T\mathbf{P}+\mathbf{R}+\mathbf{F}\mathbf{\Sigma}\mathbf{F}$.

According to Schur complement Lemma 2, $\mathbf{\Xi}+r\Pi^T\mathbf{T}\Pi<0$ is equivalent to $\mathbf{\Omega_1}<0$. From Lemma 4 and (D.8), under the condition (33), $\dot{V}(t)\le 0$ and $V(t)\le V(0)$. That is, $V(t)$ is a bounded function, and $\|\mathbf{M}x(t)\|\longrightarrow 0$. Thus the proof is completed.

REFERENCES

[1] S. H. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.

[2] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442.

[3] C. Li and G. Chen, *Stability of a neural network model with small-world connections*, Phys. Rev. E (3), 68 (2003), 052901.

[4] X. Wang and G. Chen, *Synchronization in scale-free dynamical networks: Robustness and fragility*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 54–62.

[5] X. Wang and G. Chen, *Synchronization in small-world dynamical networks*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002), pp. 187–192.

[6] W. Lu and T. Chen, *Synchronization of coupled connected neural networks with delays*, IEEE Trans. Circuits Syst. I Regul. Pap., 51 (2004), pp. 2491–2503.

[7] W. Wang and J. Cao, *Synchronization in an array of linearly coupled networks with time-varying delay*, Phys. A, 366 (2006), pp. 197–211.

[8] G. Chen, J. Zhou, and Z. Liu, *Global synchronization of coupled delayed neural networks and applications to chaotic CNN models*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 14 (2004), pp. 2229–2240.

[9] C. Wu and L. O. Chua, *Synchronization in an array of linearly coupled dynamical systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 42 (1995), pp. 430–447.

[10] C. Wu, *Synchronization in arrays of coupled nonlinear systems with delay and nonreciprocal time-varying coupling*, IEEE Trans. Circuits Systems II, 52 (2005), pp. 282–286.

[11] C. Li, S. Li, X. Liao, and J. Yu, *Synchronization in coupled map lattices with small-world delayed interactions*, Phys. A, 335 (2004), pp. 365–370.

[12] C. Li, H. Xu, X. Liao, and J. Yu, *Synchronization in small-world oscillator networks with coupling delays*, Phys. A, 335 (2004), pp. 359–364.

[13] C. Li and G. Chen, *Synchronization in general complex dynamical networks with coupling delays*, Phys. A, 343 (2004), pp. 263–278.

[14] J. Lü, X. Yu, and G. Chen, *Chaos synchronization of general complex dynamical networks*, Phys. A, 334 (2004), pp. 281–302.

[15] Z. Li and G. Chen, *Robust adaptive synchronization of uncertain dynamical networks*, Phys. Lett. A, 324 (2004), pp. 166–178.

[16] J. Cao, P. Li, and W. Wang, *Global synchronization in arrays of delayed neural networks with constant and delayed coupling*, Phys. Lett. A, 353 (2006), pp. 318–325.

[17] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[18] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, NJ, 1999.

[19] J. Chen and X. Chen, *Special Matrices*, Tsinghua University Press, Beijing, China, 2001.

[20] W. Yu and J. Cao, *Adaptive synchronization and lag synchronization of uncertain dynamical system with time delay based on parameter identification*, Phys. A, 375 (2007), pp. 467–482.

[21] W. Yu and J. Cao, *Stability and Hopf bifurcation analysis on a four-neuron BAM neural network with time delays*, Phys. Lett. A, 351 (2006), pp. 64–78.

[22] J. Cao and J. Liang, *Boundedness and stability for Cohen-Grossberg neural networks with time-varying delays*, J. Math. Anal. Appl., 296 (2004), pp. 665–685.

[23] J. Cao and J. Wang, *Global asymptotic and robust stability of recurrent neural networks with time delays*, IEEE Trans. Circuits Syst. I Regul. Pap., 52 (2005), pp. 417–426.

[24] X. F. Liao, K. W. Wong, Z. Wu, and G. Chen, *Novel robust stability criteria for interval-delayed Hopfield neural networks*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 48 (2001), pp. 1355–1359.

[25] S. Q. Hu and J. Wang, *Global exponential stability of continuous-time interval neural networks*, Phys. Rev. E (3), 65 (2002), 036133.

[26] J. Cao and J. Wang, *Absolute exponential stability of recurrent neural networks with time delays and Lipschitz-continuous activation functions*, Neural Networks, 17 (2004), pp. 379–390.

[27] J. Cao, H. X. Li, and D. W. C. Ho, *Synchronization criteria of Lur'e systems with time-delay feedback control*, Chaos Solitons Fractals, 23 (2005), pp. 1285–1298.
[28] H. Lu, *Chaotic attractors in delayed neural networks*, Phys. Lett. A, 298 (2002), pp. 109–116.
[29] K. Q. Gu, V. L. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*, Birkhäuser Boston, Boston, MA, 2003.
[30] W. Yu and J. Cao, *Adaptive Q-S (lag, anticipated, and complete) time-varying synchronization and parameters identification of uncertain delayed neural networks*, Chaos, 16 (2006), 023119.
[31] J. Cao, W. Yu, and Y. Qu, *A new complex network model and convergence dynamics for reputation computation in virtual organizations*, Phys. Lett. A, 356 (2006), pp. 414–425.
[32] J. Zhou, J. A. Lu, and J. Lü, *Adaptive synchronization of an uncertain complex dynamical network*, IEEE Trans. Automat. Control, 51 (2006), pp. 652–656.
[33] J. Lü and G. Chen, *A time-varying complex dynamical network models and its controlled synchronization criteria*, IEEE Trans. Automat. Control, 50 (2005), pp. 841–846.
[34] J. Lü, X. Yu, G. Chen, and D. Cheng, *Characterizing the synchronizability of small-world dynamical networks*, IEEE Trans. Circuits Syst. I Regul. Pap., 51 (2004), pp. 787–796.
[35] W. Yu and J. Cao, *Synchronization control of stochastic delayed neural networks*, Phys. A, 373 (2007), pp. 252–260.
[36] W. Yu, J. Cao, and G. Chen, *Robust adaptive control of unknown modified Cohen-Grossberg neural networks with delay*, IEEE Trans. Circuits Systems II, 54 (2007), pp. 502–506.
[37] L. O. Chua, *The genesis of Chua's circuit*, Arch Elektron Ubertragung, 46 (1992), pp. 250–257.
[38] W. Yu, G. Chen, J. Cao, J. Lü, and U. Parlitz, *Parameter identification of dynamical systems from time series*, Phys. Rev. E (3), 75 (2007), 067201.
[39] W. Yu, J. Cao, K. W. Wong, and J. Lü, *New communication schemes based on adaptive synchronization*, Chaos, 17 (2007), 033114.
[40] A. L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509–512.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 134–160
© 2008 Society for Industrial and Applied Mathematics

Global Phase-Locking in Finite Populations of Phase-Coupled Oscillators*

Mark Verwoerd† and Oliver Mason†

Abstract. We present new necessary and sufficient conditions for the existence of fixed points in a finite system of coupled phase oscillators on a complete graph. We use these conditions to derive bounds on the critical coupling.

Key words. synchronization, coupled oscillators, fixed points

AMS subject classifications. 34C15, 35B32, 37C25

DOI. 10.1137/070686858

1. Introduction.

The phenomenon of synchronization arises in a wide variety of application areas across neuroscience, biology, engineering, and physics [6, 17, 2, 16, 5]. As such, the identification and study of structures and mechanisms that support the onset of synchronized behavior is a key issue in the theory of interconnected dynamical systems. In particular, there has been a great deal of interest across the mathematics, physics, and engineering communities in the development and analysis of simple mathematical models of synchronization [9, 19, 20, 3, 21, 22].

To date, one of the most widely studied frameworks for the analysis of synchronization is the so-called *Kuramoto model* of phase-coupled oscillators [10, 23]. In fact, this model has been used in numerous applications in the chemical and biological sciences, and its basic properties have been analyzed using a combination of numerical and analytical techniques [11, 23, 24, 1]. The basic Kuramoto model is comprised of a system of coupled oscillators, which may have different natural frequencies, where the coupling between two oscillators is given by a weighted sinusoidal function of the difference of their phases. The weights used in the model are typically taken to be the same for all pairs of oscillators and are given by the ratio of a fixed parameter, the coupling strength, to the network size.

The aspect of the Kuramoto model that has attracted the most attention to date is the manner in which the onset of synchronization depends on the strength of coupling between the oscillators. For instance, at very low values of the coupling strength, little or no synchronization is observed. As the coupling strength is increased, some partial synchronization appears in the network up to a threshold value of the coupling strength, referred to here as the critical coupling, at which fully synchronized behavior emerges [9, 3]. The mechanism of (de)synchronization in finite populations of oscillators has been described in considerable

*Received by the editors March 30, 2007; accepted for publication (in revised form) by B. Ermentrout July 20, 2007; published electronically January 16, 2008. This work was partially supported by Science Foundation Ireland (SFI) grant 03/RP1/I382 and SFI grant 04/IN1/I478. Science Foundation Ireland is not responsible for any use of data appearing in this publication.

http://www.siam.org/journals/siads/7-1/68685.html

†Hamilton Institute, National University of Ireland, Maynooth, Co. Kildare, Ireland (mark.verwoerd@nuim.ie, oliver.mason@nuim.ie).

detail [13, 12]. In particular, when the coupling strength drops below its critical value and as it continues to decrease, the system undergoes a series of so-called *frequency-splitting bifurcations*. At each such bifurcation, the ensemble of oscillators subdivides into smaller and smaller groups of oscillators with identical average frequency, until eventually all oscillators oscillate at their own intrinsic frequency. A detailed analysis of this behavior for a system with three oscillators was given in [13]. While the aforementioned contributions focus on the behavior of the system in the *sub*critical coupling regime, the present paper studies globally phase-locked solutions, which by definition exist only in the *super*critical coupling regime.

Another aspect of the Kuramoto model that has attracted attention recently is the emergence of phase chaos [14, 18] in systems of dimension four and higher. A generic feature of coupled oscillator systems, phase chaos in the Kuramoto model is most prominent in systems with relatively low dimension (comprising between ten and fifteen oscillators) [14]. Again, this phenomenon can exist only in the subcritical coupling regime, and we shall not further consider it here.

In the original Kuramoto model, it is assumed that all pairs of oscillators in the network are connected with the same coupling strength [10]. This type of coupling is referred to as "all-to-all" coupling and corresponds to a network in which the underlying graph is complete [4]. Extensions of the Kuramoto model to lattices [8] and rings [20] have also been considered, and more recently the dynamics of coupled oscillators on networks with small-world [27, 7] and scale-free [15] topologies have started to attract a lot of interest. More generally, there are many fundamental questions relating to the interplay between a network's topology and dynamical processes taking place on it which are still unanswered. The work described in [6], which proposes an extension of group-based symmetry, using so-called groupoid formalism, as a means of classifying possible behaviors for networked dynamical systems is particularly noteworthy in this context.

Many of the recent results concerned with the dynamics and synchronization of coupled oscillators have either been based on numerical simulations or else have been derived for the limiting case of networks of infinite size. In contrast, relatively few rigorous results are available for finite-size networks [23, 9]. In this paper, we shall be concerned with synchronization in finite systems of coupled oscillators. Specifically, we shall establish (new) necessary and sufficient conditions for the existence of fixed points in a finite system of coupled oscillators (see also [25, 26]), compute bounds on the critical coupling strength for such systems, and provide insight into the number of fixed points possible under strong coupling. Our analysis is in the spirit of the work presented in [9, 3] and places particular emphasis on the *existence* of fixed points. Of course, the stability of such fixed points is also a topic of great interest and has been considered in [9, 19, 20, 3]. However, we shall not explicitly address the question of stability in the current paper.

The outline of the paper is as follows. In section 2, we introduce the Kuramoto model and review some of its basic properties. Here, we also give a formal definition of critical coupling, which is essentially the lowest value of the coupling strength for which fixed points exist. In section 3, we show that fixed points will always exist for sufficiently strong coupling (essentially proving that the critical coupling is a finite number), and then, in section 4 we provide lower bounds on the critical coupling. Section 5 contains necessary and sufficient conditions for the existence of fixed points, which are then used in section 6 to describe an

algorithm for computing the critical coupling. Section 7 contains a numerical example to illustrate the results of the paper, and finally, in section 8 we present our concluding remarks.

2. Mathematical preliminaries and the Kuramoto model.

2.1. Basic notation.

Throughout the paper, $\mathbb{R}$ ($\mathbb{C}$) denotes the field of real (complex) numbers, $\mathbb{R}^N$ ($\mathbb{C}^N$) denotes the vector space of all N-tuples of real (complex) numbers, and $\mathbb{R}^{N\times N}$ ($\mathbb{C}^{N\times N}$) denotes the space of $N\times N$ matrices with entries in $\mathbb{R}$ ($\mathbb{C}$). i is used to denote the complex number satisfying $i^2=-1$. For a vector $x\in\mathbb{R}^N$, x_i denotes the ith entry of x. Also, $\mathbf{1}_N$ denotes the vector in $\mathbb{R}^N$, all of whose entries are equal to one.

We shall use V to denote the projection matrix in $\mathbb{R}^{N\times N}$ given by

$$[V_{ij}] := \begin{cases} \frac{N-1}{N}, & j=i, \\ -\frac{1}{N}, & j\neq i, \end{cases} \qquad i,j=1,\dots,N, \tag{1}$$

and $V\mathbb{R}^N$ shall denote the image of $\mathbb{R}^N$ under V. Formally,

$$V\mathbb{R}^N := \left\{ x\in\mathbb{R}^N : \sum_{j=1}^N x_j = 0 \right\}.$$

2.2. The basic Kuramoto model.

The basic Kuramoto model of phase-coupled oscillators under the assumption of all-to-all coupling is given by

$$\dot{\theta}_i = \omega_i + \frac{k}{N}\sum_{j=1}^N \sin(\theta_j-\theta_i), \quad i=1,\dots,N. \tag{2}$$

Here, $\theta_i(\cdot)\in\mathbb{R}$ (S^1) and $\omega_i\in\mathbb{R}$ respectively denote the phase and intrinsic (or natural) frequency of oscillator i, and the constant $k\in\mathbb{R}_+$ is a global coupling coefficient.

This model can be described more compactly in vector notation as

$$\dot{\theta} = \omega + kf(\theta), \tag{3}$$

where $\theta(t) := (\theta_1(t),\dots,\theta_N(t))$, $\omega := (\omega_1,\dots,\omega_N)$, and the mapping $f:\mathbb{R}^N\mapsto\mathbb{R}^N$ is given by

$$\begin{aligned} f(\xi) &= (f_1(\xi),\dots,f_N(\xi)), \\ f_i(\xi) &:= \frac{1}{N}\sum_{j=1}^N \sin(\xi_j-\xi_i), \quad 1\le i\le N. \end{aligned} \tag{4}$$

The assumption of all-to-all coupling is naturally very restrictive and ought to be relaxed in order for this work to be more directly applicable to the modeling of biological systems, or most engineering systems for that matter. Work toward this end is underway, and we hope to be able to present some results in the near future. Meanwhile, in this paper, we shall focus exclusively on configurations with all-to-all coupling. First, we recall some fundamental notions in the theory of synchronized oscillators.

2.3. The order parameter. Let $\mathbb{D}$ denote the complex unit disc $\{z \in \mathbb{C} : |z| \leq 1\}$. Then define $r : \mathbb{R}^N \mapsto \mathbb{D}$ by

$$r(\xi) := \frac{1}{N}\sum_{j=1}^{N} e^{i\xi_j}. \tag{5}$$

Let $r^{-1}(z) := \{\xi \in \mathbb{R}^N : r(\xi) = z\}$ denote the preimage of r, and note that the preimage is nonempty for all $z \in \mathbb{D}$ provided $N \geq 2$. We introduce the notation $\mathcal{R}_0 := r^{-1}(0)$. Then, for $\xi \in \mathbb{R}^N$, we may express $r(\xi)$ in polar coordinates:

$$r(\xi) = \begin{cases} R(\xi)e^{i\psi(\xi)}, & \xi \in \mathbb{R}^N \backslash \mathcal{R}_0, \\ 0, & \xi \in \mathcal{R}_0. \end{cases} \tag{6}$$

Here, $R : \mathbb{R}^N \mapsto [0, 1]$ and $\psi : \mathbb{R}^N \backslash \mathcal{R}_0 \mapsto [0, 2\pi)$ are respectively defined as

$$R(\xi) := \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N} \sin(\xi_j)\right)^2 + \left(\frac{1}{N}\sum_{j=1}^{N} \cos(\xi_j)\right)^2} \tag{7}$$

and

$$\psi(\xi) := \arctan\left(\frac{\frac{1}{N}\sum_{j=1}^{N} \sin(\xi_j)}{\frac{1}{N}\sum_{j=1}^{N} \cos(\xi_j)}\right). \tag{8}$$

The following properties of the maps $R(\cdot)$ and $\psi(\cdot)$ follow immediately from (5):

$$R(\xi + c\mathbf{1}_N) := \left|\frac{1}{N}\sum_{j=1}^{N} e^{i(\xi_j + c)}\right| = |e^{ic}|R(\xi) = R(\xi) \quad \forall \xi \in \mathbb{R}^N; \tag{9}$$

$$\psi(\xi + c\mathbf{1}_N) = \psi(\xi) + c \mod 2\pi \quad \forall \xi \in \mathbb{R}^N \backslash \mathcal{R}_0. \tag{10}$$

In the physics literature, $r(\cdot)$ is known as the *order parameter* and is used to characterize the amount of order or synchronization in the system (2). The idea is to think of the phase θ_j of oscillator j as a unit vector $e^{i\theta_j}$ in $\mathbb{C}$; the order parameter then corresponds to the geometric centroid of the set of vectors $\{e^{i\theta_j} : j = 1, \ldots, N\}$, as illustrated in Figure 1. The magnitude of the order parameter, given by $R(\theta)$, serves as a measure of the order in the system in the sense that the closer the vectors are to being perfectly aligned, the closer $R(\theta)$ is to its maximal value 1, while vectors that are far from alignment will give rise to values of $R(\theta)$ significantly smaller than 1.

It follows from (6) that for $\xi \in \mathbb{R}^N \backslash \mathcal{R}_0$,

$$R(\xi) = e^{-i\psi(\xi)} r(\xi) \tag{11}$$

$$= \frac{1}{N}\sum_{j=1}^{N} e^{i(\xi_j - \psi(\xi))}. \tag{12}$$

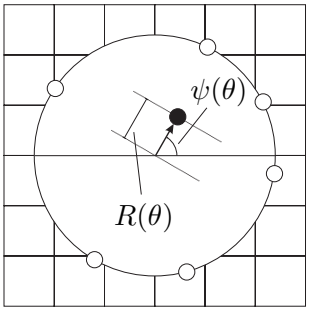


Figure 1. *The order parameter $r(\theta) := R(\theta)e^{i\psi(\theta)}$ is defined as the centroid (closed circle) of the set of unit vectors (open circles) associated with the phases of the oscillators.*

Equating real and imaginary parts in (11), we immediately see that for $\xi \in \mathbb{R}^N \backslash \mathcal{R}_0$,

$$R(\xi) = \frac{1}{N}\sum_{j=1}^{N} \cos(\psi(\xi) - \xi_j); \tag{13}$$

$$\sum_{j=1}^{N} \sin(\psi(\xi) - \xi_j) = 0. \tag{14}$$

Both of these identities shall prove useful throughout the paper.

Before proceeding, note that the function $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ given by (4) can be written in terms of the functions $R(\cdot)$ and $\psi(\cdot)$ as

$$f_i(\xi) := \begin{cases} R(\xi)\sin(\psi(\xi) - \xi_i), & \xi \in \mathbb{R}^N \backslash \mathcal{R}_0, \\ 0, & \xi \in \mathcal{R}_0, \end{cases} \tag{15}$$

for $1 \leq i \leq N$.

2.4. Fixed points and global phase-locking. Let $\langle \omega \rangle$ denote the sample mean of the natural frequencies, $\langle \omega \rangle := \frac{1}{N}\sum_{j=1}^{N} \omega_j$. Similarly, let $\langle \theta(t) \rangle$ denote the mean phase of a solution of (2) at time t. In general, $\langle \omega \rangle$ and $\langle \theta(t) \rangle$ will be nonzero. However, we shall now show that for the study of phase-locked solutions of (2), we may assume without loss of generality that $\langle \omega \rangle = 0$, $\langle \theta(t) \rangle = 0$ for $t \geq t_0$. This helps to simplify the analysis of phase-locked solutions of (2), as it allows us to transform the problem into a question of fixed point existence for a lower-dimensional system.

Consider the new coordinates

$$x_i(t) := \theta_i(t) - \langle \theta(t) \rangle, \quad i = 1, \ldots, N. \tag{16}$$

Then $x(t) := V\theta(t)$. Similarly, define $\Omega := V\omega$. In the new coordinates, the system dynamics are given by

$$\dot{x} = \Omega + kf(x), \quad x(t) \in V\mathbb{R}^N, \tag{17}$$

where $f(\cdot)$ is defined in (4).

The key point here is that as $\langle\Omega\rangle = 0$, $V\mathbb{R}^N$ is invariant under (17). To avoid confusion, we shall use $x(t)$ to denote solutions to the system (17) on $V\mathbb{R}^N$, while $\theta(t)$ shall be used to denote solutions to the original Kuramoto system (2) on $\mathbb{R}^N$. Our main concern for the remainder of the paper is to find conditions on k and Ω under which the system (17) has one or more fixed points in the sense of the following definition.

Definition 1 (fixed point). *Given $\omega \in \mathbb{R}^N$, let $\Omega := V\omega$. We say that $x \in V\mathbb{R}^N$ is a* fixed point *(of the system (17)) if*

$$kf(x) = -\Omega. \tag{18}$$

There is a natural correspondence between fixed points of (17) and phase-locked solutions of (2). In fact, for every fixed point $x^* \in V\mathbb{R}^N$ there is a 1-dimensional manifold $\mathcal{M} := \{\theta \in \mathbb{R}^N : \theta = x^* + \langle\omega\rangle t,\ t \in \mathbb{R}\}$ that is invariant under the original system dynamics (2). More precisely, let x^* be a fixed point and let $\theta^0 \in \mathbb{R}^N$ be such that $V\theta^0 = x^*$. Then the solution $\theta(t)$ of the system (2) with initial condition $\theta(t_0) = \theta^0$ satisfies

$$\theta_i(t) - \theta_j(t) = \theta_i^0 - \theta_j^0 \tag{19}$$

for all $t \geq t_0$ and all (i, j). In other words, a fixed point in the sense of Definition 1 corresponds to a situation in which each oscillator is phase-locked to every other and moves at constant speed $\dot{\theta}_i = \langle\omega\rangle$. We shall refer to this phenomenon as global phase-locking. In the literature, it is also known as full (or complete) synchronization. See also [9].

2.5. Critical coupling. We next define the notion of *critical coupling*, which is central to the work of the rest of the paper. Essentially, the critical coupling is the smallest k for which the system (17) has at least one fixed point. Formally, we have the following definition.

Definition 2. *Given $\omega \in \mathbb{R}^N$, let $\Omega := V\omega$. We define the* critical coupling*, k_c, as follows:*

$$k_c := \inf_k \left\{k \in \mathbb{R}_+ : \exists x \in V\mathbb{R}^N \text{ s.t. } kf(x) = -\Omega\right\}. \tag{20}$$

Note that this definition of the critical coupling, which is equivalent to that of K_L in [9], does not coincide with the traditional notion used in the physics literature. Indeed, the traditional notion of critical coupling is defined in terms of the lowest value of k for which there exists at least one solution $x(t)$, $t \geq t_0$, and a constant $c \in (0, 1]$ such that $R(x(t)) = c$ for all $t \geq t_0$ (so-called stationary or steady solutions [23]). Note that these solutions are not necessarily fixed points, although, in finite dimensions, the probability of finding stationary solutions that are not fixed points is vanishingly small. In his original analysis, Kuramoto showed that in the limiting case when N tends to infinity, stationary solutions always exist for large enough k, provided the distribution of natural frequencies is symmetric. Our definition, although more restrictive in a sense, does not impose any restriction on the shape of the distribution of natural frequencies other than that it should have compact support. In fact, it follows from the result of Lemma 4 below that, if the distribution of natural frequencies does not have compact support, then the critical coupling will exceed any finite number with probability tending to 1 as N tends to infinity. In this paper we shall therefore focus on distributions with compact support.

3. Existence of fixed points under strong coupling. In this section we shall show that, provided the distribution of intrinsic frequencies has compact support, the critical coupling given in Definition 2 is always finite. Following from this, in the next section, we shall derive a number of lower bounds for the value of the critical coupling. There are two steps in the derivation given here: first, we characterize the fixed points of the homogeneous system

$$\dot{x} = kf(x). \tag{21}$$

Then, in the second step, we use a perturbation argument to show that for every fixed point of the homogeneous system, we can find an open set containing it, such that, under strong enough coupling, the original system (17) has a unique fixed point on this set. As a first step, the following lemma characterizes the fixed points of the homogeneous system.

Lemma 1. *Let $f(\cdot)$ be given by (4) and $\xi \in \mathbb{R}^N$. We have that $f(\xi) = 0$ if and only if one or both of the following conditions are satisfied:*

(a) $R(\xi) = 0$;

(b) $\sin(\xi_i - \xi_j) = 0$ *for all* (i, j).

Proof. Sufficiency of conditions (a) and (b) follows from (15) and (4), respectively. To prove necessity, suppose $f(\xi) = 0$ and $R(\xi) \neq 0$ (if $R(\xi) = 0$, we are done). It follows that $\sin(\psi(\xi) - \xi_i) = 0$ for all i. This implies that there exist integers $k_i \in \mathbb{Z}$, $i = 1, 2, \ldots, N$, such that $\psi(\xi) - \xi_i = k_i\pi$ for all i, and we have that $\xi_i - \xi_j = (k_j - k_i)\pi$. We conclude that $\sin(\xi_i - \xi_j) = 0$ for all (i, j). ∎

Remark 1. *It is not hard to see that conditions* (a) *and* (b) *in Lemma 1 are mutually exclusive if and only if the dimension N is odd. We shall prove necessity. Suppose conditions* (a) *and* (b) *both hold and suppose furthermore that N is odd. Then for all (i, j) we have that either $\cos(\xi_i - \xi_j) = 1$ or $\cos(\xi_i - \xi_j) = -1$. We write $R^2(\xi) = \frac{1}{N^2}\sum_{i,j}\cos(\xi_i - \xi_j) = \frac{1}{N^2}\left(N + 2\sum_{i,j>i}\cos(\xi_i - \xi_j)\right)$. Since $R(\xi) = 0$ by assumption, it follows that $2\sum_{i,j>i}\cos(\xi_i - \xi_j) = -N$. The left-hand side evaluates to an even integer. By assumption, the number on the right-hand side is odd. We arrive at a contradiction and conclude that if N is odd, conditions* (a) *and* (b) *cannot both hold.*

Next we shall prove that the fixed points of our N-dimensional system (17) can be found by solving a system of $N - 1$ equations in as many variables. We have the following result.

Lemma 2. *Let $p \in \{1, \ldots, N\}$ and let $x^* \in V\mathbb{R}^N$. Then x^* is a fixed point of (17) if and only if $kf_i(x^*) = -\Omega_i$ for $i \neq p$.*

Proof. The proof of necessity is trivial. To prove sufficiency, recall that

$$\sum_{j=1}^{N}(\Omega_j + kf_j(x)) = 0 \quad \forall x \in \mathbb{R}^N. \tag{22}$$

Now suppose $kf_i(x) = -\Omega_i$ for all $i \neq p$. Then it follows from (22) that $\Omega_p + kf_p(x) = 0$. In other words, it follows that $kf_i(x) = -\Omega_i$ for *all* i. We conclude that x is a fixed point. ∎

Let $x^* \in V\mathbb{R}^N$ be a fixed point of the homogeneous system (21) such that $R(x^*) \neq 0$. We shall now show that *locally*, in a neighborhood of x^*, the system of equations

$$\begin{cases} -\Omega_1 &= kf_1(x_1, \ldots, x_{N-1}, -\sum_{j=1}^{N-1} x_j) \\ \vdots & \vdots \\ -\Omega_{N-1} &= kf_{N-1}(x_1, \ldots, x_{N-1}, -\sum_{j=1}^{N-1} x_j) \end{cases} \tag{23}$$

has a unique solution, provided k is large enough. It follows directly from Lemma 2 that every solution of (23) defines a fixed point and, conversely, that every fixed point satisfies (23). We proceed as follows. Let $x^* \in V\mathbb{R}^N$. We define the Jacobian $J(x^*)$ as follows:

$$[J_{ij}(x^*)] := \left.\frac{\partial f_i(x_1, \ldots, -\sum_{j=1}^{N-1} x_j)}{\partial x_j}\right|_{x=x^*}, \tag{24}$$

where $i, j = 1, \ldots, N-1$. We have the following result.

Lemma 3. *Let $f(\cdot)$ be given by (4) and suppose that $x^* \in V\mathbb{R}^N$ satisfies $f(x) = 0$ and $R(x) \neq 0$. Then* $\det(J(x^*)) \neq 0$.

Proof. Let x^* be a fixed point of the homogeneous system and suppose $R(x^*) \neq 0$. Then by Lemma 1, we have that $\sin(x_j^* - x_i^*) = 0$ for all (i, j), and it follows that

$$\begin{aligned} \cos(x_j^* - x_i^*) &= \cos((x_j^* - x_s^*) - (x_i^* - x_s^*)) \\ &= \cos(x_j^* - x_s^*)\cos(x_i^* - x_s^*) \end{aligned} \tag{25}$$

for all s and all (i, j). The claim is that $J(x^*)$ is nonsingular. To prove this, we proceed as follows. From the definition, we have that

$$J_{ij}(x^*) = \begin{cases} -\sum_{m=1,\, m\neq i}^{N-1} \cos(x_m^* - x_i^*) - 2\cos(x_N^* - x_i^*), & i = j, \\ \cos(x_j^* - x_i^*) - \cos(x_N^* - x_i^*), & i \neq j. \end{cases} \tag{26}$$

Using the aforementioned identity, setting $s = N$, we rewrite (26) as follows:

$$J_{ij}(x^*) = \begin{cases} -\left(\sum_{m=1,\, m\neq i}^{N-1} \cos(x_m^* - x_N^*) + 2\right)\cos(x_i^* - x_N^*), & i = j, \\ \left(\cos(x_j^* - x_N^*) - 1\right)\cos(x_i^* - x_N^*), & i \neq j. \end{cases} \tag{27}$$

Inspection shows that the rank of $J(x^*)$ is invariant under permutations of the components of x^*. Hence we can assume, without loss of generality, that there exists $\rho \in \{0, \ldots, N-1\}$, such that

$$\cos(x_j^* - x_N^*) = \begin{cases} -1, & 1 \leq i \leq \rho, \\ +1, & \rho + 1 \leq i \leq N. \end{cases} \tag{28}$$

Under this assumption $J(x^*)$ takes the form

$$J(x^*) = \begin{pmatrix} A & 0 \\ C & B \end{pmatrix}, \tag{29}$$

where A and B are square matrices of dimension $\rho \times \rho$ and $(N-1-\rho)\times(N-1-\rho)$, respectively. It follows that $J(x^*)$ is nonsingular if and only if A and B are nonsingular. Inspection shows that $A = (N-2\rho)I + 2\mathbf{1}\mathbf{1}^T$ and $B = (2\rho - N)I$. It follows that A or B is singular if and only if $N = 2\rho$. In case N is odd, this condition is never satisfied. In case N is even, this condition, combined with (25) and the fact that $R^2(x) = \frac{1}{N^2}\sum_{i,j}\cos(x_i - x_j)$, implies that $R(x^*) = 0$,

which contradicts our starting assumption. We conclude that, under the hypotheses of the lemma, J is nonsingular. This concludes the proof. ■

Let $\Pi : \mathbb{R}^{N-1} \mapsto V\mathbb{R}^N$ be given as

$$(\Pi(y))_i := \begin{cases} y_i & \text{for } i = 1, 2, \dots N-1; \\ -\sum_{j=1}^{N-1} y_j & \text{for } i = N, \end{cases} \tag{30}$$

and note that Π has an inverse Π^{-1} that is defined everywhere in $V\mathbb{R}^N$. We are now ready to state the main result.

Theorem 1. *Let $f(\cdot)$ be given by (4) and $x^* \in V\mathbb{R}^N$ be such that $f(x^*) = 0$ and $R(x^*) \neq 0$. Also, let $\Omega \in V\mathbb{R}^N$. Then there exist $K \in \mathbb{R}$ and an open set $U \in \mathbb{R}^{N-1}$ such that* (a) *$\Pi^{-1}(x^*)$ is an interior point of U; and* (b) *for all $k > K$, the system of equations (23) has a unique solution on U.*

Proof. Define $y^* := \Pi^{-1}(x^*)$, and let $g : \mathbb{R}^{N-1} \mapsto \mathbb{R}^{N-1}$ be given as $g_i(y) := f_i(y_1, \dots, y_{N-1}, -\sum_{j=1}^{N-1} y_j)$, $i = 1, \dots, N-1$. Note that $g(y^*) = 0$. Also, by Lemma 3, we have that $\det(\frac{\partial g}{\partial y}(y^*))| \neq 0$. Under these conditions, the inverse function theorem says that there exists an open set $U \subset \mathbb{R}^{N-1}$ containing y^* such that $g|_U : U \mapsto g(U)$ is a diffeomorphism. By continuity (and bijectivity) of g^{-1} there exists $\delta > 0$ such that for all $z \in \mathbb{R}^{N-1}$ satisfying $\|z\| < \delta$, the equation $g(y) = z$ has a unique solution on U. Now let z be given as $z_i := -\Omega_i/k$. Since, by assumption, $\max_i |\Omega_i| < \infty$, it follows that, provided k is large enough, the system of equations $\{kg_i(y) = -\Omega : i = 1, \dots, N-1\}$ has a unique solution on U. This concludes the proof. ■

As alluded to earlier, there is a unique correspondence between solutions of (23) and the fixed points of the system (17). Indeed, by Lemma 2 we have that if y is a solution of (23), then $\Pi^{-1}(y)$ is a fixed point, and conversely, if y is a fixed point, then $\Pi(y)$ is a solution of (23). Thus, an immediate consequence of Theorem 1 is that for large enough k, the system (23) will have at least one fixed point. In other words, Theorem 1 tells us that the critical coupling, k_c, is always finite.

Note furthermore that the proof of Theorem 1 does not require detailed knowledge of the coupling function g and that, as such, its applicability is not restricted to networks with all-to-all coupling. To illustrate this, consider the case of a 4-cycle, where g is given as

$$\begin{aligned} g_1(y) &= \frac{1}{4}\sin(y_2 - y_1) + \frac{1}{4}\sin(-2y_1 - y_2 - y_3), \\ g_2(y) &= \frac{1}{4}\sin(y_3 - y_2) + \frac{1}{4}\sin(y_1 - y_2), \\ g_3(y) &= \frac{1}{4}\sin(-2y_3 - y_1 - y_2) + \frac{1}{4}\sin(y_2 - y_3). \end{aligned} \tag{31}$$

We have that $g(0) = 0$ and $\det(\frac{\partial g}{\partial y}(0)) = -\frac{1}{4} \neq 0$. This implies that for k large enough, the system of equations $\{kg_i(y) = -\Omega_i : i = 1, 2, 3\}$ has a unique solution on some open set containing the origin.

Last, note that continuity of g^{-1} implies that the fixed points of the original system (17) will converge to the fixed points of the homogeneous system (21) as k tends to infinity.

4. Lower bounds on the critical coupling. In the previous section we showed that the critical coupling is finite, provided the oscillator's intrinsic frequencies are finite. In the present section we shall investigate in more detail the relation between the distribution of intrinsic frequencies and the critical coupling. In particular, we shall derive various lower bounds and discuss some of these bounds' implications for the system's dynamic behavior.

First, let us observe that k_c (Definition 2) is lower bounded by the l^∞ norm of Ω:

$$k_c \geq \|\Omega\|_\infty := \max_i |\omega_i - \langle\omega\rangle|. \tag{32}$$

This follows trivially from inspection of (17). In order to derive another lower bound, we shall need the following result.

Lemma 4. *Let $f(\cdot)$ be given by (4). Then the following hold:*

1. *For all $x \in \mathbb{R}^N$,*

$$\|f(x)\|_2 \leq \sqrt{NR^2(x)\,(1-R^2(x))}. \tag{33}$$

2. *If N is even, then for every $c \in [0,1]$ there exists $x \in V\mathbb{R}^N$ such that $R(x) = c$ and $\|f(x)\|_2 = \sqrt{NR^2(x)\,(1-R^2(x))}$.*
3. *If N is odd, then inequality (33) is strict for all $x \in \mathbb{R}^N$ such that $0 < R(x) < 1$.*

Proof. Part 1. Observe that inequality (33) is trivially satisfied when $x \in \mathcal{R}_0$. Suppose therefore that $x \in \mathbb{R}^N \backslash \mathcal{R}_0$. Then by definition

$$\begin{aligned} \|f(x)\|_2^2 &:= \sum_{j=1}^N (f_j(x))^2 \\ &= R^2(x) \sum_{j=1}^N \sin^2(\psi(x) - x_j), \end{aligned} \tag{34}$$

where $\psi(x)$ and $R(x)$ are the phase and magnitude of the order parameter, previously defined in (8) and (7), respectively. Introducing the shorthand notation $z_i(x) := \cos(\psi(x) - x_i)$ and using (13), we now rewrite (34) as follows:

$$\|f(x)\|_2^2 = \left(\frac{1}{N}\sum_{j=1}^N z_j(x)\right)^2 \sum_{j=1}^N \left(1 - z_j(x)^2\right). \tag{35}$$

To derive the desired inequality we pick a $c \in [0,1]$ and maximize $\|f(x)\|_2$ over the set $\{x \in \mathbb{R}^N : R(x) = c\}$. We shall not solve this optimization problem directly but shall take an indirect route by considering another, easier optimization problem, whose solution will then give us an upper bound on the solution to the first problem. Then we shall show that, under certain conditions, the two solutions coincide.

To this end, let $c \in (0,1]$ and consider the constrained optimization problem

OPT 1:

maximize	$\sum_{j=1}^N \left(1 - z_j(x)^2\right)$
subject to	$\frac{1}{N}\sum_{j=1}^N z_j(x) = c, \quad x \in \mathbb{R}^N \backslash \mathcal{R}_0$

Note that the constraint is feasible for all values of c in the specified interval. We shall denote the solution to OPT 1 as $s_1(c)$. Next consider a second optimization problem

$$\text{OPT 2:} \quad \boxed{\begin{array}{ll} \text{maximize} & \sum_{j=1}^{N} \left(1 - y_j^2\right) \\ \text{subject to} & \frac{1}{N}\sum_{j=1}^{N} y_j = c, \quad y \in \mathbb{R}^N \end{array}}$$

and let the solution to this problem be denoted as $s_2(c)$. We then have that $s_2(c) \geq s_1(c)$ for all $c \in (0, 1]$. In other words, the solution to OPT 1 is upper bounded by the solution to OPT 2. The solution to OPT 2 can be found by means of standard Lagrange multiplier techniques. The optimum $s_2(c) = N\left(1 - c^2\right)$ is attained when $y_i = c$ for all i. We conclude that

$$\max_{\{x \in \mathbb{R}^N : R(x) = c\}} \|f(x)\|_2^2 \leq Nc^2\left(1 - c^2\right), \tag{36}$$

and hence

$$\|f(x)\|_2 \leq \sqrt{N} R(x) \sqrt{1 - R^2(x)} \tag{37}$$

for all $x \in \mathbb{R}^N$.

Part 2. To prove the second part of the theorem, let $c \in (0, 1]$ and note that $s_1(c) = s_2(c)$ if and only there exists $x \in \mathbb{R}^N \backslash \mathcal{R}_0$ such that

$$\cos(\psi(x) - x_i) = c \tag{38}$$

for all i. Suppose N is even and let x be given as

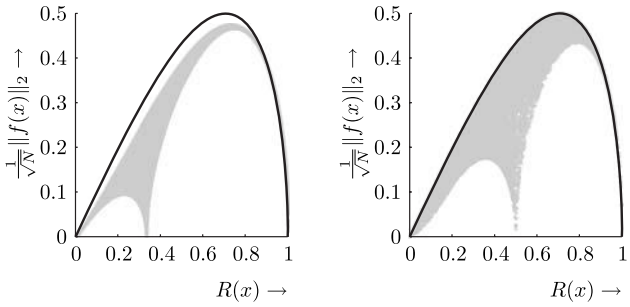
$$x_i := \begin{cases} \arccos(c), & i = 1, \ldots, \frac{N}{2}, \\ -\arccos(c), & i = \frac{N}{2}, \ldots, N. \end{cases} \tag{39}$$

Then $\sum_{j=1}^{N} x_j = 0$, and, by definition, $x \in V\mathbb{R}^N$. Moreover, $\psi(x) = 0$, and $\cos(\psi(x) - x_i) = c$ for all i. This completes the second part.

Part 3. To prove the third part, let N be odd and suppose there exists $x \in \mathbb{R}^N$ such that condition (38) is satisfied. Then it follows from the identity $\sin^2(\psi(x) - x_i) + \cos^2(\psi(x) - x_i) = 1$ that there must exist $a \in \{-1, 1\}^N$ such that $\sin(\psi(x) - x_i) = a_i\sqrt{1 - c^2}$ for all i. By identity (14), we have that $\sum_j \sin(\psi(x) - x_j) = 0$, which, assuming $c \neq 1$, implies that $\sum_{j=1}^{N} a_j = 0$. But this cannot be true unless N is even. Thus we arrive at a contradiction and we conclude that if N is odd then $s_2(c) > s_1(c)$ for all c such that $0 < c < 1$. This concludes the proof. ■

Figure 2 illustrates the result of Lemma 4. When $N = 4$ (even), the lower bound is attained at every value of $R(x)$, which shows that the given bound is the tightest possible. However, as illustrated in the left panel, when $N = 3$, the bound is never attained except on the set $\{x \in \mathbb{R}^N : R(x) \in \{0, 1\}\}$. It can be shown, however, that in the limit of large N the given bound is arbitrarily tight, even for odd N, in the sense that for every $c \in [0, 1]$,

$$\min_{\{x \in \mathbb{R}^{2m+1} : R(x) = c\}} \frac{1}{\sqrt{2m+1}} \left| \|f(x)\|_2 - \sqrt{(2m+1)c^2\left(1 - c^2\right)} \right|$$



(a) $N = 3$ (b) $N = 4$

Figure 2. *Scatter plot of* $\frac{1}{\sqrt{N}}\|f(x)\|_2$ *for* $N = 3$ *(left panel) and* $N = 4$ *(right panel). The phases* x *were drawn from a uniform distribution. The solid black line in both panels is the upper bound* $R(x)\sqrt{1 - R^2(x)}$.

tends to zero as m tends to infinity.

Lemma 4 has some interesting implications. For instance, it can provide insight into the rate at which solutions of a homogeneous system of Kuramoto oscillators ((2) with $\omega_i = 0$ for $1 \leq i \leq N$) on $\mathbb{R}^N$ converge to fixed points. To see this, consider the homogeneous system

$$\begin{cases} \dot{\theta}(t) &= kf(\theta(t)), \\ \theta(t_0) &= \theta_0, \end{cases} \tag{40}$$

where $\theta_0 \in \mathbb{R}^N$. We shall compute the time-derivative of the magnitude squared of the order parameter, $L(\cdot) := R^2(\cdot)$, and show that this derivative is (i) nonnegative along solutions of (40) and (ii) bounded from above by a certain function $D(t)$ for every t. We proceed as follows [9]. By definition,

$$\frac{\mathrm{d}L(\theta(t))}{\mathrm{d}t} := \frac{L(\theta)}{\partial\theta}\dot{\theta}(t) = \frac{L(\theta)}{\partial\theta}kf(\theta(t)).$$

Using the identity

$$\frac{\partial L(\theta)}{\partial\theta} = \frac{2}{N}\left[f(\theta)\right]^T,$$

it follows that

$$\frac{\mathrm{d}L(\theta(t))}{\mathrm{d}t} = \frac{2k}{N}\|f(\theta(t))\|_2^2, \tag{41}$$

which shows that the time-derivative is positive everywhere, except at the equilibria, where it is zero. It follows that the magnitude of the order parameter is a nondecreasing function of time. Based on the observation that the time-derivative of L is positive almost everywhere (the set of equilibria having measure zero), we formulate the following conjecture [3, 9].

Conjecture 1. *For almost all initial conditions* θ_0*, the solution* $\theta(t)$ *to the homogeneous system* (40) *has the property that* $\lim_{t\to\infty} R(\theta(t)) = 1$.

In agreement with Conjecture 1, one can prove that, for the homogeneous system, the global phase-locking manifold $\mathcal{M} := \{\theta \in \mathbb{R}^N : \theta_i = \theta_j \text{ for all } i, j\}$ is (locally) asymptotically stable. However, the existence of other invariant manifolds, not contained in $\mathcal{M}$, implies

that $\mathcal{M}$ is not globally asymptotically stable. We conjecture that $\mathcal{M}$ is "almost globally asymptotically stable" in the sense that its region of attraction is the entire space minus a set of measure zero.

For our next result, we shall need the concept of a dominating function, which is defined as follows.

Definition 3. *Let $h, g : \mathbb{R} \mapsto \mathbb{R}$ and let $\mathcal{I} \subset \mathbb{R}$ be some interval. We say that h dominates g on $\mathcal{I}$ if $h(t) \geq g(t)$ for all $t \in \mathcal{I}$. In that case we call h a* dominating function *for g on $\mathcal{I}$.*

Our next result states that $L(\theta(t))$ is dominated by a certain scalar function $D(t)$ that depends only on θ_0. In order to prove this result, we need the following two lemmas.

Lemma 5. *Let $\theta(\cdot)$ be a solution of the homogeneous system (40) and suppose $\dot{L}(\theta(t')) = 2kL(\theta(t'))\,(1 - L(\theta(t')))$ for some $t' \in \mathbb{R}$. Then*

$$\dot{L}(\theta(t)) = 2kL(\theta(t))\,(1 - L(\theta(t))) \quad \forall t \geq t'.$$

Proof. Recall that $\dot{L}(\theta(t)) = \frac{2k}{N}\|f(\theta(t))\|_2^2$. It follows from the proof of Lemma 4 that $\|f(\theta(t'))\|_2^2 = NL(\theta(t'))\,(1 - L(\theta(t')))$ for some $t' \in \mathbb{R}$ if and only if one or two of the following conditions hold: (a) $L(\theta(t')) = 0$; (b) N is even and there exists a permutation $\hat{\theta}(t')$ of $\theta(t')$ such that

$$\cos(\hat{\theta}_i(t') - \hat{\theta}_1(t')) = 1, \qquad i = 1, 2, \ldots, \frac{N}{2}, \tag{42}$$

$$\cos(\hat{\theta}_i(t') - \hat{\theta}_N(t')) = 1, \qquad i = \frac{N}{2} + 1, \ldots, N. \tag{43}$$

If $L(\theta(t')) = 0$, we have that $L(\theta(t)) = 0$ for all $t \geq t'$ and the result follows trivially. Now suppose conditions (42) and (43) hold. It follows that $\dot{\hat{\theta}}_i(t') = \dot{\hat{\theta}}_j(t')$ for $i, j \leq \frac{N}{2}$ and $i, j > \frac{N}{2}$, and hence

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\cos(\hat{\theta}_i(\tau) - \hat{\theta}_j(\tau))\right)\Big|_{\tau=t'} = 0, \quad i, j \leq \frac{N}{2}; \quad i, j > \frac{N}{2}.$$

In other words, if $\hat{\theta}(\cdot)$ satisfies conditions (42) and (43) for some t', it satisfies (42) and (43) for all $t \geq t'$. This concludes the proof. ∎

Lemma 6. *Let $h, g : \mathbb{R} \mapsto \mathbb{R}$ be such that for every $x_0 \in \mathbb{R}$, the systems*

$$\begin{cases} \dot{x} &= h(x), \\ x(0) &= x_0, \end{cases} \qquad \begin{cases} \dot{x} &= g(x), \\ x(0) &= x_0 \end{cases}$$

have unique solutions in $C^1[0, \infty)$. Let these solution be denoted $x_h(t; x_0)$ and $x_g(t; x_0)$, respectively. Suppose there exist $a, b \in \mathbb{R}$, $a \neq b$, such that

$$h(x) > g(x) \geq 0$$

for all $x \in (a, b) \subset \mathbb{R}$. Then for every $x_0 \in (a, b)$, we have that

$$x_h(t; x_0) \geq x_g(t; x_0) \tag{44}$$

for all $t \in I$, where $I \subset \mathbb{R}$ is defined as

$$I := \begin{cases} [0, \min_t\{x_h(t; x_0) = b\}) & \text{when } \{x_h(t; x_0) = b\} \neq \emptyset; \\ [0, \infty) & \text{otherwise.} \end{cases} \tag{45}$$

Proof. Under the hypotheses of the lemma we have that $h(x_0) > g(x_0)$, and it follows that for small enough t, $x_h(t) > x_g(t)$ (omitting the argument x_0 for notational convenience). Note also that x_h and x_g are increasing whenever $x_h(t; x_0) < b$ and $x_g(t; x_0) < b$, respectively. Now suppose there exists $t_2 > 0$ such that $x_g(t_2) > x_h(t_2)$ and $x_g(t_2) < b$. Then by continuity there exists $t_1 < t_2$ such that $a < x_h(t_1) = x_g(t_1) < b$ and $h(x_h(t_1)) \leq g(x_g(t_1))$. Define $x' := x_h(t_1) = x_g(t_1)$. It follows that $g(x') \geq h(x')$, which contradicts our starting hypothesis. We conclude that $x_h(t) \geq x_g(t)$ for all $t \in I$. ∎

We have the following result.

Corollary 1. *Let $\theta(\cdot)$ be a solution to the homogeneous system (40) with initial condition $\theta(t_0) = \theta_0$. Then*

$$D(t) := \frac{1}{1 - e^{-2k(t-t_0)}\left(\frac{L(\theta_0)-1}{L(\theta_0)}\right)} \tag{46}$$

is a dominating function for $L(\theta(t))$ on $[t_0, \infty)$.

Proof. By Lemma 4 we have that $\dot{L}(\theta(t)) \leq 2kL(\theta(t))\,(1 - L(\theta(t)))$ for all t. We claim that, on $[t_0, \infty)$, $L(\theta(t))$ is dominated by the solution $y(t)$ of the ODE

$$\begin{cases} \dot{y} &= 2ky(1-y), \\ y(t_0) &= L(\theta_0), \end{cases} \tag{47}$$

which is given as

$$y(t) = \frac{1}{1 - e^{-2k(t-t_0)}\left(\frac{L(\theta_0)-1}{L(\theta_0)}\right)}, \quad t \geq t_0. \tag{48}$$

To prove this, suppose $\dot{L}(\theta(t)) = 2kL(\theta(t))\,(1 - L(\theta(t)))$ for some $t \geq t_0$ and let t' denote the smallest such t (in case $\dot{L}(\theta(t)) < 2kL(\theta(t))\,(1 - L(\theta(t)))$ for all $t \geq t_0$, the result follows immediately from Lemma 6). Then by Lemma 5, we have that $\dot{L}(\theta(t)) = 2kL(\theta(t))\,(1 - L(\theta(t)))$ for all $t \geq t'$, and it follows that

$$L(\theta(t))|_{L(\theta(t'))=a} = \frac{1}{1 - e^{-2k(t-t')}\left(\frac{a-1}{a}\right)}, \quad t \geq t'. \tag{49}$$

One can easily verify that $L(\theta(t))|_{L(\theta(t'))=a}$ is nondecreasing as a function of a for all $t \geq t'$ (and $a \in [0, 1]$). Now let l be an upper bound for $L(\theta(t'))$. It follows that $\frac{1}{1-e^{-2k(t-t')}\left(\frac{l-1}{l}\right)}$ is a dominating function for $L(\theta(t))$ on the interval $[t', \infty)$. To compute an upper bound for $L(\theta(t')$, we proceed as follows. By definition of t', we have that $\dot{L}(\theta(t)) < 2kL(\theta(t))\,(1 - L(\theta(t)))$ for all $t < t'$. It follows from Lemma 6 that

$$L(\theta(t)) \leq \frac{1}{1 - e^{-2k(t-t_0)}\left(\frac{L(\theta_0)-1}{L(\theta_0)}\right)}, \quad t_0 \leq t < t'. \tag{50}$$

By continuity, we have that

$$L(\theta(t')) \leq \frac{1}{1 - e^{-2k(t'-t_0)}\left(\frac{L(\theta_0)-1}{L(\theta_0)}\right)}.$$

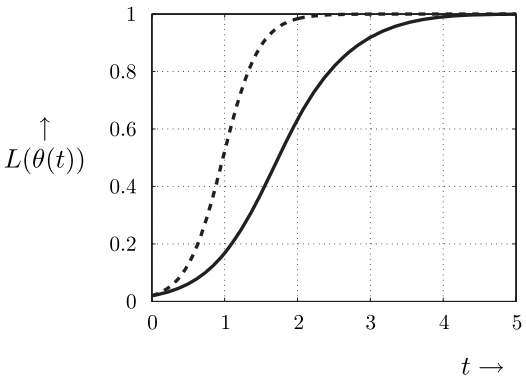


Figure 3. *Numerical simulation of the homogeneous system (40) with $N = 100$ oscillators and coupling coefficient $k = 2$: Time evolution of $L(\theta(t)) := R^2(\theta(t))$ (solid line) and the dominating function $D(t)$—(46) (dashed line).*

Using the upper bound $\frac{1}{1-e^{-2k(t'-t_0)}\left(\frac{L(\theta_0)-1}{L(\theta_0)}\right)}$ for $L(\theta(t'))$, it follows from (49) that

$$L(\theta(t)) \leq \frac{1}{1-e^{-2k(t-t_0)}\left(\frac{L(\theta_0)-1}{L(\theta_0)}\right)}, \quad t \geq t'. \tag{51}$$

Combining (50) and (51), we arrive at the desired result. This concludes the proof. ∎

Figure 3 shows the graph of $L(\theta(t))$ and that of the dominating function $D(t)$—(46) for a particular realization of the initial condition θ_0. In this example, $N = 100$ and $k = 2$. We observe that, in agreement with Conjecture 1, the solution converges to a globally phase-locked state, that is, $L(\theta(t)) \to 1$. Note that convergence can be very slow depending on the choice of initial condition. Indeed, for any $T \in \mathbb{R}$ and any $\epsilon > 0$, we can find $\delta > 0$ such that if $L(\theta_0) < \delta$ then $L(\theta(t)) < \epsilon$ for all $t \leq t_0 + T$. The upshot of this is that if the initial condition θ_0 is selected by drawing from a uniform distribution and the number of oscillators is large, then $L(\theta_0)$ is likely to be small, and as a consequence convergence to the stable equilibrium is likely to be slow. In the limit case when N tends to infinity, we have that $L(\theta_0)$ tends to zero with probability 1 and the time required for $L(\theta(t))$ to exceed some given finite threshold diverges to infinity.

Let $\sigma_\omega := \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\omega_j - \langle\omega\rangle)^2}$ denote the (sample) standard deviation associated with the vector of intrinsic frequencies ω. Using Lemma 4 we derive another lower bound on the critical coupling, as follows.

Corollary 2. *The critical coupling k_c satisfies*

$$k_c \geq 2\sigma_\omega. \tag{52}$$

Proof. Let $x^* \in V\mathbb{R}^N$ be a fixed point of the system (17). Then by definition $k\|f(x^*)\|_2 = \|V\omega\|_2 = \sqrt{N}\sigma_\omega$ and by Lemma 4 we have that

$$\|f(x^*)\|_2 \leq \sqrt{N}\sqrt{R^2(x^*)\,(1 - R^2(x^*))}. \tag{53}$$

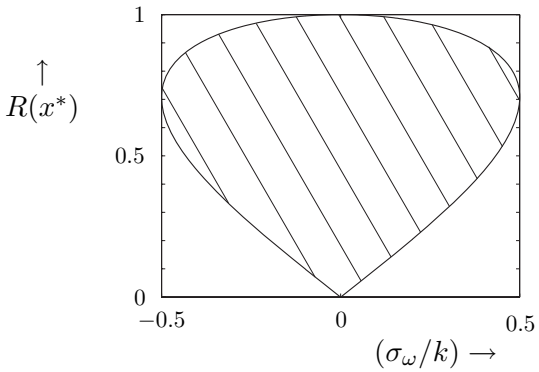


Figure 4. *Graph associated with inequality (56). For a given value of the ratio (σ_ω/k), the magnitude of the order parameter $R(\cdot)$, evaluated at a fixed point x^*, must lie within the striped region.*

It is not hard to see that the right-hand side of (53) is upper bounded by $\frac{1}{2}\sqrt{N}$. It follows that

$$k \geq \frac{\sqrt{N}\sigma_\omega}{\sqrt{N}\frac{1}{2}} = 2\sigma_\omega. \tag{54}$$

This completes the proof. ■

Note that Corollary 2 is in agreement with the intuition that greater variation in intrinsic frequencies requires stronger coupling to achieve global phase-locking.

Using Lemma 4 we can compute bounds on the value of the order parameter evaluated at the fixed points of the system, should they exist. Indeed, suppose $k > k_c$; then for any fixed point $x^* \in V\mathbb{R}^N$ we have that

$$\sqrt{R^2(x^*)\left(1 - R^2(x^*)\right)} \geq \frac{\sigma_\omega}{k}. \tag{55}$$

Solving for $R(x^*)$ gives

$$\frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\left(\frac{\sigma_\omega}{k}\right)^2} \leq R^2(x^*) \leq \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\left(\frac{\sigma_\omega}{k}\right)^2}. \tag{56}$$

The graph associated with inequality (56) is shown in Figure 4.

5. Necessary and sufficient conditions. In the last section, we derived lower bounds for the critical coupling of the system (17) which provided necessary conditions for the existence of fixed points. We next derive conditions that are both necessary and sufficient for fixed points to exist, and we shall use these results to describe an algorithm for computing the critical coupling in section 6.

Throughout this section, the function $f : \mathbb{R}^N \mapsto \mathbb{R}^N$ is given by (4) and the set $\mathcal{F}(k, \Omega)$ is defined as

$$\mathcal{F}(k, \Omega) := \left\{x \in V\mathbb{R}^N : kf(x) = -\Omega\right\}, \quad k \geq 0, \quad \Omega \in V\mathbb{R}^N.$$

On $\mathcal{F}(k, \Omega)$ we introduce a notion of equivalence as follows.

Definition 4 (equivalence on $\mathcal{F}(k,\Omega)$). *Given $\Omega \in V\mathbb{R}^N$ and $k \geq 0$, let $x, x' \in F(k,\Omega)$. We say that x and x' are equivalent ($x \simeq x'$) if $R(x) = R(x')$.*

To motivate this definition consider the following fact. Let k, Ω be given, and let $s \in \mathbb{Z}^N$ be such that $\sum_{j=1}^N s_j = 0$. If x is a fixed point of the system (17), then $x' := x + 2s\pi$ is also a fixed point of the system (17) and, in addition, $R(x') = R(x)$.

The following theorem provides a necessary and sufficient condition for the system (17) to have a fixed point, given a particular coupling strength k, and a particular realization of intrinsic frequencies, Ω. Note that the first part of this theorem is essentially identical to the result of Theorem 1 in [19].

Theorem 2. *Let $k > 0$ and $\Omega \in V\mathbb{R}^N$. Then $\mathcal{F}(k,\Omega) \neq \emptyset$ if and only if there exist $\beta \in [\frac{1}{k}\|\Omega\|_\infty, 1] \subset \mathbb{R}$ and $a \in \{-1,1\}^N$ such that*

$$\beta = \frac{1}{N}\sum_{j=1}^N a_j \sqrt{1 - \left(\frac{\Omega_j}{k\beta}\right)^2}. \tag{57}$$

Moreover, suppose (a^1, β^1) and (a^2, β^2) both satisfy (57), and let $x^1, x^2 \in F(k,\Omega)$ be such that

$$\begin{cases} k\beta^i \sin(\psi(x^i) - x^i_j) &= -\Omega_j, \\ a_j \cos(\psi(x^i) - x^i_j) &\geq 0, \end{cases} \qquad i \in \{1,2\}, \quad j = 1,2,\ldots,N. \tag{58}$$

Then $x^1 \simeq x^2$ if and only if $\beta^1 = \beta^2$ and $\sum_{j=1}^N (a_i - a_j)\sqrt{1 - \left(\frac{\Omega_j}{k\beta^1}\right)^2} = 0$.

Proof. Suppose $\Omega \neq 0$ (the case $\Omega = 0$ is easy). Let $x^* \in V\mathbb{R}^N$ be a fixed point of (17). By definition, $kf(x^*) = -\Omega$, and since $\Omega \neq 0$, we have that $f(x^*) \neq 0$, and consequently $R(x^*) \neq 0$. It follows that

$$\sin(\psi(x^*) - x^*_i) = -\frac{\Omega_i}{kR(x^*)}, \qquad i = 1,2,\ldots,N. \tag{59}$$

Let $\beta := R(x^*)$. By (59) we have that $\beta \geq \frac{1}{k}\|\Omega\|_\infty$. Recall that for all $x \in \mathbb{R}^N \backslash \mathcal{R}_0$, $R(x)$ can be written as

$$R(x) = \frac{1}{N}\sum_{j=1}^N \cos(\psi(x) - x_j), \tag{60}$$

and let a_i be given as

$$a_i := \begin{cases} -1 & \text{if } \cos(\psi(x^*) - x^*_i) \leq 0; \\ +1 & \text{otherwise.} \end{cases} \tag{61}$$

Combining (59), (60), and (61), we arrive at

$$\beta = \frac{1}{N}\sum_{j=1}^N a_j \sqrt{1 - \left(\frac{\Omega_j}{k\beta}\right)^2}. \tag{62}$$

This proves necessity. To prove sufficiency, let $a \in \{-1,1\}^N$ be given, and suppose $\beta \geq \frac{1}{k}\|\Omega\|_\infty > 0$ (again, the case $\Omega = 0$ is easy). Then for every $c \in \mathbb{R}$, the system

$$\left\{ \begin{array}{rcl} k\beta \sin(-y_i - c) & = & -\Omega_i, \\ a_i \cos(-y_i - c) & \geq & 0, \end{array} \right. \qquad i = 1, 2, \ldots, N, \tag{63}$$

has a unique solution $y^* \in [-\pi, \pi)^N$. We pick c such that $\sum_{j=1}^N y_j^* = 0$. Since $\sum_{j=1}^N \sin(y_j^* + c) = 0$, it follows that

$$R(y^*) = R(y^* + c\mathbf{1}) = \left| \sum_{j=1}^N \cos(y_j^* + c) \right|. \tag{64}$$

From (63), we have that

$$\cos(y_i^* + c) = a_i \sqrt{1 - \left(\frac{\Omega_i}{k\beta}\right)^2}, \quad i = 1, \ldots, N. \tag{65}$$

Combining (64) and (65), we arrive at

$$R(y^*) = \left| \frac{1}{N} \sum_{j=1}^N a_j \sqrt{1 - \left(\frac{\Omega_j}{k\beta}\right)^2} \right|. \tag{66}$$

The second part of the theorem follows easily after noting that if x^i satisfies (58) then $R(x^i) = \beta^i$, $i = 1, 2$. ■

Theorem 2 gives us a necessary and sufficient condition for the equation $kf(x) = -\Omega$ to have at least one solution for a given value of k. It is not clear, however, that there exists a k for which this condition is satisfied. The following corollary provides an easy sufficient condition.

Corollary 3. *Let $k > 0$ and $\Omega \in V\mathbb{R}^N$. Suppose*

$$\frac{1}{k}\|\Omega\|_\infty \leq \frac{1}{N} \sum_{j=1}^N a_j \sqrt{1 - \left(\frac{\Omega_j}{\|\Omega\|_\infty}\right)^2} \tag{67}$$

for some $a \in \{-1,1\}^N$. Then $\mathcal{F}(k, \Omega) \neq \emptyset$.

Proof. Suppose $\Omega \neq 0$ (again, the case $\Omega = 0$ is easy). Let $a \in \{-1,1\}^N$. Define $m : [\frac{1}{k}\|\Omega\|_\infty, 1] \mapsto \mathbb{R}$, $m(\beta) := \beta$, and $n : [\frac{1}{k}\|\Omega\|_\infty, 1] \times \{-1,1\}^N \mapsto \mathbb{R}$,

$$n(\beta, a) := \frac{1}{N} \sum_{j=1}^N a_j \sqrt{1 - \left(\frac{\Omega_j}{k\beta}\right)^2}. \tag{68}$$

Since $\Omega \neq 0$ we have that $m(1) > n(1, a)$. Now suppose condition (67) is satisfied. Then we have that $m(\frac{1}{k}\|\Omega\|_\infty) \leq n(\frac{1}{k}\|\Omega\|_\infty, a)$, and by the intermediate value theorem there must exist $\beta^* \in [\frac{1}{k}\|\Omega\|_\infty, 1]$ such that $m(\beta^*) = n(\beta^*, a)$. It follows from Theorem 2 that the system (17) has a fixed point. ■

Corollary 4. *Let $\Omega \in V\mathbb{R}^N$. Then* (i) *the critical coupling k_c is finite, and* (ii) *for large enough coupling, the system* (17) *has at least 2^{N-1} fixed points.*

Proof. Note that the right-hand side of (67) does not depend on k. Hence, it follows that, provided

$$\frac{1}{N}\sum_{j=1}^{N} a_j \sqrt{1-\left(\frac{\Omega_j}{\|\Omega\|_\infty}\right)^2} > 0, \tag{69}$$

condition (67) is always satisfied for large enough k. Furthermore, it follows easily that if (69) is not satisfied for given a, then it is satisfied for $a' := -a$. This implies (i) that the critical coupling k_c is always finite, and (ii) that the set $A^+ := \{a \in \{-1,1\}^N : (69) \text{ is satisfied}\}$ contains precisely 2^{N-1} elements (counting multiplicity), each of which defines a unique (up to equivalence in the sense of Definition 4) fixed point. This concludes the proof. ∎

Corollary 5. *Let $k > 0$ and $\Omega \in V\mathbb{R}^N$. Then $\mathcal{F}(k,\Omega) \neq \emptyset$ if and only if there exist $\beta \in [\frac{1}{k}\|\Omega\|_\infty, 1]$ such that*

$$\beta = \frac{1}{N}\sum_{j=1}^{N} \sqrt{1-\left(\frac{\Omega_j}{k\beta}\right)^2}.$$

Proof. The proof of Corollary 3 suggests that if the fixed point equation (57) does not have a solution, then necessarily

$$\beta > \frac{1}{N}\sum_{j=1}^{N} a_j \sqrt{1-\left(\frac{\Omega_j}{k\beta}\right)^2}$$

for all $\beta \in [\frac{1}{k}\|\Omega\|_\infty, 1]$ and all $a \in \{-1,1\}^N$. Since we have that

$$\frac{1}{N}\sum_{j=1}^{N} \sqrt{1-\left(\frac{\Omega_j}{k\beta}\right)^2} \geq \frac{1}{N}\sum_{j=1}^{N} a_j \sqrt{1-\left(\frac{\Omega_j}{k\beta}\right)^2}$$

for all $a \in \{-1,1\}^N$, it follows that the given condition is necessary and sufficient for the system (17) to have at least one fixed point. This concludes the proof. ∎

The next and final corollary gives us an upper bound on the critical coupling.

Corollary 6. *The critical coupling, k_c, satisfies*

$$k_c \leq \frac{\|\Omega\|_\infty}{\frac{1}{N}\sum_{j=1}^{N} \sqrt{1-\left(\frac{\Omega_j}{\|\Omega\|_\infty}\right)^2}}. \tag{70}$$

Proof. The proof follows directly from Corollary 3. ∎

6. An algorithm for computing k_c. In this section we present a bisection algorithm that will allow us to numerically evaluate the critical coupling with arbitrary precision. Throughout, we shall assume that $\Omega \neq 0$. Define $\mathcal{I} := (\|\Omega\|_\infty, \infty)$, and let $p_i : \mathcal{I} \mapsto (0,1]$ and

$P : \mathcal{I} \mapsto (0, 1]$ be given as

$$p_i(u) := \sqrt{1 - \left(\frac{\Omega_i}{u}\right)^2}; \qquad P(u) := \frac{1}{N}\sum_{j=1}^{N} p_j(u). \tag{71}$$

Also, define $h(u; k) : \mathcal{I} \times \mathbb{R}_+ \mapsto \mathbb{R}_+$:

$$h(u; k) := \frac{1}{k}u. \tag{72}$$

From Corollary 5 it follows that the critical coupling is the smallest k for which the equation $P(u) = h(u; k)$ has at least one solution on $\mathcal{I}$. We have the following result.

Theorem 3. *For all $\Omega \in V\mathbb{R}^N$, $\Omega \neq 0$, the equation*

$$2\frac{1}{N}\sum_{j=1}^{N}\sqrt{1 - \left(\frac{\Omega_j}{u}\right)^2} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{1 - \left(\frac{\Omega_j}{u}\right)^2}} \tag{73}$$

has a unique solution $u^* \in \mathcal{I}$, and we have that

$$k_c = \frac{u^*}{\frac{1}{N}\sum_{j=1}^{N}\sqrt{1 - \left(\frac{\Omega_j}{u^*}\right)^2}}. \tag{74}$$

Proof. Observe that, by strict concavity of P and linearity of $h(\,\cdot\,; k)$, the equation $P(u) = h(u; k)$ can have at most two solutions on $\mathcal{I}$ for any $k > 0$. We shall now show that, when $k = k_c$, it can have no more than one solution. Since, by definition of critical coupling, $P(u) = h(u; k_c)$ must have *at least* one solution, we shall conclude that it has precisely one solution. Let $k = k_c$, and suppose there exist $u^1, u^2 \in \mathcal{I}$, $u^1 \neq u^2$, such that $P(u^1) = h(u^1; k_c)$ and $P(u^2) = h(u^2; k_c)$. By strict concavity of P we have that $P(\frac{1}{2}(u^1 + u^2)) > \frac{1}{2}\big(P(u^1) + P(u^2)\big)$. Define $u' := \frac{1}{2}(u^1 + u^2)$ and note that $u' \in \mathcal{I}$. We have that $P(u') > h(u'; k_c)$. This implies that there exists $k' < k_c$ such that $P(u') = h(u'; k')$. But by definition k_c is the smallest k for which $P(u) = h(u; k)$ has a solution. We arrive at a contradiction and conclude that $u^1 = u^2$ or, in other words, that the equation $P(u) = h(u; k_c)$ has exactly one solution on $\mathcal{I}$. Denoting this solution by u^*, it is not hard to see that, at $u = u^*$, the derivative of P with respect to u and the derivative of h with respect to u (both of which are defined on the entire interval $\mathcal{I}$) must coincide. For suppose $\frac{\partial h}{\partial u}(u^*) < \frac{\partial P}{\partial u}(u^*)$; then by continuity there exists $\delta > 0$ such that $h(u; k_c) < P(u)$ for all u such that $u - u^* < \delta$. Let u' be one such u. It follows that there exists $k' < k_c$ such that $P(u') = h(u'; k')$. This leads to a contradiction, and we conclude that $h(u; k_c) \geq P(u)$. By analogy we have that $h(u; k_c) \leq P(u)$. We conclude that $\frac{\partial h}{\partial u}(u^*) = \frac{\partial P}{\partial u}(u^*)$. That is,

$$\frac{1}{k_c} = \frac{1}{u^*}\frac{1}{N}\sum_{j=1}^{N}\frac{\left(\frac{\Omega_j}{u^*}\right)^2}{\sqrt{1 - \left(\frac{\Omega_j}{u^*}\right)^2}} \tag{75}$$

or, equivalently,

$$\begin{aligned}\frac{u^*}{k_c} &= \frac{1}{N}\sum_{j=1}^{N}\frac{\left(\frac{\Omega_j}{u^*}\right)^2}{\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}} \\ &= -\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}+\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}}. \end{aligned} \tag{76}$$

Now recall that by definition of u^*, we have that

$$\frac{u^*}{k_c} = \frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}. \tag{77}$$

Equating the right-hand side of (76) with the right-hand side of (77) gives

$$2\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}}. \tag{78}$$

This shows that u^* is a solution to (73). What remains to be shown is that u^* is the only solution. Define $v, w : \mathcal{I} \mapsto \mathbb{R}$:

$$v(u) := 2\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u}\right)^2}, \quad w(u) := \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{1-\left(\frac{\Omega_j}{u}\right)^2}}.$$

Note that, on their respective domains, v is strictly monotonically decreasing, while w is strictly monotonically increasing. In addition, note that there exist $a, b \in \mathcal{I}$ such that $v(a) > w(a)$ and $v(b) < w(b)$. Hence, by continuity, there must exist a point $u' \in (a, b) \subset \mathcal{I}$ such that $v(u') = w(u')$. Monotonicity of v and w implies that this point is unique. It follows that u^* is the unique solution of (73) on $\mathcal{I}$. And by (77) we have that

$$k_c = \frac{u^*}{\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}}. \tag{79}$$

This concludes the proof. ■

Based on the result of Theorem 3, we define the map $K : V\mathbb{R}^N\backslash\{0\} \mapsto \mathbb{R}_+$:

$$K(\Omega) = \frac{u^*}{\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}}, \tag{80}$$

where, as before, u^* denotes the unique solution of (73) on $\mathcal{I}$, given Ω. Note that, given any realization of ω such that $V\omega \neq 0$, we have that $k_c = K(V\omega)$. We have the following corollary.

Corollary 7.

1. *For all* $\Omega \in V\mathbb{R}^N$, $\Omega \neq 0$, *we have that* $\|\Omega\|_\infty \leq K(\Omega) \leq 2\|\Omega\|_\infty$;
2. *there exists* $\Omega \in V\mathbb{R}^N$ *such that* $K(\Omega) = 2\|\Omega\|_\infty$ *if and only if* N *is even;*
3. *for every* $\epsilon > 0$ *there exist a positive integer* N *and* $\Omega \in V\mathbb{R}^N$ *such that* $|K(\Omega) - \|\Omega\|_\infty| < \epsilon$.

Proof. Part 1. We show that for all $\Omega \neq 0$, the solution u^* of (73) satisfies $u^* \leq \sqrt{2}\|\Omega\|_\infty$. The result then follows easily. Let $u' := \sqrt{2}\|\Omega\|_\infty$. Then we have that

$$2\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u}\right)^2} > \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{1-\left(\frac{\Omega_j}{u}\right)^2}} \quad \forall u > u'.$$

It follows that

$$K(\Omega) \leq \frac{u'}{\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u'}\right)^2}} \leq \frac{\sqrt{2}\|\Omega\|_\infty}{\frac{1}{2}\sqrt{2}} = 2\|\Omega\|_\infty$$

for all $\Omega \in V\mathbb{R}^N$. The lower bound $K(\Omega) \geq \|\Omega\|_\infty$ was obtained earlier in section 4.

Part 2. From the above it follows that $K(\Omega) = 2\|\Omega\|_\infty$ if and only if

$$2\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u'}\right)^2} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sqrt{1-\left(\frac{\Omega_j}{u'}\right)^2}}$$

or, equivalently, $\Omega_i^2 = \Omega_j^2$ for all (i,j). It is easy to see that this latter condition is never satisfied when N is odd (keeping in mind that $\sum_j \Omega_j = 0$). Now suppose N is even and pick any $c \neq 0$. Define

$$\Omega_i := \begin{cases} c, & i = 1, 2, \ldots, \frac{N}{2}; \\ -c, & i = \frac{N}{2}+1, \ldots, N. \end{cases}$$

Then we have that $\Omega \in V\mathbb{R}^N$. Moreover, $\Omega_i^2 = \Omega_j^2 = c^2$ for all (i,j). It follows that $K(\Omega) = 2\|\Omega\|_\infty$.

Part 3. Let $\epsilon > 0$ be given and suppose N is odd. Pick $c \neq 0$ and define

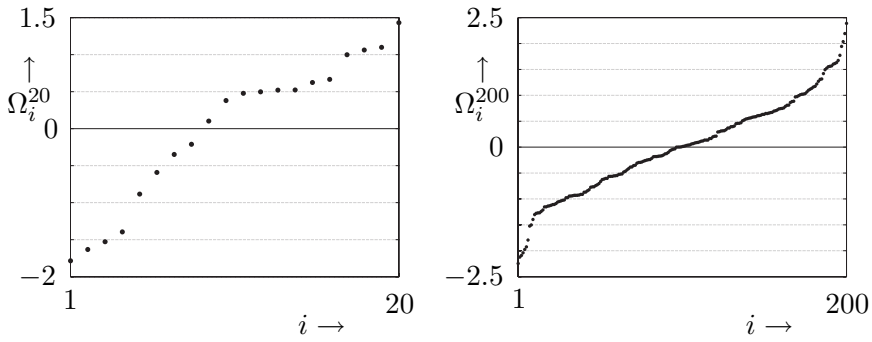
$$\Omega_i := \begin{cases} 0, & i = 1, 2, \ldots, N-1; \\ c, & i = N. \end{cases}$$

Then (73) evaluates to

$$2(N-1) + 2\sqrt{1-\left(\frac{c}{u}\right)^2} = (N-1) + \frac{1}{\sqrt{1-\left(\frac{c}{u}\right)^2}}, \tag{81}$$

and it is not hard to see that as N tends to infinity, the solution u^* of (81) tends to c. Indeed, for $N \geq 2$ we have

$$\left(\frac{c}{u^*}\right)^2 = \frac{1}{2}(N-1)\left(-\frac{1}{4}(N-1) + \frac{1}{4}\sqrt{(N-1)^2+8}\right). \tag{82}$$



(a) $N = 20$ (b) $N = 200$

Figure 5. *The vector of frequencies $\Omega_i^N := \omega_i^N - \langle \omega^N \rangle$, $N \in \{20, 200\}$, used in this example. The natural frequencies ω_i^N were sampled from a normal distribution with zero mean and unit variance and relabeled in such a way that $\omega_1^N \leq \omega_2^N \leq \cdots \leq \omega_N^N$.*

Let $\epsilon_1 > 0$, and pick N such that $\frac{1}{N} < \epsilon_1$ and $u^* < (1+\epsilon_1)c$. It follows that

$$\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2} < 1-\epsilon_1, \tag{83}$$

and hence

$$K(\Omega) := \frac{u^*}{\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{u^*}\right)^2}} < c\left(\frac{1+\epsilon_1}{1-\epsilon_1}\right) = c + 2\left(\frac{\epsilon_1}{1-\epsilon_1}\right)c. \tag{84}$$

Now let ϵ_1 be given as $\epsilon_1 := \frac{\epsilon}{2c+\epsilon}$ and choose N accordingly. It follows that $K(\Omega) < \|\Omega\|_\infty + \epsilon$. This concludes the proof. ■

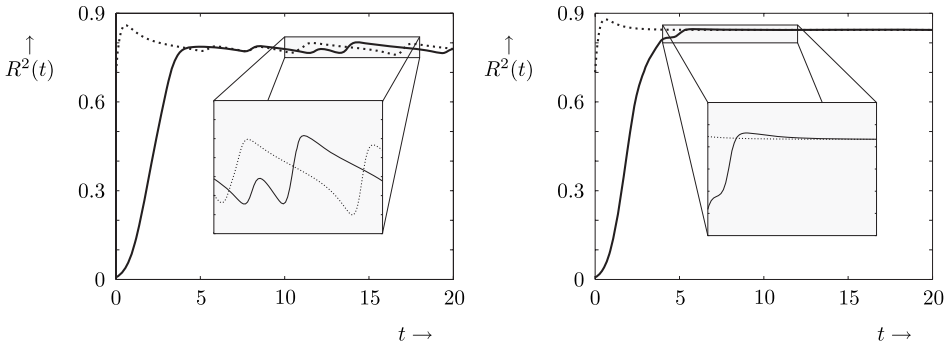
We are now ready to present our algorithm, which, given Ω, will compute u^* with user-defined precision $\epsilon > 0$ in a finite number of iterations, $n = \lceil \log_2(\frac{\|\Omega\|_\infty}{\epsilon}) \rceil + 1$.

Algorithm 1.

1. $a := \|\Omega\|_\infty$, $b := \sqrt{2}\|\Omega\|_\infty$
2. *While* $(b-a) > \epsilon$,
3. $\quad u := \frac{1}{2}(b-a)$.
4. $\quad$ *If* $\left[\sum_j \sqrt{\left(1-\frac{\Omega_j}{u}\right)^2} > \frac{1}{2}\sum_j \frac{1}{\sqrt{\left(1-\frac{\Omega_j}{u}\right)^2}}\right]$ *then* $a := u$, *else* $b := u$.
5. *End.*

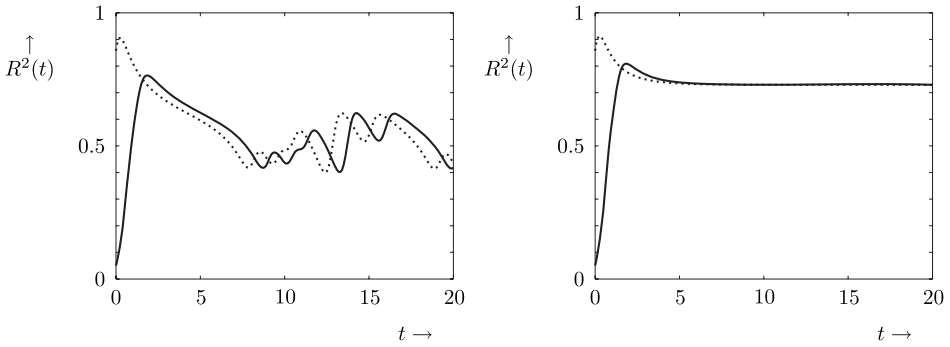
Once we have an estimate $\hat{u}$ of u^*, we can use (74), replacing u^* with $\hat{u}$, to estimate k_c.

7. Numerical example. We illustrate the results presented in this paper by means of a numerical example. We consider two systems with $N = 20$ and $N = 200$ oscillators, respectively, with frequencies $\{\Omega_i^N\}$, as depicted in Figures 5 and 6. The frequencies in this example were sampled from a normal distribution with zero mean and unit variance and relabeled such that $\omega_1^N \leq \omega_2^N \leq \cdots \leq \omega_N^N$ (note that this can be done without loss



(a) $k = 2.4$ (b) $k = 2.65$

Figure 6. *Case $N = 200$: Time evolution of the magnitude squared of the order parameter, $R^2(t)$, for two different initial conditions (indicated by a dashed and solid line, respectively) and two values of k. In the left panel, the value of k (2.4) is (well) below the known upper bound on k_c (2.6145), and the system does not converge to a fixed point; in the right panel the value of k (2.65) is slightly above the known upper bound on k_c, and the system converges to a fixed point, as expected.*



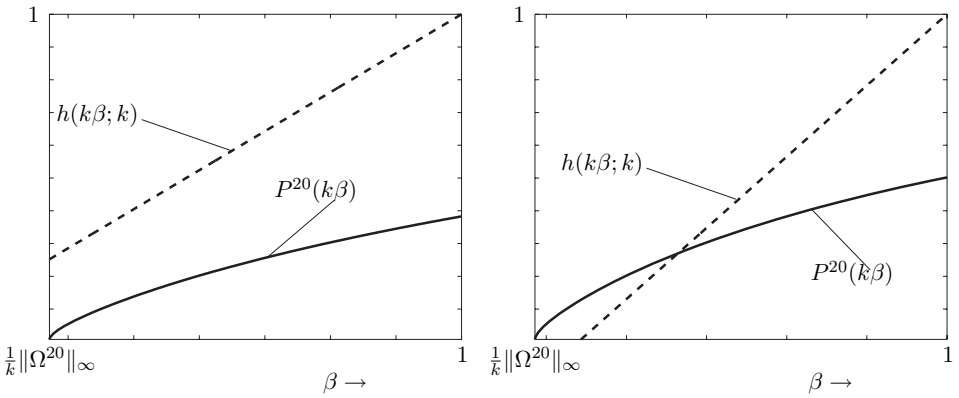
(a) $k = 2.1$ (b) $k = 2.3$

Figure 7. *Case $N = 20$: Time evolution of the magnitude squared of the order parameter, $R^2(t)$, for two different initial conditions (indicated by a dashed and solid line, respectively) and two values of k. In the left panel, the value of k (2.1) is slightly below the known upper bound on k_c (2.2281), and the system does not converge to a fixed point; in the right panel the value of k (2.3) is slightly above the known upper bound on k_c, and the system converges to a fixed point, as expected.*

of generality). For this particular realization of ω^{20} (ω^{200}), we have that $\|\Omega^{20}\|_\infty = 1.7858$ ($\|\Omega^{200}\|_\infty = 2.3893$) and

$$\frac{1}{N}\sum_{j=1}^{N}\sqrt{1-\left(\frac{\Omega_j}{\|\Omega\|_\infty}\right)^2} = 0.8015\ (0.9139).$$

It follows from Corollary 6 that $k_c \leq 2.2281$ (2.6145), and by (32), we have that $k_c \geq \|\Omega\|_\infty = 1.7858$ (2.3893). Figure 7 shows the time evolution of the magnitude squared of the order parameter, $R^2(t)$ (previously denoted as $L(t)$), for two different initial conditions and two values of the coupling coefficient, $k = 2.3$ and $k = 2.65$ ($k = 2.1$ and $k = 2.3$). We observe that when k is slightly greater than the known lower bound on k_c, the value of $R^2(t)$ converges



(a) $k = 2.1$ (b) $k = 2.3$

Figure 8. *Case $N = 20$: The graph of $P^{20}(k\beta)$ (71) versus β for $k = 2.1, 2.3$ and $\beta \in [\frac{1}{k}\|\Omega^{20}\|_\infty, 1]$. The dashed line is the graph of $h(k\beta;k) = \beta$ (72). An intersection corresponds to a solution of the fixed point equation $P(k\beta) = h(k\beta;k)$, and thus, by Theorem 2, to a fixed point of the system (17).*

to a constant and inspection shows that the solution $x(t)$ of the system (17) tends to a fixed point. On the other hand, when the coupling coefficient is slightly below the known upper bound on the critical coupling, the trajectories $x(t)$ appear not to converge. Note that in this case we do not know whether the system (17) has a fixed point or not, as the condition stated in Corollary 6 is only sufficient while at the same time the respective coupling strengths exceed their known lower bounds (1.7858 and 2.3893, respectively). To gain more insight into this situation, let us consider the case $N = 20$ in some more detail. We fix the coupling coefficient at $k = 2.1$ and numerically evaluate the function $P^{20}(k, \cdot)$,

$$P^{20}(k\beta) = \frac{1}{20}\sum_{j=1}^{20}\sqrt{1 - \left(\frac{\Omega_j^{20}}{k\beta}\right)^2}, \tag{85}$$

for several values of β in the interval $[\frac{1}{k}\|\Omega^{20}\|_\infty, 1]$. We repeat the same computation for $k = 2.3$. The result is shown in Figure 8. We observe that the equation $P^{20}(k\beta) = \beta$ does not have a solution on the interval $[\frac{1}{k}\|\Omega^{20}\|_\infty, 1]$ when $k = 2.1$ but does have a solution when $k = 2.3$.

We use Algorithm 1 to compute the "exact" value of the critical coupling to the fifth significant digit. We find that $k_c = 2.2198$ for the case $N = 20$ and $k_c = 2.6144$ for the case $N = 200$. Note that in both cases, but particularly in the latter, the upper bounds (2.2281 and 2.6145, respectively) provide good estimates of the true values of the critical coupling.

8. Conclusion. We derived necessary and sufficient conditions for the existence of fixed points in a finite system of coupled oscillators. In particular, we derived an easy sufficient condition in terms of the individual oscillator frequencies (Corollary 3), which we used to compute an upper bound on the critical coupling (Corollary 6). We showed that when no prior knowledge of the distribution of frequencies is available, we can still bound the critical coupling in terms of the infinity norm of the frequencies with their mean removed (Corollary 7).

These bounds were shown to be the tightest possible in the sense that we can find realizations of the intrinsic frequencies for which the upper bound is attained and others for which the critical coupling is arbitrarily close to the lower bound. Finally, we proposed an efficient algorithm (Algorithm 1) for computing the critical coupling to within arbitrary bounds in a finite number of steps. In future work we shall seek to extend the present analysis to complex networks of arbitrary topology and investigate more closely the impact of the shape of the distribution of intrinsic frequencies on the value of the critical coupling. We shall also consider the important question of stability and present analytical results for the limit case when the number of oscillators tends to infinity.

Acknowledgment. The authors would like to thank the anonymous reviewers for their thorough and insightful reviews which have helped to improve the paper.

REFERENCES

[1] J. Acebrón, L. Bonilla, C. Pérez Vicente, F. Ritort, and R. Spigler, *The Kuramoto model: A simple paradigm for synchronization phenomena*, Rev. Modern Phys., 77 (2005), pp. 137–185.
[2] A. Beuter, L. Glass, M. C. Mackey, and M. S. Titcombe, eds., *Nonlinear Dynamics in Physiology and Medicine*, Springer-Verlag, New York, 2003.
[3] N. Chopra and M. Spong, *On synchronization of Kuramoto oscillators*, in Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005, pp. 3916–3922.
[4] R. Diestel, *Graph Theory*, Springer-Verlag, New York, 2000.
[5] L. Glass, *Synchronization and rhythmic processes in physiology*, Nature Rev., 410 (2001), pp. 277–284.
[6] M. Golubitsky and I. Stewart, *Nonlinear dynamics of networks: The groupoid formalism*, Bull. Amer. Math. Soc. (N.S.), 43 (2006), pp. 305–364.
[7] H. Hong, M. Choi, and B. Kim, *Synchronization on small-world networks*, Phys. Rev. E (3), 65 (2002), 026139.
[8] H. Hong, H. Park, and M. Choi, *Collective synchronization in spatially extended systems of coupled oscillators with random frequencies*, Phys. Rev. E (3), 72 (2005), 036217.
[9] A. Jadbabaie, N. Motee, and M. Barahona, *On the stability of the Kuramoto model of coupled nonlinear oscillators*, in Proceedings of the American Control Conference, Boston, MA, 2004, pp. 4296–4301.
[10] Y. Kuramoto, *Self-entrainment of a population of coupled nonlinear oscillators*, in International Symposium on Mathematical Problems in Theoretical Physics, Lecture Notes in Phys. 39, H. Araki, ed., Springer-Verlag, Berlin, 1975, pp. 420–422.
[11] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, New York, 1984.
[12] Y. Maistrenko, O. Popovych, and P. Tass, *Desynchronization and Chaos in the Kuramoto Model*, Lecture Notes in Phys. 671, Springer-Verlag, Berlin, 2005.
[13] Y. Maistrenko, V. Popovych, O. Burylko, and P. Tass, *Mechanism of desynchronization in the finite-dimensional Kuramoto model*, Phys. Rev. Lett., 93 (2004), 084102.
[14] Y. Maistrenko, V. Popovych, and P. Tass, *Chaotic attractor in the Kuramoto model*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 3457–3466.
[15] Y. Moreno and A. Pacheco, *Synchronization of Kuramoto oscillators in scale-free networks*, Europhys. Lett., 68 (2004), pp. 603–609.
[16] J. D. Murray, *Mathematical Biology*, 3rd ed., Springer-Verlag, New York, 2002.
[17] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge University Press, Cambridge, UK, 2003.
[18] V. Popovych, Y. Maistrenko, and P. Tass, *Phase chaos in coupled oscillators*, Phys. Rev. E (3), 71 (2005), 065201.
[19] J. Rogge and D. Aeyels, *Existence of partial entrainment and stability of phase locking behavior of coupled oscillators*, Progr. Theoret. Phys., 112 (2004), pp. 921–942.

[20] J. Rogge and D. Aeyels, *Stability of phase locking in a ring of unidirectionally coupled oscillators*, J. Phys. A, 37 (2004), pp. 11135–11148.

[21] R. Sepulchre, *Oscillators as systems and synchrony as a design principle*, in Current Trends in Nonlinear Systems and Control, Systems Control Found. Appl., L. Menini, L. Zaccarian, and C. T. Abdallah, eds., Birkhäuser Boston, Boston, MA, 2006, pp. 123–141.

[22] R. Sepulchre, D. Paley, and N. E. Leonard, *Group coordination and cooperative control of steered particles in the plane*, in Group Coordination and Cooperative Control, Lecture Notes in Control and Inform. Sci. 336, K. Y. Pettersen, J. T. Gravdahl, and H. Nijmeijer, eds., Springer-Verlag, Berlin, 2006, pp. 217–232.

[23] S. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators*, Phys. D, 143 (2000), pp. 1–20.

[24] S. Strogatz, *Exploring complex networks*, Nature, 410 (2001), pp. 268–276.

[25] M. Verwoerd, *Fixed-point analysis of a finite system of Kuramoto oscillators*, in Proceedings of the Fourth Irish Conference on the Mathematical Foundations of Computer Science and Information Technology (MFCSIT) (Cork, 2006), University College Cork, National University of Ireland, Cork, Ireland, 2006, pp. 215–219.

[26] M. Verwoerd and O. Mason, *Conditions for the existence of fixed points in a finite system of Kuramoto oscillators*, in Proceedings of the American Control Conference (New York, 2007), American Automatic Control Council, Dayton, OH, 2007, pp. 4613–4618.

[27] D. Watts and S. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 161–185
© 2008 Society for Industrial and Applied Mathematics

Traveling Pulses and Wave Propagation Failure in Inhomogeneous Neural Media*

Zachary P. Kilpatrick†, Stefanos E. Folias‡, and Paul C. Bressloff†

Abstract. We use averaging and homogenization theory to study the propagation of traveling pulses in an inhomogeneous excitable neural network. The network is modeled in terms of a nonlocal integro-differential equation, in which the integral kernel represents the spatial distribution of synaptic weights. We show how a spatially periodic modulation of homogeneous synaptic connections leads to an effective reduction in the speed of a traveling pulse. In the case of large amplitude modulations, the traveling pulse represents the envelope of a multibump solution, in which individual bumps are nonpropagating and transient. The appearance (disappearance) of bumps at the leading (trailing) edge of the pulse generates the coherent propagation of the pulse. Wave propagation failure occurs when activity is insufficient to maintain bumps at the leading edge.

Key words. traveling waves, excitatory neural network, inhomogeneous media, homogenization, neural field theory, wave propagation failure

AMS subject classification. 92C20

DOI. 10.1137/070699214

1. Introduction.

Traveling waves of electrical activity have been observed in vivo in a number of sensory cortical areas including the somatosensory cortex of behaving rats [30], turtle and mollusk olfactory bulbs [22, 23], turtle cortex [34], and visuomotor cortices in the cat [36]. Such waves are usually seen during periods without sensory stimulation; the subsequent presentation of a stimulus then induces a switch to synchronous oscillatory behavior [13]. Traveling waves are also a characteristic feature of certain neurological disorders in humans, including epilepsy [8] and migraines [24]. Therefore, investigating the mechanisms underlying wave propagation in neural tissue is important for understanding both normal and pathological brain states. A common experimental paradigm is to record electrical activity in vitro using thin slices of cortical tissue, in which inhibition has been suppressed by blocking $GABA_A$ receptors with an antagonist such as bicuculline [7, 15, 40, 35, 32, 18]. Synchronized discharges can then be evoked by a weak electrical stimulus from any site on the cortical slice. Following rapid vertical propagation, each discharge propagates away from the stimulus in both horizontal directions at a mean velocity of about 6–9 cm/s. Although the conditions for wave propagation may differ from the intact cortex due to the removal of some long-range connections during slice preparation, the in vitro slice is more amenable to pharmacological manipulation and to multielectrode recordings.

*Received by the editors August 3, 2007; accepted for publication (in revised form) by B. Ermentrout October 5, 2007; published electronically January 23, 2008. This research was supported by NSF grants DMS-0515725 and IGERT-0217424.

http://www.siam.org/journals/siads/7-1/69921.html

†Department of Mathematics, University of Utah, 155 S 1400 E, Salt Lake City, UT 84112 (kilpatri@math.utah.edu, bressloff@math.utah.edu).

‡Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215 (sf@math.bu.edu).

Mathematical analyses of cortical wave propagation typically consider reduced one-dimensional network models. Under the additional assumption that the synaptic interactions are homogeneous, it has been shown that an excitatory neural network supports the propagation of a traveling front [12, 19, 5] or, in the presence of slow adaptation, a traveling pulse [39, 1, 31, 42, 43, 10, 11, 14, 38]. However, the patchy nature of long-range horizontal connections in superficial layers of certain cortical areas suggests that the cortex is more realistically modeled as an inhomogeneous neural medium. For example, in the primary visual cortex the horizontal connections tend to link cells with similar stimulus feature preferences such as orientation and ocular dominance [28, 41, 2]. Moreover, these patchy connections tend to be anisotropic, with the direction of anisotropy correlated with the underlying orientation preference map. Hence the anisotropic pattern of connections rotates approximately periodically across the cortex resulting in a periodic inhomogeneous medium [3, 4]. Another example of inhomogeneous horizontal connections is found in the prefrontal cortex [27, 29, 17], where pyramidal cells are segregated into stripes that are mutually connected via horizontally projecting axon collaterals; neurons within the gaps between stripes do not have such projections.

In this paper we investigate how a spatially periodic modulation of long-range synaptic weights affects the propagation of traveling pulses in a one-dimensional excitatory neural network, extending previous work on traveling fronts in neural network models [3] and reaction-diffusion systems [20, 21]. We proceed by introducing a slowly varying phase into the traveling wave solution of the unperturbed homogeneous network, and then we use perturbation theory to derive a dynamical equation for the phase, from which the mean speed of the wave can be calculated. We show that a periodic modulation of the long-range connections slows down the wave, and if the amplitude and wavelength of the periodic modulation is sufficiently large, then wave propagation failure can occur. A particularly interesting result of our analysis is that in the case of large amplitude modulations, the traveling pulse is no longer superthreshold everywhere within its interior, even though it still propagates as a coherent solitary wave. We find that the pulse now corresponds to the envelope of a multibump solution, in which individual bumps are nonpropagating and transient. The appearance (disappearance) of bumps at the leading (trailing) edge of the pulse generates the propagation of activity; propagation failure occurs when activity is insufficient to create new bumps at the leading edge.

2. Inhomogeneous network model. Consider a one-dimensional neural network model of the form [31]

$$\begin{aligned} \tau_m \frac{\partial u(x,t)}{\partial t} &= -u(x,t) + \int_{-\infty}^{\infty} w(x,x') f(u(x',t)) dx' - \beta v(x,t), \\ \frac{1}{\alpha}\frac{\partial v(x,t)}{\partial t} &= -v(x,t) + u(x,t), \end{aligned} \tag{2.1}$$

where $u(x,t)$ is the population activity at position $x \in \mathbb{R}$, τ_m is a membrane time constant, $f(u)$ is the output firing rate function, $w(x,x')$ is the excitatory connection strength from neurons at x' to neurons at x, and $v(x,t)$ is a local negative feedback mechanism, with β and α determining the relative strength and rate of feedback. This type of feedback, which could be spike frequency adaptation or synaptic depression, favors traveling waves [1, 31]. The nonlinearity f is a smooth monotonic increasing function,

$$f(u) = \frac{1}{1 + e^{-\eta(u-\kappa)}}, \tag{2.2}$$

where η is a gain parameter and κ is a threshold. As $\eta \to \infty$, $f \to H$, where $H(u) = \Theta(u - \kappa)$ and

$$\Theta[u] = \begin{cases} 0, & u \le 0, \\ 1, & u > 0. \end{cases} \tag{2.3}$$

The periodic microstructure of the cortex is incorporated by taking the weight distribution to be of the form [3, 4]

$$w(x, x') = W(|x - x'|) \left[1 + \mathcal{D}'\left(\frac{x'}{\varepsilon}\right)\right], \tag{2.4}$$

where $\mathcal{D}$ is a 2π-periodic function and ε determines the microscopic length-scale. (We consider the first-order derivative of $\mathcal{D}$ so that the zeroth-order harmonic is explicitly excluded.) It is important to note that (2.4) is a one-dimensional abstraction of the detailed anatomical structure found in the two-dimensional layers of real cortex. (See [6] for a more detailed discussion of cortical models.) However, it captures both the periodic-like nature of long-range connections and possible inhomogeneities arising from the fact that this periodicity is correlated with a fixed set of cortical feature maps.

For concreteness, we take the homogeneous weight function W to be an exponential,

$$W(x) = \frac{W_0}{2d} \mathrm{e}^{-|x|/d}, \tag{2.5}$$

where d is the effective range of the excitatory weight distribution, and set

$$\mathcal{D}(x) = \rho \sin(x), \quad 0 \le \rho < 1, \tag{2.6}$$

where ρ is the amplitude of the periodic modulation. We require that $0 \le \rho < 1$ so that the weight distribution remains nonnegative everywhere. Example plots of the resulting weight function $w(x, y)$ of (2.4) are shown in Figure 1 for fixed x. This illustrates both the periodic modulation and the associated network inhomogeneity, since the shape of the weight distribution varies periodically as the location x of the postsynaptic neuron shifts. Plotting $w(x, x')$ for fixed x' simply gives an exponential distribution whose maximum depends on x'. Finally, the temporal and spatial scales of the network are fixed by setting $\tau_m = 1$, $d = 1$, and the scale of the synaptic weights is fixed by setting $W_0 = 1$. The membrane time constant is typically around 10 ms and the length-scale of synaptic connections is typically 1 mm. Thus, in dimensionless units the speed of an experimentally measured wave will be $c = \mathcal{O}(1)$.

3. Averaging theory and homogenization. Our goal in this paper is to determine how the periodic modulation of the weight function affects properties of traveling pulses in the one-dimensional system obtained by substituting (2.4) into (2.1):

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} &= -u(x,t) + \int_{-\infty}^{\infty} W(|x - x'|) \left[1 + \mathcal{D}'\left(\frac{x'}{\varepsilon}\right)\right] f(u(x',t))dx' - \beta v(x,t), \\ \frac{1}{\alpha}\frac{\partial v(x,t)}{\partial t} &= -v(x,t) + u(x,t). \end{aligned} \tag{3.1}$$

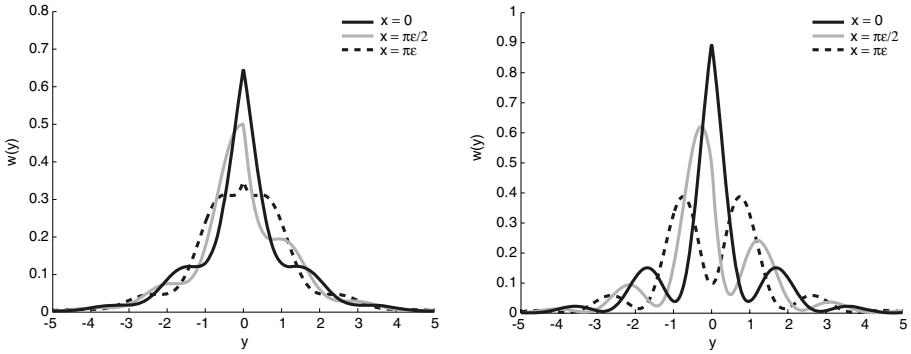


Figure 1. *Left: Weight kernel $w(y) = w(x, y)$ for a neuron centered at $x = 0, \pi\varepsilon/2, \pi\varepsilon$ when $\rho = 0.3$, and $\varepsilon = 0.3$. Right: Corresponding weight kernel when $\rho = 0.8$ and $\varepsilon = 0.3$.*

Assuming ε is a small parameter (in units of the space constant d), a zeroth-order approximation to (3.1) can be generated by performing spatial averaging with respect to the periodic weight modulation, leading to the homogeneous system given by (2.1) with weight distribution $w(x, x') = W(|x - x'|)$. Suppose that the homogeneous network supports the propagation of a traveling pulse of constant speed c. That is, $u(x, t) = U(\xi)$, $v(x, t) = V(\xi)$, where $\xi = x - ct$ is the traveling wave coordinate, and $U(\xi), V(\xi) \to 0$ as $\xi \to \pm\infty$. Substituting such a solution into (2.1) with $w(x, x') = W(|x - x'|)$ gives

$$-cU'(\xi) = -U(\xi) + \int_{-\infty}^{\infty} W(\xi - \xi')f(U(\xi'))d\xi' - \beta V(\xi),$$

$$-\frac{c}{\alpha}V'(\xi) = -V(\xi) + U(\xi). \tag{3.2}$$

Assuming the existence of a solution $(U(\xi), V(\xi))$ to (3.2), we would like to determine whether or not a traveling wave persists in the presence of the periodic weight modulation. A crucial requirement for trajectories of the averaged homogeneous system to remain sufficiently close to trajectories of the exact inhomogeneous system for sufficiently small ε is that solutions of the averaged system be structurally stable [16]. However, traveling pulses correspond to homoclinic trajectories within a dynamical systems framework and are thus not structurally stable. Therefore, one must go beyond lowest-order averaging to resolve differences between the homogeneous and inhomogeneous systems. We will proceed by carrying out a perturbation expansion in ε, extending previous work on traveling fronts in reaction-diffusion systems [20, 21] and excitable neural networks [3].

We begin by performing an integration by parts on the first equation in the system (3.1) so that

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} W(x - x')f(u(x',t))\mathrm{d}x' + \varepsilon \int_{-\infty}^{\infty} \mathcal{D}\left(\frac{x'}{\varepsilon}\right)\left[W'(x - x')f(u(x',t)) - W(x - x')\frac{\partial f(u(x',t))}{\partial x'}\right]\mathrm{d}x',$$

$$\frac{1}{\alpha}\frac{\partial v(x,t)}{\partial t} = -v(x,t) + u(x,t). \tag{3.3}$$

Although the inhomogeneous system is not translationally invariant, we can assume that perturbations about the homogeneous system will provide us with nearly translationally invariant solutions [20]. Thus, we perform the change of variables $\xi = x - \phi(t)$ and $\tau = t$ so that (3.3) becomes

$$\begin{aligned}\frac{\partial u(\xi,\tau)}{\partial \tau} &= -u(\xi,\tau) + \int_{-\infty}^{\infty} W(\xi-\xi')f(u(\xi',\tau))\mathrm{d}\xi' + \phi'\frac{\partial u(\xi,\tau)}{\partial \xi} \\ &\quad + \varepsilon\int_{-\infty}^{\infty}\mathcal{D}\left(\frac{\xi'+\phi}{\varepsilon}\right)\left[W'(\xi-\xi')f(u(\xi',\tau)) - W(\xi-\xi')\frac{\partial f(u(\xi',\tau))}{\partial \xi'}\right]\mathrm{d}\xi', \\ \frac{1}{\alpha}\frac{\partial v(\xi,\tau)}{\partial \tau} &= -v(\xi,\tau) + u(\xi,\tau) + \frac{\phi'}{\alpha}\frac{\partial v(\xi,\tau)}{\partial \xi}. \end{aligned} \tag{3.4}$$

Next perform the perturbation expansions

$$u(\xi,\tau) = U(\xi) + \varepsilon u_1(\xi,\tau) + \varepsilon^2 u_2(\xi,\tau) + \cdots, \tag{3.5}$$

$$v(\xi,\tau) = V(\xi) + \varepsilon v_1(\xi,\tau) + \varepsilon^2 v_2(\xi,\tau) + \cdots, \tag{3.6}$$

$$\phi'(\tau) = c + \varepsilon\phi_1'(\tau) + \varepsilon^2\phi_2'(\tau) + \cdots, \tag{3.7}$$

where $(U(\xi), V(\xi))^{\mathrm{T}}$ is a traveling pulse solution of the corresponding homogeneous system (see (3.2)) and c is the speed of the unperturbed pulse. The first-order terms u_1, v_1 satisfy

$$-\frac{\partial}{\partial\tau}\begin{pmatrix} u_1(\xi,\tau) \\ v_1(\xi,\tau)/\alpha \end{pmatrix} + \mathcal{L}\begin{pmatrix} u_1(\xi,\tau) \\ v_1(\xi,\tau) \end{pmatrix} = -\phi_1'(\tau)\begin{pmatrix} U'(\xi) \\ V'(\xi)/\alpha \end{pmatrix} + \begin{pmatrix} h_1(\xi,\frac{\phi}{\varepsilon}) \\ 0 \end{pmatrix}, \tag{3.8}$$

where

$$\mathcal{L}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} c\frac{du}{d\xi} - u + \int_{-\infty}^{\infty} W(\xi-\xi')f'(U(\xi'))u(\xi')d\xi' - \beta v \\ \frac{c}{\alpha}\frac{dv}{d\xi} - v + u \end{pmatrix} \tag{3.9}$$

for $u, v \in \mathcal{C}^1(\mathbb{R},\mathbb{C})$ and

$$h_1\left(\xi,\frac{\phi}{\varepsilon}\right) = -\int_{-\infty}^{\infty}\mathcal{D}\left(\frac{\xi'+\phi}{\varepsilon}\right)\left[W'(\xi-\xi')f(U(\xi')) - W(\xi-\xi')\frac{\mathrm{d}f(U(\xi'))}{\mathrm{d}\xi'}\right]\mathrm{d}\xi'. \tag{3.10}$$

The linear operator $\mathcal{L}$ has a one-dimensional null-space spanned by $(U'(\xi), V'(\xi))^{\mathrm{T}}$. The existence of $(U'(\xi), V'(\xi))^{\mathrm{T}}$ as a null-vector follows immediately from differentiating the homogeneous equation (3.2) and is a consequence of the translation invariance of the homogeneous system. Uniqueness can be shown using properties of positive linear operators. A bounded solution to (3.9) then exists if and only if the right-hand side is orthogonal to all elements of the null-space of the adjoint operator $\mathcal{L}^*$. The latter is defined according to the inner product relation

$$\int_{-\infty}^{\infty}(a(\xi)\ \ b(\xi))\mathcal{L}\begin{pmatrix} u(\xi) \\ v(\xi) \end{pmatrix}d\xi = \int_{-\infty}^{\infty}(u(\xi)\ \ v(\xi))\mathcal{L}^*\begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix}, \tag{3.11}$$

where $u(\xi)$, $v(\xi)$, $a(\xi)$, and $b(\xi)$ are arbitrary integrable functions. It follows that

$$\mathcal{L}^*\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -c\frac{da}{d\xi} - a + b + f'(U(\xi))\int_{-\infty}^{\infty} W(\xi-\xi')a(\xi')d\xi' \\ -\frac{c}{\alpha}\frac{db}{d\xi} - \beta a - b \end{pmatrix}. \tag{3.12}$$

The adjoint operator also has a one-dimensional null-space spanned by some vector $(A,B)^T$. (An explicit construction of this null-vector in the case of a Heaviside nonlinearity will be presented in section 4.) Therefore, for (3.9) to have a solution, it is necessary that

$$K\phi_1'(\tau) = \int_{-\infty}^{\infty} A(\xi) h_1\left(\xi, \frac{\phi}{\varepsilon}\right) d\xi, \tag{3.13}$$

where

$$K = \int_{-\infty}^{\infty} [A(\xi)U'(\xi) + \alpha^{-1}B(\xi)V'(\xi)]d\xi. \tag{3.14}$$

Substituting for h_1 using (3.10) leads to a first-order differential equation for the phase ϕ:

$$\frac{d\phi}{d\tau} = c - \varepsilon\Phi_1\left(\frac{\phi}{\varepsilon}\right), \tag{3.15}$$

where

$$\Phi_1\left(\frac{\phi}{\varepsilon}\right) = \frac{1}{K}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\xi)\mathcal{D}\left(\frac{\xi'+\phi}{\varepsilon}\right) \times \left[W'(\xi-\xi')f(U(\xi')) - W(\xi-\xi')\frac{df(U(\xi'))}{d\xi'}\right] d\xi' d\xi. \tag{3.16}$$

If the right-hand side of (3.15) is strictly positive, then there exists a traveling pulse of the approximate form $U(x-\phi(t))$ and of average speed $\bar{c} = 2\pi\varepsilon/T$ with

$$T = \int_0^{2\pi\varepsilon} \frac{d\phi}{c - \varepsilon\Phi_1(\phi/\varepsilon)}. \tag{3.17}$$

However, if the right-hand side of (3.15) vanishes for some ϕ, then the first-order analysis predicts wave propagation failure.

4. Calculation of average wave speed.

In this section we use (3.17) to calculate the average wave speed $\bar{c}$ as a function of ε in the limiting case of a Heaviside nonlinearity. Note that since derivatives of f always appear inside integral terms, the high gain limit $\eta \to \infty$ is well defined. One advantage of using a Heaviside nonlinearity is that all calculations can be carried out explicitly. Moreover, as previously shown for traveling fronts [3], in the case of smooth nonlinearities it is necessary to develop the perturbation analysis to $\mathcal{O}(\varepsilon^2)$ since the $\mathcal{O}(\varepsilon)$ terms may be exponentially small; see also section 4.3.

4.1. Homogeneous network with Heaviside nonlinearity.

The existence (and stability) of single bump traveling pulse solutions in the homogeneous network obtained by setting $f = H$ and $w(x,x') = W(|x-x'|)$ in (2.1) has been studied by a number of authors [31, 33, 42, 10, 14]. A single bump solution $(U(\xi), V(\xi))$ is one for which U is above threshold over a domain of length a, corresponding to the width of the bump, and subthreshold everywhere else. In other words, the activity U crosses threshold at only two points, which by translation invariance can be taken to be $\xi = -a, 0$:

$$U(-a) = U(0) = \kappa; \qquad U(\xi) \longrightarrow 0 \ \text{ as } \ \xi \longrightarrow \pm\infty;$$
$$U(\xi) > \kappa, \ \ -a < \xi < 0; \qquad U(\xi) \ < \ \kappa \qquad \text{otherwise.}$$

It follows from (3.2) with $f = H$ that

$$-c\,U_\xi = -U - \beta V + \int_{-a}^{0} W(\xi-\eta)d\eta,$$
$$-\frac{c}{\alpha}\,V_\xi = -V \;+\; U. \tag{4.1}$$

One way to solve this pair of equations is to use variation of parameters [42, 14]. For completeness, we present the details of this calculation here, since some of the results will be used in our subsequent analysis.

Let $\mathbf{s} = (U, V)^T$ and rewrite the system as

$$\mathcal{L}\mathbf{s} \equiv \begin{pmatrix} cU'(\xi) - U(\xi) - \beta V(\xi) \\ cV'(\xi) + \alpha U(\xi) - \alpha V(\xi) \end{pmatrix} = -\begin{pmatrix} N_e(\xi) \\ 0 \end{pmatrix}, \tag{4.2}$$

where

$$N_e(\xi) = \Omega(\xi + a) - \Omega(\xi), \quad \Omega(\xi) = \int_{-\infty}^{\xi} W(\xi')d\xi'. \tag{4.3}$$

We solve (4.2) using variation of parameters. The homogeneous problem $\mathcal{L}\mathbf{s} = \mathbf{0}$ has two linearly independent solutions,

$$\mathbf{S}_+(\xi) = \begin{pmatrix} \beta \\ m_+ - 1 \end{pmatrix} \exp(\mu_+\xi), \qquad \mathbf{S}_-(\xi) = \begin{pmatrix} \beta \\ m_- - 1 \end{pmatrix} \exp(\mu_-\xi),$$

where

$$\mu_\pm = \frac{m_\pm}{c}, \qquad m_\pm = \frac{1}{2}(1 + \alpha \pm \sqrt{(1-\alpha)^2 - 4\mu\beta}). \tag{4.4}$$

We shall work in the parameter regime where $\mu_\pm$ are real, though interesting behavior can arise when $\mu_\pm$ is complex [38]. Thus, set

$$\mathbf{s}(\xi) = [\mathbf{S}_+|\mathbf{S}_-] \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix},$$

where $a, b \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$, and $[A|B]$ denotes the matrix whose first column is A and whose second column is B. Since $\mathcal{L}\mathbf{S}_\pm = 0$, (4.2) becomes

$$[\mathbf{S}_+|\mathbf{S}_-] \frac{\partial}{\partial\xi} \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = -\frac{1}{c} \begin{pmatrix} N_e(\xi) \\ 0 \end{pmatrix}. \tag{4.5}$$

Since $[\mathbf{S}_+|\mathbf{S}_-]$ is invertible, we find that

$$\frac{\partial}{\partial\xi} \begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = -\frac{1}{c\beta(m_+ - m_-)} [\mathbf{Z}_+|\mathbf{Z}_-]^T \begin{pmatrix} N_e(\xi) \\ 0 \end{pmatrix},$$

where

$$\mathbf{Z}_+(\xi) = \begin{pmatrix} 1 - m_- \\ \beta \end{pmatrix} \exp(-\mu_+\xi), \qquad \mathbf{Z}_-(\xi) = -\begin{pmatrix} 1 - m_+ \\ \beta \end{pmatrix} \exp(-\mu_-\xi).$$

For $c > 0$, we can integrate from ξ to ∞ and find

$$\begin{pmatrix} a(\xi) \\ b(\xi) \end{pmatrix} = \begin{pmatrix} a_\infty \\ b_\infty \end{pmatrix} + \frac{1}{c\beta(m_+ - m_-)} \int_\xi^\infty [\mathbf{Z}_+|\mathbf{Z}_-]^T \begin{pmatrix} N_e(\xi') \\ 0 \end{pmatrix} d\xi',$$

where a_∞, b_∞ are the values of $a(\xi)$, $b(\xi)$ as $\xi \to \infty$. Thus

$$\mathbf{s}(\xi) = [\mathbf{S}_+|\mathbf{S}_-] \begin{pmatrix} a_\infty \\ b_\infty \end{pmatrix} + \frac{1}{c\beta(m_+ - m_-)} [\mathbf{S}_+|\mathbf{S}_-] \int_\xi^\infty [\mathbf{Z}_+|\mathbf{Z}_-]^T \begin{pmatrix} N_e(\xi') \\ 0 \end{pmatrix} d\xi'. \tag{4.6}$$

Using Hölder's inequality and that $N_e \in \mathcal{C}^0(\mathbb{R}, \mathbb{R})$, we can show that the integral in (4.6) is bounded for all $\xi \in R$. Thus, a bounded solution $\mathbf{s}$ exists if $a_\infty = b_\infty = 0$. Our general traveling pulse solution is given by

$$\mathbf{s}(\xi) = \frac{1}{c\beta(m_+ - m_-)} [\mathbf{S}_+|\mathbf{S}_-] \int_\xi^\infty [\mathbf{Z}_+|\mathbf{Z}_-]^T \begin{pmatrix} N_e(\xi') \\ 0 \end{pmatrix} d\xi'.$$

Furthermore, if we define

$$\mathcal{M}_\pm(\xi) = \frac{1}{c(m_+ - m_-)} \int_\xi^\infty \mathrm{e}^{\mu_\pm(\xi - \xi')} N_e(\xi') d\xi',$$

we can express our solution (U, V) as

$$U(\xi) = (1 - m_-)\mathcal{M}_+(\xi) - (1 - m_+)\mathcal{M}_-(\xi), \tag{4.7}$$

$$V(\xi) = \beta^{-1}(m_+ - 1)(1 - m_-)\left[\mathcal{M}_+(\xi) - \mathcal{M}_-(\xi)\right]. \tag{4.8}$$

Since $N_e(\xi)$ is dependent upon the pulse width a, the threshold conditions $U(-a) = U(0) = \kappa$ lead to the following consistency conditions for the existence of a traveling pulse:

$$\kappa = (1 - m_-)\mathcal{M}_+(-a) - (1 - m_+)\mathcal{M}_-(-a), \tag{4.9}$$

$$\kappa = (1 - m_-)\mathcal{M}_+(0) - (1 - m_+)\mathcal{M}_-(0). \tag{4.10}$$

This pair of nonlinear equations determines the pulse width a and wave speed c of a single bump traveling wave solution as a function of the various parameters of the network. For a given weight distribution $W(x)$, existence of such a solution is established if a solution for a, c can be found, and provided that U does not cross threshold at any other points besides $\xi = -a, 0$. Recently the existence (and stability) of single bump traveling waves has been examined for quite a general class of weight distributions [32] which includes both exponential and Gaussian distributions. For concreteness, we consider the exponential weight function (2.5) with $W_0 = d = 1$. Numerically solving (4.9) and (4.10) for a and c as a function of the adaptation rate α yields the existence curves shown in Figure 2. This figure illustrates the well-known result that for sufficiently slow adaptation (small α) there exists a pair of traveling pulses with the fast/wide pulse stable and the slow/narrow pulse unstable [31]. Also shown in the figure is the stability of the various solution branches, which can be determined analytically using an Evans function approach [42, 10, 14, 32].

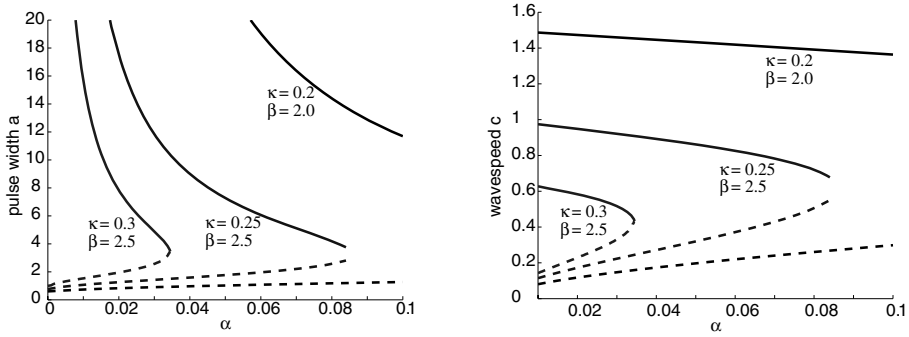


Figure 2. *Existence curves for a single bump traveling pulse solution of* (2.1) *in the case of a homogeneous network with an exponential weight distribution* $W(x) = \mathrm{e}^{-|x|/2}$. *Left: Existence curves in the* (α, a) *plane. Right: Existence curves in the* (α, c) *plane. Pulses exist only for small enough* α *(sufficiently slow adaptation). For each parameter set, there exists a stable branch (solid) of wide/fast pulses and an unstable branch (dashed) of narrow/slow pulses. In the case* $\kappa = 0.3$ *the branches annihilate at a saddle-node bifurcation at a critical value* α_c. *In the other two cases, the branches end abruptly due to the fact that the eigenvalues* $\mu_\pm$ *become complex-valued* [38].

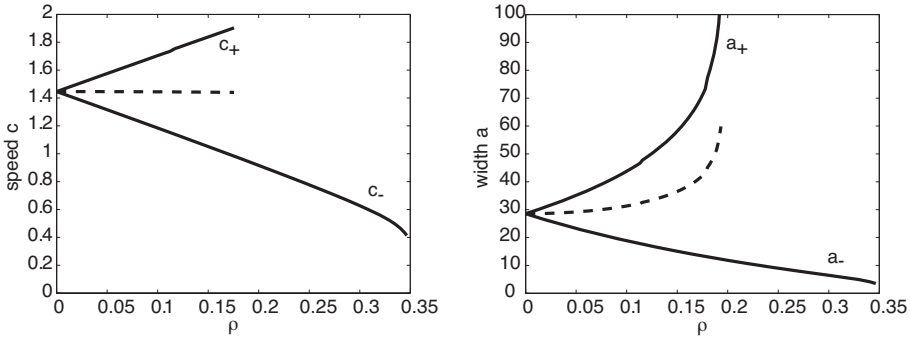


Figure 3. *Existence curves for a single bump traveling pulse solution of* (2.1) *in the case of a homogeneous network with an exponential weight distribution* $W(x) = W_\pm \mathrm{e}^{-|x|/2}$ *with* $W_\pm = 1 \pm \rho$. *Left: Plot of wave speed* $c_\pm$ *as a function of* ρ *for weight amplitudes* $W_\pm$. *The dashed curve indicates the arithmetic mean of pair* $c_\pm$. *The slower branch terminates at around* $\rho = 0.35$ *due to a saddle-node bifurcation. The faster branch terminates due to a blow-up of the pulse width. Right: Plot of pulse width* $a_\pm$ *as a function of* ρ *for weight amplitudes* $W_\pm$.

The analysis of existence in a homogeneous network also provides some insight into what happens when we include a periodic modulation of the weights according to (2.4) and (2.6). Such a modulation induces a periodic variation in the amplitude W_0 of the exponential weight distribution (2.5) between the limiting values $W_\pm = (1 \pm \rho)W_0$. This suggests that the speed of a wave in the inhomogeneous network will be bounded by the speeds $c_\pm$ of a traveling wave in the corresponding homogeneous network obtained by taking $W_0 \to W_\pm$. Note that rescaling the weight distribution in (4.9) and (4.10) is equivalent to rescaling the threshold according to $\kappa \to \kappa/(1 \pm \rho)$. In Figure 3 we plot the speeds $c_\pm$ and the corresponding pulse widths $a_\pm$ as a function of ρ. For sufficiently small ρ, the wave speed c_+ increases with ρ at approximately the same rate as c_- decreases so that their arithmetic mean remains constant. Therefore, one might expect that a periodic variation in weights would lead to a corresponding periodic

variation in wave speed such that the mean wave speed is approximately independent of ρ. However, when a pulse enters a region of enhanced synaptic weights, the resulting increase in wave speed coincides with a rapid increase in pulse width as a function of ρ. Thus, the pulse will tend to extend into neighboring regions of reduced synaptic weights and the resulting spatial averaging will counteract the speeding up of the wave. On the other hand, when a pulse enters a region of reduced synaptic weights, the reduction in wave speed coincides with a reduction in pulse width so that spatial averaging can no longer be carried out as effectively. (The effectiveness of spatial averaging will depend on the ratio of the pulse width a to the periodicity $2\pi\varepsilon$ of the weight modulation.) Hence, we expect regions where the weights are reduced to have more effect on wave propagation than regions where the weights are enhanced, suggesting that a periodic weight modulation leads to slower, narrower waves. This is indeed found to be the case, both in our perturbation analysis (section 4.2) and in our numerical simulations (section 5). Interestingly, we also find that traveling waves persist for larger values of ρ than predicted by our analysis of single bumps in homogeneous networks, although such waves tend to consist of multiple bumps (see section 5).

4.2. Inhomogeneous network with Heaviside nonlinearity. Suppose that the homogeneous network with a Heaviside nonlinearity supports a stable traveling wave solution $(U(\xi), V(\xi))^T$ of wave speed c. As shown in section 4.1, a stable/unstable pair of traveling waves exists for sufficiently slow adaptation. In order to calculate the average wave speed $\bar{c}$ for nonzero ε and ρ (see (3.17)), we first need to compute the null-vector $(A(\xi), B(\xi))^T$ of the adjoint operator $\mathcal{L}^*$ defined by (3.12). In the case of a Heaviside nonlinearity,

$$\begin{aligned} -c\frac{dA(\xi)}{d\xi} - A(\xi) + B(\xi) + \frac{\delta(\xi)}{|U'(0)|}\int_{-\infty}^{\infty} W(\xi-\xi')A(\xi')d\xi' \\ + \frac{\delta(\xi+a)}{|U'(-a)|}\int_{-\infty}^{\infty} W(\xi-\xi')A(\xi')d\xi' = 0, \\ -\frac{c}{\alpha}\frac{dB(\xi)}{d\xi} - \beta A(\xi) - B(\xi) = 0. \end{aligned} \tag{4.11}$$

For $\xi \neq 0, -a$, this has solutions of the form $(A(\xi), B(\xi)^T) = \mathbf{u}e^{-\lambda\xi}$ with associated characteristic equation $\mathbf{M}\mathbf{u} = c\lambda\mathbf{u}$ and

$$\mathbf{M} = \begin{pmatrix} 1 & -1 \\ \beta\alpha & \alpha \end{pmatrix}. \tag{4.12}$$

The eigenvalues are $\lambda = \mu_\pm = m_\pm/c$ with $m_\pm$ given by (4.4). The corresponding eigenvectors are

$$\mathbf{u}_\pm = \begin{pmatrix} 1 \\ 1 - m_\pm \end{pmatrix}. \tag{4.13}$$

The presence of the Dirac delta functions at $\xi = 0, -a$ then suggests that we take the null-solution to be of the form

$$\begin{aligned} \mathbf{V}^*(\xi) = \gamma_+\mathbf{u}_+\left[e^{-\mu_+\xi}\Theta(\xi) + \chi e^{-\mu_+(\xi+a)}\Theta(\xi+a)\right] \\ + \gamma_-\mathbf{u}_-\left[e^{-\mu_-\xi}\Theta(\xi) + \chi e^{-\mu_-(\xi+a)}\Theta(\xi+a)\right] \end{aligned} \tag{4.14}$$

with the coefficients $\gamma_\pm$ chosen such that the Dirac delta function terms that come from differentiating the null-vector appear only in the $A(\xi)$ term,

$$\gamma_+ \mathbf{u}_+ + \gamma_- \mathbf{u}_- = \begin{pmatrix} \Gamma \\ 0 \end{pmatrix}, \tag{4.15}$$

and χ is a constant yet to be determined. Taking

$$\gamma_\pm = \pm(1 - m_\mp), \tag{4.16}$$

we have $\Gamma = m_+ - m_-$.

In order to determine χ, substitute (4.14) into (4.11) to obtain the pair of equations

$$c(m_+ - m_-) = \frac{1}{|U'(0)|}(\Lambda(0) + \chi\Lambda(-a)) \tag{4.17}$$

and

$$\chi c(m_+ - m_-) = \frac{1}{|U'(-a)|}(\Lambda(a) + \chi\Lambda(0)) \tag{4.18}$$

with

$$\Lambda(\zeta) = \int_0^\infty [(1 - m_-)W(\xi + \zeta)e^{-\mu_+\xi} - (1 - m_+)W(\xi + \zeta)e^{-\mu_-\xi}]d\xi. \tag{4.19}$$

We require that (4.17) and (4.18) be consistent with the formula for $U'(\xi)$ obtained by differentiating (4.7) with respect to ξ:

$$\begin{aligned} U'(\xi) = &\frac{1 - m_-}{c(m_+ - m_-)} \int_\xi^\infty e^{\mu_+(\xi - \xi')}[W(\xi' + a) - W(\xi')]d\xi' \\ &- \frac{1 - m_+}{c(m_+ - m_-)} \int_\xi^\infty e^{\mu_-(\xi - \xi')}[W(\xi' + a) - W(\xi')]d\xi'. \end{aligned} \tag{4.20}$$

It follows that

$$|U'(0)| = -U'(0) = \frac{\Lambda(0) - \Lambda(a)}{c(m_+ - m_-)}, \quad |U'(-a)| = U'(-a) = \frac{\Lambda(0) - \Lambda(-a)}{c(m_+ - m_-)},$$

which, together with (4.17) and (4.18), imply

$$\Lambda(0) - \Lambda(a) = \Lambda(0) + \chi\Lambda(-a), \quad \chi(\Lambda(0) - \Lambda(-a)) = \Lambda(a) + \chi\Lambda(0).$$

Hence, (4.14) is a solution provided that

$$\chi = -\frac{\Lambda(a)}{\Lambda(-a)}. \tag{4.21}$$

This is also a constructive proof that the adjoint linear operator $\mathcal{L}^*$ for a Heaviside nonlinearity has a one-dimensional null-space spanned by $\mathbf{V}^*$.

Having found the null-solution (4.14), we now determine the phase function Φ_1 given by (3.16) with $f = H$. First, the constant K of (3.14) is evaluated by substituting for $(A(\xi), B(\xi))$ using (4.14) and substituting for $(U(\xi), V(\xi))$ using (4.7) and (4.8). The rather lengthy expression for K is given in the appendix. Next, we evaluate the double integral on the right-hand side of (3.16) by setting $\mathcal{D}(x) = \mathrm{e}^{ix}$ and using Fourier transforms. This gives

$$K\Phi_1\left(\frac{\phi}{\varepsilon}\right) = \frac{i}{\varepsilon}\mathrm{e}^{i\phi/\varepsilon}\int_{-\infty}^{\infty} W(x)\left[\int_{-\infty}^{\infty}\mathrm{e}^{iqx}\widetilde{A}^*(q)\widetilde{f(U)}(q+\varepsilon^{-1})\frac{\mathrm{d}q}{2\pi}\right]dx, \tag{4.22}$$

where * denotes complex conjugate and

$$\widetilde{A}(q) = \int_{-\infty}^{\infty}\mathrm{e}^{iqx}A(x)\mathrm{d}x. \tag{4.23}$$

In the case of a Heaviside nonlinearity and a pulse of width a, $f(U(\xi)) = \Theta(\xi + a) - \Theta(\xi)$, and $A(x)$ is given explicitly by the first component of the null-vector in (4.14). Taking Fourier transforms of these expressions shows that

$$\widetilde{A}(q) = -\left(1+\chi\mathrm{e}^{-iqa}\right)\left[\frac{\gamma_+}{iq-\mu_+} + \frac{\gamma_-}{iq-\mu_-}\right], \quad \widetilde{f(U)}(q) = \frac{1-\mathrm{e}^{-iqa}}{iq-0^+}. \tag{4.24}$$

If these Fourier transforms are now substituted into (4.22), we have

$$\begin{aligned} K\Phi_1\left(\frac{\phi}{\varepsilon}\right) = \frac{\mathrm{e}^{i\phi/\varepsilon}}{\varepsilon}\int_{-\infty}^{\infty} W(x)\Bigg[\int_{-\infty}^{\infty}\Bigg\{&\frac{\gamma_+(1-\mathrm{e}^{-i(q+\varepsilon^{-1})a}+\chi\mathrm{e}^{iqa}-\chi\mathrm{e}^{-ia/\varepsilon})\mathrm{e}^{iqx}}{(q+\varepsilon^{-1}+i0^+)(q-i\mu_+)} \\ &+\frac{\gamma_-(1-\mathrm{e}^{-i(q+\varepsilon^{-1})a}+\chi\mathrm{e}^{iqa}-\chi\mathrm{e}^{-ia/\varepsilon})\mathrm{e}^{iqx}}{(q+\varepsilon^{-1}+i0^+)(q-i\mu_-)}\Bigg\}\frac{\mathrm{d}q}{2\pi i}\Bigg]\mathrm{d}x. \end{aligned} \tag{4.25}$$

The resulting integral over q can be evaluated by closing the contour in the upper-half or lower-half complex q-plane depending on the sign of $x, x \pm a$. We find that there are only contributions from the poles at $q = i\mu_\pm$ with $\mu_\pm > 0$, whereas there is a removable singularity at $q = -\varepsilon^{-1} - i0^+$. Thus

$$\begin{aligned} K\Phi_1\left(\frac{\phi}{\varepsilon}\right) = &\frac{\gamma_+\mathrm{e}^{i\phi/\varepsilon}}{\varepsilon(\varepsilon^{-1}+i\mu_+)}\left[\left(1-\chi\mathrm{e}^{-ia/\varepsilon}\right)\widehat{\Omega}_+(0)+\chi\widehat{\Omega}_+(-a)-\mathrm{e}^{-ia/\varepsilon}\widehat{\Omega}_+(a)\right] \\ &+\frac{\gamma_-\mathrm{e}^{i\phi/\varepsilon}}{\varepsilon(\varepsilon^{-1}+i\mu_-)}\left[\left(1-\chi\mathrm{e}^{-ia/\varepsilon}\right)\widehat{\Omega}_-(0)+\chi\widehat{\Omega}_-(-a)-\mathrm{e}^{-ia/\varepsilon}\widehat{\Omega}_-(a)\right], \end{aligned} \tag{4.26}$$

with

$$\widehat{\Omega}_\pm(s) = \int_0^\infty W(x+s)\mathrm{e}^{-\mu_\pm x}\mathrm{d}x. \tag{4.27}$$

Taking the imaginary part of the above equation then determines the phase function $K\Phi_1$ for $\mathcal{D}(x) = \rho\sin(x)$. After a straightforward calculation, we find that

$$\begin{aligned} \frac{K}{\rho}\Phi_1\left(\frac{\phi}{\varepsilon}\right) = &(\Xi_+ + \Xi_-)\sin\left(\frac{\phi}{\varepsilon}\right) + (\Pi_+ + \Pi_-)\sin\left(\frac{\phi-a}{\varepsilon}\right) \\ &+ (\Upsilon_+ + \Upsilon_-)\cos\left(\frac{\phi}{\varepsilon}\right) + (\Psi_+ + \Psi_-)\cos\left(\frac{\phi-a}{\varepsilon}\right), \end{aligned} \tag{4.28}$$

where the explicit expressions for $\Xi_\pm, \Pi_\pm, \Upsilon_\pm, \Psi_\pm$ are given in the appendix.

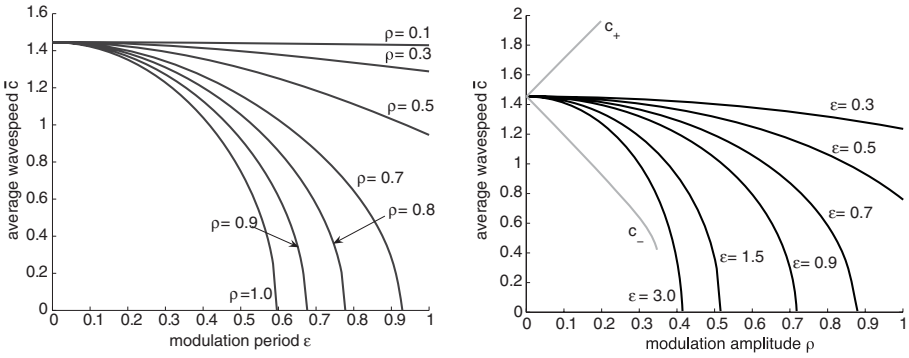


Figure 4. *Left: Average wave speed $\bar{c}$ versus ε for various values of the modulation amplitude ρ. The critical value of ε for wave propagation failure decreases as ρ increases. Right: Average wave speed $\bar{c}$ versus ρ for various values of the modulation period ε. For the sake of comparison, the speed curves previously plotted in Figure 3 for a homogeneous network are also shown (gray curves). Other parameters are $\kappa = 0.2$, $\alpha = 0.04$, and $\beta = 2.0$.*

Finally, we numerically calculate the average wave speed $\bar{c}$ by substituting (4.28) into (3.17). Note that we use the exact expression for Φ_1 that includes all higher-order terms in ε, rather than keeping only the $\mathcal{O}(1)$ term, since this gives a better estimate of the wave speed. In Figure 4 we show some example plots of $\bar{c}$ as a function of ε and ρ. It can be seen that for each choice of parameters, $\bar{c}$ is a monotonically decreasing function of ε and ρ, with $\bar{c}$ approaching the speed c of the homogeneous wave in the limits $\varepsilon \to 0$ and $\rho \to 0$. Hence, although the periodic modulation enhances the strength of connections in some regions and reduces them in others compared to the homogeneous case (see Figure 1), the net result is an effective reduction in wave speed. This is consistent with our discussion of Figure 3 in section 4.1, where we used a spatial averaging argument combined with the observation that faster waves are wider to infer that regions of reduced synaptic weights affect wave propagation more than regions of enhanced weights. Figure 4 also suggests that for sufficiently small ε there exists a traveling wave solution for all ρ, $0 \leq \rho < 1$, whereas for larger values of ε there is a critical value ρ_c beyond which propagation failure occurs. That is, $\bar{c} \to 0$ as $\rho \to \rho_c$, and this critical value decreases as the periodicity ε of the inhomogeneity increases. Similarly, for sufficiently large ρ there exists a critical period ε_c such that $\bar{c} \to 0$ as $\varepsilon \to \varepsilon_c$. Analogous results were previously obtained for traveling fronts in a scalar equation [3]. It is important to bear in mind that the calculation of $\bar{c}$ is based on the $\mathcal{O}(\varepsilon)$ perturbation analysis of section 3, although we do include higher-order terms in the calculation of Φ_1. This raises the important question as to whether or not our analysis correctly predicts wave propagation failure in the full system, given that $\bar{c}$ tends to approach zero at relatively large values of ε and ρ. Moreover, the perturbation analysis does not determine the stability of the wave so that propagation failure could occur due to destabilization of the wave for $\rho < \rho_c$ or $\varepsilon < \varepsilon_c$. This will indeed turn out to be the case as we show in section 5, where we present numerical solutions of (2.1) and provide further insight into the mechanism for propagation failure.

4.3. Smooth nonlinearities and higher-order corrections. In the case of smooth nonlinearities, the Fourier transforms $\widetilde{A}(q)$ and $\widetilde{f(U)}(q)$ appearing in (4.22) no longer have simple

poles, and in general Φ_1 will consist of exponentially small terms. It follows that Φ_1 may be less significant than the $\mathcal{O}(\varepsilon^2)$ terms ignored in the perturbation expansion of (3.4). Therefore, following the treatment of traveling fronts [3], we carry out a perturbation expansion of system (3.4) to $\mathcal{O}(\varepsilon^2)$. This yields an equation for (u_2, v_2) of the form

$$-\frac{\partial}{\partial\tau}\begin{pmatrix} u_2(\xi,\tau) \\ v_2(\xi,\tau)/\alpha \end{pmatrix} + \mathcal{L}\begin{pmatrix} u_2(\xi,\tau) \\ v_2(\xi,\tau) \end{pmatrix} = -\phi_2'(\tau)\begin{pmatrix} U'(\xi) \\ V'(\xi)/\alpha \end{pmatrix} - \phi_1'(\tau)\begin{pmatrix} u_1'(\xi) \\ v_1'(\xi)/\alpha \end{pmatrix} + \begin{pmatrix} h_2(\xi,\frac{\phi}{\varepsilon}) \\ 0 \end{pmatrix}, \tag{4.29}$$

where $\mathcal{L}$ is defined by (3.9) and

$$\begin{aligned} h_2\left(\xi,\frac{\phi}{\varepsilon}\right) &= -\frac{1}{2}\int_{-\infty}^{\infty} W(\xi-\xi')f''(U(\xi'))[u_1(\xi')]^2 \mathrm{d}\xi' \\ &\quad - \int_{-\infty}^{\infty} \mathcal{D}\left(\frac{[\xi'+\phi]}{\varepsilon}\right) W'(\xi-\xi')f'(U(\xi'))u_1(\xi')\mathrm{d}\xi' \\ &\quad + \int_{-\infty}^{\infty} \mathcal{D}\left(\frac{[\xi'+\phi]}{\varepsilon}\right) W(\xi-\xi')[f'(U(\xi'))u_1'(\xi') + f''(U(\xi'))U'(\xi')u_1(\xi)]\mathrm{d}\xi'. \end{aligned} \tag{4.30}$$

The existence of a bounded solution requires the solvability conditions (3.13) and

$$K\phi_2'(\tau) + L(\tau)\phi_1'(\tau) = \int_{-\infty}^{\infty} A(\xi)h_2\left(\xi,\frac{\phi}{\varepsilon}\right)\mathrm{d}\xi, \tag{4.31}$$

where

$$L(\tau) = \int_{-\infty}^{\infty}\left[A(\xi)\frac{\partial u_1(\xi,\tau)}{\partial\xi} + \alpha^{-1}B(\xi)\frac{\partial v_1(\xi,\tau)}{\partial\xi}\right]\mathrm{d}\xi. \tag{4.32}$$

In order to evaluate the solvability condition (4.31), we must first determine $u_1(\xi,\phi/\varepsilon)$ from (3.8). If we choose $\mathcal{D}(x)$ to be a sinusoid, then $u_1(\xi,\phi/\varepsilon)$ will include terms that are proportional to $\sin(\phi/\varepsilon)$ and $\cos(\phi/\varepsilon)$. Thus substituting $u_1(\xi/\phi/\varepsilon)$ into (4.30) will generate terms of the form $\sin^2(\phi/\varepsilon)$ and $\cos^2(\phi/\varepsilon)$ due to the quadratic term in u_1. Using the identities $2\sin^2(x) = 1-\cos(2x)$ and $2\cos^2(x) = 1+\cos(2x)$ implies that there will be an ε-independent contribution to ϕ_2'. Thus for smooth nonlinearities we find that

$$\frac{\mathrm{d}\phi}{\mathrm{d}\tau} = c + \varepsilon^2 C_2(c) + D_2\left(c,\frac{\phi}{\varepsilon}\right), \tag{4.33}$$

where C_2 is independent of ε and D_2 is exponentially small in ε. Equation (4.33) is the second-order version of the phase equation (3.15) in cases where the first-order term is exponentially small. Again, the condition for wave propagation failure is that the right-hand side of (4.33) vanishes for some ϕ.

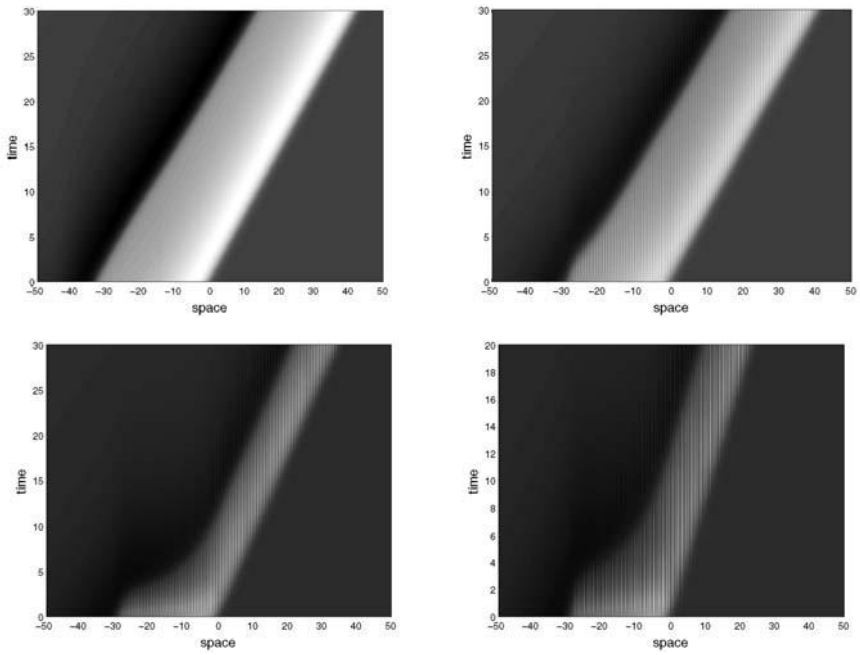


Figure 5. *Top left: Stable traveling pulse for a homogeneous network with exponential weight function (2.5) and fixed parameters $\kappa = 0.2$, $\beta = 2.0$, and $\alpha = 0.04$ (for all plots). Top right: Corresponding traveling pulse for an inhomogeneous network with weight distribution (2.4) and a sinusoidal modulation with $\varepsilon = 0.1$ and $\rho = 0.3$. We see rippling in the interior of the pulse. Bottom left: Using a more severe inhomogeneity, $\rho = 0.8$, leads to rippling in the active region of the pulse such that now the interior crosses below threshold periodically. Bottom right: For $\rho = 1$, the effect is even more severe.*

5. Numerical results. Our perturbation analysis suggests that as ρ increases, the mean speed of a traveling pulse decreases, and, at least for sufficiently large periods ε of the weight modulation, wave propagation failure can occur. However, one of the simplifying assumptions of our analysis is that the perturbed solution is still a traveling pulse; that is, at each time t there is a single bounded interval over which the solution is above threshold, which is equal to the pulse width a of the homogeneous pulse in the limit $\varepsilon \to 0$. The inclusion of a periodic modulation of a monotonically decreasing weight function suggests that the assumption of a single pulse solution may break down as ρ increases toward unity. In this section we explore this issue by numerically solving the full system of equations (2.1) in the case of a Heaviside nonlinearity ($f = H$), and we show that wave propagation can persist in the presence of multiple bumps. Numerical simulations of propagating pulses are carried out using MATLAB. Initial conditions are taken to be solutions to the homogeneous problem given by (4.7) and (4.8). We then apply backward Euler to the linear terms and forward Euler with a Riemann sum to the convolution operator. Space and time discretizations are taken to be $\Delta t = 0.01$ and $\Delta x = 0.01$. The numerical results are stable with respect to reductions in the mesh size provided that $\Delta x \ll 2\pi\varepsilon$. Finally, boundary points evolve freely, rather than by prescription, and the domain size is wide enough so that pulses are unaffected by boundaries.

In Figure 5 we show some examples of traveling pulse solutions in an inhomogeneous

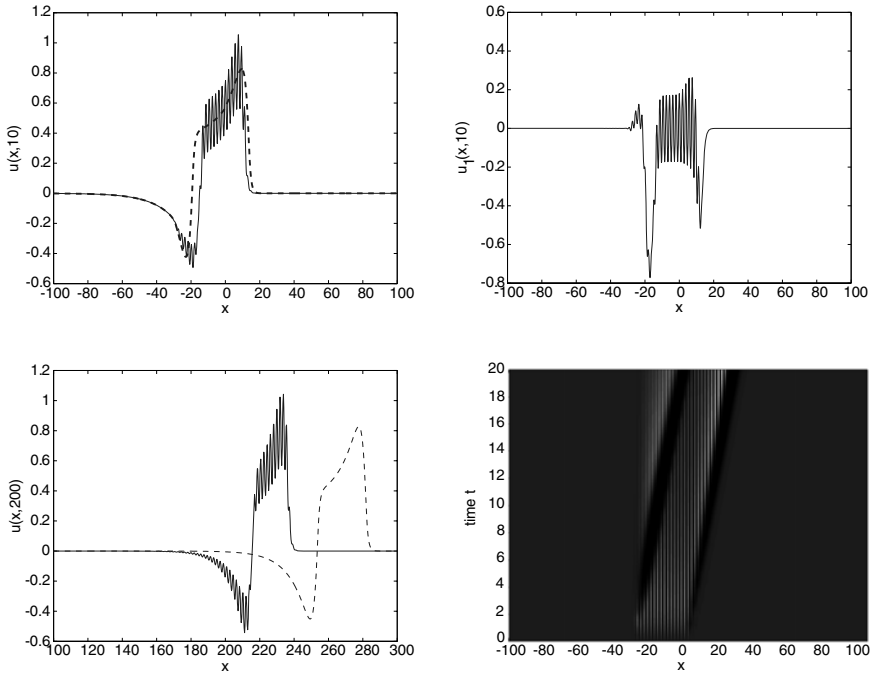


Figure 6. *Top left: Comparison of the wave profiles at time $t = 10$ for the homogeneous (dashed line) and the inhomogeneous (solid line) cases. Here, the parameters are $\kappa = 0.2$, $\beta = 2.0$, $\alpha = 0.04$, $\varepsilon = 0.3$, and $\rho = 0.3$. Including periodic modulation clearly thins the pulse as we see its profile fits within that of the homogeneous medium. Top right: Subtraction of the homogeneous solution from the inhomogeneous solution at time $t = 10$. We see here an approximation of $u_1(x,t)$, from our perturbation analysis. The dominant detail is the oscillations with period $2\pi\varepsilon$. Bottom left: Profile comparison at $t = 200$. The homogeneous solution has moved well ahead of the inhomogeneous solution due to speed difference. Bottom right: Pseudocolor plot of $u_1(x,t)$, obtained by subtracting the homogeneous solution from the inhomogeneous solution. The dark bands delineate the underlying homogeneous solution.*

network with weight distribution given by (2.4), (2.5), and (2.6). The period of the modulation is taken to be relatively small ($\varepsilon = 0.1$). We take as initial conditions the invariant profile for the corresponding homogeneous case, obtained by solving in traveling wave coordinates for the $\varepsilon = 0$ case, which gives (U, V) in (4.7) and (4.8). It can be seen from Figure 5 that as the amplitude ρ of the periodic modulation increases the wave slows down and narrows, which is consistent with our perturbation analysis. Moreover, the network activity develops a rippling within the interior of the pulse, as can be seen more clearly in Figure 6, where we directly compare the numerical solution of the homogeneous network with that of a corresponding inhomogeneous network. Superimposing the two wave profiles at an early time ($t = 10$) illustrates the thinning of the pulse and shows that the difference between the two wave profiles is an oscillatory component of approximately zero mean, which would correspond to u_1 in our perturbation analysis. Similarly, comparing the two wave profiles at a later time ($t = 200$) illustrates the slowing down of the pulse. As ρ increases, the amplitude of the ripples also increases such that, for sufficiently large ρ, activity at any given time t alternates between superthreshold and subthreshold domains. This is illustrated in Figure 7. A closer

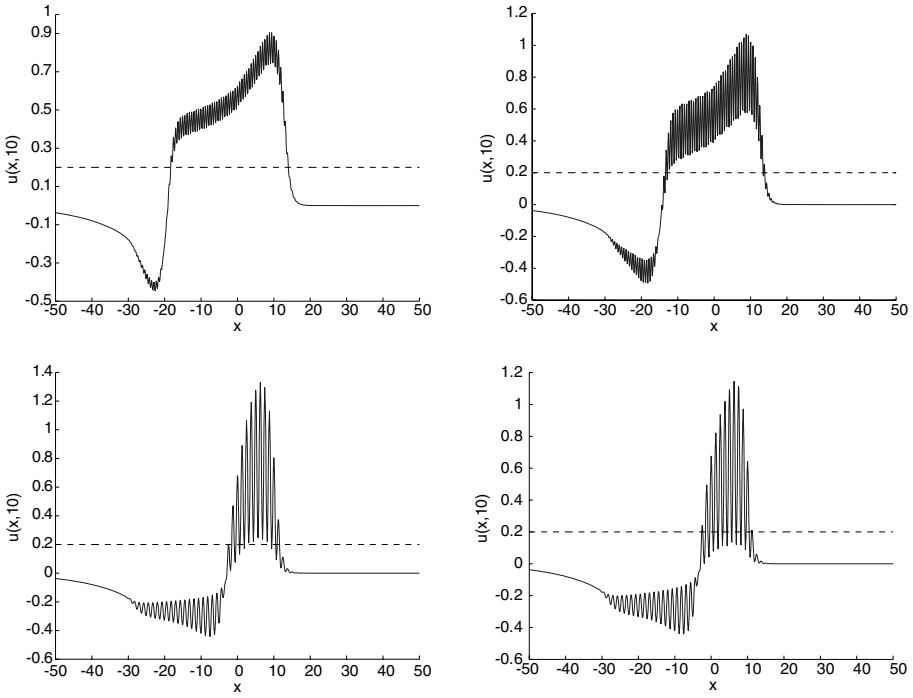


Figure 7. *A collection of traveling wave profiles taken at time $t = 10$ for various amplitudes ρ and $\varepsilon = 0.1$. Other parameters are as in Figure 5. Top left: $\rho = 0.1$. Notice that rippling of the activity does not dip below threshold within the pulse interior. Top right: $\rho = 0.3$. Rippling crosses below threshold at the edges of the pulse creating a couple of bumps. Bottom left: $\rho = 0.7$. Rippling now generates a multiple bump solution. Bottom right: $\rho = 0.8$.*

look at the time evolution of the wave profile when the rippling is above threshold within the interior of the pulse shows that individual ripples are nonpropagating and transient, with new ripples appearing at the leading edge of the wave and subsequently disappearing at the trailing edge; see Figure 8. Interestingly, such behavior persists for large ρ when the ripples cross below threshold within the interior of the pulse; see Figure 9. Now the pulse actually consists of multiple bumps, each of which is nonpropagating but only exists for a finite length of time. The sequence of events associated with the emergence and disappearance of these bumps generates a wave envelope that behaves very much like a single coherent traveling pulse. Hence, for sufficiently short wavelength oscillatory modulations of the weight distribution, the transient multiple bump solution can be homogenized and treated as a single traveling pulse. However, the wave speed of the multiple bump solution differs from that predicted using perturbation theory. This is shown in Figure 10, where we compare the $\bar{c}$ versus ε curves obtained using perturbation theory with data obtained by directly simulating the full system (2.1). In the case of small ρ, a stable (single bump) traveling pulse persists for all ε, $0 \leq \varepsilon < 1$, and $\bar{c}$ is a monotonically decreasing function of ε. Moreover, the numerically calculated value of the average wave speed agrees quite well with the first-order perturbation analysis. On the other hand, for large values of ρ, such agreement no longer holds, and we find that the traveling pulse destabilizes at a critical value of ε that is well below the value predicted from

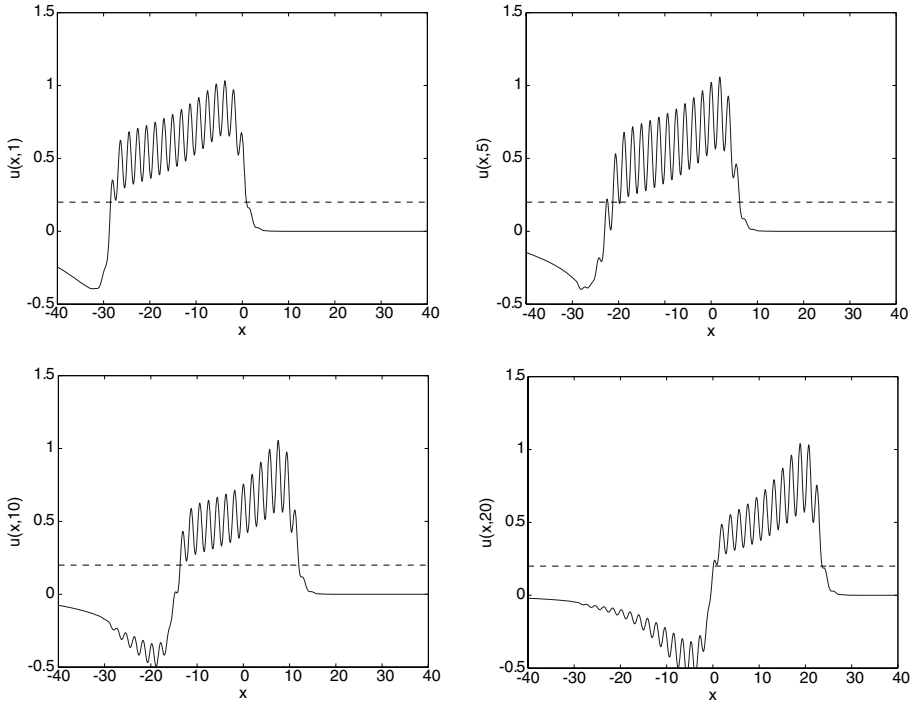


Figure 8. *A series of snapshots in time of a traveling pulse for* $\kappa = 0.2$, $\beta = 2.0$, $\alpha = 0.04$, $\rho = 0.3$, $\varepsilon = 0.3$. *The interior of the pulse consists of nonpropagating, transient ripples. The disappearance of ripples at one end and the emergence of new ripples at the other end generate the propagation of activity. Notice that the solitary wave profile is not invariant, reflecting the underlying inhomogeneity. Top left:* $t = 1$. *Top right:* $t = 5$. *Bottom left:* $t = 10$. *Bottom right:* $t = 20$. *Clicking on the above images displays the accompanying movie* (69921_01.avi [1.7MB]).

the perturbation analysis.

In Figure 11 we compare the behavior of traveling pulses for short wavelength ($\varepsilon = 0.2$) and long wavelength ($\varepsilon = 0.9$) periodic modulation. The amplitude is taken to be relatively large, $\rho = 0.8$, so that multiple bump solutions occur. We see that for long wavelength modulation, the initial pulse transitions into a nonpropagating multiple bump solution, with successive bumps disappearing sequentially and no additional bumps being created; the failure to generate new bumps means that activity cannot propagate. We can see this more clearly when examining a series of snapshots of the pulse/bump profiles in Figure 12. In conclusion, one way to understand wave propagation failure for large ρ is to note that a large amplitude periodic weight modulation can generate a pinned multiple bump solution. However, in the absence of any inhibition, such a multiple bump solution is unstable [26, 37]. In the case of small ε, destabilization of the bumps generates new bumps at the leading edge of the bump such that activity can propagate in a coherent fashion. Increasing ε prevents the creation of new bumps and propagation failure occurs.

The effect of the periodic weight modulation on a different type of solution is illustrated in Figure 13, where, motivated by a prior numerical study of multiple bumps [25], the initial condition of the network is taken to consist of three bumps,

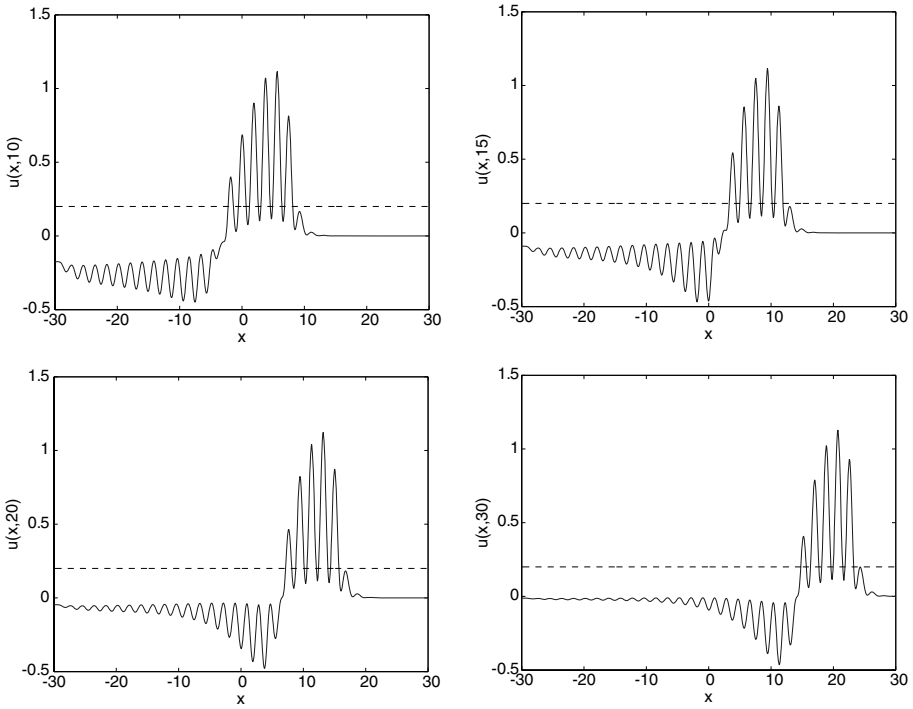


Figure 9. *A series of snapshots in time of the "pulse" profile for* $\kappa = 0.2$, $\beta = 2.0$, $\alpha = 0.04$, $\rho = 0.8$, $\varepsilon = 0.3$. *The solitary pulse corresponds to the envelope of a multiple bump solution, in which individual bumps are nonpropagating and transient. The disappearance of bumps at one end and the emergence of new bumps at the other end generate the propagation of activity. Notice that the solitary wave profile is not invariant, reflecting the underlying inhomogeneity. Top left:* $t = 10$. *Top right:* $t = 15$. *Bottom left:* $t = 20$. *Bottom right:* $t = 30$. *Clicking on the above images displays the accompanying movie* (69921_02.avi [1.9MB]).

$$u(x,0) = \sum_{j=-1}^{1} \cos\left(\frac{x}{\varepsilon}\right) \exp\left(-\left(\frac{0.1(x - j \cdot 20)}{\varepsilon}\right)^2\right). \tag{5.1}$$

Each initial bump generates a pair of left and right moving fronts. In the homogeneous case, we see that collision of left and right moving waves results in a bidirectional front. That is, the region within the interior of the boundary formed by the two outermost fronts becomes superthreshold. In the inhomogeneous case, the collision of the waves is insufficient to maintain activity across this region, and one finds a pair of counterpropagating pulses.

6. Discussion. In this paper we analyzed wave propagation in an excitatory neural network treated as a periodic excitable medium. The periodicity was introduced as an inhomogeneous periodic modulation in the long-range synaptic connections and was motivated by the existence of patchy horizontal connections in the cerebral cortex. We showed that for small amplitude, short wavelength periodic modulation the main effect of the inhomogeneity is to slow down a traveling pulse, and the mean speed of the pulse can be estimated quite well using perturbation theory. In the case of large amplitude modulation, a stable traveling pulse still exists for sufficiently small ε, but now the pulse is the envelope of a multiple bump

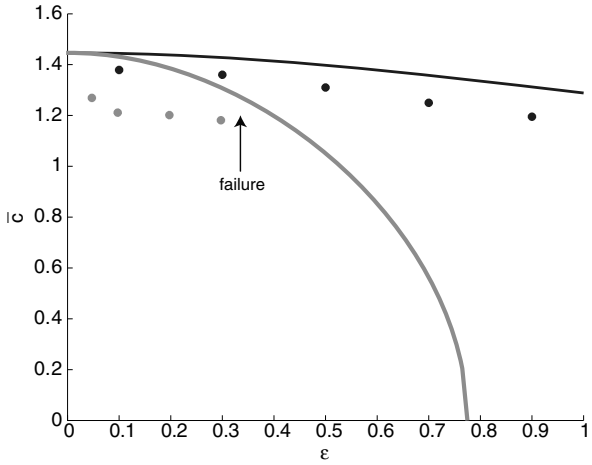


Figure 10. *Comparison of perturbation theory with direct numerical simulations. Continuous curves show average wave speed $\bar{c}$ as a function of ε obtained using perturbation theory. Data points are the corresponding wave speeds determined from numerically solving (2.1). In the case of low amplitude modulation ($\rho = 0.3$, dark curve) a stable traveling pulse persists for all ε, $\varepsilon < 1$, whereas for large amplitude modulation ($\rho = 0.8$, light curve), wave propagation failure occurs as ε increases.*

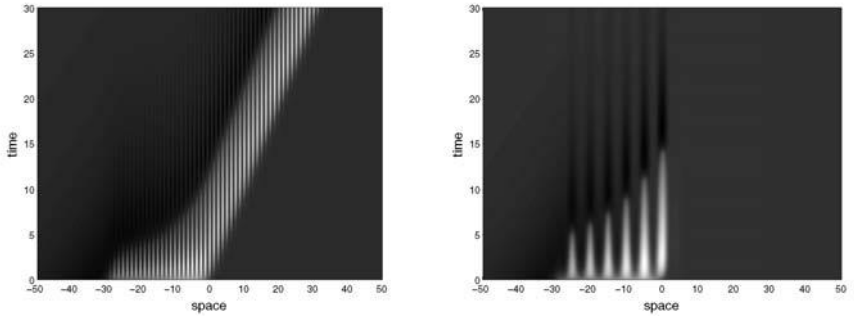


Figure 11. *Comparison of traveling pulses in the case of short and long wavelength periodic modulation with $\rho = 0.8$ and all other parameters as in Figure 5. Left: For short wavelength modulation ($\varepsilon = 0.2$) the traveling pulse shrinks and slows but does not annihilate. Right: For long wavelength modulation ($\varepsilon = 0.9$) wave propagation failure occurs. The initial pulse transitions into a collection of multiple equal width stationary bumps which are unstable.*

solution in which individual bumps are unstable and transient. Wave propagation arises via the appearance (disappearance) of bumps at the leading (trailing) edge of the pulse. As ε increases, wave propagation failure occurs due to the fact that there is insufficient activity to generate new bumps.

Although the existence of multiple bump traveling "pulses" is interesting from a dynamical systems perspective, it is less clear whether such solutions can be observed in real neural tissue. One of the biological limitations of the integrodifferential equations used in this and other studies is that, although these equations support traveling waves that have speeds consistent with neurophysiology, the pulses tend to be too wide. That is, taking the range of synaptic connections to be 1 mm, the width of a stable pulse tends to vary between 5–30

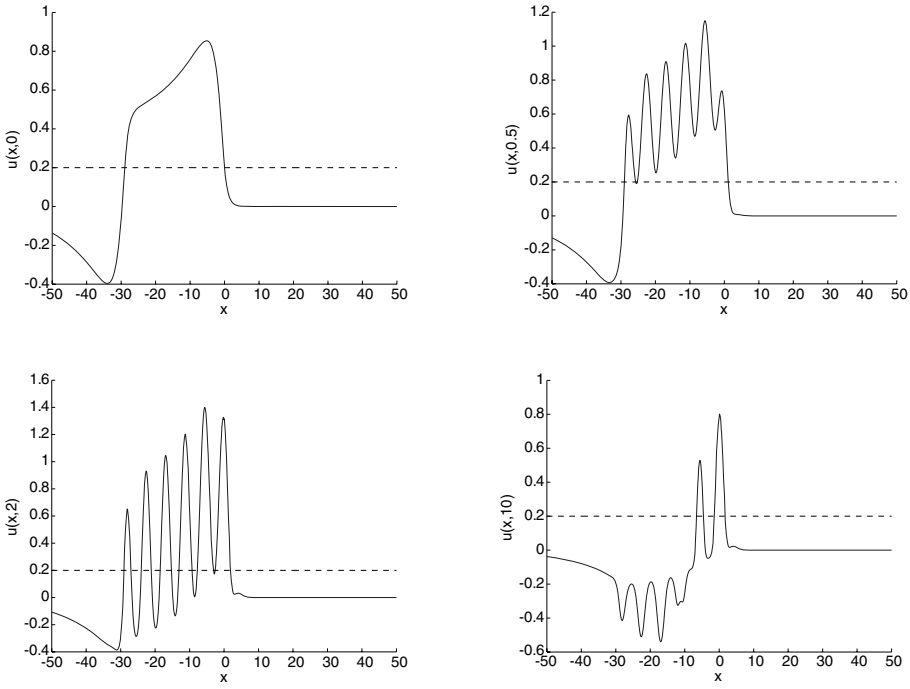


Figure 12. *A series of snapshots in time of the "pulse" profile for the inhomogeneous network with* $\kappa = 0.2$, $\beta = 2.0$, $\alpha = 0.04$, $\rho = 0.8$, $\varepsilon = 0.9$. *Top left: The initial wave profile, which is taken to be the invariant wave profile* U *of the homogeneous network. Top right: Shortly after the simulation begins* ($t = 0.5$), *the interior of the pulse develops ripples such that the active region contains a subregion for which activity is subthreshold. Bottom left: At time* $t = 2$ *a multiple bump profile has emerged. We can really see here how a multiple bump solution, as defined by multiple neighboring standing profiles, emerges from the pulse profile of* $t = 0$. *Bottom right: Collapse of the pulse interior occurs due to the disappearance of the unstable bumps. Since no new bumps emerge, there is no propagating activity. Clicking on the above images displays the accompanying movie* (69921_03.avi [1.1MB]).

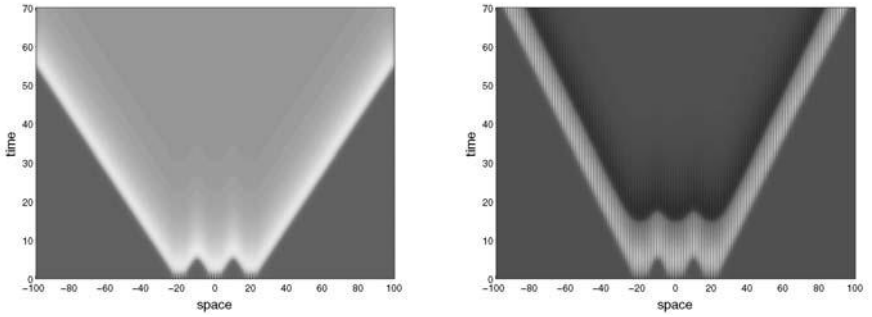


Figure 13. *Left: In the case of a homogeneous network, a three bump initial condition evolves into a bidirectional front following the collision of left and right traveling waves. The parameters are* $\kappa = 0.2$, $\beta = 2.0$, $\alpha = 0.04$. *Right: In the corresponding inhomogeneous network with* $\varepsilon = 0.2$ *and* $\rho = 0.8$, *the collision of left and right traveling waves results in a pair of counterpropagating pulses. Here the modulated synaptic interactions are insufficient to maintain activity in the region between the two pulses.*

mm (see Figure 2), whereas waves in slices tend to be only 1 mm wide [31]. More realistic widths and wave speeds could be generated by taking the effective range of synaptic connections to be a few hundred μ m, that is, by assuming that the predominant contribution to synaptic excitability is via local circuitry rather than via long-range patchy horizontal connections. However, inhomogeneities occurring at smaller spatial scales are unlikely to exhibit any periodic structure.

Irrespective of these particular issues, our analysis raises a more general point that would be interesting to pursue experimentally; namely, is it possible to detect the effects of network inhomogeneities by measuring the properties of traveling waves? Signatures of such inhomogeneities would include time-dependent rippling of the wave profile and variations in wave speed. However, such features may not be detectable given the current resolution of microelectrode recordings.

Appendix. In this appendix we present the explicit parameter-dependent expressions for the various coefficients appearing in the solution of the phase function Φ_1, (4.28). First, the constants premultiplying the periodic functions on the right-hand side of (4.28) are as follows:

$$\Xi_\pm = \frac{\gamma_\pm}{1+\mu_\pm^2\varepsilon^2}\left[\frac{1}{2(1+\mu_\pm)}+\frac{\chi}{2}\left(\frac{\mathrm{e}^{-a}-\mathrm{e}^{-\mu_\pm a}}{\mu_\pm-1}+\frac{\mathrm{e}^{-\mu_\pm a}}{\mu_\pm+1}\right)\right],$$

$$\Pi_\pm = \frac{\gamma_\pm}{1+\mu_\pm^2\varepsilon^2}\left[-\frac{\chi}{2(1+\mu_\pm)}-\frac{\mathrm{e}^{-a}}{2(\mu_\pm+1)}\right],$$

$$\Upsilon_\pm = \frac{\gamma_\pm}{1+\mu_\pm^2\varepsilon^2}\left[-\frac{\mu_\pm\varepsilon}{2(1+\mu_\pm)}-\frac{\chi\mu_\pm\varepsilon}{2}\left(\frac{\mathrm{e}^{-a}-\mathrm{e}^{-\mu_\pm a}}{\mu_\pm-1}+\frac{\mathrm{e}^{-\mu_\pm a}}{\mu_\pm+1}\right)\right],$$

$$\Psi_\pm = \frac{\gamma_\pm}{1+\mu_\pm^2\varepsilon^2}\left[\frac{\chi\mu_\pm\varepsilon}{2(1+\mu_\pm)}+\frac{\mu_\pm\varepsilon\mathrm{e}^{-a}}{2(1+\mu_\pm)}\right].$$

Second, the constant scaling factor K on the left-hand side of (4.28) is determined by substituting (4.7), (4.8), and (4.14) into (3.14). Using the fact that the null-vector is zero for $\xi<-a$, we can expand the integral in terms of definite integrals of exponential products with the $\mathcal{M}_\pm(\xi)$ functions

$$\begin{aligned}
K = {} & [\gamma_+(1-m_-)(1+\chi e^{-\mu_+ a})(1-\alpha^{-1}\beta^{-1}(1-m_+)^2)]\int_0^\infty e^{-\mu_+\xi}\mathcal{M}'_+(\xi)\mathrm{d}\xi \\
& + [\gamma_-(1-m_-)(1+\chi e^{-\mu_- a})(1+\alpha^{-1}\beta^{-1}(m_+-1)(1-m_-))]\int_0^\infty e^{-\mu_-\xi}\mathcal{M}'_+(\xi)\mathrm{d}\xi \\
& - [\gamma_+(1-m_+)(1+\chi e^{-\mu_+ a})(1+\alpha^{-1}\beta^{-1}(m_--1)(1-m_+))]\int_0^\infty e^{-\mu_+\xi}\mathcal{M}'_-(\xi)\mathrm{d}\xi \\
& - [\gamma_-(1-m_+)(1+\chi e^{-\mu_- a})(1-\alpha^{-1}\beta^{-1}(1-m_-)^2)]\int_0^\infty e^{-\mu_-\xi}\mathcal{M}'_-(\xi)\mathrm{d}\xi \\
& + \chi[\gamma_+ e^{-\mu_+ a}(1-m_-)(1-\alpha^{-1}\beta^{-1}(1-m_+)^2)]\int_{-a}^0 e^{-\mu_+\xi}\mathcal{M}'_+(\xi)\mathrm{d}\xi \\
& + \chi[\gamma_- e^{-\mu_- a}(1-m_-)(1+\alpha^{-1}\beta^{-1}(m_+-1)(1-m_-))]\int_{-a}^0 e^{-\mu_-\xi}\mathcal{M}'_+(\xi)\mathrm{d}\xi
\end{aligned}$$

$$- \chi[\gamma_+ e^{-\mu_+ a}(1-m_+)(1+\alpha^{-1}\beta^{-1}(m_- - 1)(1-m_+))] \int_{-a}^{0} e^{-\mu_+ \xi}\mathcal{M}'_-(\xi)\mathrm{d}\xi$$
$$- \chi[\gamma_- e^{-\mu_- a}(1-m_+)(1-\alpha^{-1}\beta^{-1}(1-m_-)^2)] \int_{-a}^{0} e^{-\mu_- \xi}\mathcal{M}'_-(\xi)\mathrm{d}\xi.$$

The individual integrals can be computed as follows:

$$\int_0^\infty e^{-\mu_\pm \xi}\mathcal{M}'_\pm(\xi)\mathrm{d}\xi = \frac{e^{-a}-1}{2c(m_+-m_-)(\mu_\pm+1)^2},$$
$$\int_0^\infty e^{-\mu_+ \xi}\mathcal{M}'_-(\xi)\mathrm{d}\xi = \frac{e^{-a}-1}{2c(m_+-m_-)(\mu_-+1)(\mu_++1)},$$
$$\int_0^\infty e^{-\mu_- \xi}\mathcal{M}'_+(\xi)\mathrm{d}\xi = \frac{e^{-a}-1}{2c(m_+-m_-)(\mu_++1)(\mu_-+1)},$$

and

$$\int_{-a}^{0} e^{-\mu_\pm \xi}\mathcal{M}'_\pm(\xi)\mathrm{d}\xi = \frac{1}{2c(m_+-m_-)}\left\{\frac{a}{(\mu_\pm-1)} + \frac{1-e^{(\mu_\pm-1)a}}{(\mu_\pm-1)^2} + \frac{e^{-a}(e^{(\mu_\pm+1)a}-1)}{(\mu_\pm+1)^2} - \frac{a}{2(\mu_\pm+1)}\right\},$$

$$\int_{-a}^{0} e^{-\mu_+ \xi}\mathcal{M}'_-(\xi)\mathrm{d}\xi = \frac{1}{2c(m_+-m_-)}\left\{\frac{1-e^{-(\mu_- - \mu_+)a}}{(\mu_- - \mu_+)(\mu_- - 1)} - \frac{e^{(\mu_+-1)a}-1}{(\mu_+-1)(\mu_--1)} + \frac{e^{\mu_+ a}-e^{-a}}{(\mu_++1)(\mu_-+1)} - \frac{1-e^{-(\mu_- - \mu_+)a}}{(\mu_-+1)(\mu_- - \mu_+)}\right\},$$

$$\int_{-a}^{0} e^{-\mu_- \xi}\mathcal{M}'_+(\xi)\mathrm{d}\xi = \frac{1}{2c(m_+-m_-)}\left\{\frac{1-e^{-(\mu_+ - \mu_-)a}}{(\mu_+-1)(\mu_+ - \mu_-)} - \frac{e^{(\mu_- -1)a}-1}{(\mu_+-1)(\mu_--1)} + e^{-a}\frac{e^{(\mu_-+1)a}-1}{(\mu_++1)(\mu_-+1)} - \frac{1-e^{-(\mu_+ - \mu_-)a}}{(\mu_++1)(\mu_+ - \mu_-)}\right\}.$$

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral inhibition type neural fields*, Biol. Cybernet., 27 (1977), pp. 77–87.

[2] W. H. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick, *Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex*, J. Neurosci., 17 (1997), pp. 2112–2127.

[3] P. C. Bressloff, *Traveling fronts and wave propagation failure in an inhomogeneous neural network*, Phys. D, 155 (2001), pp. 83–100.

[4] P. C. Bressloff, *Spatially periodic modulation of cortical patterns by long-range horizontal connections*, Phys. D, 185 (2003), pp. 131–157.

[5] P. C. Bressloff and S. E. Folias, *Front bifurcations in an excitatory neural network*, SIAM J. Appl. Math., 65 (2004), pp. 131–151.
[6] P. C. Bressloff, *Pattern formation in visual cortex*, in Methods and Models in Neurophysics, C. C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, eds., Les Houches Lectures in Neurophysics, Springer-Verlag, Berlin, 2004, pp. 477–574.
[7] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Periodicity and directionality in the propagation of excitation in neural network model*, J. Neurophysiol., 60 (1988), pp. 1695–1713.
[8] B. W. Connors and Y. Amitai, *Generation of epileptiform discharge by local circuits of neocortex*, in Epilepsy: Models, Mechanisms and Concepts, P. A. Schwartkroin, ed., Cambridge University Press, Cambridge, UK, 1993, pp. 388–423.
[9] S. Coombes and M. R. Owen, *Bumps, breathers, and waves in a neural network with spike frequency adaptation*, Phys. Rev. Lett., 94 (2005), 148102.
[10] S. Coombes and M. R. Owen, *Evans functions for integral neural field equations with Heaviside firing rate function*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 574–600.
[11] S. Coombes, *Waves, bumps, and patterns in neural field theories*, Biol. Cybernet., 93 (2005), pp. 91–108.
[12] G. B. Ermentrout and J. B. McLeod, *Existence and uniqueness of travelling waves for a neural network*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 461–478.
[13] G. B. Ermentrout and D. Kleinfeld, *Traveling electrical waves in cortex: Insights from phase dynamics and speculation on a computational role*, Neuron, 29 (2001), pp. 33–44.
[14] S. E. Folias and P. C. Bressloff, *Stimulus-locked traveling waves and breathers in an excitatory neural network*, SIAM J. Appl. Math., 65 (2005), pp. 2067–2092.
[15] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiol., 78 (1997), pp. 1199–1211.
[16] J. Guckenheimer and P. J. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[17] B. S. Gutkin, G. B. Ermentrout, and J. O'Sullivan, *Layer 3 patchy recurrent excitatory connections may determine the spatial organization of sustained activity in the primate prefrontal cortex*, Neurocomputing, 32–33 (2000), pp. 391–400.
[18] X. Huang, W. C. Troy, Q. Yang, H. Ma, C. Laing, S. Schiff, and J. Y. Wu, *Spiral waves in disinhibited mammalian cortex*, J. Neurosci., 24 (2004), pp. 9897–9902.
[19] M. A. P. Idiart and L. F. Abbott, *Propagation of excitation in neural network models*, Network, 4 (1993), pp. 285–294.
[20] J. P. Keener, *Propagation of waves in an excitable medium with discrete release sites*, SIAM J. Appl. Math., 61 (2000), pp. 317–334.
[21] J. P. Keener, *Homogenization and propagation in the bistable equation*, Phys. D, 136 (2000), pp. 1–17.
[22] D. Kleinfeld, K. R. Delaney, M. S. Fee, J. A. Flores, D. W. Tank, and A. Galperin, *Dynamics of propagating waves in the olfactory network of a terrestrial mollusk: An electrical and optical study*, J. Neurophysiol., 72 (1994), pp. 1402–1419.
[23] Y. W. Lam, L. B. Cohen, M. Wachowiak, and M. R. Zochowski, *Odors elicit three different oscillations in the turtle olfactory bulb*, J. Neurosci., 20 (2000), pp. 749–762.
[24] J. W. Lance, *Current concepts of migraine pathogenesis*, Neurology, 43 (1993), pp. 11–15.
[25] C. R. Laing, W. C. Troy, B. Gutkin, and G. B. Ermentrout, *Multiple bumps in a neuronal model of working memory*, SIAM J. Appl. Math., 63 (2002), pp. 62–97.
[26] C. R. Laing and W. C. Troy, *Two-bump solutions of Amari-type models of neuronal pattern formation*, Phys. D, 178 (2003), pp. 190–218.
[27] J. B. Levitt, D. A. Lewis, T. Yoshioka, and J. S. Lund, *Topography of pyramidal neuron intrinsic connections in macaque prefrontal cortex*, J. Comp. Neurol., 338 (1993), pp. 360–376.
[28] R. Malach, Y. Amir, M. Harel, and A. Grinvald, *Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex*, Proc. Natl. Acad. Sci. USA, 90 (1993), pp. 10469–10473.
[29] D. S. Melchitzky, S. R. Sesack, M. L. Pucak, and D. A. Lewis, *Synaptic targets of pyramidal neurons providing intrinsic horizontal connections in monkey prefrontal cortex*, J. Comp. Neurol., 390 (1998), pp. 211–224.

[30] M. A. L. Nicolelis, L. A. Baccala, R. C. S. Lin, and J. K. Chapin, *Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system*, Science, 268 (1995), pp. 1353–1358.

[31] D. J. Pinto and G. B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* I. *Traveling fronts and pulses*, SIAM J. Appl. Math., 62 (2001), pp. 206–225.

[32] D. J. Pinto, S. L. Patrick, W. C. Huang, and B. W. Connors, *Initiation, propagation, and termination of epileptiform activity in rodent neocortex in vitro involve distinct mechanisms*, J. Neurosci., 25 (2005), pp. 8131–8140.

[33] D. J. Pinto, R. K. Jackson, and C. E. Wayne, *Existence and stability of traveling pulses in a continuous neuronal network*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 954–984.

[34] J. C. Prechtl, L. B. Cohen, B. Pasaram, P. P. Mitra, and D. Kleinfeld, *Visual stimuli induce waves of electrical activity in turtle cortex*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 7621–7626.

[35] K. A. Richardson, S. J. Schiff, and B. J. Gluckman, *Control of traveling waves in the mammalian cortex*, Phys. Rev. Lett., 94 (2005), 028103.

[36] P. R. Roelfsema, A. K. Engel, P. Konig, and W. Singer, *Visuomotor integration is associated with zero time-lag synchronization among cortical areas*, Nature, 385 (1997), pp. 157–161.

[37] J. E. Rubin and W. C. Troy, *Sustained spatial patterns of activity in neuronal populations without recurrent excitation*, SIAM J. Appl. Math., 64 (2004), pp. 1609–1635.

[38] W. C. Troy and V. Shusterman, *Patterns and features of families of traveling waves in large-scale neuronal networks*, SIAM J. Appl. Dyn. Syst., 6 (2007), pp. 263–292.

[39] H. R. Wilson and J. D. Cowan, *A mathematical theory of functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.

[40] J.-Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neurosci., 19 (1999), pp. 5005–5015.

[41] T. Yoshioka, G. G. Blasdell, J. B. Levitt, and J. S. Lund, *Relation between patterns of intrinsic lateral connectivity, ocular dominance, and cytochrome oxidase-reactive regions in macaque monkey striate cortex*, Cerebral Cortex, 6 (1996), pp. 297–310.

[42] L. Zhang, *On stability of traveling wave solutions in synaptically coupled neuronal networks*, Differential Integral Equations, 16 (2003), pp. 513–536.

[43] L. Zhang, *Existence, uniqueness, and exponential stability of traveling wave solutions of some integral differential equations arising from neuronal networks*, J. Differential Equations, 197 (2004), pp. 162–196.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 186–206
© 2008 Society for Industrial and Applied Mathematics

Localized Pattern Formation with a Large-Scale Mode: Slanted Snaking*

J. H. P. Dawes†

Abstract. Steady states of localized activity appear naturally in uniformly driven, dissipative systems as a result of subcritical instabilities. In the usual setting of an infinite domain, branches of such localized states bifurcate at the subcritical "pattern-forming" instability and intertwine in a manner often referred to as "homoclinic snaking." In this paper we consider an extension of this paradigm where, in addition to the pattern-forming instability (with nonzero wavenumber), a large-scale neutral mode exists, having zero growth rate at zero wavenumber. Such a situation naturally arises in the presence of a conservation law; we give examples of physical systems in which this arises, in particular, thermal convection in a horizontal fluid layer with a vertical magnetic field. We introduce a novel scaling that allows the derivation of a nonlocal Ginzburg–Landau equation to describe the formation of localized states. Our results show that the existence of the large-scale mode substantially enlarges the region of parameter space where localized states exist and are stable.

Key words. homoclinic snaking, pattern formation, bifurcation

AMS subject classifications. 76E25, 34E13, 35B32

DOI. 10.1137/06067794X

1. Introduction.

Recent decades have seen a sustained level of interest in systems whose response is spatially localized despite a spatially uniform applied forcing. One broad class of systems which display such localized states of activity is both driven and strongly dissipative, and displays "pattern-forming" (also called Turing) instabilities at which spatially homogeneous solutions become unstable and spatial structure appears [11, 19, 25]. It is widely appreciated that, while supercritical pattern-forming instabilities lead to spatially extended (and almost periodic) structures, subcritical instabilities robustly lead to localized patterns [21, 37, 8, 30, 5]. Localized steady-state pattern formation has been observed in a huge variety of experiments and models for physical, chemical, and biological systems, for example, neural dynamics [23], elastic buckling [20], and nonlinear optics [1, 35]. In other cases, for example, vertically vibrated granular media [34], the localized pattern is oscillatory in nature. In many problems the existence of the localized states can be heuristically explained by an energetic argument: at a critical parameter value, often called the "Maxwell point," the system has no energetic preference between the "ground state," corresponding to no pattern, and the patterned state. Once a localized state has been formed there is a locking between the phase of the pattern and the phase of the envelope which allows the localized state to persist over a finite range of parameter values, as first remarked on by Pomeau [26].

*Received by the editors December 18, 2006; accepted for publication (in revised form) by M. Silber November 2, 2007; published electronically February 8, 2008. This research was supported by Newnham College, Cambridge, and by the Royal Society.
http://www.siam.org/journals/siads/7-1/67794.html
†DAMTP, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WA, UK (j.h.p.dawes@damtp.cam.ac.uk).

Mathematically, much has been done to investigate canonical model problems displaying subcritical pattern-forming instabilities and therefore localized patterns. One such model equation is the one-dimensional Swift–Hohenberg equation

$$w_t = [r - (1 + \partial_{xx}^2)^2]w + N(w) \tag{1.1}$$

(writing ∂_{xx}^2 as a convenient abbreviation for $\partial^2/\partial x^2$), which has been studied in detail in the particular cases $N(w) = sw^2 - w^3$ and $N(w) = sw^3 - w^5$, where $s > 0$ controls the subcritical nature of the pattern-forming instability. These choices for $N(w)$ are commonly referred to as the "quadratic-cubic" and "cubic-quintic" cases, respectively. Careful investigations of these localized states have been carried out by several authors including Sakaguchi and Brand [30] and Burke and Knobloch [5]. Many different localized and front-like solutions between steady states have been reported. For our present purposes, we note that the most obvious consequence of a subcritical bifurcation is the existence of stable localized states in a small subregion between the linear instability of the state $w(x, t) \equiv 0$ at $r = 0$ and the saddle-node bifurcation at $r = r_{sn} < 0$ on the primary branch of uniform amplitude pattern. The width of this region is exponentially small in the small amplitude parameter ε employed in the standard Ginzburg–Landau multiple-scales expansion near $r = 0$ [22]. Moreover, since the usual "spatial dynamics" analysis assumes an infinitely wide domain $-\infty < x < \infty$, localized states containing arbitrary numbers of pattern bumps are simultaneously stable over almost all of this region.

Concentrating on the cubic-quintic case, a final widely appreciated, and well-understood, point is that two distinct pairs of branches of localized states persist over the locking region. One pair corresponds to locking at relative phases of $\phi = 0$ and $\phi = \pi$ (these are related by the $w \to -w$ symmetry present in the cubic-quintic case) and the other pair to the relative phases $\phi = \pi/2$ and $\phi = -\pi/2$, similarly related by symmetry. The first pair corresponds to localized states that are even about the midpoint, and the second pair corresponds to states that are odd about the midpoint. This robust phase-locking is exactly the "locking" intuitively understood by Pomeau.

Intriguingly, the spatial dynamics analysis shows that these branches of localized states, that generically bifurcate from $r = 0$, exist in $r < 0$ down to saddle-node bifurcations slightly below the Maxwell point. They then undergo a sequence of repeated and intertwined saddle-node bifurcations on alternate sides of the Maxwell point; after each pair of saddle-node bifurcations the localized state gains an extra pair of bumps. In a finite domain the branches terminate in bifurcations from the uniform amplitude pattern located near $r = 0$ and near the saddle-node bifurcation at $r = r_{sn}$. In an infinite domain the process of saddle-node bifurcations and gaining extra bumps continues ad infinitum. This sequence of repeated saddle-node bifurcations in an infinite domain is known as "homoclinic snaking" since all the states approach $w = 0$ as $x \to \pm\infty$. A brief explanation of the "homoclinic snaking" phenomenon is that at some parameter value $r = r_{mx}$, $r_{sn} < r_{mx} < 0$, there exists a "Maxwell point" where there exists a stationary front between the trivial solution $w = 0$ and the uniform spatially periodic pattern. In the corresponding spatial dynamical system this heteroclinic connection becomes a heteroclinic tangle when normal form symmetry is broken. Within the heteroclinic tangle we can identify intersections of the stable and unstable manifolds of $w = 0$;

these correspond to spatial homoclinic orbits that comprise the "homoclinic snaking" and they can be shown to persist near $r = r_{mx}$, rather than existing only exactly at $r = r_{mx}$.

In a large but finite domain (with periodic boundary conditions), it appears that localized states persist and continue to exhibit intertwined wiggles around the Maxwell point [18]. In a finite domain they are not truly "localized" since the domain is bounded, but they clearly converge to the infinite domain solutions as the domain size $L \to \infty$. A detailed discussion of persistence and the evolution of the bifurcation structure with domain size will be presented elsewhere.

In this paper we consider a class of pattern-forming systems that differs from (1.1) in a fundamental way; a neutrally stable long-wavelength mode exists in addition to the pattern-forming instability at wavenumber $k = 1$. Our broad motivation comes from three physical situations: thermal convection in the presence of a vertical magnetic field [13, 10], vertically vibrated granular media [17, 34], and thin films [16]. In all three cases the large-scale mode arises due to the existence of a conserved quantity in the dynamics. In magnetoconvection this is the total flux of magnetic field through the fluid layer; in the granular and thin film cases the conserved quantity is the total mass. We remark that such a conserved quantity makes sense only for *finite* experimental domains, and our analysis takes this into account.

The treatment of the large-scale mode in magnetoconvection differs in one respect from the granular media and thin film cases; in the former problem the large-scale mode (the horizontally averaged vertical component of the magnetic field) can take either sign; the dynamics is unchanged by this, so the dynamical equations must also remain unchanged. In the latter two cases the large-scale mode quantity is a scalar, density-like, quantity and no sign-change symmetry exists. We focus in this paper on the symmetric case relevant to magnetoconvection; the density-like case is very similar and we summarize a few calculations in an appendix. As we will show, the presence of a large-scale mode stretches the snaking behavior out over a substantial region of parameter space and enables localized states to exist both below the saddle-node bifurcation on the subcritical uniform amplitude branch at $r = r_{sn}$ and far above the point of linear instability of the trivial state.

Pattern formation in the presence of a long-wavelength neutral mode of this kind was considered by Matthews and Cox [24], who carried out a weakly nonlinear analysis of a Swift–Hohenberg-type equation modified by applying ∂^2_{xx} to the right-hand side of (1.1) to produce a dispersion relation that tended to zero (i.e., indicated neutral stability) as $k \to 0$. These authors considered the pattern amplitude to be $O(\varepsilon)$ and the large-scale mode to deviate by only an $O(\varepsilon^2)$ amount from the initially homogeneous state. While asymptotically correct, these scalings are unable to capture solutions where the fluctuations in the large-scale mode are large. As a result, numerical solutions often blow up, and higher-order stabilizing terms were found to be required [10]. Similar difficulties were noted by Golovin, Davis, and Voorhees [16].

In this paper we present a modified multiple-scales analysis that uses the diffusivity of the large-scale mode as the small parameter. We show that this enables the asymptotics for small-amplitude patterns to nevertheless capture the effect of $O(1)$ fluctuations in the large-scale mode. In consequence, this asymptotic treatment avoids the singularities found by earlier authors. The layout of the paper is as follows. Section 2 proposes the extension of (1.1) which we study; details linking it directly to the governing equations for magnetoconvection are deferred to Appendix A. Appendix B summarizes the multiple-scales derivation in the very

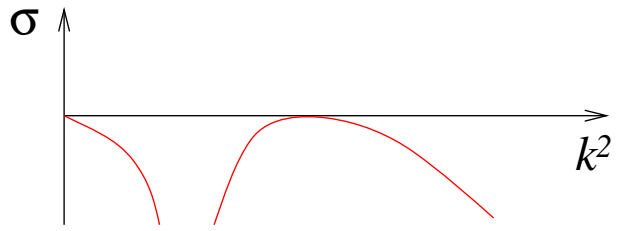


Figure 1. *Sketch of the growth rate $\sigma(k^2)$ for a reflection-symmetric pattern-forming instability at wavenumber $k = 1$, in the presence of a large-scale mode that is neutrally stable as $k \to 0$.*

similar case of a density-like large-scale mode, appropriate to the vibrated granular medium and thin film cases. In section 3 we return to the Swift–Hohenberg ansatz and compare the results of the multiple-scales approach with results from numerical continuation. Section 4 concludes the paper.

2. Ginzburg–Landau asymptotics.

In this section we propose a model equation for pattern formation coupled to a large-scale mode, appropriate for magnetoconvection. We assume that the pattern-forming domain $0 \le x \le L$ is large, but, crucially, finite, and we carry out a multiple-scales analysis to derive an amplitude equation similar to the Ginzburg–Landau equation, but containing a nonlocal term, which captures the influence of the large-scale mode. From the Ginzburg–Landau equation we deduce the existence of modulational instabilities that lead to localized states and investigate scaling laws governing the location of the bifurcation points.

2.1. Model equations.

Suppose that a one-dimensional pattern-forming system is described by the pattern amplitude $w(x,t)$ and the large-scale mode $B(x,t)$. We consider the dispersion curves of the linearized growth rate as a function of perturbation wavenumber k to take the form shown in Figure 1, with quadratic maxima at $k = 0$ and $k = 1$.

We further assume that the system is translationally invariant and reflection-symmetric (i.e., $x \to -x$). Hence a conservation law for $B(x,t)$ contains only even numbers of derivatives. In the absence of the large-scale mode we assume that the pattern-forming instability is supercritical; this is appropriate for thermal convection. An additional symmetry requirement is appropriate for magnetoconvection: that the dynamics is invariant under a change in the sign of the large-scale mode $B(x,t)$. Model equations constrained to have these properties are

$$w_t = [r - (1 + \partial_{xx}^2)^2]w - w^3 - QB^2 w, \tag{2.1}$$

$$B_t = \varepsilon B_{xx} + \frac{c}{\varepsilon}(Bw^2)_{xx}, \tag{2.2}$$

where natural (and analytically tractable) forms of the coupling terms with coefficients Q and c/ε have been taken. The factor of ε^{-1} in the second equation enables an asymptotic balance between the nonlinear and the diffusion terms to occur when the pattern amplitude $w(x,t)$ is $O(\varepsilon)$. As shown in Appendix B, this factor of ε^{-1} appears naturally in magnetoconvection.

Integrating (2.2) over the domain $0 \le x \le L$ and applying periodic boundary conditions imply that $\frac{1}{L}\int_0^L B(x,t)dx \equiv \langle B \rangle$ is constant in time; by rescaling $B(x,t)$ we may take it to be

unity. Essentially, this rescaling absorbs the original mean value $\langle B\rangle$ into the coupling parameter Q. Such a rescaling corresponds exactly to the usual nondimensionalization employed in magnetoconvection, where Q is known as the Chandrasekhar number [7, 27].

We now restrict our attention to steady solutions, setting $\partial_t \equiv 0$, and integrating (2.2) twice to obtain

$$B = \frac{P}{1+cw^2/\varepsilon^2}, \tag{2.3}$$

where P is a constant of integration that corresponds to the value of the large-scale mode away from the localized pattern. Using $\langle B\rangle = 1$ we can now compute P as an integral over the domain; P therefore becomes a nonlocal functional of the pattern amplitude $w(x)$:

$$\frac{1}{P} = \left\langle \frac{1}{1+cw^2/\varepsilon^2} \right\rangle. \tag{2.4}$$

Substituting into (2.1) and looking for steady states, we obtain the nonlocal Swift–Hohenberg equation

$$0 = [r-(1+\partial_{xx}^2)^2]w - w^3 - \frac{QP^2w}{(1+cw^2/\varepsilon^2)^2}. \tag{2.5}$$

The study of bifurcations and solution stability in nonlocal equations is an area of substantial current interest; see, for example, [4, 12] and the references therein. In this paper such difficulties are largely bypassed since, although (2.5) is nonlocal, the linear stability analysis of $w(x) \equiv 0$ remains a local problem. As a result, the usual approaches to small-amplitude solutions of the Swift–Hohenberg equation can be applied, as we now show.

We now introduce the multiple-scales ansatz

$$w(x,t) = \varepsilon A(X)\sin x + \varepsilon^2 w_2 + \varepsilon^3 w_3 + \cdots, \tag{2.6}$$

defining the long lengthscale $X = \varepsilon x$. We rescale the parameters $r = \varepsilon^2\mu$ and $Q = \varepsilon^2 q$ since we are focusing on small-amplitude patterns. The amplitude $A(X)$ can be taken to be real since the instability which generates localized states occurs in the pattern amplitude and not in its phase. At third order in the expansion an amplitude equation for $A(X)$ is obtained by multiplying by $\sin x$ and integrating over the short lengthscale, denoting the average as $\frac{1}{2\pi}\int_0^{2\pi} f(x)dx \equiv \langle f(x)\rangle_x$. We obtain

$$0 = \mu A + 4A_{XX} - 3A^3 - 2qP^2\left\langle \frac{A\sin^2 x}{(1+cA^2\sin^2 x)^2}\right\rangle_x, \tag{2.7}$$

where P now becomes

$$\begin{aligned}\frac{1}{P} &= \frac{1}{\varepsilon L}\int_0^{\varepsilon L}\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{1+cA^2\sin^2 x}\,dx\,dX \\ &\equiv \left\langle \left\langle \frac{1}{1+cA^2\sin^2 x}\right\rangle_x\right\rangle_X = \left\langle \frac{1}{\sqrt{1+cA^2}}\right\rangle_X.\end{aligned} \tag{2.8}$$

Carrying out the x-integral in (2.7), we obtain

$$0 = \mu A + 4A_{XX} - 3A^3 - \frac{qP^2A}{(1+cA^2)^{3/2}}. \tag{2.9}$$

It follows that the trivial state $A(X) \equiv 0$ (for which $B = 1$) undergoes a pitchfork bifurcation at $\mu = q$. The pitchfork bifurcation is supercritical for small q but becomes subcritical for $cq > 6$, as can be checked by expanding both the nonlinear term and P for small A in (2.9).

2.2. Modulational instabilities. As is typical in subcritical bifurcations of this kind, secondary bifurcations occur close to $\mu = 0$, and close to the saddle-node point on the uniform branch, which results in localized states. Interestingly, for this problem these secondary modulational instabilities exist also for $2 < cq < 6$ where the primary bifurcation is supercritical. On the primary branch, where $A = A_0$ constant, we find $P = \sqrt{1 + cA_0^2}$, and hence from (2.9)

$$q = (\mu - 3A_0^2)\sqrt{1 + cA_0^2}, \tag{2.10}$$

which, on simplifying, yields

$$9cA_0^6 + (9 - 6c\mu)A_0^4 + (c\mu^2 - 6\mu)A_0^2 + \mu^2 - q^2 = 0.$$

To locate the bifurcation points indicating modulational instability we set $A = A_0(1 + ae^{iKX})$. Substituting this ansatz into (2.9) and linearizing in a, we obtain

$$0 = (\mu - 4K^2 - 9A_0^2)a - (\mu - 3A_0^2)\frac{1 - 2cA_0^2}{1 + cA_0^2}a,$$

where we have used (2.10) to eliminate q. Simplifying further, we find that modulational instability occurs when

$$15cA_0^4 + (6 - 3c\mu + 4cK^2)A_0^2 + 4K^2 = 0.$$

In the limit of large domains, $K = 2\pi/(\varepsilon L) \ll 1$, we therefore expect instabilities when

$$A_0^2 = \frac{4K^2}{3c\mu - 6} + O(K^4) \qquad \text{and} \qquad A_0^2 = \frac{c\mu - 2}{5c} + O(K^2). \tag{2.11}$$

Clearly, no modulational instability is possible if $c\mu < 2$. The first of the conditions in (2.11) indicates that instability occurs at small A_0, near the primary bifurcation at $\mu = q$. The second condition indicates that instability also occurs at large amplitudes.

The continuation software AUTO [14] was used to solve (2.9) as a boundary-value problem in a finite domain. Neumann boundary conditions $A_X = A_{XXX} = 0$ at $X = 0, \varepsilon L$ were imposed to avoid numerical difficulties arising from the continuous translational symmetry implied by periodic boundary conditions. Bifurcation diagrams for the supercritical and subcritical cases are shown in Figure 2. Of particular note in Figure 2(b) is that the branch of localized states both extends further into $\mu < q$ than the uniform branch, i.e., $\mu_{sn1} < \mu_{sn2} < q$, and also extends substantially into $\mu > q$ before rejoining the primary branch at large amplitudes.

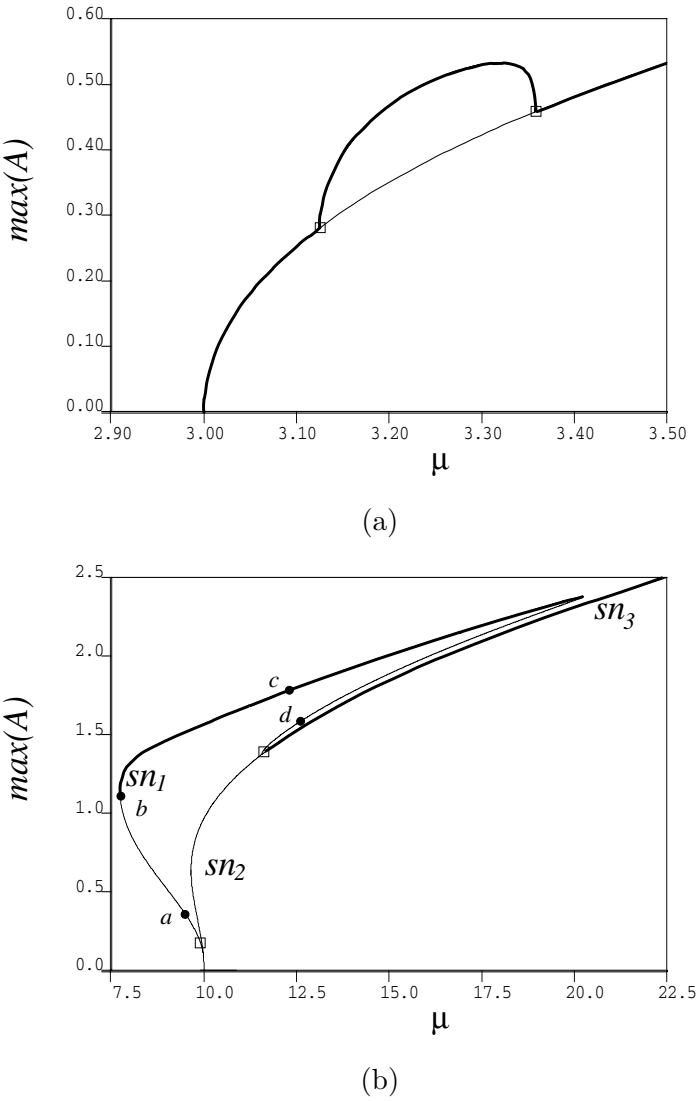


Figure 2. *Bifurcation diagrams in the* $(\mu, \max(A))$ *plane. Thick and thin lines denote stable and unstable branches, respectively. A primary branch of uniform amplitude pattern bifurcates from* $A(X) \equiv 0$ *when* $\mu = q$*. We fix* $c = 1$ *and the domain size* $\varepsilon L = 10\pi$*.* $\square$ *denotes bifurcation points from the uniform branch to the branch of modulated states.* (a) $q = 3$*, for which the primary bifurcation is supercritical.* (b) $q = 10$*, for which the primary bifurcation is subcritical. Note in both cases the existence of a secondary instability leading to a branch of spatially localized states. Labels in* (b) *correspond to the different parts of Figure 3.*

Figure 3 shows solutions to (2.9) at the four points indicated on the localized branch in Figure 2(b). Close to the ends of the branch the solution takes on the usual sech-like profile; at the center it resembles a pair of tanh-like fronts between the trivial state $A = 0$ and a nonzero constant value A_0. Stable fronts are possible when the two states have the same "energy"; in the standard description of localized states the energies are equal at a single value of the driving parameter μ, known as the "Maxwell point." In the present case we

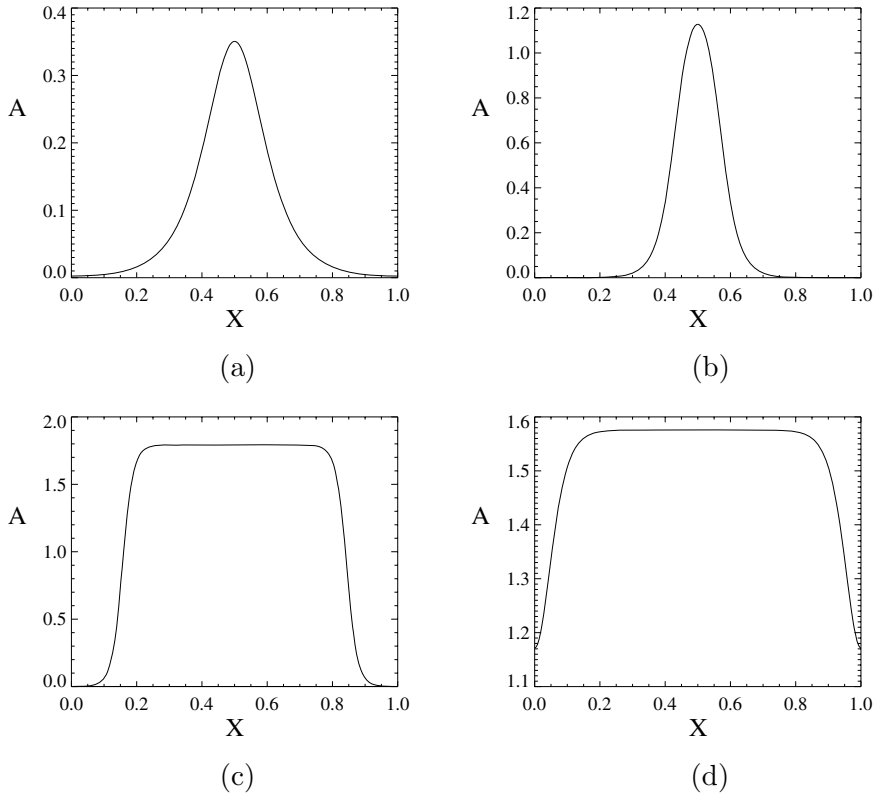


Figure 3. *Localized solutions $A(X)$ of (2.9) at the four points on the secondary branch of localized states indicated on Figure 2(b). Parameter values:* (a) $\mu = 9.523$; (b) $\mu = 7.737$; (c) $\mu = 12.373$; (d) $\mu = 12.534$. *Domain size $\varepsilon L = 10\pi$, $c = 1$, $q = 10$. Note that the horizontal axis is rescaled to $[0, 1]$ in the figures.*

expect that the value A_0 depends on the driving parameter μ but not on the domain size or the overall magnitude q of the large-scale mode. These intuitions can be demonstrated with a straightforward, but surprisingly accurate, calculation to estimate the relation between A_0 and μ, as we now show.

Equation (2.9) has a first integral, obtained by multiplying by A_X and integrating:

$$E = \frac{\mu}{2}A^2 + 2A_X^2 - \frac{3}{4}A^4 + \frac{qP^2}{c}\frac{1}{\sqrt{1+cA^2}}. \tag{2.12}$$

Assuming that the localized solution resembles Figure 3(c) and is nearly piecewise constant, we may neglect the A_X term in (2.12); this turns out not to affect the accuracy of the following calculation in any significant way, while considerably simplifying the computation. Supposing that the solution is $A = A_0 \neq 0$ over a proportion ℓ/L of the domain, and zero on the remainder, from (2.8) we obtain $P = P_\ell$, where

$$\frac{1}{P_\ell} = 1 + \frac{\ell}{L}\left(\frac{1}{\sqrt{1+cA_0^2}} - 1\right). \tag{2.13}$$

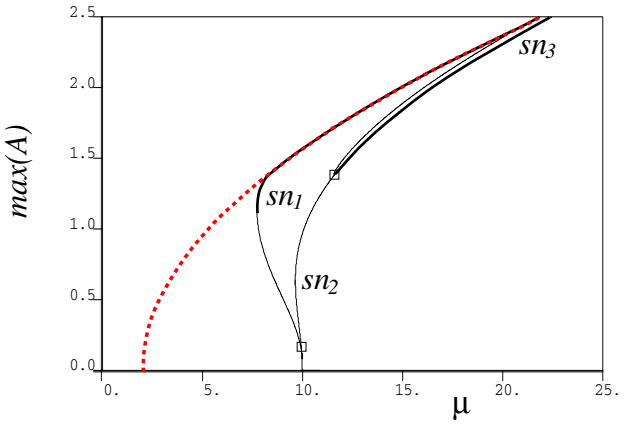


Figure 4. *Bifurcation diagram in the* $(\mu, \max(A))$ *plane for* $q = 10$, $c = 1$, $\varepsilon L = 10\pi$, *as in Figure* 2(b). *Thick and thin lines denote stable and unstable branches, respectively.* □ *denotes bifurcation points from the uniform branch to the branch of modulated states. The "Maxwell curve" given by* (2.15) *is the dashed red curve. The central part of the stable branch of localized states is indistinguishable from it.*

Comparing the values of E above and below a front connecting the trivial and nontrivial states in different parts of the domain, we obtain

$$E|_{A=0} = \frac{qP_\ell^2}{c},$$
$$E|_{A=A_0} = \frac{\mu}{2}A_0^2 - \frac{3}{4}A_0^4 + \frac{qP_\ell^2}{c}\frac{1}{\sqrt{1+cA_0^2}},$$

and from (2.9) we also have

$$qP_\ell^2 = (\mu - 3A_0^2)(1+cA_0^2)^{3/2}. \tag{2.14}$$

Equating $E|_{A=0} = E|_{A=A_0}$ and eliminating qP_ℓ^2 using (2.14), we obtain the following relation between μ and A_0:

$$cA_0^2\left(\mu - \frac{3}{2}A_0^2\right) = 2(\mu - 3A_0^2)(1+cA_0^2)\left(\sqrt{1+cA_0^2} - 1\right).$$

This can be simplified to the cubic polynomial in A_0^2:

$$144c^2A_0^6 + (207c - 96c^2\mu)A_0^4 + (72 + 16c^2\mu^2 - 108c\mu)A_0^2 + 12\mu(c\mu - 2) = 0. \tag{2.15}$$

For $\mu = 12$ we find that this analytic result predicts $A_0 = 1.7590818$, compared to the numerical value from (2.9), corresponding to Figure 3(c), of $A_0 = 1.759082$. Moreover, instead of a Maxwell point we have a "Maxwell curve" along which stable fronts, and therefore localized states, exist; see Figure 4. It is worth remarking that (2.15) relates the amplitude A_0 only to the linear driving parameter μ and does not contain the coupling parameter q or the proportion ℓ/L of the domain that contains the localized pattern. As a result, the localized pattern

amplitude depends only on μ. This is physically reasonable in the magnetoconvection case: when the magnetic field has been expelled from one part of the fluid domain, the amplitude of the thermal convection that results is due to the thermal driving alone.

Our final remark in this section is that on combining (2.13) and (2.14) we can relate q and ℓ/L, assuming that A_0 is determined from μ using (2.15):

$$q = (\mu - 3A_0^2)(1 + cA_0^2)^{3/2}\left(1 + \frac{\ell}{L}\left(\frac{1}{\sqrt{1+cA_0^2}} - 1\right)\right)^2.$$

Approximating this expression in the limit of "strong coupling and strong driving," where $q \sim \mu \gg 3A_0^2 \gg 1$, it can be written in the form

$$\frac{q}{\mu} \approx A_0^3\left(1 - \frac{\ell}{L}\right)^2,$$

which is reminiscent of (4.5) in [13] and indicates that in the strong coupling and strong driving regime it is some combination of q and μ that determines the width of the localized state. Note that the amplitude A_0 scales in some unspecified way with μ and q, and so this relation does not indicate a simple power-law scaling exponent, as we discuss further in the next section.

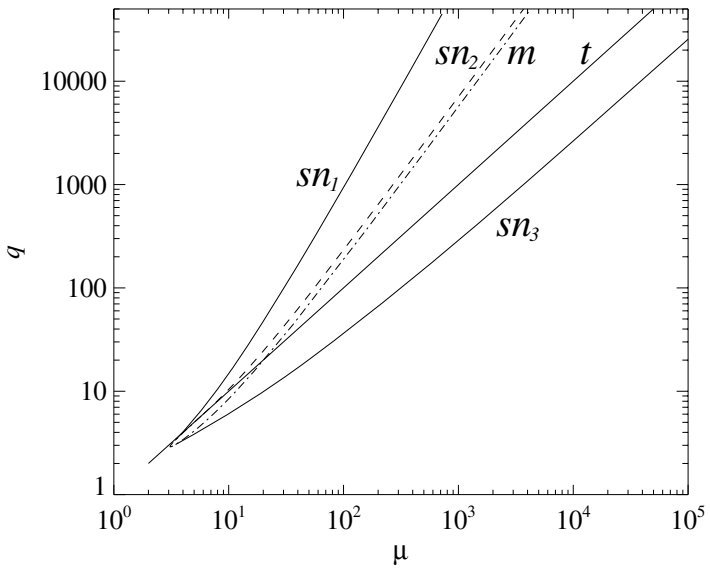
2.3. Scaling laws in the nonlocal Ginzburg–Landau equation. The location and shape of the primary and secondary branches evolve continuously between Figures 2(a) and 2(b) as q is increased, first by the introduction of saddle-node points on the secondary branch, and then by the subcriticality of the primary branch. Figure 5 displays the bifurcation structure in the (μ, q) plane; for $q \gg 10$ the bifurcation points appear to scale as power-laws with increasing q, and the region of existence of stable localized states increases in size rapidly. Figure 5 (in which $c = 1$) shows that the localized states exist subcritically (i.e., for $\mu - q < 0$) even for $q < 6$, where the primary bifurcation is still supercritical.

The scaling law for the saddle-node bifurcation sn_2 on the uniform amplitude branch can be located by eliminating A_0^2 between the second condition of (2.11) and (2.10). This yields the curve

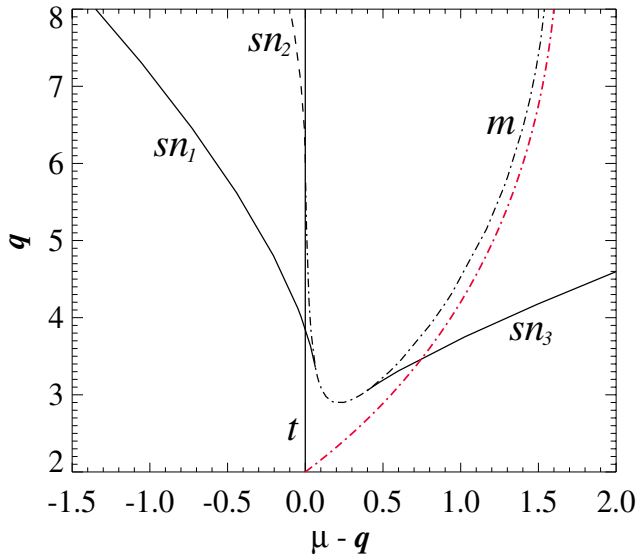
$$\left(cq\frac{5\sqrt{5}}{2}\right)^2 = (3 + c\mu)^3, \tag{2.16}$$

which agrees closely with numerical calculations and is shown as the red dot-dashed curve in Figure 5(b). At large q, (2.16) agrees exactly with the numerical results in Figure 5(a).

Fitting the same functional form to the saddle-node curves sn_1 and sn_3 results in the scaling laws sn_1: $q \approx 0.0927(\mu + 3.55)^{1.987}$, sn_3: $q \approx 0.298(\mu + 27.9)^{0.986}$ to three significant figures. These exponents are intriguing, and, while they fit the data extremely well at large q, they differ significantly from ratios of small integers. It is possible that they can be deduced using the properties of solutions of (2.9) involving the snoidal and cnoidal special functions. At small q we note that there is systematic deviation from power-law scalings.



(a)



(b)

Figure 5. (a) *Bifurcation diagram in the* (μ, q) *plane showing the power-law behavior of the bifurcation curves at large* q. $c = 1$, *domain size* $\varepsilon L = 10\pi$. (b) *Enlargement of* (a) *for small* q, *plotted in the* $(\mu - q, q)$ *plane for clarity.* sn_1 *and* sn_3 *refer to saddle-node bifurcations on the branch of localized states.* sn_2 *is the saddle-node on the primary, uniform amplitude branch.* t *refers to the linear instability point* $\mu = q$. m *labels the modulational instability of the primary branch above the saddle-node point, with the red dot-dashed line indicating the analytic result* (2.16).

3. Numerical results for slanted snaking. Returning to the Swift–Hohenberg model (2.1)–(2.2), we find, corresponding to the Ginzburg–Landau analysis, that the modulational instability leads to branches of localized states for which the periodic pattern is locked to the modulating envelope. As for the standard snaking scenario, two distinct pairs of branches persist—one pair where the phase difference between pattern and envelope is 0 or π and one pair where it is $\pm\pi/2$. The numerical procedure, again using AUTO [14], was as follows.

Starting at the trivial solution $w(x) \equiv 0$, either "Dirichlet" $w(0) = w_{xx}(0) = w(L) = w_{xx}(L) = 0$ or "Neumann" $w_x(0) = w_{xxx}(0) = w_x(L) = w_{xxx}(L) = 0$ boundary conditions were used to avoid the neutral eigenvalue associated with the translational invariance when locating the linear instability at $r = Q$ and switch easily onto the uniform amplitude pattern branch. By continuation along the uniform amplitude branch we located the first modulational instability (indicated by the lower solid square symbol in Figure 6), and AUTO was able to switch onto the modulated branch. We now replace the Dirichlet or Neumann boundary conditions in the numerics with periodic ones, supplemented with a global integral constraint to fix the overall phase of the solution and ensure that there is no drift in the direction of the neutrally stable translation mode. The implementation of the integral constraint followed Rademacher, Sandstede, and Scheel [29]; it was found to be numerically very robust. This approach also accurately detects the "cross-link" or "ladder" branches [5, 22] of asymmetric localized states that complete the bifurcation structure. As in the standard homoclinic snaking scenario, these cross-link branches are never stable; they are omitted from Figure 6 to aid clarity.

Two continuations, one beginning with Dirichlet boundary conditions and one with Neumann boundary conditions, were carried out. The full snaking bifurcation diagram is obtained by superimposing the results. This procedure provides an additional check on the numerical accuracy. Figure 6 indicates that, as expected, these branches are intertwined (around the "Maxwell curve") and stretch both below and above the bifurcation points from the uniform amplitude branch. We call this behavior "slanted snaking." Figure 7 illustrates the evolution of the localized states along the snaking branches.

Another feature of Figure 6 not captured by the Ginzburg–Landau reduction is that the wavelength along the branch of localized states in Figure 6 decreases as r increases; in consequence the branches of localized states terminate on a different uniform amplitude primary branch, in this case one with wavelength $L/8$. This aspect of the bifurcation diagram is brought out clearly in Figure 8, where the vertical axis is $\max(w)$ rather than the L^2 norm.

Decreasing r from, say, $r = 3.5$ on the localized branch therefore results in a stepwise decrease in the number of bumps of localized pattern, as the successive saddle-node bifurcations are passed. As ε decreases the localization becomes increasingly pronounced.

Alternatively we may consider the driving parameter r to be fixed and consider the effect of increasing the strength of the coupling Q between the large-scale field and the pattern mode. For the Ginzburg–Landau system this corresponds to a vertical section through Figure 5(a); the same stepwise series of saddle-node bifurcations is seen in (2.1)–(2.2) when Q is decreased at fixed r. Figure 9 (for which $r = 1$) illustrates the location of the lowest two saddle-node bifurcations on the snaking branch as ε is decreased. Localized states are born in a modulational instability close to $Q = 1$ and exist in $Q > 1$ up to the curve sn_1 on which they undergo the first saddle-node bifurcation on the snaking curve. The branch then turns around

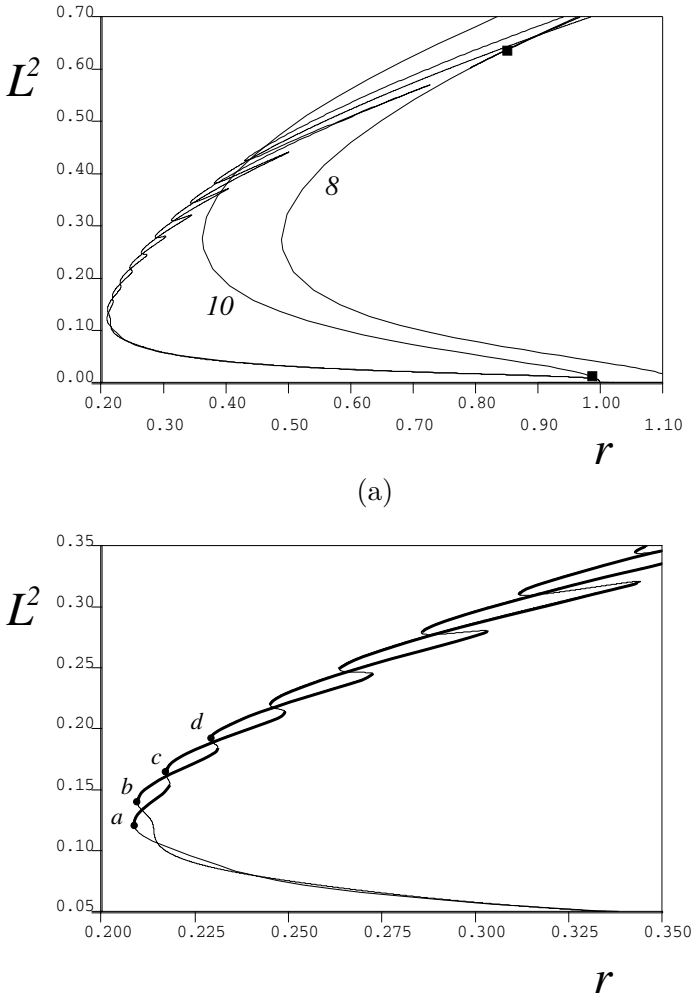


Figure 6. *Slanted snaking in* (2.1)–(2.2)*,* $c = Q = 1$*, domain size* $L = 20\pi$*.* (a) *Branches* 10 *and* 8 *denote uniform spatial patterns with wavelengths* $L/10$ *and* $L/8$*, respectively. Snaking branches bifurcate from the* 10 *branch near* $r = 1$ *(at the solid square) and extend down to* $r = 0.21$*. They then continue up to* $r \approx 3.8$ *(not shown) before returning to terminate on the* 8 *branch, indicated by a solid square.* (b) *Enlargement of* (a) *showing the characteristic intertwining of the two snaking branches. Thick and thin lines indicate stable and unstable solutions, respectively. Labels a–d refer to Figure* 7*.*

and continues as Q decreases, up to the second solid line at which the second saddle-node bifurcation on the snaking curve takes place. Subsequent pairs of saddle-node bifurcations are not shown; numerical results indicate that the next pair, and possibly others, rather curiously follows the same power-law scaling with ε as the first pair. As ε increases, the region of stable localized states shrinks until it disappears for $\varepsilon > 0.46$. For comparison, Figure 9 also shows the location of the saddle-node bifurcation of the uniform amplitude pattern sn_2. The power-law for the curve sn_2 follows that derived in (2.16) for the location of the modulational instability near the saddle-node bifurcation sn_2: the scaling $q^2 \propto \mu^3$ for large q and μ implies

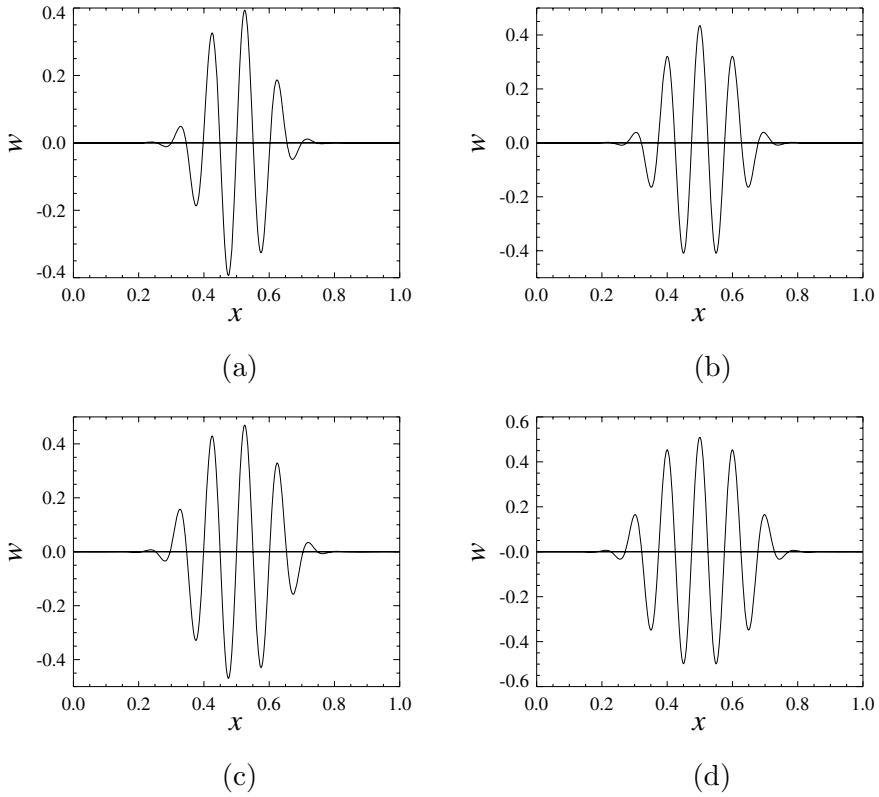


Figure 7. *Localized states at four saddle-node bifurcation points in Figure* 6(b). (a) *and* (c) *lie on the* $\phi = \pi/2$ *(odd) branch;* (b) *and* (d) *lie on the* $\phi = 0$ *(even) branch.* $c = Q = 1$*;* $L = 20\pi$*. Note that the domain has been rescaled to* $[0, 1]$.

$(Q/\varepsilon^2)^2 \propto (r/\varepsilon^2)^3$, i.e., $Q \propto \varepsilon^{-1} r^{3/2}$. This scaling agrees with Figure 9.

For sn_1 we found in the Ginzburg–Landau approximation (shown in Figure 5) that $q \propto \mu^{1.987}$ for large q, μ. This implies $Q/\varepsilon^2 \propto (r/\varepsilon^2)^{1.987}$, which yields $Q \propto \varepsilon^{-1.974}$. For comparison, Figure 9 indicates the different scaling $Q \propto \varepsilon^{-2.03}$. This difference is too large to be explained purely in terms of numerical errors. We believe that the difference is due to the "beyond-all-orders" terms that determine the width of the homoclinic snaking wiggles in Figure 6 and that are neglected in the multiple-scales analysis of section 2.

For completeness we note that the lower solid line in Figure 9 follows the power-law $Q \propto \varepsilon^{-1.31}$, which has no counterpart in the multiple-scales asymptotics of section 2; clearly the width of the snake, and therefore this power law exponent also, is influenced by the beyond-all-orders asymptotic scalings.

4. Discussion. In this paper we have examined a very simple model equation for the influence of a neutrally stable long-wavelength mode on steady-state pattern formation. Such a situation is motivated by several physical problems, and we have fixed on model equations containing symmetries appropriate to the onset of thermal convection in an imposed vertical magnetic field, a long-studied problem in the literature [7, 6, 27].

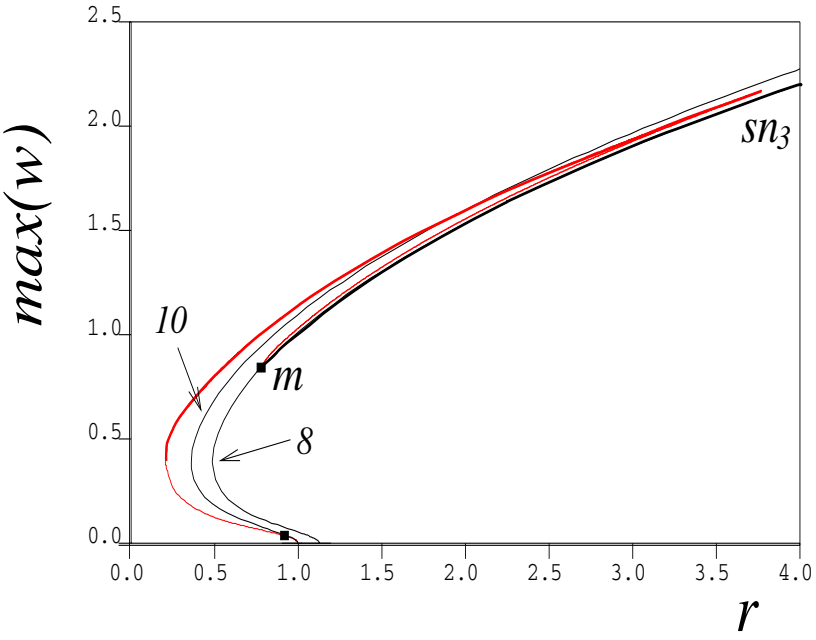


Figure 8. *Bifurcation behavior in (2.1)–(2.2) for $c = Q = 1$ and domain size $L = 20\pi$. The bifurcation curves are as for Figure 6; plotting $\max(w)$ against r shows more clearly the excursion of the secondary branch up to $r \approx 3.8$, and the termination of the snaking branches (now superimposed) on the 8 branch. Solid squares denote the bifurcations from uniform periodic patterns to localized patterns. Thick and thin lines denote stable and unstable branches, respectively. Red and black lines denote localized and spatially periodic solution branches, respectively.*

We introduce a new approach to the usual weakly nonlinear multiple-scales analysis, taking the diffusivity of the large-scale mode as the small parameter while allowing order unity fluctuations in amplitude. We then carry out the multiple-scales expansion to third order and deduce a nonlocal Ginzburg–Landau equation that describes the dynamics. It is found that the subcriticality induced by the large-scale mode distorts the usual homoclinic snaking picture and allows localized states to exist over a much larger region of parameter space than is possible in its absence. We refer to this distortion of homoclinic snaking as "slanted snaking." As a result, the existence of a large-scale mode provides a far more robust physical mechanism for the stabilization of localized solutions than the "locking" or "pinning" mechanism that explains their existence in an exponentially small wedge of parameter space in the standard picture.

In other respects the localized states qualitatively resemble standard homoclinic snaking. For example, the secondary branches of asymmetric localized states, found by Burke and Knobloch [5] (called "ladders" in that paper), which link the two snaking branches exist also in this problem, although for clarity we have omitted them from the figures. For the parameter values we investigate in detail ($\varepsilon = 0.1$, $Q = 1$), it is interesting that the lowest saddle-node point on the snake does not correspond to a single isolated period of the pattern. This indicates that our coupling terms, of a simple kind that admit analytic investigation, do

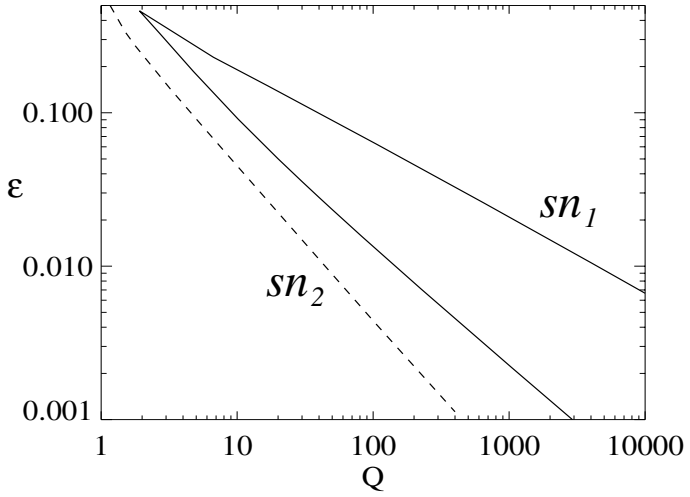


Figure 9. *Scaling of the saddle-node bifurcations on the primary and secondary branches of solutions to (2.1)–(2.2) in the (ε, Q) plane for fixed $r = c = 1$ and domain size $L = 20\pi$. The trivial solution $w(x,t) \equiv 0$ is stable for $Q > 1$. A single localized cell exists between $Q = 1$ and the saddle-node curve sn_1; it is stable between the two solid curves. Uniform amplitude states exist (subcritically) below the dashed line sn_2.*

not force the pattern to be quite as localized as convection cells appear to be in simulations of the full fluid equations [2, 13]. It is quite possible that more complicated coupled terms, involving extra derivatives of either $B(x,t)$ or $w(x,t)$, deform the snake still further, allowing only a single convection cell to persist at small r or, equivalently, large Q. Possible forms for these coupling terms are discussed further in Appendix A.

Overall, these results provide a convincing explanation of the link between homoclinic snaking and the stepwise reduction in the number of cells in the localized states with increasing Q as found first by Blanchflower [3] and reproduced in [13] (see Figure 8 therein). Various features of the full magnetoconvection problem, such as the destabilization of localized steady states by an oscillatory instability in the quiescent region, are, of course, not reproduced by this model. But steady-state features are well reproduced, for example, the evidence, from Figure 9(a) of [13], that the two curves of saddle-node bifurcations that bound the region of the existence of a single-roll localized state scale in different ways with the small parameter $\varepsilon \equiv \zeta$. This is certainly true for the solid lines in Figure 9.

The exponents of the power-law scalings that we present, in both the multiple-scales analysis and the full Swift–Hohenberg model, seem to be strongly dependent on the exact form of the nonlinear coupling terms. A detailed investigation of such dependencies and possible explanations through the properties of analytic solutions in terms of Jacobi elliptic functions are left for future work.

The corresponding calculation for systems for which the large-scale mode is a density is very similar to the analysis presented in detail here. Indeed, the only change required to the model equations is to take the coupling term to be QBw in (2.1). Brief details of the resulting calculations are contained in Appendix B. A further extension is to examine systems where it

appears that the large-scale mode promotes localized activity but the short-scale dynamics is not "pattern-forming." A clear example of this is the vertically and horizontally shaken granular layer experiments of Götzendorfer et al. [17]. In these experiments the vertical excitation of a granular layer does not result in uniform excitation of the material, but rather in a patch of highly energetic particles, while the remainder of the domain remains inactive. Götzendorfer et al. refer to this, slightly fancifully perhaps, as the "sublimation" of the "solid" phase of the material into a "gaseous" form. The general mechanism is, however, physically clear: there is a balance between the (rapid) flux of particles from the active part of the layer into the quiet part as energetic particles are propelled upward and outward, and the natural (slow) trickling of particles from the quiet part back into the active one. These two processes correspond, respectively, to the two terms on the right-hand side of (B.2). Moreover, when the layer height locally exceeds a critical value, the vertical excitation cannot excite particles directly in that part of the bed; this effect is captured by the damping term $Q\rho w$ in (B.1). Model equations capturing exactly these effects were written down by Tsimring and Aranson [32].

There are close connections between this work and that of Matthews and Cox and others [24, 9, 10, 16, 36]; these authors studied systems essentially equivalent to (2.1)–(2.2) obtained by applying ∂_{xx}^2 to the right-hand side of the standard Swift–Hohenberg equation (1.1), including both quadratic and cubic nonlinearities in $N(w)$. In these papers the weakly nonlinear analysis proceeds by looking for small distortions of the large-scale mode, i.e., a small parameter δ and a long lengthscale $X = \delta x$ are introduced, before expanding $w = \delta a(X)\sin x + O(\delta^2)$ and $B = 1 + \delta^2 b(X) + O(\delta^3)$. The resulting amplitude equations for $a(X)$ and $b(X)$ enable the detection of secondary modulational instabilities, as happens here, even in the case that the initial pattern-forming instability is supercritical, which is analogous to Figure 2(a). However, these scalings implicitly restrict our attention only to cases of small disturbances to the distribution of the large-scale mode; the present analysis can, in this sense, go further.

Other physical systems that show related phenomena, and which we intend to examine in future work, include the numerical results of Tsitverblit and Kit [33] on natural double-diffusive convection (see, for example, their Figure 1, which appears to show snaking behavior that does not have the saddle-node points aligned to only two values of the bifurcation parameter). Moreover, model equations for dielectric gas discharge (due to Purwins and collaborators [31, 28]) and for optical cavity lasers [15] have been proposed which include integral terms. These integral terms appear to play the same role in enhancing localization in these systems as the nonlinear diffusion equation for the large-scale mode does in this paper.

Appendix A. Magnetoconvection. In this appendix we briefly sketch the derivation of an evolution equation for the large-scale mode in the magnetoconvection case, starting from the governing equations. This justifies the form of our model equation (2.2).

For thermal convection in a vertical magnetic field, the appropriate governing equations for the fluid velocity $\mathbf{u}(x,y,z,t)$, the temperature perturbation $\theta(x,y,z,t)$, and the magnetic field $\mathbf{B}(x,y,z,t)$ are the momentum, temperature, and induction equations:

$$\begin{aligned}\partial_t\mathbf{u} + \mathbf{u}\cdot\nabla\mathbf{u} &= -\nabla p + \sigma R\theta\hat{\mathbf{z}} + \sigma\zeta Q\mathbf{B}\cdot\nabla\mathbf{B} + \sigma\nabla^2\mathbf{u},\\ \partial_t\theta + \mathbf{u}\cdot\nabla\theta &= w + \nabla^2\theta,\\ \partial_t\mathbf{B} &= \nabla\times(\mathbf{u}\times\mathbf{B}) + \zeta\nabla^2\mathbf{B}.\end{aligned}$$

For simplicity we restrict our attention to two-dimensional solutions, ignoring the y coordinate; the extension of these calculations to three dimensions would appear to be straightforward. The temperature variable $\theta(x,z,t)$ is the perturbation to the conduction profile $T = 1 - z$. The velocity and magnetic fields are solenoidal: $\nabla \cdot \mathbf{u} = \nabla \cdot \mathbf{B} = 0$. Note that $\mathbf{B}$ is the full magnetic field—not the perturbation to an initially vertical field of strength unity.

The dimensionless groups are the Rayleigh number R, Chandrasekhar number Q, Prandtl number $\sigma = \nu/\kappa$, and magnetic Prandtl number $\zeta = \eta/\kappa$.

We suppose that the boundaries are stress-free and are held at fixed temperatures, and we constrain the field to be vertical there. These boundary conditions allow a simple analytical treatment. The linear theory for the onset of thermal convection in a vertical field is well known [7, 27, 10, 13]. We look for small-amplitude solutions in one horizontal dimension, taking ζ as our small parameter (i.e., $\varepsilon = \zeta$). We propose the solution ansatz

$$\begin{aligned}\mathbf{u} &= (u_1 \cos \pi z, 0, w_1 \sin \pi z),\\ \theta &= \theta_1 \sin \pi z,\\ \mathbf{B} &= (B_x \sin \pi z, 0, B_z \cos \pi z + B_0),\end{aligned}$$

where u_1, w_1, B_x, B_z, and B_0 are functions of x and t only. Computing $\hat{\mathbf{z}} \cdot \nabla \times (\mathbf{u} \times \mathbf{B})$ we obtain equations for the z independent terms and, separately, those that depend on $\cos \pi z$:

$$\partial_t B_0 = \frac{1}{2}(w_1 B_x - u_1 B_z)' + \zeta B_0'', \tag{A.1}$$

$$\partial_t B_z = -(u_1 B_0)' + \zeta(B_z'' - \pi^2 B_z), \tag{A.2}$$

where primes $'$ denote ∂_x. We assume that the quantities u_1, w_1, B_x, B_z all vary on a short spatial scale with wavenumber k, so that we may, at leading order, replace $\partial_{xx}^2 \to -k^2$, where it acts on these variables. As a result the solenoidal conditions $\nabla \cdot \mathbf{u} = 0$ and $\nabla \cdot \mathbf{B} = 0$ imply $u_1 = \pi/k^2 \partial_x w_1$ and $B_x = -\pi/k^2 B_z$, which allows us to eliminate u_1 and B_x from (A.1)–(A.2). We obtain

$$\partial_t B_0 = -\frac{\pi}{2k^2}(w_1 B_z)'' + \zeta B_0'', \tag{A.3}$$

$$\partial_t B_z = -\frac{\pi}{k^2}(w_1' B_0)' + \zeta(B_z'' - \pi^2 B_z). \tag{A.4}$$

Since the linear eigenfunction for the onset of weakly nonlinear convection involves u_1, w_1, B_x, and B_z, it is clear also that B_z does not evolve independently of w_1. Looking for steady states of (A.4), we expect, therefore, that $B_z = -\pi/(k^2\beta^2\zeta)(w_1' B_0)'$ near onset, where $\beta^2 = k^2 + \pi^2$. Substituting this into (A.3) yields an evolution equation for the large-scale field $B_0(x,t)$, coupling it to the weakly nonlinear convection pattern amplitude $w_1(x,t)$:

$$\partial_t B_0 = \frac{\pi^2}{2k^4\beta^2\zeta}\left(w_1(w_1' B_0)'\right)'' + \zeta B_0''.$$

This is of the same form as (2.2), setting $c = \pi^2/(2k^4\beta^2)$, except for two extra derivatives in the first term. Since near onset w_1 involves only the single lengthscale $2\pi/k$, and $B_0 \approx 1$, it is clear that these derivatives will not qualitatively change the behavior. However, it is quite

possible that they will alter the exponents of various scaling laws, for example, those shown in Figure 9. A similar evaluation of the $\hat{\mathbf{z}}$ component of the $\mathbf{B}\cdot\nabla\mathbf{B}$ term, and subsequent use of the above approximations of B_x and B_z in terms of B_0 and w_1, yield, heuristically, the term that couples the large-scale mode back to the vertical velocity:

$$\partial_t w_1 = \cdots + \frac{\sigma Q\pi^2}{k^2\beta^2}\left(\frac{B_0'}{k^2}+B_0\right)(w_1'B_0)'.$$

Appendix B. A density-like large-scale mode. Suppose that a one-dimensional pattern-forming system is described by the pattern amplitude $w(x,t)$ and a density-like large-scale mode $\rho(x,t)$, for example, the local layer height in a granular medium or thin film. In this case the symmetry $\rho\to-\rho$ is absent, and the corresponding model equations are

$$w_t = [r-(1+\partial_{xx}^2)^2]w - w^3 - Q\rho w, \tag{B.1}$$

$$\rho_t = \varepsilon\rho_{xx} + \frac{c}{\varepsilon}(\rho w^2)_{xx}. \tag{B.2}$$

As before, we set $\frac{1}{L}\int_0^L \rho(x,t)dx \equiv \langle\rho\rangle = 1$. We restrict our attention to steady solutions, setting $\partial_t\equiv 0$, and integrate (B.2) twice to obtain

$$\rho = \frac{P}{1+cw^2/\varepsilon^2}, \qquad \text{where} \qquad \frac{1}{P}\equiv\left\langle\frac{1}{1+cw^2/\varepsilon^2}\right\rangle_x. \tag{B.3}$$

Substituting this into (B.1) and looking for steady states, we find

$$0 = [r-(1+\partial_{xx}^2)^2]w - w^3 - \frac{QPw}{1+cw^2/\varepsilon^2}.$$

We now introduce the multiple-scales ansatz $w(x,t) = \varepsilon A(X)\sin x + \varepsilon^2 w_2 + \varepsilon^3 w_3 + \cdots$, introducing the long lengthscale $X=\varepsilon x$. We rescale the parameters $r=\varepsilon^2\mu$ and $Q=\varepsilon^2 q$ in the standard way. At third order in the expansion an amplitude equation for $A(X)$ is obtained by multiplying by $\sin x$ and integrating over the short lengthscale, denoting the average $\frac{1}{2\pi}\int_0^{2\pi}\cdot\,dx \equiv \langle\,\cdot\,\rangle$. This yields

$$0 = \mu A + 4A_{XX} - 3A^3 - 2qP\left\langle\frac{A\sin^2 x}{1+cA^2\sin^2 x}\right\rangle_x, \tag{B.4}$$

where the constant P is defined as before:

$$\frac{1}{P} = \frac{1}{\varepsilon L}\int_0^{\varepsilon L}\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{1+cA^2\sin^2 x}\,dx\,dX \equiv \left\langle\left\langle\frac{1}{1+cA^2\sin^2 x}\right\rangle_x\right\rangle_X = \left\langle\frac{1}{\sqrt{1+cA^2}}\right\rangle_X.$$

Carrying out the x-integral in (B.4), we obtain

$$0 = \mu A + 4A_{XX} - 3A^3 - \frac{2qP(\sqrt{1+cA^2}-1)}{cA\sqrt{1+cA^2}}. \tag{B.5}$$

Consideration of the last term in the limit $A \to 0$ indicates that there is no singularity at small A and that the trivial solution $A = 0$ is linearly unstable when $\mu > q$.

It can be seen that (B.5) has a first integral,

$$E = \frac{\mu}{2} + 2(A_X)^2 - \frac{3}{4}A^4 - \frac{2qP}{c}\left[\log A + \tanh^{-1}\left(\frac{1}{\sqrt{1+cA^2}}\right)\right],$$

in which, after some manipulation, the last term on the right-hand side can be rewritten to give

$$E = \frac{\mu}{2} + 2(A_X)^2 - \frac{3}{4}A^4 - \frac{qP}{c}\log F(A),$$

where $F(A) = A^2 + \frac{2}{c}\left[1 + \left(cA^2+1\right)^{1/2}\right]$. We may also eliminate qP to yield the expression corresponding to (2.15) that describes the amplitude of a localized state as a function of μ:

$$(\mu - 3A_0^2)\sqrt{1+cA_0^2}\left[\frac{1}{2}\log F(A_0) - \log\frac{2}{\sqrt{c}}\right] = \left(\frac{\mu}{2} - \frac{3A_0^2}{4}\right)\left(\sqrt{1+cA_0^2} - 1\right).$$

Despite the more complicated functional form, this curve is qualitatively very similar to that defined by (2.15) and provides an analytic estimate of the "Maxwell curve" in this case.

Acknowledgments. I would like to acknowledge constructive suggestions from the referees. This work has benefited from instructive and useful conversations with Alan Champneys, Steve Cox, Paul Matthews, and Mike Proctor. It was begun at the Isaac Newton Institute, Cambridge.

REFERENCES

[1] N. Akhmediev and A. Ankiewicz, eds., *Dissipative Solitons*, Lecture Notes in Phys. 661, Springer-Verlag, Berlin, 2005.
[2] S. M. Blanchflower, *Magnetohydrodynamic convectons*, Phys. Lett. A, 261 (1999), pp. 74–81.
[3] S. M. Blanchflower, *Modelling Photospheric Magnetoconvection*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1999.
[4] A. Bose and G. A. Kriegsmann, *Stability of localized structures in non-local reaction-diffusion equations*, Methods Appl. Anal., 5 (1998), pp. 351–366.
[5] J. Burke and E. Knobloch, *Localized states in the generalized Swift–Hohenberg equation*, Phys. Rev. E (3), 73 (2006), 056211.
[6] F. H. Busse, *Nonlinear interaction of magnetic field and convection*, J. Fluid Mech., 71 (1975), pp. 193–206.
[7] S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability*, Clarendon, Oxford, UK, 1961; republished by Dover, New York, 1981.
[8] P. Coullet, C. Riera, and C. Tresser, *Stable static localized structures in one dimension*, Phys. Rev. Lett., 84 (2000), pp. 3069–3072.
[9] S. M. Cox and P. C. Matthews, *Instability and localisation of patterns due to a conserved quantity*, Phys. D, 175 (2003), pp. 196–219.
[10] S. M. Cox, P. C. Matthews, and S. L. Pollicott, *Swift–Hohenberg model for magnetoconvection*, Phys. Rev. E (3), 69 (2004), 066314.
[11] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.
[12] F. A. Davidson and N. Dodds, *Spectral properties of non-local differential operators*, Appl. Anal., 85 (2006), pp. 717–734.

[13] J. H. P. Dawes, *Localized states in thermal convection with an imposed vertical magnetic field*, J. Fluid Mech., 570 (2007), pp. 385–406.
[14] E. Doedel, A. R. Champneys, T. Fairgrieve, Y. Kuznetsov, B. Sandstede, and X. Wang, *AUTO97: Continuation and Bifurcation Software for Ordinary Differential Equations*, 1997, http://indy.cs.concordia.ca/auto/.
[15] W. J. Firth, L. Columbo, and A. J. Scroggie, *Proposed resolution of theory-experiment discrepancy in homoclinic snaking*, Phys. Rev. Lett., 99 (2007), 104503.
[16] A. A. Golovin, S. H. Davis, and P. W. Voorhees, *Self-organisation of quantum dots in epitaxially strained solid films*, Phys. Rev. E (3), 68 (2003), 056203.
[17] A. Götzendorfer, J. Kreft, C. A. Kruelle, and I. Rehberg, *Sublimation of a vibrated granular monolayer: Coexistence of gas and solid*, Phys. Rev. Lett., 95 (2005), 135704.
[18] Y. Hiraoka and T. Ogawa, *Rigorous numerics for localized patterns to the quintic Swift–Hohenberg equation*, Japan J. Indust. Appl. Math., 22 (2005), pp. 57–75.
[19] R. B. Hoyle, *Pattern Formation: An Introduction to Methods*, Cambridge University Press, Cambridge, UK, 2006.
[20] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. A. Wadee, C. J. Budd, and G. J. Lord, *Cellular buckling in long structures*, Nonlinear Dynam., 21 (2000), pp. 3–29.
[21] G. Iooss and M. C. Pérouème, *Perturbed homoclinic solutions in reversible* 1 : 1 *resonance vector fields*, J. Differential Equations, 102 (1993), pp. 62–88.
[22] G. Kozyreff and S. J. Chapman, *Asymptotics of large bound states of localized structures*, Phys. Rev. Lett., 97 (2006), 044502.
[23] C. R. Laing, W. C. Troy, B. Gutkin, and G. B. Ermentrout, *Multiple bumps in a neuronal model of working memory*, SIAM J. Appl. Math., 63 (2002), pp. 62–97.
[24] P. C. Matthews and S. M. Cox, *Pattern formation with a conservation law*, Nonlinearity, 13 (2000), pp. 1293–1320.
[25] L. M. Pismen, *Patterns and Interfaces in Dissipative Dynamics*, Springer Ser. Synergetics 15, Springer-Verlag, Berlin, 2006.
[26] Y. Pomeau, *Front motion, metastability and subcritical bifurcations in hydrodynamics*, Phys. D, 23 (1986), pp. 3–11.
[27] M. R. E. Proctor and N. O. Weiss, *Magnetoconvection*, Rep. Progr. Phys., 45 (1982), pp. 1317–1379.
[28] H.-G. Purwins, H. U. Bödeker, and A. W. Liehr, *Dissipative solitons in reaction-diffusion systems*, in Dissipative Solitons, Lecture Notes in Phys. 661, Springer-Verlag, Berlin, 2005, pp. 267–308.
[29] J. D. M. Rademacher, B. Sandstede, and A. Scheel, *Computing absolute and essential spectra using continuation*, Phys. D, 229 (2007), pp. 166–183.
[30] H. Sakaguchi and H. R. Brand, *Stable localized solutions of arbitrary length for the quintic Swift–Hohenberg equation*, Phys. D, 97 (1996), pp. 274–285.
[31] C. Strümpel, H.-G. Purwins, and Y. A. Astrov, *Spatiotemporal filamentary patterns in a dc-driven planar gas discharge system*, Phys. Rev. E (3), 63 (2001), 026409.
[32] L. S. Tsimring and I. S. Aranson, *Localized and cellular patterns in a vibrated granular layer*, Phys. Rev. Lett., 79 (1997), pp. 213–216.
[33] N. Tsitverblit and E. Kit, *The multiplicity of steady flows in confined double-diffusive convection with lateral heating*, Phys. Fluids A, 5 (1993), pp. 1062–1064.
[34] P. B. Umbanhowar, F. Melo, and H. L. Swinney, *Localized excitations in a vertically vibrated granular layer*, Nature, 382 (1996), pp. 793–796.
[35] A. G. Vladimirov, J. M. McSloy, D. V. Skryabin, and W. J. Firth, *Two-dimensional clusters of solitary structures in driven optical cavities*, Phys. Rev. E (3), 65 (2002), 046606.
[36] D. M. Winterbottom, P. C. Matthews, and S. M. Cox, *Pattern formation in a model of a vibrated granular layer*, SIAM J. Appl. Dynam. Systems, 7 (2008), pp. 63–78.
[37] P. D. Woods and A. R. Champneys, *Heteroclinic tangles and homoclinic snaking in the unfolding of a degenerate reversible Hamiltonian–Hopf bifurcation*, Phys. D, 129 (1999), pp. 147–170.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 207–219
© 2008 Society for Industrial and Applied Mathematics

Numerical Experiments on Noisy Chains: From Collective Transitions to Nucleation-Diffusion*

Mario Castro† and Grant Lythe‡

Abstract. We consider chains of particles with nearest-neighbor coupling, independently subjected to noise, all initially in the same well of a symmetric double-well potential. If there are sufficiently few particles, transitions from one well to another are "collective"; i.e., all particles remain close together as they make the passage from one well to the other. In longer chains, only a fraction of the particles make an initial transition, creating a nucleated region that may grow or collapse by diffusion of its boundaries. Numerical experiments are used to explore the change of the scaling of the passage time as a function of the length of the chain, which distinguishes the two regimes. A suitable relationship between the noise amplitude, coupling, and number of particles in the chain yields convergence to the continuum ϕ^4 or Allen–Cahn stochastic partial differential equations in one space dimension. We estimate the characteristic width of newly nucleated regions and construct a numerical effective potential describing the dynamics in the nucleation-diffusion regime.

Key words. stochastics, nucleation, passage time

AMS subject classifications. 60H15, 82C05

DOI. 10.1137/070695514

1. Introduction.

Many spatially extended systems exhibit two locally stable states that coexist in the sense that, at any one time, different parts of the system are in different states. Nucleation events are fluctuation-driven transitions of part of the system from one state to another, creating domains in which the system is in one state. The domain boundaries are coherent structures [1, 2] that subsequently move about due to fluctuations [3]. The model system we study here is homogeneous and symmetric; i.e., there is no preferred part of the system, and, in the long run, each of the two states is present in equal proportion on average. These properties also mean that there is no a priori natural lengthscale of a newly nucleated region. Behind the apparent simplicity of the energy landscape lies a challenging problem of determining the most probable nucleation pathways [4, 5, 6].

In this article, we consider the model system of a chain of overdamped particles, each coupled to its two neighbors, subject to noise and to the double-well potential with minima at $\phi = \pm 1$:

$$V(\phi) = -\frac{1}{2}\phi^2 + \frac{1}{4}\phi^4. \tag{1.1}$$

*Received by the editors June 26, 2007; accepted for publication (in revised form) by A. Hagberg November 19, 2007; published electronically March 19, 2008. This work has been partially supported by the DGI of the Ministerio de Educación y Ciencia, Spain, through grant FIS2006-12253-C06-06.
http://www.siam.org/journals/siads/7-1/69551.html
†GISC, Escuela Técnica Superior de Ingeniería (ICAI), Universidad Pontificia Comillas, E-28015 Madrid, Spain (mariocastro73@gmail.com).
‡Department of Applied Mathematics, University of Leeds, Leeds LS2 9JT, U.K. (grantlythe@gmail.com).

The position of the ith particle at time t is a real-valued random variable $\mathbf{\Phi}_t(i)$, and the stochastic differential equation for the ith particle is

$$\mathrm{d}\mathbf{\Phi}_t(i) = \left(\mathbf{\Phi}_t(i) - \mathbf{\Phi}_t^3(i) + k(\mathbf{\Phi}_t(i-1) + \mathbf{\Phi}_t(i+1) - 2\mathbf{\Phi}_t(i))\right)\mathrm{d}t + (2/\beta)^{1/2}\,\mathrm{d}\mathbf{B}_t(i), \tag{1.2}$$

where $\mathbb{E}(\mathrm{d}\mathbf{B}_t(i)\mathrm{d}\mathbf{B}_{t'}(j)) = \delta_{i-j}\mathrm{d}t$. The index i runs from 1 to N, and we shall always use periodic boundaries. In our numerical experiments, the whole chain is initially in the left-hand well of the potential. We record the first $t > 0$ at which the chain is in the right-hand well. As a function of N, with k and β fixed, we find that the mean of this time, which we call the complete passage time, increases exponentially until a certain value and then increases less rapidly. This change corresponds to a transition from "collective" behavior, where all the particles surmount the potential barrier together, to "nucleation-diffusion" behavior, where only a subset of the chain makes the initial transition and the domain subsequently grows by diffusion of its boundaries. A Java applet that permits interactive numerical experiments is at http://www.maths.leeds.ac.uk/Applied/stochastic/chain.htm.

Although a discrete system is an appropriate model in many situations, an important reason for interest in (1.2) is as a finite-difference approximation of a continuum stochastic partial differential equation (SPDE) in one space dimension. Let

$$k = \Delta x^{-2}, \qquad N = \frac{L}{\Delta x}, \qquad \text{and} \qquad \beta = \frac{\Delta x}{\Theta}. \tag{1.3}$$

As $\Delta x \to 0$, the limit of the set of equations (1.2) is the overdamped ϕ^4 SPDE [7, 8, 1, 9]

$$\frac{\partial}{\partial t}\boldsymbol{\phi}_t(x) = \frac{\partial^2}{\partial x^2}\boldsymbol{\phi}_t(x) + \boldsymbol{\phi}_t(x) - \boldsymbol{\phi}_t^3(x) + (2\Theta)^{1/2}\,\boldsymbol{\xi}_t(x), \tag{1.4}$$

where $\boldsymbol{\phi}_t(i\Delta x) = \mathbf{\Phi}_t(i)$ and $x \in [0, L]$. The last term in (1.4) is space-time white noise:

$$\mathbb{E}(\boldsymbol{\xi}_t(x)\boldsymbol{\xi}_{t'}(x')) = \delta(x - x')\delta(t - t'), \tag{1.5}$$

where Θ is thought of as proportional to temperature. We will often refer to (1.3) as it provides scaling relations among the parameters of the problem and it will help us to understand the nucleation process in terms of adimensional relations. If, instead of (1.3), we let $N = \Delta x^{-1}$ with β and k fixed then, as $\Delta x \to 0$, we obtain the stochastic Allen–Cahn equation:

$$\frac{\partial}{\partial t}\boldsymbol{\phi}_t(x) = \epsilon^2\frac{\partial^2}{\partial x^2}\boldsymbol{\phi}_t(x) + \boldsymbol{\phi}_t(x) - \boldsymbol{\phi}_t^3(x) + \sigma\,\boldsymbol{\xi}_t(x), \tag{1.6}$$

where $\epsilon^2 = k/N^2$, $\sigma = (2\Delta x/\beta)^{1/2}$, and $x \in [0, 1]$. In the SPDEs, a configuration is a continuous function of x, $\boldsymbol{\phi}_t(x)$, obtained by fixing t in one realization. At most values of x, $\boldsymbol{\phi}_t(x)$ is close to either -1 or $+1$. A narrow region where the configuration crosses through 0 from below is called a kink; one where it crosses from above is called an antikink. The width of a kink in the ϕ^4 SPDE is order 1; in the Allen–Cahn case it is proportional to ϵ [10, 11, 12]. We shall concentrate on the ϕ^4 SPDE below.

After a sufficiently long time, in both the continuum SPDEs and the discrete system, a statistically steady state is attained and maintained by a balance between continual nucleation of new domains and the diffusion and annihilation of existing ones [13, 14, 3]. Many steady-state quantities, such as the mean number of kinks per unit length, can be calculated from

the invariant density of the SPDE, by evaluating the partition function [15, 16, 17]. Further insight has recently been obtained by demonstrating the equivalence between the invariant density of paths of the SPDE, on the spatial domain $[0, L]$, and the density of paths of a suitable bridge process [18, 19], with time in the interval $[0, L]$.

Simple models, in which kinks and antikinks are nucleated with a fixed separation, diffuse and annihilate on collision and give insight into the dynamics that produces and maintains the stationary density [20, 21, 22, 23]. If kink-antikink dynamics in our symmetric system is described in terms of a potential as a function of separation, a nucleation event is a fluctuation-induced escape from a well to a flat region. Büttiker and Christen calculated the nucleation rate by introducing a parameter analogous to a nucleus size s [20]. The mean kink lifetime is proportional to s and the nucleation rate is inversely proportional to s, so that the steady-state kink density is independent of s. More detailed models also calculate the distribution of kink lifetimes [3, 21, 23]. The initial separation at the instant of nucleation is an input parameter in these simple models; in order to calculate the appropriate value in the SPDE dynamics, it is necessary to return the focus to the details of the nucleation process.

1.1. The energy landscape. In chains sufficiently short that transitions are collective, the complete passage time is most conveniently calculated by considering the energy function of the discretized system:

$$E(\mathbf{\Phi}_t(1), \ldots, \mathbf{\Phi}_t(N)) = \sum_{i=1}^{N} \left(V(\mathbf{\Phi}_t(i)) + \frac{1}{2} k \left(\mathbf{\Phi}_t(i) - \mathbf{\Phi}_t(i-1) \right)^2 \right). \tag{1.7}$$

The initial condition is at the energy minimum: $E(-1, \ldots, -1) = -N/4$. If the transition is collective, crossing the saddle point on the energy surface at the origin $E(0, \ldots, 0) = 0$ involves surmounting an energy barrier $\frac{1}{4}N$; the mean time for such a transition is proportional to $\exp(\frac{1}{4}\beta N)$. On the other hand, if only a part of the system makes the initial transition, the energy barrier is less than $\frac{1}{4}N$; the nucleation event produces a region with two boundaries. In the latter case, which we call the "nucleation-diffusion regime," it is convenient to calculate the nucleation rate working in the continuum limit, (1.4), where the analogue of (1.7) is the energy functional [4]

$$\mathcal{E}[\boldsymbol{\phi}_t] = \int \left(V(\boldsymbol{\phi}_t) + \frac{1}{2} \left(\frac{\partial}{\partial x} \boldsymbol{\phi}_t \right)^2 \right) \mathrm{d}x. \tag{1.8}$$

The energy of a kink is $E_k = \mathcal{E}[\psi] = \sqrt{8/9}$ [1]. A kink or antikink has energy E_0; a transition creates one of each and therefore has energy barrier $2E_0$ and characteristic time $\exp(\frac{2E_0}{\Theta})$ [24, 25, 26, 27, 15, 4]. The numerical value of the kink energy depends on the precise potential used, but the qualitative dynamics requires only a double-well potential with wells of equal depth. It is also possible to study SPDEs where $V(\phi)$ has two (or more) wells of unequal depth, so that there are stable and metastable states. Then, one can base calculations of nucleation rates on the idea of a critical nucleus or droplet [27, 28, 29, 30, 31], an extremum of (1.8), whose length diverges as the asymmetry between wells vanishes.

In an interval containing only one kink, centered at $\mathbf{X}_t$, the configuration can be written as $\boldsymbol{\phi}_t(x) = \psi(x - \mathbf{X}_t) + (\Theta/E_\mathrm{k})^{1/2} \chi(x - \mathbf{X}_t)$, where ψ is a smooth function, satisfying

$V'(\psi) + \psi''(x) = 0$, with $\psi(x) \to \pm 1$ as $x \to \pm\infty$, that gives the shape of an unperturbed kink. The fluctuating stochastic field χ then has stationary statistical properties [32, 33].

The idea that collective transitions are favored in small domains but another mechanism is at work in larger domains has also emerged from numerical and analytical studies of minimum-action transition paths [4, 34, 6, 5]. Optimal transitions consist of nucleation events followed by propagation of domain walls, with behavior depending on L and on the (periodic, Dirichlet, or Neumann) boundary conditions.

In this work we locate, by simple numerical experiments, the crossover from collective transitions to nucleation-diffusion behavior. In the latter regime, the characteristic width of newly nucleated regions is $b = 8E_0$ and the complete passage time is proportional to $\exp(2E_0/\Theta)$. This value of b, obtained from the slope of the logarithm of the complete passage time versus Θ^{-1}, is consistent with numerical observations of the critical value of L at which the crossover from collective to nucleation-diffusion behavior is found, and with the long-held hypothesis that the separation between a kink and an antikink at nucleation is several times the kink width.

Systematic computational studies of the ϕ^4 SPDE, in the limit $L \to \infty$, require low temperatures in order to unambiguously identify kinks; they are computationally costly because the steady-state density of kinks decreases exponentially with temperature (necessitating very long chains) and the equilibration time increases exponentially with temperature (necessitating very long runs). Our aim in this paper is complementary to such studies. We focus on the dynamics of small- to medium-length chains in order to distinguish the regimes of collective transitions and of nucleation-diffusion. Although a first-principles theory of nucleation is the most difficult theoretical challenge, numerical studies that focus only on the measurement of the nucleation rate have the advantage of not needing to first attain a steady state.

In section 2 we define the complete passage time, our adopted measure of the mean time for the whole chain to make the transition from one well to another. The key observation is that there is a critical number of particles in the chain, below which the complete passage time increases exponentially and above which it increases more slowly. In the course of deriving theoretical expressions for the complete passage time, we are led to consider the short-range interaction between kinks and antikinks. This is examined in a second set of numerical experiments, differing from the first in the initial conditions, which now have a kink and an antikink relatively close together. Also, in section 2, we use numerical results to make an estimate of the typical separation, b, of a kink-antikink pair at nucleation. In section 3 we report on numerical experiments where the distribution of the "center-of-mass" of the chain is measured by means of long numerical runs and displayed in terms of an "effective potential" that has the Büttiker–Christen form of two wells separated by a long flat region.

2. Complete passage time.

Our first set of numerical experiments measures the time taken for all particles to make the transition from one minimum to the other. We choose the initial condition $\mathbf{\Phi}_0(i) = -1$, $i = 1, \ldots, N$, and denote

$$\mathbf{h} = \inf\left\{ t > 0 : \sum_{i=1}^{N} \mathbf{\Phi}_t(i) = N \right\}. \tag{2.1}$$

The complete passage time, τ, is defined as the mean of $\mathbf{h}$: $\tau = \mathbb{E}(\mathbf{h})$.

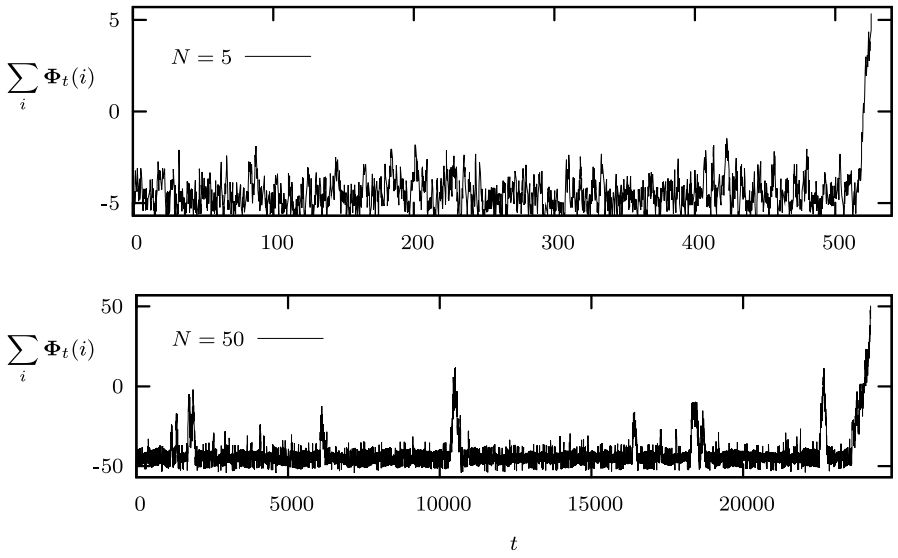


Figure 1. *Realizations of the coupled-particle system with $\beta = 6$ and $k = 1$, illustrating the two routes to a complete passage of the chain from one well to another. Upper timeseries: $N = 5$, collective regime; Lower timeseries: $N = 50$, nucleation-diffusion regime.*

In Figure 1, the sum $\sum_i \mathbf{\Phi}_t(i)$ is plotted as a function of time for two realizations. In the top part of the figure, $N = 5$, $\beta = 6$, and $k = 1$. The transition from one well to the other is collective; all particles in the chain make the transition close together. The second realization, shown in the lower part of the figure, has $N = 50$, $\beta = 6$, and $k = 1$. Several episodes are visible where the sum increases and then falls back to its starting level. These correspond to nucleation-diffusion episodes: a group of nearby particles makes the transition to the upper well, creating a region with two boundaries. The boundaries diffuse until the region either disappears or encompasses the whole chain.

Figure 2 shows τ versus N for $k = 1$ and four values of β. A data point typically corresponds to 10^3 realizations. We observe, in numerical experiments of this type, that the critical number of particles, at which the crossover from collective to nucleation-diffusion behavior is found, is independent of β if k is fixed. In Figure 3, three graphs of τ against N are displayed with different values of k; all have $\beta = 6$. The critical number of particles is seen to be an increasing function of k.

The existence of the SPDE limit implies that the passage time should approach a limit as $k \to \infty$ ($\Delta x \to 0$), with $L = N/\sqrt{k}$ fixed. In Figure 4, the mean passage time is plotted against L, using the scaled variables (1.3). Each data set has the same value of $\Theta = (\beta\sqrt{k})^{-1}$. The figure shows a much greater degree of universality than we expected: the data corresponding to large and small values of k are difficult to distinguish on the scale of the figure.

2.1. Analytical expressions. We develop our analytical approximations in the continuum limit using the scaled variables (1.3). Then the complete passage time, $\tau(\Theta, L)$, is a function of the temperature and the length of the domain.

Transitions are collective in sufficiently short chains because there is a single, well-defined, saddle point on the energy surface, with the whole chain at 0, separating two global minima,

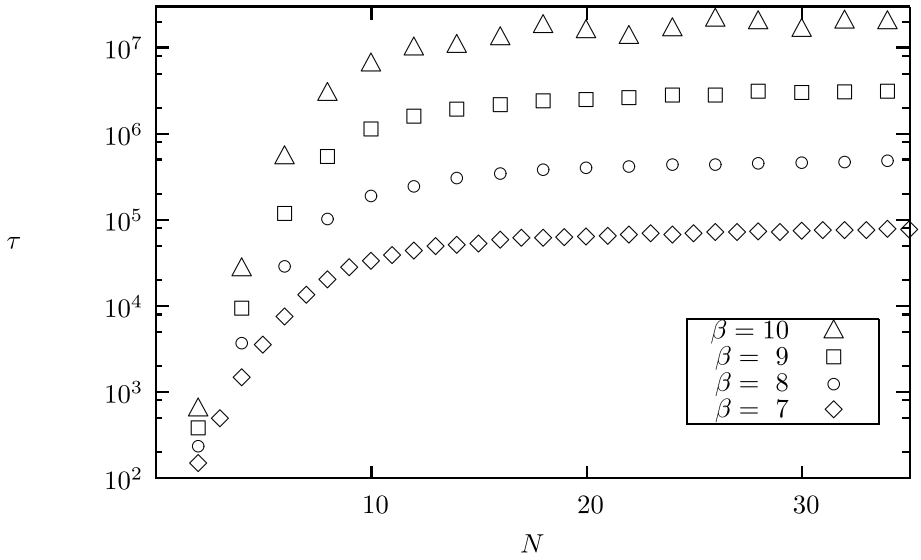


Figure 2. *Complete passage time versus number of particles for $k = 1$. The critical number of particles is independent of β. (Statistical errors are approximately the symbol sizes.)*

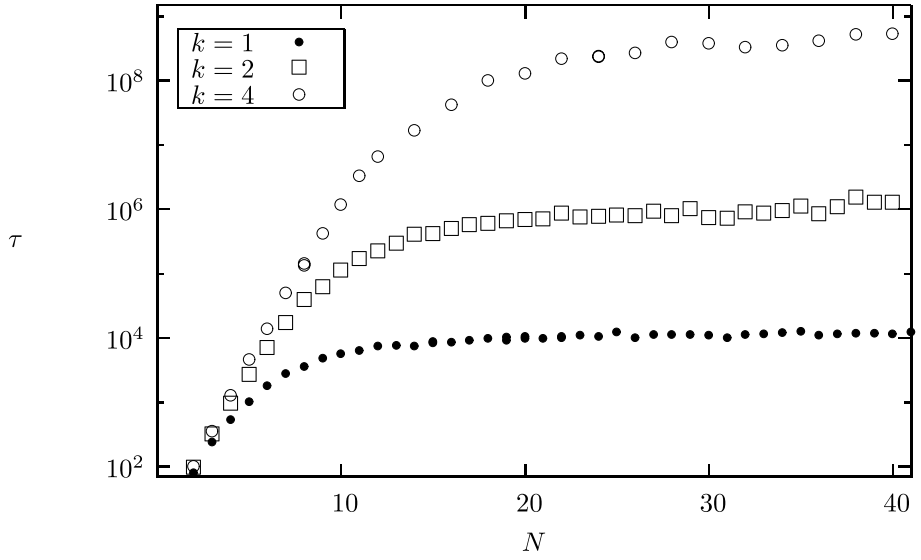


Figure 3. *Complete passage time versus number of particles for $\beta = 6$. The critical number of particles is an increasing function of k.*

with the whole chain at ± 1. The probability per unit time of a transition over the saddle point is calculated by constructing a nonequilibrium steady-state density which is the equilibrium density multiplied by a function of the distance along a line connecting the minima via the saddle [24, 25, 26]. As $\Theta \to 0$ and $L \to 0$, $\tau(\Theta, L) \to \tau_{\mathrm{c}}(\Theta, L)$, where

$$\tau_{\mathrm{c}}(\Theta, L) = \left(A_{\mathrm{c}}(L) + \mathcal{O}(\Theta^{-1})\right) \exp\left(\frac{1}{4}\frac{L}{\Theta}\right). \tag{2.2}$$

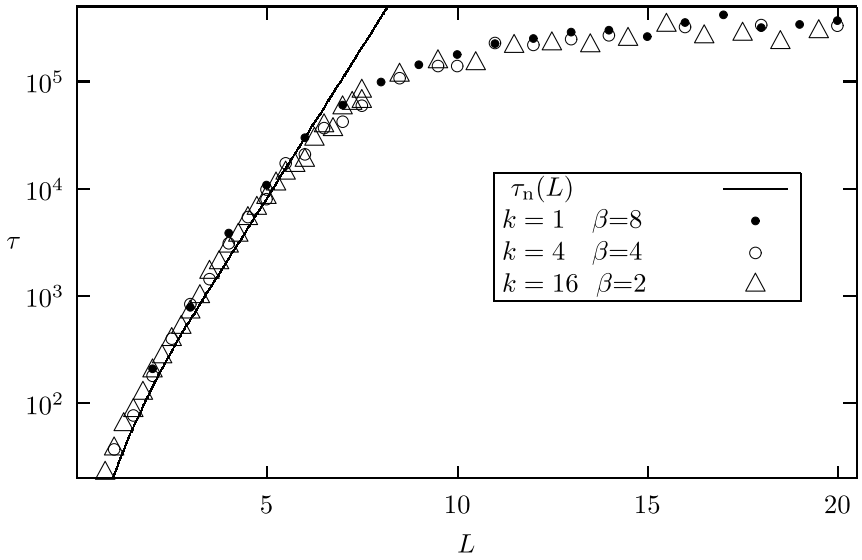


Figure 4. *Complete passage time versus L for $\Theta = \frac{1}{8}$. The values of $k = 1$, $k = 4$, and $k = 16$ correspond to $\Delta x = 1$, $\Delta x = 0.5$, and $\Delta x = 0.25$. The solid line is $4\pi \exp(-\frac{L}{\sqrt{2}}) \exp(\frac{1}{4}\frac{L}{\Theta})$, valid for collective transitions.*

The exponent in (2.2) is the difference between the energy at the saddle point and that at the minima; the prefactor $A_c(L)$ is sometimes referred to as a frequency factor. Its square is the ratio of the products of the eigenvalues corresponding to the N stable directions at a minimum and the $N-1$ stable directions at the saddle point. Interestingly, the ratio diverges at $L = 2\pi$ [34, 12, 35], corresponding to a breakdown of the collective-transition hypothesis. For our purposes, let us write

$$\tau(\Theta, L) = \left(A(L) + \mathcal{O}(\Theta^{-1})\right) \exp\left(\frac{f(L)}{4\Theta}\right), \tag{2.3}$$

with $A(L)$ and $f(L)$ to be determined, in the collective regime, the nucleation-diffusion regime, and the crossover region between the two. According to (2.2), as $L \to 0$, $f(L)/L \to 1$. We conjecture that, as L increases, $f(L) \to b$ and $A(L) \to A_\infty$ for constants b and A_∞ to be determined. This choice of functional form will be justified a posteriori by comparison with numerical data.

Next, consider the situation for $L > b$. Here, we make the hypothesis that nucleation events, equally likely to happen anywhere, always produce a region of width b. In other words, there is a constant probability per unit length and time, $\Gamma = (b\tau_n(\Theta, b))^{-1}$, of a nucleation event. Let the probability that a region of width x grows to encompass the whole domain be $q(x, L)$. Then

$$\tau(\Theta, L) = \frac{b}{L}\frac{1}{q(b, L)}\left(\tau_n(\Theta, b) + \tau_d(\Theta, b, L)\right), \quad L > b,$$

where $\tau_d(\Theta, b, L)$ is the mean time spent during a diffusion episode. It is easy to calculate an upper limit on $\tau_d(\Theta, b, L)$ by ignoring the short-range kink-antikink attraction, assuming that each diffuses with diffusivity Θ/E_0 [32, 36]:

$$\tau_{\mathrm{d}}(\Theta, b, L) < \frac{bLE_0}{4\Theta}.$$

The strict limit $L \to \infty$ requires a statistical theory of multiple kinks and antikinks and a different stopping criterion than (2.1); it is not the subject of this article. We concentrate on small- to medium-length domains that do not contain multiple kink-antikink pairs. In particular, we choose L sufficiently small that the probability of a second nucleation event during a diffusion episode is negligible, which is precisely the condition that $\tau_{\mathrm{n}}(\Theta, b) \gg \tau_{\mathrm{d}}(\Theta, b, L)$. The complete passage time then simplifies to

$$\tau(\Theta, L) \simeq \frac{b}{L} \frac{\tau_{\mathrm{n}}(\Theta, b)}{q(b, L)}, \quad b \ll L \ll L^*, \tag{2.4}$$

where $L^* = \frac{4\Theta}{bE_0} \exp(b/4\Theta)$. At the temperature ranges of interest here, L^* is many orders of magnitude greater than b.

2.2. Short-range kink-antikink interaction. To proceed in the calculation of the complete passage time, we need to consider the function $q(b, L)$. If we were to ignore the attraction between kinks and antikinks, their motion would be purely diffusive and the probability that a region of width x grows to encompass the entire domain of length L would be equal to the probability that a Brownian motion, started at $x > 0$, reaches L before 0. Thus, at first approximation, $q(x, L) \simeq q_0(x, L)$, where

$$q_0(x, L) = \frac{x}{L}. \tag{2.5}$$

In order to ascertain the effect on $q(x, L)$ of the short-term interaction between a kink and an antikink, we performed a set of numerical experiments with initial conditions corresponding to a domain of width x:

$$\mathbf{\Phi}_0(i) = \begin{cases} 1, & i \le n, \\ -1, & n < i \le N, \end{cases}$$

where $n = x/\sqrt{k}$. Realizations were stopped the first time that either $\sum_{i=1}^{N} \mathbf{\Phi}_t(i) = 0$ or $\sum_{i=1}^{N} \mathbf{\Phi}_t(i) = N$. The probability of the latter outcome is plotted as a function of L in Figure 5. Based on these numerical data, we devised the following approximation:

$$q(b, L) \simeq \frac{b - a}{L - 2a}. \tag{2.6}$$

The intuitive motivation for the form (2.6) is that two domain walls experience a strong attraction at separations less than a. The lines in Figure 5 are (2.5) and (2.6), using the least-squares best fit value of a, which yielded $a = 5.4 \pm 0.1$.

With (2.6), we have the following expression for the complete passage time in the nucleation-diffusion regime as $\Theta \to 0$:

$$\tau(\Theta, L) \to \frac{1 - 2a/L}{1 - a/b} A_\infty \exp\left(\frac{1}{4}\frac{b}{\Theta}\right). \tag{2.7}$$

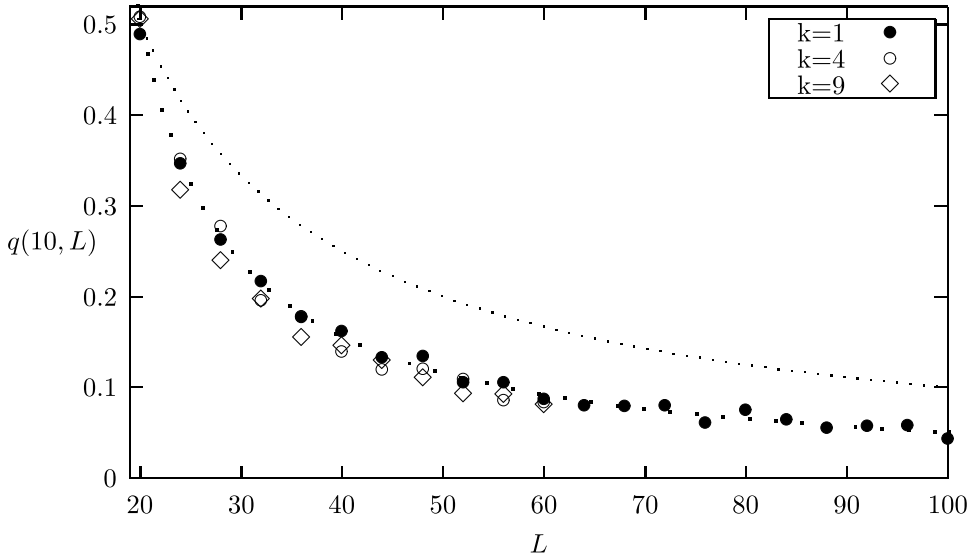


Figure 5. *The effect of the short-range kink-antikink attraction. The probability that an initial region of width $x = 10$ grows to encompass the entire domain of length L is plotted against L for $k = 1$, $k = 4$, and $k = 9$. In all cases $\Theta = 0.01$. The dotted lines are (2.5), which would hold if there were no attraction, and (2.6).*

2.3. Estimating the width *b*. By comparing the expression (2.7) with numerical data, we can estimate the value of the parameter b, the characteristic width of nucleation regions. In particular,

$$\tau(\Theta, L) \to \frac{A_\infty}{1 - a/b} \exp\left(\frac{1}{4}\frac{b}{\Theta}\right), \quad b \ll L \ll L^*. \tag{2.8}$$

That is, the logarithm of the complete passage time is a linear function of Θ^{-1}. The slope is proportional to b as L/b increases.

Figure 6 contains four plots summarizing numerical data with $L = 5$, $L = 10$, $L = 20$, and $L = 30$. The numerical runs were carried out with $k = 4$. For sufficiently small L, transitions are collective and (2.2) holds, so that the slope of a graph of $\ln \tau$ versus Θ^{-1} is proportional to L. The first set of numerical data, with $L = 5$, falls into this regime. As a function of L, the slope does not increase indefinitely but approaches a well-defined limit. A least-squares fit of the numerical data for large L gives $b = 7.4 \pm 0.1$. This value is consistent with the knee in our numerical curves of complete passage time versus L and with the following argument.

The complete passage time is proportional to $\exp(\frac{1}{4}\frac{L}{\Theta})$ for sufficiently small L and proportional to $\exp(\frac{2E_0}{\Theta})$ for long chains. With the ansatz $\tau \propto \exp(\frac{1}{4}\frac{f(L)}{\Theta})$,

- $f(L)/L \to 1$ as $L \to 0$, and
- $f(L) \to 8E_0$ when $b \ll L \ll L^*$,

thus identifying the characteristic mean width of newly nucleated regions as $b = 8E_0$. (In the Allen–Cahn scaling (1.6), the characteristic width of a newly nucleated region is $8E_0\epsilon$.)

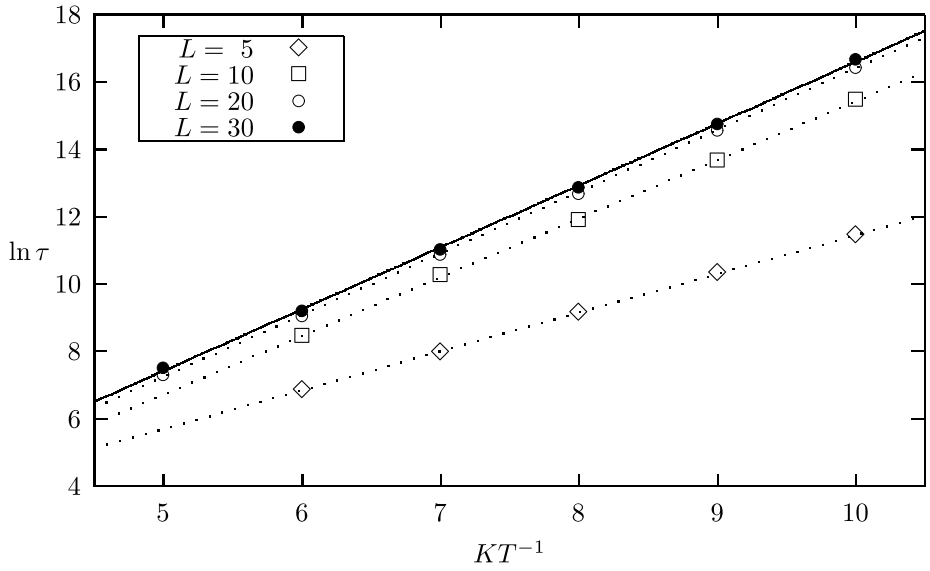


Figure 6. *Estimating the width, b, from numerical data. For sufficiently small L, the complete passage time of a domain length L is proportional to* $\exp(\frac{1}{4}\frac{L}{\Theta})$*; when* $L \gg b$*, it is proportional to* $\exp(\frac{1}{4}\frac{b}{\Theta})$*. The solid line, a least-squares fit to the* $L = 40$ *data, has slope corresponding to* $b = 7.4$*.*

3. Effective potential. Chains, whatever their length, eventually reach a stationary state where transitions are equally frequent in either direction. In the final set of numerical experiments reported here, we exploit the stationarity of the long-term dynamics by considering the "center-of-mass" process, defined as

$$\boldsymbol{\phi}_t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\Phi}_t(i), \tag{3.1}$$

and we construct its stationary density. In practice, the time for the system to explore all regions of state space thoroughly is several orders of magnitude greater than the complete passage time. The stationary density of the center-of-mass is denoted

$$\rho(\phi) = \lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}x}\mathcal{P}[\boldsymbol{\phi}_t < \phi]; \tag{3.2}$$

the effective potential, $V_{\mathrm{eff}}(\phi)$, is defined via $\rho(\phi) = \mathrm{e}^{-V_{\mathrm{eff}}(\phi)}$.

At parameter values such that transitions are collective, the effective potential is not dissimilar to $V(\phi)$. In the regime of nucleation-diffusion behavior, on the other hand, a long plateau appears in the effective potential, terminated at either end by a narrow well. See Figure 7. The wells are approximately quadratic near the minima, at $\boldsymbol{\phi} = \pm(1 - \frac{1}{2}\frac{\Theta}{E_0})$ [37, 38, 39]; the width of each well is $(\Theta/(3N))^{1/2}$ [37].

The effective potential takes a pleasingly universal form under the scaling (1.3). In Figure 8, we plot numerical results, all obtained on chains with $L = 100$ and $\Theta = \frac{1}{8}$. The three values of k correspond to $N = 100$, $N = 200$, and $N = 400$.

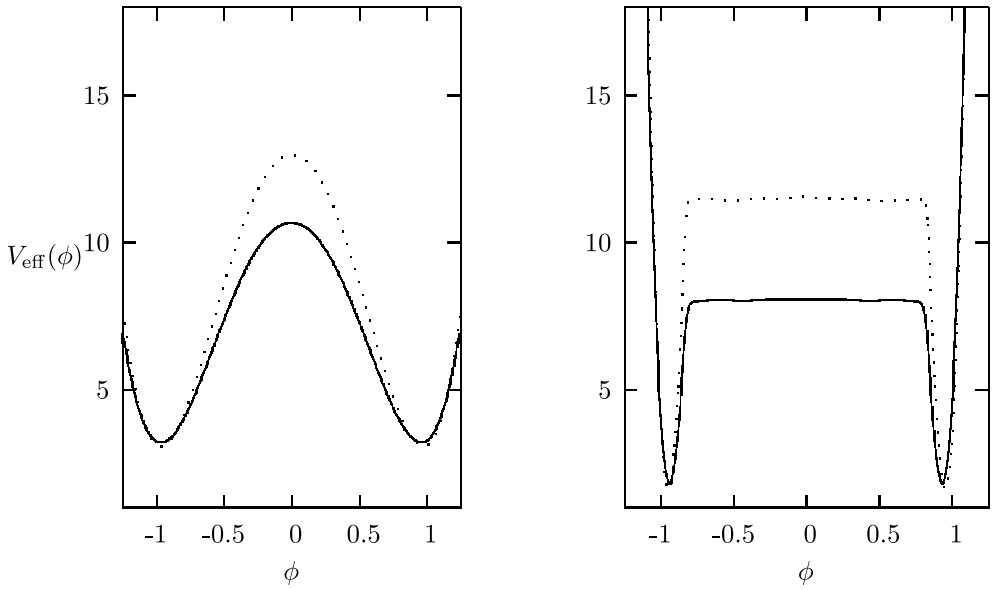


Figure 7. *Effective potential: Numerical results with $k = 1$. Solid lines: $\beta = 8$. Dotted lines: $\beta = 10$. On the left, $N = 5$ and transitions are collective. On the right, $N = 100$ and the chain exhibits nucleation-diffusion behavior.*

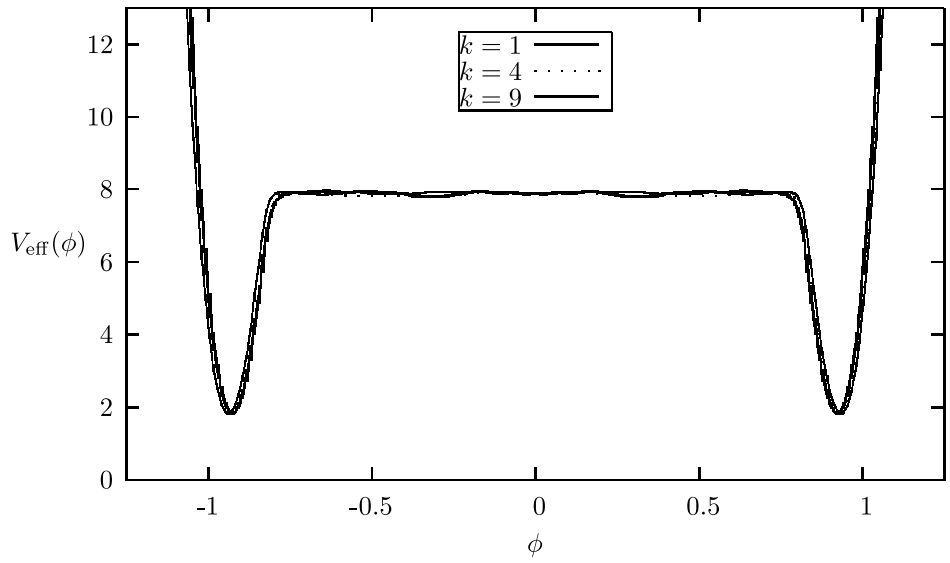


Figure 8. *The effective potential for three values of k, with $\Theta = \frac{1}{8}$ and $L = 100$. The wide central plateau corresponds to the freedom of a kink-antikink separation to wander over the majority of the length of the chain without noticeably affecting the total energy.*

The effective potentials, in Figure 8 and on the right in Figure 7, bear a striking resemblance to those of Büttiker and Christen [14, 20], introduced to provide an intuitive understanding of the dynamics of the nucleation-diffusion regime. A nucleation event is imagined as an exit from a potential well to a long flat region, which corresponds to the freedom of the

kink and antikink, once they have attained a certain separation, to wander over large parts of the chain without affecting the total energy of the configuration. The potential well at the opposite end of the flat region corresponds to the possibility that the nucleated region grows to encompass the whole domain; the kink and antikink meet again and annihilate. Our numerical experiments construct, in a systematic manner, a potential with these useful illustrative properties.

REFERENCES

[1] A. R. Bishop, J. A. Krumhansl, and J. R. Schrieffer, *Solitons in condensed matter: A paradigm*, Phys. D, 1 (1980), pp. 1–44.

[2] T. Dauxois and M. Peyrard, *Physics of Solitons*, Cambridge University Press, Cambridge, UK, 2006.

[3] S. Habib and G. Lythe, *Dynamics of kinks: Nucleation, diffusion and annihilation*, Phys. Rev. Lett., 84 (2000), pp. 1070–1073.

[4] W. G. Faris and G. Jona-Lasinio, *Large fluctuations for a nonlinear heat equation with noise*, J. Phys. A, 15 (1982), pp. 3025–3055.

[5] H. C. Fogedby, J. Hertz, and A. Svane, *Soliton-dynamical approach to a noisy Ginzburg-Landau model*, Europhys. Lett., 62 (2003), pp. 795–800.

[6] W. E, W. Ren, and E. Vanden-Eijnden, *Minimum action for the study of rare events*, Comm. Pure Appl. Math., 57 (2004), pp. 637–656.

[7] C. R. Doering, *Nonlinear parabolic stochastic differential equations with additive colored noise on* $\mathbb{R}^d \times \mathbb{R}_+$*: A regulated stochastic quantization*, Comm. Math. Phys., 109 (1987), pp. 537–561.

[8] T. Funaki, *Random motion of strings and related stochastic evolution equations*, Nagoya Math. J., 89 (1983), pp. 129–193.

[9] G. Lythe and S. Habib, *Kink stochastics*, Computing in Science and Engineering, 8 (2006), pp. 10–15.

[10] M. A. Katsoulakis, G. T. Kossioris, and O. Lakkis, *Noise regularization and computations for the* 1*-dimensional stochastic Allen–Cahn problem*, Interfaces Free Bound., 9 (2007), pp. 1–30.

[11] T. Shardlow, *Stochastic perturbations of the Allen-Cahn equation*, Electron. J. Differential Equations, 47 (2000).

[12] R. V. Kohn, F. Otto, M. G. Reznikoff, and E. Vanden-Eijnden, *Action minimization and sharp-interface limits for the stochastic Allen-Cahn equation*, Comm. Pure Appl. Math., 59 (2006), pp. 1–46.

[13] J. Lothe and J. P. Hirth, *Dislocation dynamics at low temperatures*, Phys. Rev., 115 (1959), pp. 543–550.

[14] M. Büttiker and T. Christen, *Nucleation of weakly driven kinks*, Phys. Rev. Lett., 75 (1995), pp. 1895–1898.

[15] A. Seeger and P. Schiller, *Kinks in dislocation lines and their effects on the internal friction in crystals*, in Physical Acoustics: Principles and Methods, W. P. Mason, ed., Academic Press, New York, 1966, pp. 361–495.

[16] D. J. Scalapino, M. Sears, and R. A. Ferrell, *Statistical mechanics of one-dimensional Ginzburg-Landau fields*, Phys. Rev. B, 6 (1972), pp. 3409–3416.

[17] F. J. Alexander and S. Habib, *Statistical mechanics of kinks in* 1+1 *dimensions*, Phys. Rev. Lett., 71 (1993), pp. 955–958.

[18] A. M. Stuart, J. Voss, and P. Wiberg, *Conditional path sampling of SDEs and the Langevin MCMC method*, Comm. Math. Sci., 2 (2004), pp. 585–697.

[19] M. G. Reznikoff and E. Vanden-Eijnden, *Invariant measures of stochastic partial differential equations and conditioned diffusions*, C. R. Math. Acad. Sci. Paris, 340 (2005), pp. 305–308.

[20] M. Büttiker and T. Christen, *Diffusion controlled initial recombination*, Phys. Rev. E, 58 (1998), pp. 1533–1548.

[21] S. Habib, K. Lindenberg, G. Lythe, and C. Molina-París, *Diffusion-limited reaction in one dimension: Paired and unpaired nucleation*, J. Chem. Phys., 115 (2001), pp. 73–89.

[22] T. O. Masser and D. ben Avraham, *Method of intervals for the study of diffusion-limited annihilation*, Phys. Rev. E, 63 (2001), 006108.
[23] G. Lythe, *Diffusion-limited reaction in one dimension*, Phys. D, 222 (2006), pp. 159–163.
[24] G. H. Vineyard, *Frequency factors and isotope effects in solid state rate processes*, J. Physics and Chemistry of Solids, 3 (1957), pp. 121–127.
[25] R. Landauer and J. A. Swanson, *Frequency factors in the thermally activated process*, Phys. Rev., 121 (1961), pp. 1668–1674.
[26] M. Büttiker and R. Landauer, *Nucleation theory of overdamped soliton motion*, Phys. Rev. Lett., 43 (1979), pp. 1453–1456.
[27] M. Büttiker and R. Landauer, *Nucleation theory of overdamped soliton motion*, Phys. Rev. A, 23 (1981), pp. 1397–1410.
[28] A. J. Graham and W. C. Kerr, *Solution of Kramers' problem for a moderately to heavily damped elastic string*, Phys. Rev. E, 65 (2001), 016106.
[29] W. C. Kerr and A. J. Graham, *Nucleation rate of critical droplets on an elastic string in a ϕ^6 potential*, Phys. Rev. E, 70 (2004), 066103.
[30] C. Sagui and M. Grant, *Theory of nucleation and growth during phase separation*, Phys. Rev. E, 59 (1999), pp. 4175–4187.
[31] M. Castro, *Phase-field approach to heterogeneous nucleation*, Phys. Rev. B, 67 (2003), 035412.
[32] D. J. Kaup, *Thermal corrections to overdamped soliton motion*, Phys. Rev. B, 27 (1983), pp. 6787–6795.
[33] G. Lythe and F. Mertens, *Brownian motion of ϕ^4 kinks*, in preparation.
[34] R. S. Maier and D. L. Stein, *Droplet nucleation and domain wall motion in a bounded interval*, Phys. Rev. Lett., 87 (2001), 270601.
[35] M. Castro and G. Lythe, *Nucleation: From one particle to the continuum*, in preparation.
[36] G. Lythe and F. Mertens, *Rice's ansatz for overdamped ϕ^4 kinks at finite temperature*, Phys. Rev. E, 67 (2003), 027601.
[37] L. Bettencourt, S. Habib, and G. Lythe, *Controlling one-dimensional Langevin dynamics on the lattice*, Phys. Rev. D, 60 (1999), 105039.
[38] G. Lythe and S. Habib, *Stochastic PDEs: Convergence to the continuum?*, Comput. Phys. Comm., 142 (2001), pp. 29–35.
[39] F. J. Alexander, S. Habib, and A. Kovner, *Statistical mechanics of kinks in (1+1)-dimensions: Numerical simulations and double Gaussian approximation*, Phys. Rev. E, 48 (1993), pp. 4284–4296.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 1, pp. 220–248
© 2008 Society for Industrial and Applied Mathematics

Hopf Bifurcation in Coupled Cell Networks with Interior Symmetries*

Fernando Antoneli†, Ana Paula S. Dias‡†, and Rui C. Paiva‡†

Abstract. We consider an important class of nonsymmetric networks that lies between the class of general networks and the class of symmetric networks, where group theoretic methods still apply—namely, networks admitting "interior symmetries." The main result of this paper is the full analogue of the equivariant Hopf theorem for networks with symmetries. We extend the result of Golubitsky, Pivato, and Stewart [*Dyn. Syst.*, 19 (2004), pp. 389–407] to obtain states whose linearizations on certain subsets of cells, near bifurcation, are superpositions of synchronous states with states having spatio-temporal symmetries.

Key words. interior symmetries, coupled cell networks, bifurcation theory

AMS subject classifications. 34C15, 34C25, 37G40

DOI. 10.1137/070684604

1. Introduction.

Recently, a new framework for the dynamics of networks has been proposed, with particular attention to patterns of synchrony and associated bifurcations. See Stewart, Golubitsky, and Pivato [11], Golubitsky, Pivato, and Stewart [4], Golubitsky, Nicol, and Stewart [3], and Golubitsky, Stewart, and Török [9]. Here, a network is represented by a directed graph whose nodes and edges are classified according to associated labels or "types." The nodes (or "cells") of a network $\mathcal{G}$ represent dynamical systems, and the edges ("arrows") represent couplings. Cells with the same label have "identical" internal dynamics; arrows with the same label correspond to identical couplings. The "input set" of a cell is the set of edges directed to that cell. Label-preserving bijections between input sets of cells are called "input isomorphisms," and they capture the "local" symmetries of the network. The set of all these local symmetries has the structure of a groupoid. (A groupoid is an algebraic structure similar to a group, except that products of elements may not always be defined.)

Coupled cell systems are dynamical systems compatible with the architecture or topology of a directed graph representing the network. Formally, they are defined in the following way. Each cell c is equipped with a phase space P_c, and the total phase space of the network is the Cartesian product $P = \prod_c P_c$. A vector field f is called "admissible" if its component f_c for cell c depends only on variables associated with the input set of c (domain condition) and

*Received by the editors March 7, 2007; accepted for publication (in revised form) by B. Sandstede August 8, 2007; published electronically March 28, 2008.
http://www.siam.org/journals/siads/7-1/68460.html

†Centro de Matemática da Universidade do Porto, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal (antoneli@fc.up.pt). CMUP is supported by FCT through POCTI and POSI of Quadro Comunitário de Apoio III (2000-2006) with FEDER and national fundings. The research of the first author was supported by Centro de Matemática da Universidade do Porto; he thanks the University of Porto and the Isaac Newton Institute, Cambridge, for hospitality and additional financial support.

‡Departamento de Matemática Pura, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal (apdias@fc.up.pt, rui.paiva@fc.up.pt). The second author thanks the Isaac Newton Institute, Cambridge, for hospitality and additional financial support.

if its components for cells c, d that have isomorphic input sets are identical up to a suitable permutation of the relevant variables (pull-back condition).

In the study of network dynamics there is an important class of networks, namely, networks that possess a group of symmetries. In this context there is a group of permutations of the cells (and arrows) that preserves the network structure (including cell-types and arrow-types), and its action on P is by permutation of cell coordinates. Moreover, the coupled cell systems (ODEs) are of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x),$$

where the vector field f is smooth (C^∞) and satisfies

$$f(\gamma x) = \gamma f(x) \qquad \forall\, x \in P,\, \gamma \in \Gamma.$$

That is, f is "equivariant" under the action of the group Γ on phase space P.

The theory of equivariant dynamical systems (see Golubitsky and coworkers [6, 8]) can be applied to such dynamical systems. In this theory, a central role is played by the "fixed-point spaces" of subgroups $\Sigma \subseteq \Gamma$, defined by

$$\mathrm{Fix}(\Sigma) = \{x \in P : \sigma x = x \ \ \forall\, \sigma \in \Sigma\}.$$

Fixed-point spaces have the important property of flow invariance: they are invariant under *every* smooth equivariant vector field f, so that

$$f(\mathrm{Fix}(\Sigma)) \subseteq \mathrm{Fix}(\Sigma).$$

See [6, Lemma XIII 2.1] or [8, Theorem 1.17] for the simple proof and the implications for symmetry-breaking. In this context, there are two main local bifurcation theorems. The *equivariant branching lemma* (see Golubitsky, Stewart, and Schaeffer [8, Theorem XIII 3.3]) proves the existence of certain branches of symmetry-breaking steady states; the *equivariant Hopf theorem* (see [8, Theorem XVI 4.1]) proves the existence of certain branches of spatio-temporal symmetry-breaking time-periodic states.

In between the class of general networks and the class of symmetric networks lies an interesting class of nonsymmetric networks, where group theoretic methods still apply, namely, networks admitting "interior symmetries." In this case there is a group of permutations of a subset $\mathcal{S}$ of the cells (and edges directed to $\mathcal{S}$) that partially preserves the network structure (including cell-types and edge-types), and its action on P is by permutation of cell coordinates. In other words, the cells in $\mathcal{S}$ together with all the edges directed to them form a subnetwork which possesses a nontrivial group of symmetry $\Sigma_{\mathcal{S}}$. For example, network $\mathcal{G}_1$ (Figure 1 (left)) has exact $\mathbf{S}_3$-symmetry, whereas network $\mathcal{G}_2$ (Figure 1 (right)) has $\mathbf{S}_3$-interior symmetry. This notion was introduced and investigated by Golubitsky, Pivato, and Stewart [4]. The presence of interior symmetries places some restrictions on the structure of the network.

The local bifurcations from a synchronous equilibrium can be classified into two types: "synchrony-breaking" bifurcations and "synchrony-preserving" bifurcations. The synchrony-breaking bifurcations occur when a synchronous state loses stability and bifurcates to a state with less synchrony. Such bifurcations can be considered to be generalizations of symmetry-breaking bifurcations in symmetric coupled cell systems. Golubitsky, Pivato, and Stewart [4]

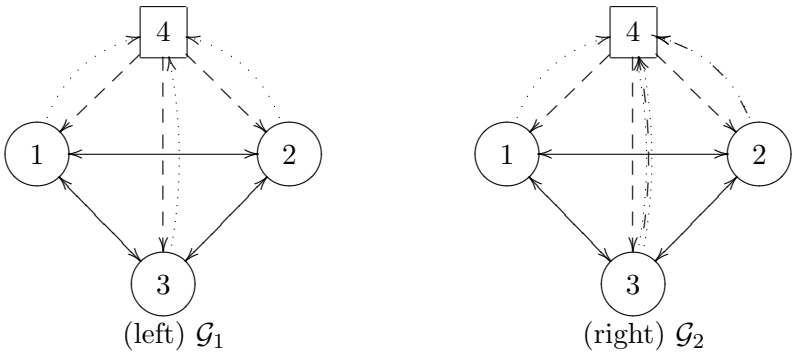


Figure 1. *(Left) Network $\mathcal{G}_1$ with exact $\mathbf{S}_3$-symmetry. (Right) Network $\mathcal{G}_2$ with $\mathbf{S}_3$-interior symmetry.*

provided analogues of the equivariant branching lemma and the equivariant Hopf theorem for coupled cell systems with interior symmetries. The analogue of the equivariant branching lemma is a natural generalization of the symmetric case, but the analogue of the equivariant Hopf theorem has novel and rather restrictive features. In particular, instead of proving the existence of states with certain spatio-temporal symmetries, they prove the existence of states whose linearizations on certain subsets of cells, near bifurcation, are superpositions of synchronous states with states having spatial symmetries.

The main result of this paper is the full analogue of the equivariant Hopf theorem for networks with symmetries (Theorem 4.8). We extend the result of Golubitsky, Pivato, and Stewart [4] to obtain states whose linearizations on certain subsets of cells, near bifurcation, are superpositions of synchronous states with states having *spatio-temporal symmetries*, that is, corresponding to "interiorly" $\mathbf{C}$-axial subgroups of $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$. This new version of the Hopf theorem with interior symmetries includes the previous version as a special case and is in complete analogy with the equivariant Hopf theorem (see Theorem 4.8). Our proof uses a modification of the Lyapunov–Schmidt reduction to arrive at a situation where the proof of the standard Hopf bifurcation theorem can be applied. This completes the program of generalizing the two main results from equivariant bifurcation theory to the class of networks with interior symmetries.

Structure of the paper. Section 2 recalls the formal definition of a coupled cell network and the associated dynamical systems, and states some basic features, including the concept of a balanced equivalence relation (coloring). We also discuss the symmetry group of a network. Section 3 recalls the definition of interior symmetry given by Golubitsky, Pivato, and Stewart [4] and gives an equivalent condition, in terms of symmetries of a subnetwork, which in some cases (no multiple edges and no self-connections) amounts to finding the symmetries of the subnetwork. We also analyze the structure of these networks and discuss some features of the admissible vector fields associated with such a class of networks. Section 4 recalls the notion of synchrony-breaking bifurcation in coupled cell networks. Then we specialize to networks with interior symmetries where group theoretic concepts play a significant role, focusing on the important case of codimension-one synchrony-breaking bifurcations. The main part of this section gives the statement and proof of the interior symmetry-breaking Hopf bifurcation theorem (Theorem 4.8) for networks with interior symmetries. We illustrate all the concepts

and results by a running example of the simplest network with $\mathbf{S}_3$-interior symmetry and the closely related network with exact $\mathbf{S}_3$-symmetry (see Figure 1). Finally, we present a numerical simulation of the states provided by Theorem 4.8 in the case of our running example.

2. Network formalism. First, we recall the formal definition of a coupled cell network and the associated dynamical systems. For a survey, overview, and examples, see [7]. The initial definition of coupled cell network [11] was modified in [9] to permit multiple arrows and self-connections, which turns out to have major advantages. More recently, Stewart [10] extended the formalism introduced in [9] to include a large class of infinite networks—the so-called *networks of finite type*.

2.1. Coupled cell networks. In this paper we consider *finite networks* and so employ the "finite multiarrow" formalism for consistency with the existing literature.

Definition 2.1 (see [9]). *A* coupled cell network $\mathcal{G}$ *comprises the following:*

(a) *A finite set* $\mathcal{C}$ *of* nodes *or* cells.

(b) *An equivalence relation* $\sim_C$ *on cells in* $\mathcal{C}$*, called* cell-equivalence. *The* type *or* cell label *of cell* c *is its* $\sim_C$*-equivalence class.*

(c) *A finite set* $\mathcal{E}$ *of* edges *or* arrows.

(d) *An equivalence relation* $\sim_E$ *on edges in* $\mathcal{E}$*, called* edge-equivalence *or* arrow-equivalence. *The* type *or* coupling label *of edge* e *is its* $\sim_E$*-equivalence class.*

(e) *Two maps* $\mathcal{H} : \mathcal{E} \to \mathcal{C}$ *and* $\mathcal{T} : \mathcal{E} \to \mathcal{C}$*. For* $e \in \mathcal{E}$ *we call* $\mathcal{H}(e)$ *the* head *of* e *and* $\mathcal{T}(e)$ *the* tail *of* e*.*

We also require a consistency condition:

(f) *Equivalent arrows have equivalent tails and heads:*

$$\mathcal{H}(e_1) \sim_C \mathcal{H}(e_2), \qquad \mathcal{T}(e_1) \sim_C \mathcal{T}(e_2)$$

for all $e_1, e_2 \in \mathcal{E}$ *with* $e_1 \sim_E e_2$*.*

Example 2.2. We can represent abstract networks by labeled directed graphs. Figure 1 shows two examples. Here the node labels, drawn as the three circles and the square, indicate the cells; the symbols show that cells $1, 2, 3$ have the same type, whereas cell 4 is different, in both cases. In the network $\mathcal{G}_1$ there are three types of edge label, whereas in the network $\mathcal{G}_2$ there are five types of edge label, drawn as different styles of arrows. The tail and head of each edge is, respectively, indicated by the absence or presence of a tip on one end of the arrow. When an arrow between cells c and d is drawn with tips at both ends, then it represents two arrows of the same type with opposite orientation between cells c and d.

2.2. Input sets and the symmetry groupoid. Associated with each cell $c \in \mathcal{C}$ is a canonical set of edges, namely, those that represent couplings into cell c, as described next.

Definition 2.3 (see [9]). *If* $c \in \mathcal{C}$*, then the* input set *of* c *is the finite set of edges directed to* c*,*

$$I(c) = \{e \in \mathcal{E} : \mathcal{H}(e) = c\}. \tag{2.1}$$

Definition 2.4 (see [9]). *The relation* $\sim_I$ *of* input equivalence *on* $\mathcal{C}$ *is defined by* $c \sim_I d$ *if and only if there exists a bijection*

$$\beta : I(c) \to I(d) \tag{2.2}$$

such that for every $i \in I(c)$,

$$i \sim_E \beta(i). \tag{2.3}$$

Any such bijection β *is called an* input isomorphism *from cell* c *to cell* d. *The set* $B(c,d)$ *denotes the collection of all input isomorphisms from cell* c *to cell* d. *The union*

$$\mathcal{B}_{\mathcal{G}} = \bigcup_{c,d\in\mathcal{C}} B(c,d) \tag{2.4}$$

is the symmetry groupoid *of the network* $\mathcal{G}$. *A coupled cell network is* homogeneous *if all input sets are isomorphic.*

The groupoid operation on $\mathcal{B}_{\mathcal{G}}$ is composition of maps, and in general the composition $\beta\alpha$ is defined only when $\alpha \in B(a,b)$ and $\beta \in B(b,c)$ for cells a, b, c. This is why $\mathcal{B}_{\mathcal{G}}$ need not be a group.

Example 2.5. In our running examples, shown in Figure 1, it is easy to see that both networks have only two input isomorphism classes of cells: $\{1,2,3\}$ and $\{4\}$. The input sets of cells $1, 2, 3$ are isomorphic, since each one of them contains three edges, two of them drawn as a solid arrow with a circle in the tail and one of them drawn as a dashed arrow with a square in the tail.

2.3. Admissible vector fields. We now explain how to interpret such diagrams as Figure 1 as being representative of a class of vector fields.

For each cell in $\mathcal{C}$ choose a *cell phase space* P_c, which we assume to be a nonzero finite-dimensional real vector space. We require

$$c \sim_C d \quad \Rightarrow \quad P_c = P_d,$$

and in this case we employ the same coordinate systems on P_c and P_d. The *total phase space* is then

$$P = \prod_{c\in\mathcal{C}} P_c$$

with a cell-based coordinate system

$$x = (x_c)_{c\in\mathcal{C}}.$$

If $\mathcal{D} \subseteq \mathcal{C}$ is any finite set of cells, then we write

$$P_{\mathcal{D}} = \prod_{d\in\mathcal{D}} P_d$$

and

$$x_{\mathcal{D}} = (x_{c_1}, \ldots, x_{c_\ell}),$$

where $x_c \in P_c$.

For any $\beta \in B(c,d)$ we define the *pull-back map*

$$\beta^* : P_{\mathcal{T}(I(d))} \to P_{\mathcal{T}(I(c))}$$

by

$$(\beta^* z)_{\mathcal{T}(i)} = z_{\mathcal{T}(\beta(i))} \tag{2.5}$$

for all $i \in I(c)$ and $z \in P_{\mathcal{T}(I(d))}$.

We use pull-back maps to relate different components of a vector field associated with a given coupled cell network. Specifically, the class of vector fields that are encoded by a coupled cell network is given by the following definition.

Definition 2.6 (see [9]). *A map $f : P \to P$ is* $\mathcal{G}$-admissible *if the following hold:*

(a) Domain condition: *For all $c \in \mathcal{C}$ the component $f_c(x)$ depends only on the internal phase space variables x_c and the coupling phase space variables $x_{\mathcal{T}(I(c))}$; that is, there exists $\hat{f}_c : P_c \times P_{\mathcal{T}(I(c))} \to P_c$ such that*

$$f_c(x) = \hat{f}_c(x_c, x_{\mathcal{T}(I(c))}). \tag{2.6}$$

(b) Pull-back condition: *For all $c, d \in \mathcal{C}$ and $\beta \in B(c, d)$*

$$\hat{f}_d(x_d, x_{\mathcal{T}(I(d))}) = \hat{f}_c(x_d, \beta^* x_{\mathcal{T}(I(d))}) \tag{2.7}$$

for all $x \in P$.

Example 2.7. For the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1 the cell phase spaces P_1, P_2, and P_3 are identical and equal to $\mathbf{R}^k$, whereas $P_4 = \mathbf{R}^l$. The general form of the admissible vector fields (ODEs) encoded by the network $\mathcal{G}_1$ is

$$\begin{aligned} \dot{x_1} &= f(x_1, \overline{x_2, x_3}, x_4), \\ \dot{x_2} &= f(x_2, \overline{x_3, x_1}, x_4), \\ \dot{x_3} &= f(x_3, \overline{x_1, x_2}, x_4), \\ \dot{x_4} &= g(x_4, \overline{x_1, x_2, x_3}), \end{aligned} \tag{2.8}$$

where $x_i \in \mathbf{R}^k$ $(i = 1, 2, 3)$, $x_4 \in \mathbf{R}^l$, $f : \mathbf{R}^{3k} \times \mathbf{R}^l \to \mathbf{R}^k$ is a smooth map invariant under permutation of the second and third arguments, and $g : \mathbf{R}^{3k} \times \mathbf{R}^l \to \mathbf{R}^l$ is a smooth map invariant under any permutation of the last three arguments. The general form of the admissible vector fields (ODEs) associated with the network $\mathcal{G}_2$ is

$$\begin{aligned} \dot{x_1} &= f(x_1, \overline{x_2, x_3}, x_4), \\ \dot{x_2} &= f(x_2, \overline{x_3, x_1}, x_4), \\ \dot{x_3} &= f(x_3, \overline{x_1, x_2}, x_4), \\ \dot{x_4} &= g(x_4, x_1, x_2, x_3), \end{aligned} \tag{2.9}$$

where $x_i \in \mathbf{R}^k$ $(i = 1, 2, 3)$, $x_4 \in \mathbf{R}^l$, $f : \mathbf{R}^{3k} \times \mathbf{R}^l \to \mathbf{R}^l$ is a smooth map invariant under permutation of the second and third arguments, and $g : \mathbf{R}^{3k} \times \mathbf{R}^l \to \mathbf{R}^l$ is a general smooth map.

2.4. Balanced equivalence relations. An equivalence relation $\bowtie$ on $\mathcal{C}$ determines a unique partition of $\mathcal{C}$ into $\bowtie$-equivalence classes, which can be interpreted as a coloring of $\mathcal{C}$ in which $\bowtie$-equivalent cells receive the same color. Conversely, any partition (coloring) determines a unique equivalence relation. The corresponding *polydiagonal* is

$$\triangle_{\bowtie} = \{x \in P : c \bowtie d \Rightarrow x_c = x_d\}. \tag{2.10}$$

A subspace V of P is called *admissibly flow-invariant* if $f(V) \subset V$ for all admissible vector fields f on P.

Definition 2.8 (see [9]). *An equivalence relation $\bowtie$ on $\mathcal{C}$ is* balanced *if for every $c, d \in \mathcal{C}$ with $c \bowtie d$ there exists $\beta \in B(c,d)$ such that $\mathcal{T}(i) \bowtie \mathcal{T}(\beta(i))$ for all $i \in I(c)$. The associated coloring is called a* balanced coloring. *In particular, $B(c,d) \neq \emptyset$ implies $c \sim_I d$. Hence, balanced equivalence relations refine input equivalence.*

A crucial property of balanced equivalence relations is that they define admissibly flow-invariant subspaces, and conversely the following holds.

Theorem 2.9 (Stewart, Golubitsky, and Pivato [11]). *Let $\bowtie$ be an equivalence relation on a coupled cell network. Then $\triangle_{\bowtie}$ is admissibly flow-invariant if and only if $\bowtie$ is balanced.*

The proof of the above result for finite networks is given in [9, 11] and for networks of finite type in [10]. The dynamical implication of such flow-invariance is that $\bowtie$ determines a *robust pattern of synchrony*: there exist trajectories $x(t)$ of the ODE such that

$$c \bowtie d \;\Rightarrow\; x_c(t) = x_d(t) \quad \forall\, t \in \mathbf{R}.$$

Such trajectories arise when initial conditions $x(0)$ lie in $\triangle_{\bowtie}$. Then the entire trajectory, for all positive and negative time, lies in $\triangle_{\bowtie}$ and is a trajectory of the restriction $f|_{\triangle_{\bowtie}}$. The associated dynamics can be steady-state, periodic, even chaotic, depending on f and its restriction to $\triangle_{\bowtie}$. An example of synchronized chaos generated by this mechanism can be found in [7].

Since there is always a canonical balanced relation $\sim_I$ on every network, let $\triangle_I$ denote the polydiagonal subspace of P associated with the input equivalence relation $\sim_I$, that is,

$$\triangle_I = \{x \in P : c \sim_I d \Rightarrow x_c = x_d\}.$$

Then $\triangle_I$ is a flow invariant subspace. Solutions of admissible vector fields contained in $\triangle_I$ represent the states of highest degree of synchrony allowed by the network.

Remark 2.10. Whenever self-connections or multiple arrows do not occur it will be convenient to revert to the formalism of [11], but now considered as a specialization of the multiarrow formalism. Since no two distinct arrows have the same head and tail, we can identify an arrow e with the pair of cells $(\mathcal{T}(e), \mathcal{H}(e))$. Now the set $\mathcal{E}$ of arrows identifies with a subset of $\mathcal{C} \times \mathcal{C} \setminus \{(c,c) : c \in \mathcal{C}\}$. Similarly the input set $I(c)$ can be identified with the set of all tail cells of arrows e that have c as a head cell.

Example 2.11. We continue with our running examples, the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1. There is an equivalence relation $\bowtie$ for which $1 \bowtie 2$; its equivalence classes are $\{1,2\}$, $\{3\}$, and $\{4\}$. The corresponding polydiagonal is

$$\triangle_{\bowtie} = \{x \in P : x_1 = x_2\} = \{(x,x,y,z)\}.$$

On this subspace the differential equations become

$$\begin{aligned} \dot{x} &= f(x, \overline{x, y}, z), \\ \dot{x} &= f(x, \overline{y, x}, z), \\ \dot{y} &= f(y, \overline{x, x}, z), \\ \dot{z} &= g(z, x, x, y). \end{aligned} \tag{2.11}$$

Since the first two equations are identical (recall that the bar over x, y means that they can be interchanged), $\triangle_{\bowtie}$ is invariant under all admissible vector fields. The relation $\bowtie$ is balanced. The only condition to verify is that cells 1 and 2, which are $\bowtie$-equivalent but distinct, have input sets that are isomorphic by an isomorphism that preserves $\bowtie$-equivalence classes for both networks. In both networks the input sets are

$$I(1) = \{(2,1),(3,1),(4,1)\} \qquad \text{and} \qquad I(2) = \{(1,2),(3,2),(4,2)\},$$

where (c,d) denotes an arrow with tail c and head d (see Remark 2.10). The bijection $\beta : I(1) \to I(2)$ with $\beta((2,1)) = (1,2)$, $\beta((3,1)) = (3,2)$, and $\beta((4,1)) = (4,2)$ is an input isomorphism that preserves $\bowtie$-equivalence classes since $1 \bowtie 2$, $3 \bowtie 3$, and $4 \bowtie 4$. That is, $\bowtie$ is a balanced relation as claimed. There are two other balanced equivalence relations (different from $\sim_I$) on the networks $\mathcal{G}_1$ and $\mathcal{G}_2$. In one of them the equivalence classes are $\{2,3\}$, $\{1\}$, and $\{4\}$. In the other the equivalence classes are $\{1,3\}$, $\{2\}$, and $\{4\}$.

2.5. Symmetry groups of networks. We now consider symmetries of networks in the group theoretic ("global") sense.

Definition 2.12 (see [1]). *Let $\mathcal{G}$ be a network. A* symmetry *of $\mathcal{G}$ consists of a pair of bijections $\gamma_C : \mathcal{C} \to \mathcal{C}$ and $\gamma_E : \mathcal{E} \to \mathcal{E}$, where γ_C preserves input equivalence and γ_E preserves edge equivalence; that is, for all $c \in \mathcal{C}$ and $e \in \mathcal{E}$,*

$$\gamma_C(c) \sim_I c \qquad \text{and} \qquad \gamma_E(e) \sim_E e. \tag{2.12}$$

In addition, the two bijections must satisfy the consistency conditions

$$\gamma_C(\mathcal{H}(e)) = \mathcal{H}(\gamma_E(e)) \qquad \text{and} \qquad \gamma_C(\mathcal{T}(e)) = \mathcal{T}(\gamma_E(e)) \tag{2.13}$$

for all $e \in \mathcal{E}$. The set of all $\gamma = (\gamma_C, \gamma_E)$ forms a finite group $\mathrm{Aut}(\mathcal{G})$ *called the* symmetry group *of the network of $\mathcal{G}$.*

Observe that a symmetry γ preserves input sets in a natural sense. Because of the way input sets are defined in the multiarrow formalism, the precise relation is

$$\gamma_E(I(c)) = I(\gamma_C(c)),$$

where $\gamma = (\gamma_C, \gamma_E) \in \mathrm{Aut}(\mathcal{G})$.

Remark 2.13. When the network $\mathcal{G}$ has no self-connections and multiarrows there is a simplification of the notion of symmetry due to the following observation. Given a vertex permutation γ_C, there is a unique edge permutation γ_E satisfying the consistency condition (2.13); that is, γ_E is implicitly defined by γ_C since, by Remark 2.10, each arrow e can be identified with a pair of cells $(\mathcal{T}(e), \mathcal{H}(e))$. Thus a *symmetry* of $\mathcal{G}$ is given by a permutation γ of $\mathcal{C}$ such that

(a) $\gamma(c) \sim_I c$ for all $c \in \mathcal{C}$.
(b) $(\gamma(a), \gamma(b)) \in \mathcal{E} \Leftrightarrow (a, b) \in \mathcal{E}$.
(c) $(\gamma(a), \gamma(b)) \sim_E (a, b)\ \forall\ (a, b) \in \mathcal{E}$.

In this case, the group $\mathrm{Aut}(\mathcal{G})$ of symmetries of the network $\mathcal{G}$ is a subgroup of the group $\mathrm{Sym}(\mathcal{C})$ of permutations on the set of cells of the network. We shall adopt this convention throughout the remainder of the paper whenever the network under consideration has no self-connections and multiarrows.

Example 2.14. Since the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of our running example of Figure 1 do not have multiple arrows and self-connections, Remark 2.13 applies. The group $\mathbf{S}_3 \subset \mathbf{S}_4$ consisting of the transpositions $(1\,2)$, $(1\,3)$, $(2\,3)$, the 3-cycle permutations $(1\,2\,3)$, $(1\,3\,2)$, and the identity is the symmetry group of the network $\mathcal{G}_1$. Observe that cell 4 is fixed by the symmetry group. On the other hand, the network $\mathcal{G}_2$ has only the identity permutation as a symmetry because the arrows $(1, 4)$, $(2, 4)$, and $(3, 4)$ are all different amongst each other.

This last example shows that the definition of symmetry of a network is very rigid. In the next section we will generalize the definition of symmetry of a network by introducing the notion of *interior symmetry*. In this new context the network $\mathcal{G}_2$ of our example admits an action of the permutation group $\mathbf{S}_3$ as a group of interior symmetries. This corresponds to the symmetry group of the subnetwork of $\mathcal{G}_2$ obtained by ignoring the arrows $(1, 4)$, $(2, 4)$, and $(3, 4)$ of $\mathcal{G}_2$.

3. Interior symmetry.

We present the notion of interior symmetry following [4] and give an alternative characterization in terms of the symmetries of a subnetwork.

3.1. Interior symmetry groups of networks.

Definition 3.1 (see [4]). *Let $\mathcal{G}$ be a coupled cell network. Let $\mathcal{S} \subseteq \mathcal{C}$ be a subset of cells, and let $I(\mathcal{S}) = \{e \in \mathcal{E} : \mathcal{H}(e) \in \mathcal{S}\}$. A pair of bijections $\sigma_C : \mathcal{C} \to \mathcal{C}$ and $\sigma_E : \mathcal{E} \to \mathcal{E}$ is an* interior symmetry *of $\mathcal{G}$ (on the subset $\mathcal{S}$) if the following hold:*

(a) *$\sigma_C : \mathcal{C} \to \mathcal{C}$ is an input equivalence-preserving permutation which is the identity map on the complement $\mathcal{C} \setminus \mathcal{S}$ of $\mathcal{S}$ in $\mathcal{C}$,*
(b) *$\sigma_E : \mathcal{E} \to \mathcal{E}$ is an edge equivalence-preserving permutation which is the identity map on the complement $\mathcal{E} \setminus I(\mathcal{S})$ of $I(\mathcal{S})$ in $\mathcal{E}$,*
(c) *the consistency condition*

$$\sigma_C(\mathcal{H}(e)) = \mathcal{H}(\sigma_E(e)) \qquad \text{and} \qquad \sigma_C(\mathcal{T}(e)) = \mathcal{T}(\sigma_E(e)) \tag{3.1}$$

is satisfied for every $e \in I(\mathcal{S})$.

The set of all interior symmetries of $\mathcal{G}$ (on the subset $\mathcal{S}$) forms a finite group $\Sigma_{\mathcal{S}}$ called the group of interior symmetries *of $\mathcal{G}$ (on the subset $\mathcal{S}$).*

Note that in Definition 3.1 if $\mathcal{S} = \mathcal{C}$, then $\Sigma_{\mathcal{S}} = \mathrm{Aut}(\mathcal{G})$. Hence, the definition of interior symmetry of a network is a generalization of a symmetry of a network. That is why we refer to the elements of $\mathrm{Aut}(\mathcal{G})$ as *global symmetries* of $\mathcal{G}$. The most interesting case is when $\mathrm{Aut}(\mathcal{G})$ is trivial but $\Sigma_{\mathcal{S}}$ is nontrivial for some $\mathcal{S}$.

Example 3.2. We continue with our running example, the two networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1. We have seen that the network $\mathcal{G}_1$ is $\mathbf{S}_3$-symmetric and the network $\mathcal{G}_2$ has only the trivial symmetry. However, the group of permutations

$$\mathbf{S}_3 = \{\mathrm{id}, (1\,2), (1\,3), (2\,3), (1\,2\,3), (1\,3\,2)\}$$

is the group of interior symmetries of the network $\mathcal{G}_2$ on the subset $\mathcal{S} = \{1, 2, 3\}$. Observe that all elements of $\mathbf{S}_3$ fix cell 4 and

$$I(\mathcal{S}) = \{(1,2),(2,1),(1,3),(3,1),(2,3),(3,2),(4,1),(4,2),(4,3)\}.$$

If we assume that the permutations in $\mathbf{S}_3$ act as identity on the set of arrows

$$\mathcal{E} \setminus I(\mathcal{S}) = \{(1,4),(2,4),(3,4)\},$$

then $\mathbf{S}_3$ is the group of interior symmetries of the network $\mathcal{G}_2$ on the subset $\mathcal{S} = \{1, 2, 3\}$.

There is an alternative characterization of interior symmetries using the notion of symmetry of a network. The main idea is the following: by "ignoring" some arrows we find a subnetwork whose symmetry group is the group of interior symmetries of the original network.

Let us be more precise. Given a coupled cell network $\mathcal{G}$ and a subset $\mathcal{S} \subset \mathcal{C}$ of cells, define $\mathcal{G}_\mathcal{S} = (\mathcal{C}, I(\mathcal{S}), \sim_C, \sim_E)$ to be the subnetwork of $\mathcal{G}$ whose set of cells is $\mathcal{C}$ (together with its cell-equivalence $\sim_C$) and whose set of arrows is $I(\mathcal{S})$. The edge-equivalence on $I(\mathcal{S})$ is obtained by the restriction of the edge-equivalence $\sim_E$ on $\mathcal{E}$.

Proposition 3.3. *Let $\mathcal{G}$ be a coupled cell network and $\mathcal{S} \subset \mathcal{C}$ be a subset of cells of the set of cells of $\mathcal{G}$. Consider the network $\mathcal{G}_\mathcal{S}$ as defined above. Then the group of interior symmetries of the network $\mathcal{G}$ (on the subset $\mathcal{S}$) can be canonically identified with the group of symmetries of the network $\mathcal{G}_\mathcal{S}$:*

$$\Sigma_\mathcal{S} \cong \mathrm{Aut}(\mathcal{G}_\mathcal{S}).$$

Proof. We start by proving that $\Sigma_\mathcal{S}$ can be canonically identified with a subset of $\mathrm{Aut}(\mathcal{G}_\mathcal{S})$. Let $\sigma = (\sigma_C, \sigma_E) \in \Sigma_\mathcal{S}$ be an interior symmetry of $\mathcal{G}$ (on the subset S), as in Definition 3.1. Then, because σ_C and σ_E are the identity maps on $\mathcal{C} \setminus \mathcal{S}$ and $\mathcal{E} \setminus I(S)$, respectively, it follows that σ is a symmetry of $\mathcal{G}_\mathcal{S}$, according to Definition 2.12. Now we show that the above identification is surjective. Let $\gamma = (\gamma_C, \gamma_E) \in \mathrm{Aut}(\mathcal{G}_\mathcal{S})$ be a symmetry of $\mathcal{G}_\mathcal{S}$ (in the sense of Definition 2.12); that is, γ_E is a permutation on the set $I(\mathcal{S})$. Now we can extend γ_E to a permutation σ_E on $\mathcal{E}$ which acts as identity on $\mathcal{E} \setminus I(\mathcal{S})$. The pair $\sigma = (\sigma_C, \sigma_E)$, where $\sigma_C = \gamma_C$ is an interior symmetry of $\mathcal{G}$ (on the subset $\mathcal{S}$) according to Definition 3.1. ∎

The characterization of interior symmetry provided by Proposition 3.3 is particularly useful when the network does not have multiple arrows and/or self-connections, since by Remark 2.13, a symmetry is simply a permutation on the set vertices of the underlying graph.

Example 3.4. Consider the two networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1. Let $\mathcal{S} = \{1, 2, 3\}$. Note that the network $\mathcal{G}_\mathcal{S}$ obtained from $\mathcal{G}_1$ is the same as the one obtained from $\mathcal{G}_2$. In Figure 2 we show these three networks. Observe that for the three networks the sets of arrows coming from the set $\mathcal{S} = \{1, 2, 3\}$ and directed to the complement $\mathcal{C} \setminus \mathcal{S} = \{4\}$ are different.

Let $\mathcal{G}$ be a network, and fix a phase space P. Suppose that $\mathcal{G}$ admits nontrivial interior symmetries $\Sigma_\mathcal{S}$ on a subset of cells $\mathcal{S}$. Then we can decompose the phase space P as a Cartesian product $P = P_\mathcal{S} \times P_{\mathcal{C}\setminus\mathcal{S}}$, where

$$P_\mathcal{S} = \prod_{s\in\mathcal{S}} P_s \qquad \text{and} \qquad P_{\mathcal{C}\setminus\mathcal{S}} = \prod_{c\in\mathcal{C}\setminus\mathcal{S}} P_c.$$

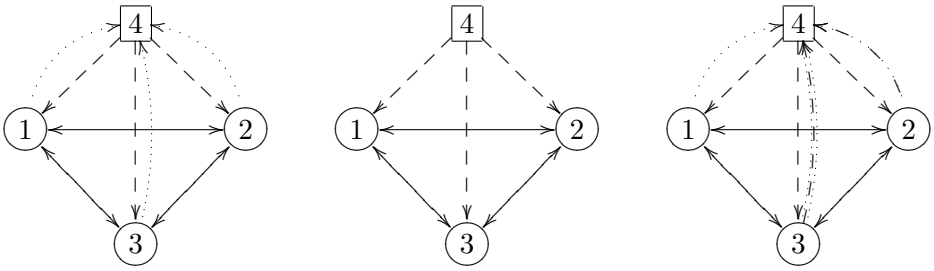


Figure 2. *(Left) Network $\mathcal{G}_1$. (Center) Network $\mathcal{G}_\mathcal{S}$, where $\mathcal{S} = \{1, 2, 3\}$. (Right) Network $\mathcal{G}_2$.*

For any $x \in P$ we write $x = (x_\mathcal{S}, x_{\mathcal{C}\setminus\mathcal{S}})$, where $x_\mathcal{S} \in P_\mathcal{S}$ and $x_{\mathcal{C}\setminus\mathcal{S}} \in P_{\mathcal{C}\setminus\mathcal{S}}$. If $\sigma = (\sigma_C, \sigma_E) \in \Sigma_\mathcal{S}$, then σ_C permutes the cells of $\mathcal{S}$ and induces an action of $\Sigma_\mathcal{S}$ on P by permuting the cell coordinates

$$\sigma(x_c)_{c\in\mathcal{C}} = (x_{\sigma_C^{-1}(c)})_{c\in\mathcal{C}}.$$

Since $\Sigma_\mathcal{S}$ fixes all cells in $\mathcal{C} \setminus \mathcal{S}$ we can write

$$\sigma(x_\mathcal{S}, x_{\mathcal{C}\setminus\mathcal{S}}) = (\sigma x_\mathcal{S}, x_{\mathcal{C}\setminus\mathcal{S}}). \tag{3.2}$$

As in the case of symmetric networks, we can construct (some) balanced equivalence relations on a network $\mathcal{G}$ from subgroups of the interior symmetry group. Suppose that $K \subseteq \Sigma_\mathcal{S}$ is a subgroup. Then

$$\mathrm{Fix}_P(K) = \{(x_\mathcal{S}, x_{\mathcal{C}\setminus\mathcal{S}}) : \sigma x_\mathcal{S} = x_\mathcal{S} \ \ \forall\, \sigma \in K\}.$$

Define the relation $\bowtie_K$ on the cells in $\mathcal{C}$ by

$$c \bowtie_K d \quad \Leftrightarrow \quad \exists\, \sigma = (\sigma_C, \sigma_E) \in K : \sigma_C(c) = d.$$

Then the $\bowtie_K$-classes are the K-orbits on the cells in $\mathcal{S}$, and the corresponding polydiagonal is

$$\triangle_K = \triangle_{\bowtie_K} = \mathrm{Fix}_P(K).$$

The following proposition from [4, Proposition 1, p. 397] is fundamental in the study of coupled cell networks with interior symmetries.

Proposition 3.5 (Golubitsky, Pivato, and Stewart [4]). *Let $\mathcal{G}$ be a network admitting a non-trivial interior symmetry group $\Sigma_\mathcal{S}$ and fix a phase space P. Let K be any subgroup of $\Sigma_\mathcal{S}$. Then $\bowtie_K$ is a balanced relation on $\mathcal{G}$. In particular, $\mathrm{Fix}_P(K)$ is a flow invariant subspace for all $\mathcal{G}$-admissible vector fields.*

Proof. Let s_1 and s_2 be two cells on the same K-orbit. Then there exists an element $\sigma = (\sigma_C, \sigma_E)$ of K such that $\sigma_C(s_1) = s_2$, and by the consistency condition (3.1) it follows that the restriction

$$\sigma_E|_{I(s_1)} : I(s_1) \to I(s_2)$$

is an input isomorphism. Since the $\bowtie_K$-equivalence classes are exactly the K-orbits on $\mathcal{C}$ it follows that the input isomorphism $\sigma_E|_{I(s_1)}$ preserves the $\bowtie_K$ equivalence relation. Hence, by

Theorem 2.9, it follows that $\triangle_H = \mathrm{Fix}_P(K)$ is a flow invariant subspace for all $\mathcal{G}$-admissible vector fields. ■

Example 3.6. Consider the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1 and fix a phase space P for both networks. There are two nontrivial conjugacy classes of subgroups of $\mathbf{S}_3$. The first conjugacy class is represented, for example, by the subgroup generated by a 3-cycle,

$$\mathbf{Z}_3 = \langle (1\,2\,3) \rangle.$$

The associated balanced relation has two equivalence classes $\{1,2,3\}$ and $\{4\}$ given by the three orbits of $\mathbf{Z}_3$ on the set of cells $\mathcal{C}$. The fixed-point subspace of $\mathbf{Z}_3$ is

$$\mathrm{Fix}_P(\mathbf{Z}_3) = \{(x,x,x,y) : x \in P_{\mathcal{S}},\, y \in P_{\mathcal{C}\setminus\mathcal{S}}\} = \mathrm{Fix}_P(\mathbf{S}_3).$$

The second conjugacy class of subgroups is represented, for example, by the subgroup generated by a transposition

$$\mathbf{Z}_2 = \langle (1\,2) \rangle.$$

The associated balanced relation has three equivalence classes $\{1,2\}$, $\{3\}$, and $\{4\}$ given by the three orbits of $\mathbf{Z}_2$ on the set of cells $\mathcal{C}$. The fixed-point subspace of $\mathbf{Z}_2$ is

$$\mathrm{Fix}_P(\mathbf{Z}_2) = \{(x,x,y,z) : x,y \in P_{\mathcal{S}},\, z \in P_{\mathcal{C}\setminus\mathcal{S}}\}.$$

The other two subgroups in the conjugacy class of $\langle (1\,2) \rangle$ are the ones generated by $(1\,3)$ and $(2\,3)$. Observe that these three balanced equivalence relations given by orbits of subgroups are exactly the same balanced equivalence relations previously found by direct methods (Example 2.11). Therefore, in our running example all flow-invariant subspaces can be given as fixed-point subspaces of subgroups.

Remark 3.7. It is not true, even for symmetric networks, that all balanced equivalence relations are given by orbits of subgroups of the symmetry group of the network. Balanced equivalence relations that are not of this type are called *exotic*. For examples of exotic balanced relations, see Antoneli and Stewart [1, 2].

3.2. Admissible vector fields with interior symmetry. Let $\mathcal{G}$ be a network with a nontrivial interior symmetry group $\Sigma_{\mathcal{S}}$ on a subset of cells $\mathcal{S}$, and fix a phase space P. We have a natural decomposition

$$P = P_{\mathcal{S}} \oplus P_{\mathcal{C}\setminus\mathcal{S}} \tag{3.3}$$

with coordinates $(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}})$. If $f : P \to P$ is a $\mathcal{G}$-admissible vector field, then we can write $f = (f_{\mathcal{S}}, f_{\mathcal{C}\setminus\mathcal{S}})$, where $f_{\mathcal{S}} : P \to P_{\mathcal{S}}$ and $f_{\mathcal{C}\setminus\mathcal{S}} : P \to P_{\mathcal{C}\setminus\mathcal{S}}$. Groupoid-equivariance of the coupled cell system implies that

$$\sigma f_{\mathcal{S}}(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) = f_{\mathcal{S}}(\sigma x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \tag{3.4}$$

for all $\sigma \in \Sigma_{\mathcal{S}}$.

A $\mathcal{G}$-admissible vector field f can be written as

$$f(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) = \begin{bmatrix} f_{\mathcal{S}}(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \\ \tilde{f}_{\mathcal{C}\setminus\mathcal{S}}(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \end{bmatrix} + \begin{bmatrix} 0 \\ h(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \end{bmatrix}, \tag{3.5}$$

where $\tilde{f}_{\mathcal{C}\setminus\mathcal{S}}, h : P \to P_{\mathcal{C}\setminus\mathcal{S}}$ and $f_{\mathcal{C}\setminus\mathcal{S}} = \tilde{f}_{\mathcal{C}\setminus\mathcal{S}} + h$. The vector field $\tilde{f} = (f_{\mathcal{S}}, \tilde{f}_{\mathcal{C}\setminus\mathcal{S}})$ is the $\Sigma_{\mathcal{S}}$-*equivariant part* of f; that is, for all $\sigma \in \Sigma_{\mathcal{S}}$,

$$\sigma\tilde{f}(x) = \tilde{f}(\sigma x),$$

or more explicitly,

$$\begin{bmatrix} \sigma f_{\mathcal{S}}(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \\ \tilde{f}_{\mathcal{C}\setminus\mathcal{S}}(x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \end{bmatrix} = \begin{bmatrix} f_{\mathcal{S}}(\sigma x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \\ \tilde{f}_{\mathcal{C}\setminus\mathcal{S}}(\sigma x_{\mathcal{S}}, x_{\mathcal{C}\setminus\mathcal{S}}) \end{bmatrix}, \tag{3.6}$$

since $\Sigma_{\mathcal{S}}$ acts trivially on $P_{\mathcal{C}\setminus\mathcal{S}}$. Equation (3.5) can be seen as a decomposition of the vector field f as the sum of a $\Sigma_{\mathcal{S}}$-equivariant vector field and a nonequivariant "perturbation" with null components in $\mathcal{S}$.

Example 3.8. Consider the network $\mathcal{G}_2$ of Figure 1. Recall from Example 2.7 the general form of the ODEs associated with the network $\mathcal{G}_2$. Using the decomposition (3.3), we have $x_{\mathcal{S}} = (x_1, x_2, x_3)$ and $x_{\mathcal{C}\setminus\mathcal{S}} = (x_4)$, where $x_i \in \mathbf{R}^k$ $(i = 1, 2, 3)$, $x_4 \in \mathbf{R}^l$. Then by (3.5) we can write a general ODE for the network $\mathcal{G}_2$ as

$$\begin{aligned} \dot{x_1} &= f(x_1, \overline{x_2, x_3}, x_4), \\ \dot{x_2} &= f(x_2, \overline{x_3, x_1}, x_4), \\ \dot{x_3} &= f(x_3, \overline{x_1, x_2}, x_4), \\ \dot{x_4} &= g(x_4, \overline{x_1, x_2, x_3}) + h(x_4, x_1, x_2, x_3), \end{aligned}$$

where $f : \mathbf{R}^{3k} \times \mathbf{R}^l \to \mathbf{R}^k$ is a smooth map invariant under permutation of the second and third argument, $g : \mathbf{R}^l \times \mathbf{R}^{3k} \to \mathbf{R}^l$ is $\mathbf{S}_3$-invariant with respect to (x_1, x_2, x_3), and $h : \mathbf{R}^l \times \mathbf{R}^{3k} \to \mathbf{R}^l$ is a general smooth map.

Now we introduce another set of coordinates on P, adapted to the action of the interior symmetry group. By Proposition 3.5 the subspace $\mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ is flow invariant. Since $\mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ is $\Sigma_{\mathcal{S}}$-invariant and $\Sigma_{\mathcal{S}}$ acts trivially on the cells in $\mathcal{C} \setminus \mathcal{S}$ we have that $P_{\mathcal{C}\setminus\mathcal{S}} \subset \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$. Let

$$U = \mathrm{Fix}_P(\Sigma_{\mathcal{S}}). \tag{3.7}$$

The action of the group $\Sigma_{\mathcal{S}}$ decomposes the set $\mathcal{S}$ as

$$\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k,$$

where the sets $\mathcal{S}_i$ $(i = 1, \ldots, k)$ are the orbits of the $\Sigma_{\mathcal{S}}$-action. Let

$$W = \left\{ x \in P \ : \ x_c = 0 \ \ \forall\, c \in \mathcal{C} \setminus \mathcal{S} \ \text{ and } \ \sum_{s \in \mathcal{S}_i} x_s = 0 \ \text{ for } 1 \leqslant i \leqslant k \right\}. \tag{3.8}$$

Since W is a $\Sigma_{\mathcal{S}}$-invariant subspace of $P_{\mathcal{S}}$ and $W \cap U = \{0\}$ we can decompose the phase space P as a direct sum of $\Sigma_{\mathcal{S}}$-invariant subspaces,

$$P = W \oplus U. \tag{3.9}$$

In particular, (3.8) implies that vectors in W, when written in coupled cell coordinates, have zero components on all cells in $\mathcal{C} \setminus \mathcal{S}$.

We can choose coordinates (w, u) with $w \in W$ and $u \in U$ adapted to the decomposition (3.9) and write any admissible vector field f as

$$f(w,u) = \begin{bmatrix} f_W(w,u) \\ f_U(w,u) \end{bmatrix} + \begin{bmatrix} 0 \\ h(w,u) \end{bmatrix}, \tag{3.10}$$

where $f_U, h : P \to U$ and $f_W : P \to W$ satisfies

$$\sigma f_W(w,u) = f_W(\sigma w, u) \qquad \forall\, \sigma \in \Sigma_{\mathcal{S}}.$$

With respect to the decomposition (3.9), the equivariant part of f is written as $\tilde{f}(w,u) = (f_W(w,u), f_U(w,u))$, and for all $\sigma \in \Sigma_{\mathcal{S}}$ we have

$$\begin{bmatrix} \sigma f_W(w,u) \\ f_U(w,u) \end{bmatrix} = \begin{bmatrix} f_W(\sigma w, u) \\ f_U(\sigma w, u) \end{bmatrix}, \tag{3.11}$$

since $\Sigma_{\mathcal{S}}$ acts trivially on $U = \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$.

Example 3.9. Consider the network $\mathcal{G}_2$ of Figure 1. With respect to the decomposition (3.3) adapted to the network structure, the total phase space P has coordinates $x_{\mathcal{S}} = (x_1, x_2, x_3)$ and $x_{\mathcal{C}\setminus\mathcal{S}} = (x_4)$, where $x_i \in \mathbf{R}^k$ $(i = 1, 2, 3)$, $x_4 \in \mathbf{R}^l$. Now with respect to the decomposition (3.9) adapted to the $\mathbf{S}_3$-action on P we have that

$$W = \{(w_1, w_2, -w_1 - w_2, 0) : w_1, w_2 \in \mathbf{R}^k\}$$

and

$$U = \mathrm{Fix}_P(\mathbf{S}_3) = \{(u_1, u_1, u_1, u_2) : u_1 \in \mathbf{R}^k,\ u_2 \in \mathbf{R}^l\}.$$

In the linear case, we may choose a basis of P adapted to the decomposition (3.9), and then a $\mathcal{G}$-admissible linear vector field L can be written as

$$L = \begin{bmatrix} A & 0 \\ C & B \end{bmatrix}, \tag{3.12}$$

where $B = L|_U : U \to U$, $C : W \to U$, and $A : W \to W$ satisfies (by (3.11))

$$A\,\sigma = \sigma A \qquad \forall\, \sigma \in \Sigma_{\mathcal{S}}.$$

The spectral properties of L in (3.12) are given by Golubitsky, Pivato, and Stewart [4, Lemma 1, p. 399]. Since we will use these results several times we reproduce them here.

Lemma 3.10 (see [4]). *Let $\mathcal{G}$ be a network admitting a nontrivial group of interior symmetries $\Sigma_{\mathcal{S}}$, and fix a total phase space P. Let $L : P \to P$ be a $\mathcal{G}$-admissible linear vector field, and consider the decomposition of L given by (3.12). Then the following hold:*

(i) *The eigenvalues of L are the eigenvalues of A together with the eigenvalues of B.*

(ii) *A vector $u \in U = \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ is an eigenvector of B with eigenvalue ν if and only if u is an eigenvector of L with eigenvalue ν.*

(iii) *If $w \in W$ is an eigenvector of A with eigenvalue μ, then there exists an eigenvector v of L with eigenvalue μ of the form*

$$v = w + u,$$

where $u \in U = \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$.

(iv) *All eigenspaces of A are $\Sigma_{\mathcal{S}}$-invariant.*

Proof. Parts (i), (ii), and (iii) are consequences of the block form (3.12) of L. Part (iv) follows from the $\Sigma_{\mathcal{S}}$-equivariance of A. ∎

Example 3.11. We continue our running example, the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1. The general form of the admissible linear mappings associated with the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1 are (in cell coordinates)

$$L_1 = \begin{pmatrix} a & b & b & d \\ b & a & b & d \\ b & b & a & d \\ e & e & e & c \end{pmatrix} \quad \text{and} \quad L_2 = \begin{pmatrix} a & b & b & d \\ b & a & b & d \\ b & b & a & d \\ e_1 & e_2 & e_3 & c \end{pmatrix},$$

where a, b are $k \times k$ matrices, c is an $l \times l$ matrix, d is a $k \times l$ matrix, and e, e_1, e_2, e_3 are $l \times k$ matrices. Choosing adequate bases for W and U, the linear mappings L_1 and L_2 can be written as

$$L_1 = \begin{pmatrix} a-b & 0 & 0 & 0 \\ 0 & a-b & 0 & 0 \\ 0 & 0 & a+2b & d \\ 0 & 0 & 3e & c \end{pmatrix}$$

and

$$L_2 = \begin{pmatrix} a-b & 0 & 0 & 0 \\ 0 & a-b & 0 & 0 \\ 0 & 0 & a+2b & d \\ e_1-e_3 & e_2-e_3 & e_1+e_2+e_3 & c \end{pmatrix}.$$

4. Synchrony-breaking bifurcations. Now we study local bifurcations in coupled cell networks with nontrivial interior symmetries. We are interested in codimension-one synchrony-breaking bifurcations. Steady-state and Hopf bifurcations in coupled cell networks with interior symmetries were studied by Golubitsky, Pivato, and Stewart [4].

4.1. Local bifurcations in coupled cell systems. Let $\mathcal{G}$ be a coupled cell network, and fix a phase space P. Let $f : P \times \mathbf{R}^k \to P$ be a smooth k-parameter family of $\mathcal{G}$-admissible vector fields in P, and assume that the ODE

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x, \lambda) \tag{4.1}$$

has a synchronous equilibrium x_0 in $\triangle_I$ (the polydiagonal subspace of P associated with the input equivalence relation $\sim_I$). In the present context we may assume that

$$f(x_0, \lambda) \equiv 0$$

and that a bifurcation occurs at $\lambda = 0$. Let $L = (\mathrm{d}f)_{(x_0,0)}$ be the linearization of f at $(x_0, 0)$, and denote by E^c the center subspace of L.

Local bifurcations in coupled cell networks can be divided into two types according to whether or not E^c is contained in the flow-invariant subspace $\triangle_I$.

Definition 4.1. *We say that a coupled cell system* (4.1) *undergoes a* synchrony-preserving *bifurcation at a synchronous equilibrium in* $\triangle_I$ *if* $E^c \subset \triangle_I$ *and that* (4.1) *undergoes a* synchrony-breaking *bifurcation if* $E^c \not\subset \triangle_I$.

Now we specialize to codimension-one bifurcations; that is, $f : P \times \mathbf{R} \to P$ is a smooth 1-parameter family of $\mathcal{G}$-admissible vector fields in P. These bifurcations fall into two classes: *steady-state* bifurcations ($L|_{E^c}$ has a zero eigenvalue) and *Hopf* bifurcations ($L|_{E^c}$ has a pair of purely imaginary eigenvalues). The new steady states and periodic solutions that emanate from the synchrony-preserving bifurcations are themselves synchronous solutions. For the remainder of this paper we will focus on codimension-one synchrony-breaking bifurcations from a synchronous equilibrium.

4.2. Local bifurcations with interior symmetry. Interior symmetries introduce genuine restrictions on the form of the linearization, and this structure can be used to study certain kinds of synchrony-breaking bifurcations, namely, the bifurcations that break the interior symmetry.

Let $\mathcal{G}$ be a network admitting a nontrivial group of interior symmetries $\Sigma_{\mathcal{S}}$ on $\mathcal{S}$, and fix a phase space P. First, note that the polydiagonal subspace $\triangle_I$ associated with the input equivalence relation $\sim_I$ satisfies

$$\triangle_I \subseteq \mathrm{Fix}_P(\Sigma_{\mathcal{S}}).$$

Since we are interested in synchrony-breaking bifurcations that also break the interior symmetry we may assume that $x_0 \in \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ and that the center subspace $E^c(L)$ associated with the critical eigenvalues satisfies

$$E^c(L) \not\subset \mathrm{Fix}_P(\Sigma_{\mathcal{S}}). \tag{4.2}$$

However, this is not enough to exclude the possibility of having critical eigenvectors in $\mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ in a synchrony-breaking bifurcation. That is, we could have a situation where some critical eigenvectors belong to $\mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ and the others are outside $\mathrm{Fix}_P(\Sigma_{\mathcal{S}})$. Indeed, it is well known [3] that (nonsymmetric) coupled cell systems generically can exhibit mode interaction in codimension-one bifurcations. In this paper we make a stronger assumption. We assume

$$E^c(L) \cap \mathrm{Fix}_P(\Sigma_{\mathcal{S}}) = \{0\}, \tag{4.3}$$

and so we exclude the possibility of having eigenvectors in $\mathrm{Fix}_P(\Sigma_{\mathcal{S}})$. This situation corresponds to a synchrony-breaking bifurcation that "breaks *only* the interior symmetry."

Definition 4.2. *Let* $f : P \to P$ *be a* $\mathcal{G}$*-admissible vector field, and let* $L = (\mathrm{d}f)_{(x_0)}$ *be the linearization of* f *at* x_0. *Consider the decomposition* (3.9) *of* P *adapted to the* $\Sigma_{\mathcal{S}}$*-action, and write* L *in block form as*

$$L = \begin{bmatrix} A & 0 \\ C & B \end{bmatrix}.$$

Then the matrix A is called the $\Sigma_{\mathcal{S}}$-equivariant subblock *of L.*

If we write f using coordinates (w,u) adapted to the decomposition $P = W \oplus U$ as

$$f(w,u) = \begin{bmatrix} f_W(w,u) \\ f_U(w,u) \end{bmatrix} + \begin{bmatrix} 0 \\ h(w,u) \end{bmatrix},$$

then

$$A = \big(\mathrm{d}_{(1)} f_W\big)_{(w_0)},$$

where $x_0 = (w_0, u_0)$ and

$$\big(\mathrm{d}_{(1)} f_W\big)_{(w_0)} \cdot w = \big(\mathrm{d} f_W\big)_{(w_0,u_0)} \cdot (w,0)$$

for all $w \in W$.

Remark 4.3. It can be shown that the following three conditions are equivalent:

(a) $E^c(L) \cap \mathrm{Fix}_P(\Sigma_{\mathcal{S}}) = \{0\}$.

(b) $\dim E^c(L) = \dim E^c(A)$.

(c) All the critical eigenvalues of L come from the $\Sigma_{\mathcal{S}}$-equivariant subblock A of L.

It is obvious that (a) implies both (b) and (c). On the other hand, to prove that (b) implies (a), we observe that by Lemma 3.10(iii), we always have $\dim E^c(A) \leqslant \dim E^c(L)$. Finally, to prove that (c) implies (a), we observe that the block form of L guarantees that no generalized eigenvector associated with an eigenvalue coming from subblock A belongs to $\mathrm{Fix}_P(\Sigma_{\mathcal{S}})$.

In general f is not $\Sigma_{\mathcal{S}}$-equivariant and L does not commute with $\Sigma_{\mathcal{S}}$. In particular, $E^c(L) \not\subset W$. However, the block matrix A does commute with $\Sigma_{\mathcal{S}}$, and thus $E^c(A) \subset W$ is $\Sigma_{\mathcal{S}}$-invariant. Moreover, if A has purely imaginary eigenvalues, there is a natural action of $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$ on $E^c(A)$, where $\mathbf{S}^1$ acts by $\exp(sA^{\mathrm{t}})$.

Definition 4.4. *Consider a* 1*-parameter family of coupled cell systems* (4.1) *with interior symmetry group $\Sigma_{\mathcal{S}}$ on $\mathcal{S}$ undergoing a codimension-one synchrony-breaking bifurcation at a synchronous equilibrium x_0 when $\lambda = 0$. We say that f undergoes a* codimension-one interior symmetry-breaking bifurcation *if the following conditions hold:*

(a) *All the critical eigenvalues μ of L come from the $\Sigma_{\mathcal{S}}$-equivariant subblock A of L.*

(b) *The critical eigenvalues μ extend uniquely and smoothly to eigenvalues $\mu(\lambda)$ of $(\mathrm{d}f)_{(x_0,\lambda)}$ for λ near 0.*

(c) *The* eigenvalue crossing condition:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \mathrm{Re}(\mu(\lambda))\Big|_{\lambda=0} \neq 0. \tag{4.4}$$

More specifically, the bifurcation problem (4.1) *is called*

- *a* codimension-one interior symmetry-breaking steady-state bifurcation *if, in addition to the conditions* (a), (b), (c) *above, the matrix A has a zero eigenvalue and the associated center subspace is given by*

$$E_0(A) = \ker(A); \tag{4.5}$$

- *a* codimension-one interior symmetry-breaking Hopf bifurcation *if, in addition to the conditions* (a), (b), (c) *above, the matrix A is nonsingular, all the critical eigenvalues (after rescaling time if necessary) have the form $\pm i$, and the associated center subspace is given by*

$$E_i(A) = \{x \in W : (A^2 + 1)x = 0\}. \tag{4.6}$$

Example 4.5. Consider the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1. Suppose that for all cells c we choose the internal phase space to be $P_c = \mathbf{C}$, and so the total phase space is $P = \mathbf{C}^4$. Consider the decomposition of $P = W \oplus U$ adapted to the $\mathbf{S}_3$-action. Then

$$W = \{(w_1, w_2, -w_1 - w_2, 0) : w_1, w_2 \in \mathbf{C}\}$$

and

$$U = \mathrm{Fix}_P(\mathbf{S}_3) = \{(u_1, u_1, u_1, u_2) : u_1, u_2 \in \mathbf{C}\},$$

and W is an $\mathbf{S}_3$-simple representation ($W = W_1 \oplus W_2$, where W_1, W_2 are two isomorphic $\mathbf{S}_3$-absolutely irreducible spaces). Now consider a 1-parameter family $f : P \times \mathbf{R} \to P$ of $\mathcal{G}$-admissible vector fields on P undergoing a codimension-one interior symmetry-breaking Hopf bifurcation at an equilibrium point x_0 when $\lambda = 0$. Since W is an $\mathbf{S}_3$-simple representation, one necessarily has that $E^c(A) = W$. Moreover, the action of the circle group $\mathbf{S}^1$ defined by $\exp(sA^{\mathrm{t}})$ is equivalent to the standard action of $\mathbf{S}^1$ on $\mathbf{C}^2$, that is,

$$\theta \cdot (z_1, z_2) = (\mathrm{e}^{i\theta} z_1, \mathrm{e}^{i\theta} z_2)$$

for all $\theta \in \mathbf{S}^1$ and $z_1, z_2 \in \mathbf{C}$.

4.3. Interior symmetry-breaking Hopf theorem. The Hopf bifurcation theorem concerns periodic solutions to differential equations near a point where the linearization has purely imaginary eigenvalues.

Let $\mathcal{G}$ be a coupled cell network admitting a nontrivial group of interior symmetries $\Sigma_{\mathcal{S}}$ on a subset $\mathcal{S}$ of cells, and choose a total phase space P. Consider a smooth 1-parameter family $f : P \times \mathbf{R} \to P$ of $\mathcal{G}$-admissible vector fields on P, and assume that

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x, \lambda) \tag{4.7}$$

has an equilibrium x_0 such that for $\lambda = 0$ the linearization $L = (\mathrm{d}f)_{(x_0,0)}$ of f at $(x_0, 0)$ is nonsingular but has purely imaginary eigenvalues.

Before stating the next theorem let us introduce an important concept which generalizes the notion of a $\mathbf{C}$-axial subgroup from equivariant bifurcation theory.

Definition 4.6. *Let $\mathcal{G}$ be a coupled cell network admitting a nontrivial group of interior symmetries $\Sigma_{\mathcal{S}}$ on a subset $\mathcal{S}$. Let P denote the total phase space, and consider the decomposition* (3.9) *of P adapted to the $\Sigma_{\mathcal{S}}$-action. Suppose that there is an action of circle group $\mathbf{S}^1$ on W which commutes with the action of $\Sigma_{\mathcal{S}}$. Let $E \subset W$ be a $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$-invariant subspace. An isotropy subgroup $\Delta \subseteq \Sigma_{\mathcal{S}} \times \mathbf{S}^1$ is called* interiorly $\mathbf{C}$-axial (on E) *if*

$$\dim_{\mathbf{R}} \mathrm{Fix}_E(\Delta) = 2.$$

Now suppose that the family (4.7) undergoes a codimension-one interior symmetry-breaking Hopf bifurcation at the equilibrium x_0 when $\lambda = 0$. Then the center subspace $E^c(A)$ of the $\Sigma_{\mathcal{S}}$-equivariant subblock of the linearization $L = (\mathrm{d}f)_{(x_0,0)}$ of f at $(x_0, 0)$ is a $\Sigma_{\mathcal{S}}$-invariant subspace of W. Therefore, the action of the circle group $\mathbf{S}^1$ defined by $\exp(sA^{\mathrm{t}})$ commutes with the action of $\Sigma_{\mathcal{S}}$, and so there is a well-defined action of $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$ on W and $E^c(A)$ is a $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$-invariant subspace.

Example 4.7. Consider the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of Figure 1. Suppose that for all cells c we choose the internal phase space to be $P_c = \mathbf{C}$, and so the total phase space is $P = \mathbf{C}^4$. Suppose that a smooth 1-parameter family $f : P \times \mathbf{R} \to P$ of $\mathcal{G}$-admissible vector fields on P undergoes a codimension-one interior symmetry-breaking Hopf bifurcation at the equilibrium $x_0 = 0$ when $\lambda = 0$. Then $E_i(A) = W$, where A is the $\Sigma_{\mathcal{S}}$-equivariant subblock of the linearization $L = (\mathrm{d}f)_{(0,0)}$ of f at $(0, 0)$. In Example 4.5 we observed that the action of $\mathbf{S}^1$ on W, given by $\exp(sA^{\mathrm{t}})$, can be identified with the standard action of $\mathbf{S}^1$ on $\mathbf{C}^2$. There are three nontrivial conjugacy classes of isotropy subgroups of $\mathbf{S}_3 \times \mathbf{S}^1$ acting on W. The first conjugacy class of subgroups is represented, for example, by the subgroup

$$\mathbf{Z}_2 = \big\langle ((1\,2), \mathbf{1}) \big\rangle.$$

The fixed-point subspace of $\mathbf{Z}_2$ is

$$\mathrm{Fix}_W(\mathbf{Z}_2) = \{(-w, -w, 2w, 0) : w \in \mathbf{C}\}.$$

The second conjugacy class of subgroups is represented, for example, by the subgroup

$$\tilde{\mathbf{Z}}_2 = \big\langle ((1\,2), \pi) \big\rangle.$$

The fixed-point subspace of $\tilde{\mathbf{Z}}_2$ is

$$\mathrm{Fix}_W(\tilde{\mathbf{Z}}_2) = \{(w, -w, 0, 0) : w \in \mathbf{C}\}.$$

The third conjugacy class of subgroups is represented, for example, by the subgroup

$$\tilde{\mathbf{Z}}_3 = \big\langle ((1\,2\,3), \tfrac{2\pi}{3}) \big\rangle.$$

The fixed-point subspace of $\tilde{\mathbf{Z}}_3$ is

$$\mathrm{Fix}_W(\tilde{\mathbf{Z}}_3) = \{(w, \mathrm{e}^{i\frac{2\pi}{3}} w, \mathrm{e}^{i\frac{4\pi}{3}} w, 0) : w \in \mathbf{C}\}.$$

The main result of this paper is the interior symmetry-breaking Hopf bifurcation theorem.

Theorem 4.8. *Let $\mathcal{G}$ be a coupled cell network admitting a nontrivial group of interior symmetries $\Sigma_{\mathcal{S}}$ relative to a subset $\mathcal{S}$ of cells, and fix a phase space P. Consider (4.7), where $f : P \times \mathbf{R} \to P$ is a smooth 1-parameter family of $\mathcal{G}$-admissible vector fields on P. Suppose that a codimension-one interior symmetry-breaking Hopf bifurcation (see Definition 4.4) occurs at an equilibrium point $x_0 \in \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ when $\lambda = 0$. Let $\Delta \subset \Sigma_{\mathcal{S}} \times \mathbf{S}^1$ be an interiorly $\mathbf{C}$-axial subgroup (on $E^c(A)$). Then generically there exists a family of small amplitude periodic*

solutions of (4.7) *bifurcating from* $(x_0, 0)$ *and having period near* 2π. *Moreover, to lowest order in the bifurcation parameter* λ, *the solution* $x(t)$ *is of the form*

$$x(t) \approx w(t) + u(t), \tag{4.8}$$

where $w(t) = \exp(tL)w_0$ *(*$w_0 \in \mathrm{Fix}_W(\Delta)$*) has exact spatio-temporal symmetry* Δ *on the cells in* $\mathcal{S}$ *and* $u(t) = \exp(tL)u_0$ *(*$u_0 \in \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$*) is synchronous on the* $\Sigma_{\mathcal{S}}$*-orbits of cells in* $\mathcal{S}$.

We call such a state a *synchronously modulated* Δ*-symmetric wave* on $\mathcal{S}$.

Remarks 4.9.

(a) The above theorem asserts no restriction on $u_j(t)$ when $j \in \mathcal{C} \setminus \mathcal{S}$.

(b) Theorem 4.8 generalizes the interior symmetry Hopf theorem of Golubitsky, Pivato, and Stewart [4, Theorem 3]. Given a subgroup $\Delta \subseteq \Sigma_{\mathcal{S}} \times \mathbf{S}^1$, we define the *spatial subgroup* of Δ to be $K = \Delta \cap \Sigma_{\mathcal{S}}$. A subgroup Δ is called *spatially* $\mathbf{C}$*-axial* if

$$\dim_{\mathbf{R}} \mathrm{Fix}_{E_i(A)}(\Delta) = \dim_{\mathbf{R}} \mathrm{Fix}_{E_i(A)}(K) = 2,$$

where K is the spatial subgroup of Δ. Obviously every spatially $\mathbf{C}$-axial subgroup is interiorly $\mathbf{C}$-axial. Since the Hopf theorem of [4] is proved for all spatially $\mathbf{C}$-axial subgroups, it is a special case of Theorem 4.8.

(c) Theorem 4.8 holds if the assumption (4.6) is generalized to: the matrix A is nonsingular and semisimple, and (after rescaling time if necessary) all the critical eigenvalues have the form $k_l i$ $(k_l \in \mathbf{Z})$.

The proof of Theorem 4.8 follows from a couple of lemmas that we state and prove below. We start by setting up the framework.

Let $C^0_{2\pi}(P)$ be the space consisting of all continuous 2π-periodic mappings from $\mathbf{R}$ to P endowed with the C^0 norm, and let $C^1_{2\pi}(P)$ be the space consisting of all continuous differentiable 2π-periodic mappings from $\mathbf{R}$ to P endowed with the C^1 norm.

By introducing a perturbed period parameter τ we can rescale time again, from t to $s(1+\tau)t$, and consider the operator $\mathcal{F} : C^1_{2\pi}(P) \times \mathbf{R} \times \mathbf{R} \to C^0_{2\pi}(P)$ given by

$$\mathcal{F}(x, \lambda, \tau) = (1+\tau)\frac{\mathrm{d}x}{\mathrm{d}s}(s) - f(x(s), \lambda). \tag{4.9}$$

The 2π-periodic solutions of the equation $\mathcal{F}(x, \lambda, \tau) = 0$ near $(0, 0, 0)$ correspond bijectively to the small amplitude periodic solutions of (4.7) near x_0 and with period near 2π. As is well known, the operator $\mathcal{F}$ is $\mathbf{S}^1$-equivariant with respect to the *phase shift action* of $\mathbf{S}^1$ on the spaces $C^1_{2\pi}(P)$ and $C^0_{2\pi}(P)$; that is, if $x \in C^0_{2\pi}(P)$ and $\theta \in \mathbf{S}^1$, then

$$\big(\theta \cdot x\big)(s) = x(s + \theta)$$

and thus

$$\theta \cdot \mathcal{F}(x, \tau, \lambda) = \mathcal{F}(\theta \cdot x, \tau, \lambda).$$

The linearization of $\mathcal{F}$ about the origin is

$$\mathcal{L}(x) = \frac{\mathrm{d}x}{\mathrm{d}s}(s) - L\,x(s), \tag{4.10}$$

and $\ker(\mathcal{L})$ consists of all functions $\mathrm{Re}(e^{is}v)$, where v is an eigenvector of L associated with the eigenvalue i.

In the standard Hopf bifurcation theorem [5, Theorem VIII 3.1] $\ker(\mathcal{L})$ is two-dimensional, and Lyapunov–Schmidt reduction in the presence of symmetry leads to a reduced equation that can be solved for a unique branch of 2π-periodic solutions as long as the eigenvalues crossing condition is valid. In the equivariant context, $\ker(\mathcal{L})$ may be higher-dimensional—generically $\ker(\mathcal{L})$ is a Γ-simple representation. The proof of the equivariant Hopf bifurcation theorem [8, Theorem XVI 4.1] proceeds by restricting the Lyapunov–Schmidt reduced equation to the fixed-point subspace $\mathrm{Fix}_{E_i(L)}(\Delta)$ of a $\mathbf{C}$-axial subgroup Δ, which is two-dimensional. Then the proof is completed as in the standard Hopf bifurcation theorem.

That approach does not work in the context of interior symmetries since in general there is no action of $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$ on $E_i(L)$, because the original vector field f (and its linearization L) is not $\Sigma_{\mathcal{S}}$-equivariant. Nevertheless, we shall introduce a "modified Lyapunov–Schmidt procedure" that does work in the context of interior symmetries.

The decomposition in (3.9) induces the decompositions

$$C^0_{2\pi}(P) = C^0_{2\pi}(W) \oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$$

and

$$C^1_{2\pi}(P) = C^1_{2\pi}(W) \oplus C^1_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}})).$$

In our modification of the standard Lyapunov–Schmidt procedure we consider the following action of the group $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$ on the spaces $C^0_{2\pi}(P)$ and $C^1_{2\pi}(P)$. Let us write $x \in C^0_{2\pi}(P)$ as $x(s) = (w(s), u(s))$, where $w \in C^0_{2\pi}(W)$ and $u \in C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$. Then

$$(\delta, \theta) \cdot x(s) = (\delta, \theta) \cdot (w(s), u(s)) = \big(\delta w(s+\theta), u(s)\big). \tag{4.11}$$

The difference from the usual action on the loop space (see, for example, [8]) is that in (4.11) the group $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$ acts trivially on $C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$ and $C^1_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$, respectively. A straightforward consequence of the above definition is stated in the next lemma for convenience.

Lemma 4.10. *For any subgroup Δ of $\Sigma_{\mathcal{S}} \times \mathbf{S}^1$ we have the decompositions*

$$\mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta) = C^0_{2\pi}(\mathrm{Fix}_W(\Delta)) \oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$$

and

$$\mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta) = C^1_{2\pi}(\mathrm{Fix}_W(\Delta)) \oplus C^1_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}})).$$

Proof. Let $x \in C^0_{2\pi}(P)$ be written as $x(s) = (w(s), u(s))$, where $w \in C^0_{2\pi}(W)$ and $u \in C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$. Then (4.11) implies that

$$C^0_{2\pi}(\mathrm{Fix}_W(\Delta)) \oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}})) \subseteq \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta).$$

Now let $(\delta, \theta) \in \Delta$ and suppose that $x \in \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta)$. Then

$$(\delta, \theta) \cdot x(s) = x(s).$$

The decomposition $x(s) = (w(s), u(s))$ yields

$$\big((\delta w)(s+\theta), u(s)\big) = (w(s), u(s));$$

that is, $w(s) \in \mathrm{Fix}_W(\Delta)$ and $u(s) \in \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$ for all $s \in \mathbf{R}$. Hence

$$\mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta) \subseteq C^0_{2\pi}(\mathrm{Fix}_W(\Delta)) \oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}})).$$

Therefore,

$$\mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta) = C^0_{2\pi}(\mathrm{Fix}_W(\Delta)) \oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}})).$$

The same argument with $C^1_{2\pi}$ instead of $C^0_{2\pi}$ gives the other equality. ■

Lemma 4.11. *Let $L : P \to P$ be a $\mathcal{G}$-admissible linear mapping. Let $\mathcal{L} : C^1_{2\pi}(P) \times \mathbf{R} \times \mathbf{R} \to C^0_{2\pi}(P)$ be the linear operator given by (4.10) and $\Delta \subset \Sigma_{\mathcal{S}} \times \mathbf{S}^1$ be a subgroup. Then we have that*

$$\begin{aligned}
\mathcal{L}\big(C^1_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))\big) &\subseteq C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}})),\\
\mathcal{L}\big(C^1_{2\pi}(\mathrm{Fix}_W(\Delta))\big) &\subseteq \big(C^0_{2\pi}(\mathrm{Fix}_W(\Delta)) \oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))\big)
\end{aligned}$$

and

$$\mathcal{L}\big(\mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)\big) \subseteq \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta). \tag{4.12}$$

In particular, we can define a linear operator

$$\mathcal{L}_\Delta : \mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta) \longrightarrow \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta) \tag{4.13}$$

by restriction.

Proof. Note that since the circle group $\mathbf{S}^1$ acts on the domain of the mappings, all the decompositions above are $\mathbf{S}^1$-invariant.

First suppose $x(s) = (0, u(s))$ with $u(s) \in C^1_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$. Then

$$\mathcal{L}(x) = \frac{\mathrm{d}u}{\mathrm{d}s}(s) - L(0, u(s)).$$

If $\sigma \in \Sigma_{\mathcal{S}}$, then

$$\begin{aligned}
\sigma\mathcal{L}(x) &= \sigma\frac{\mathrm{d}u}{\mathrm{d}s}(s) - \sigma L\,(0, u(s))\\
&= \frac{\mathrm{d}\sigma u}{\mathrm{d}s}(s) - L\,(0, u(s))\\
&= \frac{\mathrm{d}u}{\mathrm{d}s}(s) - L\,(0, u(s))\\
&= \mathcal{L}(x).
\end{aligned}$$

The second equality above follows from the fact that

$$\sigma\big(L\,(0, u)\big) = L\,(0, u)$$

for all $\sigma \in \Sigma_{\mathcal{S}}$. Therefore, we have $\mathcal{L}(x(s)) \in C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))$.

Next suppose that $x(s) = (w(s), 0)$ with $w(s) \in C^1_{2\pi}(\mathrm{Fix}_W(\Delta))$. Since $w(s) \in W$ for all $s \in \mathbf{R}$, we have that

$$(\delta, \theta) \cdot w(s) = \delta w(s + \theta) = w(s)$$

for all $(\delta, \theta) \in \Delta$, $s \in \mathbf{R}$. Write

$$\mathcal{L}(x) = \big([\mathcal{L}(x)]_1(s), [\mathcal{L}(x)]_2(s)\big)$$

with

$$[\mathcal{L}(x)]_1(s) \in W \quad \forall\, s \in \mathbf{R}$$

and

$$[\mathcal{L}(x)]_2(s) \in \mathrm{Fix}_P(\Sigma_{\mathcal{S}}) \quad \forall\, s \in \mathbf{R}.$$

Then

$$[\mathcal{L}(x)]_1(s) = \frac{\mathrm{d}w}{\mathrm{d}s}(s) - A\, w(s)$$

and

$$[\mathcal{L}(x)]_2(s) = -[C\, w(s) + B\, 0] = -Cw(s).$$

Clearly, $[\mathcal{L}(x)]_2(s) \in \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$. Let $(\delta, \theta) \in \Delta$; then

$$\begin{aligned}
(\delta, \theta) \cdot [\mathcal{L}(x)]_1(s) &= (\delta, \theta) \cdot \frac{\mathrm{d}w}{\mathrm{d}s}(s) - (\delta, \theta) \cdot A\, w(s) \\
&= \delta\, \frac{\mathrm{d}w}{\mathrm{d}s}(s + \theta) - \delta A\, w(s + \theta) \\
&= \frac{\mathrm{d}\delta w}{\mathrm{d}s}(s + \theta) - A\delta\, w(s + \theta) \\
&= \frac{\mathrm{d}w}{\mathrm{d}s}(s) - A\, w(s) \\
&= [\mathcal{L}(x)]_1(s)
\end{aligned}$$

and thus $[\mathcal{L}(x)]_1(s) \in \mathrm{Fix}_W(\Delta)$. Therefore

$$\mathcal{L}(x) \in C^0_{2\pi}(\mathrm{Fix}_W(\Delta)) \oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}})).$$

Thus by linearity of $\mathcal{L}$ and Lemma 4.10 we have

$$\mathcal{L}\big(\mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)\big) \subseteq \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta). \qquad \blacksquare$$

Consider now a 1-parameter family of $\mathcal{G}$-admissible vector fields $f(x, \lambda)$ such that $L = (\mathrm{d}f)_{(x_0,0)}$ satisfies the conditions of the definition of interior symmetry-breaking Hopf bifurcation (Definition 4.4), where A is the $\Sigma_{\mathcal{S}}$-equivariant subblock of L.

Lemma 4.12. *Let $\Delta \subset \Sigma_{\mathcal{S}} \times \mathbf{S}^1$ be a subgroup. Let $\mathcal{L}_\Delta : \mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta) \to \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta)$ be the operator given by (4.13) with $L = (\mathrm{d}f)_{(x_0,0)}$. Then*

$$\dim_{\mathbf{R}} \ker(\mathcal{L}_\Delta) = \dim_{\mathbf{R}} \mathrm{Fix}_{E_i(A)}(\Delta).$$

Proof. By Lemma 3.10 and assumption (4.6), $\ker(\mathcal{L}_\Delta)$ consists of all functions $\mathrm{Re}(\mathrm{e}^{is}v_0)$, where v_0 is an eigenvector of L associated with the eigenvalue i which can be decomposed as

$$v_0 = w_0 + u_0,$$

where $u_0 \in \mathrm{Fix}_P(\Sigma_\mathcal{S})$ is uniquely determined by an eigenvector $w_0 \in \mathrm{Fix}_W(\Delta)$ of A with purely imaginary eigenvalue and

$$(\delta,\theta)\cdot \mathrm{Re}(\mathrm{e}^{is}w_0) = \mathrm{Re}(\mathrm{e}^{i(s+\theta)}\,\delta w_0) = \mathrm{Re}(\mathrm{e}^{is}w_0)$$

for all $(\delta,\theta)\in\Delta$. Hence

$$w_0 \in \mathrm{Fix}_W(\Delta)\cap E_i(A) = \mathrm{Fix}_{E_i(A)}(\Delta).$$

By uniqueness of the decomposition $v_0 = w_0 + u_0$ and the dimension condition (b) of Remark 4.3 we have

$$\dim_{\mathbf{R}}\ker(\mathcal{L}_\Delta) = \dim_{\mathbf{R}}\mathrm{Fix}_{E_i(A)}(\Delta). \qquad \blacksquare$$

Lemma 4.13. *Let us write the 1-parameter family of admissible vector fields $f(x,\lambda)$ in the form*

$$f(x,\lambda) = \begin{bmatrix} f_\mathcal{S}(x,\lambda) \\ \tilde{f}_{\mathcal{C}\setminus\mathcal{S}}(x,\lambda)\end{bmatrix} + \begin{bmatrix} 0 \\ h(x,\lambda)\end{bmatrix}, \tag{4.14}$$

where

$$\tilde{f}(x,\lambda) = \begin{bmatrix} f_\mathcal{S}(x,\lambda) \\ \tilde{f}_{\mathcal{C}\setminus\mathcal{S}}(x,\lambda)\end{bmatrix}$$

is the $\Sigma_\mathcal{S}$-equivariant part of f. Let $\mathcal{F}$, $\tilde{\mathcal{F}}$ be operators on $C^1_{2\pi}(P)\times\mathbf{R}\times\mathbf{R}\to C^0_{2\pi}(P)$ defined by formula (4.9) using f and $\tilde{f}$, respectively. Define

$$\mathcal{H}(x,\tau,\lambda) = h(x(s),\lambda),$$

so that

$$\mathcal{F}(x,\tau,\lambda) = \tilde{\mathcal{F}}(x,\tau,\lambda) - \mathcal{H}(x,\tau,\lambda).$$

Then

$$\mathcal{F}\big(\mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)\times\mathbf{R}\times\mathbf{R}\big)\subseteq \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta).$$

In particular, we may define the operator

$$\mathcal{F}_\Delta : \mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)\times\mathbf{R}\times\mathbf{R} \longrightarrow \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta) \tag{4.15}$$

by restriction, and the linearization of $\mathcal{F}_\Delta$ about the origin is the linear operator $\mathcal{L}_\Delta$ given by the formula (4.13), where $L = (\mathrm{d}f)_{(x_0,0)}$.

Proof. The $\Sigma_\mathcal{S}$-equivariance of $\tilde{f}$ implies that $\tilde{\mathcal{F}}$ is $\Sigma_\mathcal{S}\times\mathbf{S}^1$-equivariant (see [8, Lemma XVI 3.2]). It follows then that

$$\tilde{\mathcal{F}}\big(\mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)\times\mathbf{R}\times\mathbf{R}\big)\subseteq \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta).$$

Then it is enough to show that

$$\mathcal{H}\big(\mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)\times\mathbf{R}\times\mathbf{R}\big)\subseteq \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta).$$

Now let $x(s)\in \mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)$. Recall that $h: P\to P_{\mathcal{C}\setminus\mathcal{S}}$ and $P_{\mathcal{C}\setminus\mathcal{S}}\subset \mathrm{Fix}_P(\Sigma_{\mathcal{S}})$. Therefore,

$$\mathcal{H}(x,\tau,\lambda)(s)=h(x(s),\lambda)\in \mathrm{Fix}_P(\Sigma_{\mathcal{S}})\quad (s\in\mathbf{R})$$

for all $\lambda,\tau\in\mathbf{R}$. By Lemma 4.10 we have that

$$C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))\subset C^0_{2\pi}(\mathrm{Fix}_W(\Delta))\oplus C^0_{2\pi}(\mathrm{Fix}_P(\Sigma_{\mathcal{S}}))=\mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta),$$

and the result follows. ■

Remark 4.14. Equation (4.12) of Lemma 4.11 can derived directly from the above lemma.

Proof of Theorem 4.8. Consider the operator

$$\mathcal{F}_\Delta : \mathrm{Fix}_{C^1_{2\pi}(P)}(\Delta)\times\mathbf{R}\times\mathbf{R}\longrightarrow \mathrm{Fix}_{C^0_{2\pi}(P)}(\Delta).$$

The linearization of $\mathcal{F}_\Delta$ about the origin is the linear operator $\mathcal{L}_\Delta$. Now we invoke the assumption that Δ is $\mathbf{C}$-axial for the natural $\Sigma_{\mathcal{S}}\times\mathbf{S}^1$-action on $E_i(A)$, which together with Lemma 4.12 implies that

$$\dim_{\mathbf{R}}\ker(\mathcal{L}_\Delta)=2.$$

Now we may proceed as in the proof of the standard Hopf bifurcation theorem. If we identify $\ker(\mathcal{L}_\Delta)\cong\mathbf{C}$, then the action of $\mathbf{S}^1$ on $\ker(\mathcal{L}_\Delta)$ is equivalent to the standard the action of $\mathbf{S}^1$ on $\mathbf{C}$. The Lyapunov–Schmidt reduction applied to $\mathcal{F}_\Delta$ yields a $\mathbf{S}^1$-equivariant bifurcation equation,

$$\phi:\mathbf{C}\times\mathbf{R}\times\mathbf{R}\to\mathbf{C}.$$

Moreover, the assumptions of the definition of codimension-one interior symmetry-breaking bifurcation are exactly the conditions necessary to carry out the proof. ■

Table 1

Branches of synchronously modulated Δ-symmetric waves supported by the network $\mathcal{G}_2$ and the associated subgroup. The hat over a variable indicates that $\hat{w}$ has twice the frequency.

Subgroup	Form of solution to lowest order in λ
$\mathbf{Z}_2$	$(w_1(t)+u(t), w_1(t)+u(t), w_2(t)+u(t), v(t))$
$\tilde{\mathbf{Z}}_2$	$(w_1(t)+u(t), w_1(t+\frac{1}{2})+u(t), \hat{w}(t)+u(t), v(t))$
$\tilde{\mathbf{Z}}_3$	$(w_1(t)+u(t), w_1(t+\frac{1}{3})+u(t), w_1(t+\frac{2}{3})+u(t), v(t))$

Example 4.15. Consider the network $\mathcal{G}_2$ of Figure 1. Suppose that for all cells c we choose the internal phase space to be $P_c=\mathbf{C}$. Then the total phase space is $P=\mathbf{C}^4$. Suppose that a smooth 1-parameter family $f: P\times\mathbf{R}\to P$ of $\mathcal{G}$-admissible vector fields on P undergoes a codimension-one interior symmetry-breaking Hopf bifurcation at the equilibrium $x_0=0$ when $\lambda=0$. Then $E_i(A)=W$, where A is the $\Sigma_{\mathcal{S}}$-equivariant subblock of the linearization $L=(\mathrm{d}f)_{(0,0)}$ of f at $(0,0)$. By Theorem 4.8 there are three branches of synchronously modulated Δ-symmetric waves associated with the three conjugacy classes of interiorly $\mathbf{C}$-axial subgroups of $\Sigma_{\mathcal{S}}\times\mathbf{S}^1$ (see Table 1). Observe that the first periodic state of Table 1 is associated with a spatially $\mathbf{C}$-axial subgroup, as is predicted by [4, Theorem 3]. The third periodic state of Table 1 is an approximate rotating wave.

4.4. Numerical simulation. In this last section we illustrate the conclusions of Example 4.15 with a numerical simulation. In order to write down an explicit coupled cell system associated with network $\mathcal{G}_2$ we choose the internal phase space of all four cells to be $P_c = \mathbf{C} \cong \mathbf{R}^2$.

Consider the coupled cell system

$$\begin{aligned} \dot{x_1} &= g(x_1, x_2, x_3) + 2x_4, \\ \dot{x_2} &= g(x_2, x_3, x_1) + 2x_4, \\ \dot{x_3} &= g(x_3, x_1, x_2) + 2x_4, \\ \dot{x_4} &= -x_4 + e_1 x_1 + e_2 x_2 + e_3 x_3, \end{aligned} \tag{4.16}$$

where $g : (\mathbf{R}^2)^3 \to \mathbf{R}^2$ is given by

$$\begin{aligned} g(x, y, z) = -x + (a - 2b_2)\, x\|x\|^2 + b_1\,(y + z) + b_2\,(y\|y\|^2 + z\|z\|^2) \\ + a\,(x\|y\|^2 + x\|z\|^2) + b_3\,(y\|y\|^4 + z\|z\|^4) \end{aligned}$$

and $a, b_1(\lambda), b_2, b_3, e_1, e_2, e_3$ are 2×2 matrices with b_1 depending smoothly on a parameter λ. Let f be the vector field defined by (4.16). Observe that the origin is an equilibrium point for all λ,

$$f(0, \lambda) \equiv 0.$$

The linearization of f at $(0, \lambda)$ is given by (as 2×2 block matrix)

$$L(\lambda) = \begin{pmatrix} -1 & b_1 & b_1 & 2 \\ b_1 & -1 & b_1 & 2 \\ b_1 & b_1 & -1 & 2 \\ e_1 & e_2 & e_3 & -1 \end{pmatrix},$$

where $\pm c$ represents $\pm\left(\begin{smallmatrix} c & 0 \\ 0 & c \end{smallmatrix}\right)$.

We need to choose the coefficients b_1 and e_1, e_2, e_3 in order to have purely imaginary eigenvalues for some λ coming from the subblock A when L is written in the form (3.12). The following values will do the job:

$$b_1(\lambda) = \begin{pmatrix} -1 - \lambda & -1.5 \\ 1.5 & -1 \end{pmatrix}$$

and any values between -1 and 1 for the entries of the matrices e_1, e_2, e_3.

The spectrum of the matrix $L(\lambda)$ has the following properties:

(1) For $\lambda < 0$ all eigenvalues of $L(\lambda)$ have negative real parts.
(2) For $\lambda = 0$ the matrix $L = L(0)$ has two pairs of eigenvalues $\pm i$, and the remaining eigenvalues have negative real parts. Moreover, the eigenvectors associated with the purely imaginary eigenvalues are not in $\mathrm{Fix}(\mathbf{S}_3)$.
(3) For $\lambda > 0$ all eigenvalues of $L(\lambda)$ whose associated eigenvectors are in $\mathrm{Fix}(\mathbf{S}_3)$ have negative real parts, and the remaining eigenvalues have positive real parts.

Thus (4.16) undergoes an interior symmetry-breaking Hopf bifurcation when $\lambda = 0$, giving rise to one branch of periodic solutions for each one of the three interiorly $\mathbf{C}$-axial subgroups of $\mathbf{S}_3 \times \mathbf{S}^1$, as in Table 1, when $\lambda > 0$. However, depending on the choice of the coefficients a, b_2, and b_3 of g, one can make at least one of these periodic solutions be stable. In our simulations we have chosen the following coefficients:

$$a = \begin{pmatrix} -0.5 & 0 \\ 0 & -0.5 \end{pmatrix},$$

(1) for a solution with (interior) symmetry $\tilde{\mathbf{Z}}_3$:

$$b_2 = \begin{pmatrix} 0.6 & 2 \\ 2 & 0.6 \end{pmatrix}, \qquad b_3 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

(2) for a solution with (interior) symmetry $\tilde{\mathbf{Z}}_2$:

$$b_2 = \begin{pmatrix} -0.6 & 1 \\ 1 & -0.6 \end{pmatrix}, \qquad b_3 = \begin{pmatrix} 0.2 & -0.7 \\ -0.7 & 0.2 \end{pmatrix}.$$

(3) for a solution with (interior) symmetry $\mathbf{Z}_2$:

$$b_2 = \begin{pmatrix} -0.6 & 0 \\ 0 & -0.6 \end{pmatrix}, \qquad b_3 = \begin{pmatrix} 0 & 0.7 \\ 0.7 & 0 \end{pmatrix}.$$

The coefficients e_1, e_2, and e_3 represent the coupling that breaks the $\mathbf{S}_3$-symmetry. If $e_1 = e_2 = e_3$, then the coupled cell system (4.16) is admissible for the network $\mathcal{G}_1$ of Figure 1, and so it is $\mathbf{S}_3$-symmetric. On the other hand, if $e_1 \neq e_2 \neq e_3$, then the coupled cell system (4.16) is admissible for the network $\mathcal{G}_2$ of Figure 1 and has genuine $\mathbf{S}_3$-interior symmetry.

In the following we present the results of numerical simulations obtaining the three types of periodic solutions mentioned above, for both of the networks $\mathcal{G}_1$ and $\mathcal{G}_2$ of our running example. In Figures 3, 4, and 5 we superimpose the time series of all four cells, which are identified by colors:

$$1 = \text{blue}, \qquad 2 = \text{red}, \qquad 3 = \text{green}, \qquad 4 = \text{black}.$$

The upper panels show the first components, and the lower panels show the second components. The left panels refer to network $\mathcal{G}_1$ with exact $\mathbf{S}_3$-symmetry, and the panels on the right refer to network $\mathcal{G}_1$ with $\mathbf{S}_3$-interior symmetry. Finally, in Figure 6 we present the solution with interior symmetry $\tilde{\mathbf{Z}}_3$ of network $\mathcal{G}_2$, i.e., the approximate rotating wave from Figure 3 (right), viewed in difference coordinates:

$$x_1 - x_2 = \text{blue}, \qquad x_2 - x_3 = \text{green}, \qquad x_3 - x_1 = \text{red}.$$

Acknowledgment. We thank Marty Golubitsky for useful discussions.

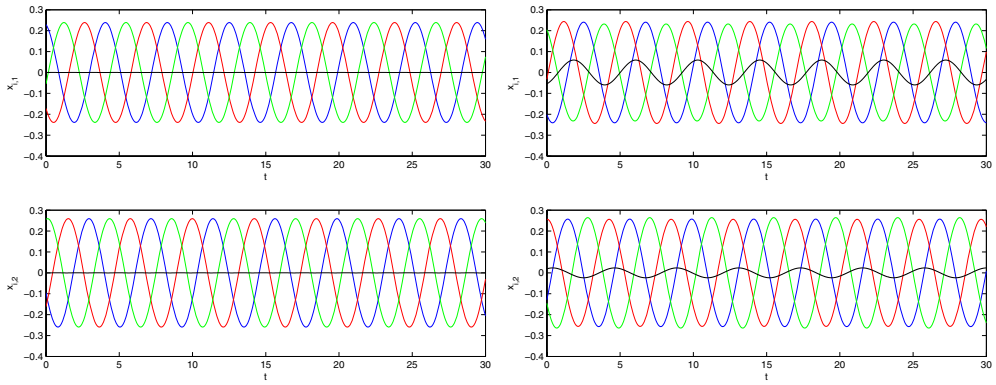


Figure 3. *Solutions with $\tilde{\mathbf{Z}}_3$ (interior) symmetry. (Left) Network $\mathcal{G}_1$ with exact $\mathbf{S}_3$-symmetry. (Right) Network $\mathcal{G}_2$ with $\mathbf{S}_3$-interior symmetry.*

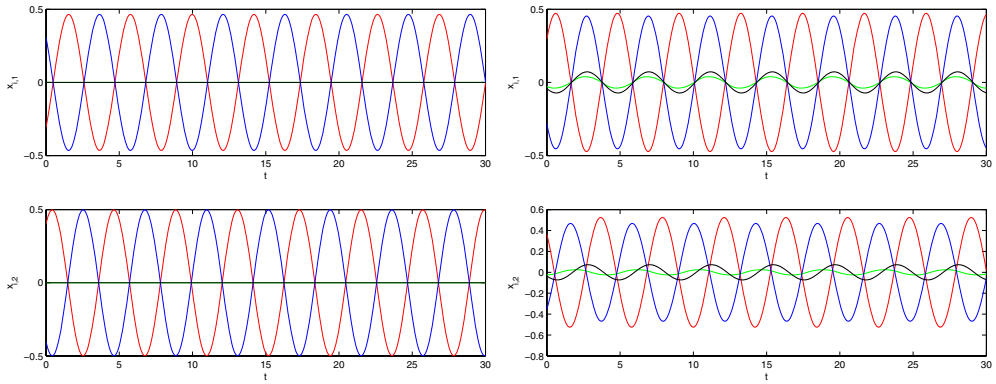


Figure 4. *Solutions with $\tilde{\mathbf{Z}}_2$ (interior) symmetry. (Left) Network $\mathcal{G}_1$ with exact $\mathbf{S}_3$-symmetry. (Right) Network $\mathcal{G}_2$ with $\mathbf{S}_3$-interior symmetry.*

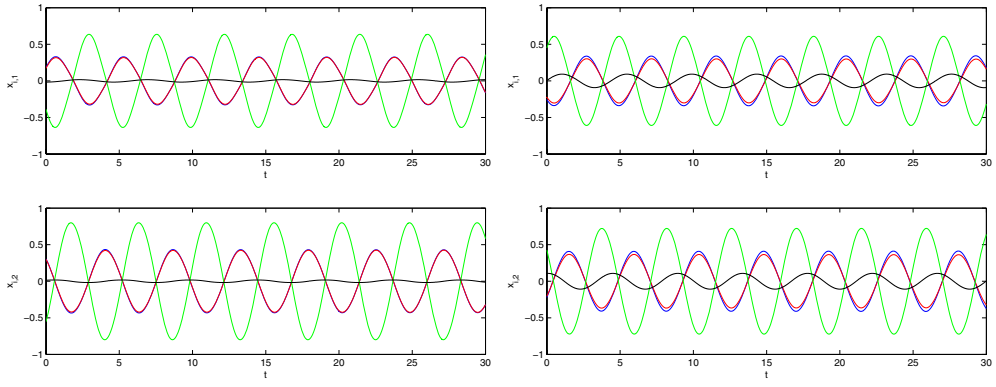


Figure 5. *Solutions with $\mathbf{Z}_2$ (interior) symmetry. (Left) Network $\mathcal{G}_1$ with exact $\mathbf{S}_3$-symmetry. (Right) Network $\mathcal{G}_2$ with $\mathbf{S}_3$-interior symmetry.*

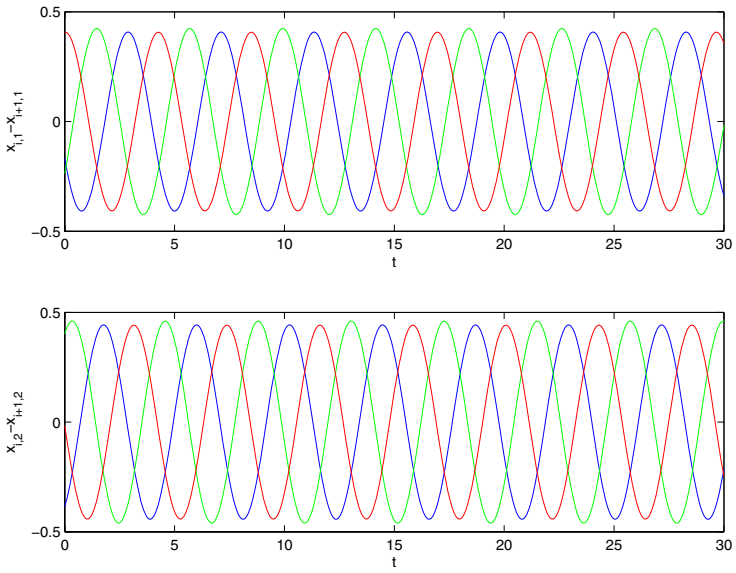


Figure 6. *Approximate rotating wave in network $\mathcal{G}_2$, viewed in difference coordinates: $x_1 - x_2$, $x_2 - x_3$, and $x_3 - x_1$.*

REFERENCES

[1] F. Antoneli and I. Stewart, *Symmetry and synchrony in coupled cell networks* 1: *Fixed-point spaces*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 16 (2006), pp. 559–577.

[2] F. Antoneli and I. Stewart, *Symmetry and synchrony in coupled cell networks* 2: *Group networks*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 935–951.

[3] M. Golubitsky, M. Nicol, and I. Stewart, *Some curious phenomena in coupled cell systems*, J. Nonlinear Sci., 14 (2004), pp. 119–236.

[4] M. Golubitsky, M. Pivato, and I. Stewart, *Interior symmetry and local bifurcation in coupled cell networks*, Dyn. Syst., 19 (2004), pp. 389–407.

[5] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Vol. 1, Appl. Math. Sci. 51, Springer-Verlag, New York, 1985.

[6] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, Progr. Math. 200, Birkhäuser, Basel, 2002.

[7] M. Golubitsky and I. Stewart, *Nonlinear dynamics of networks: The groupoid formalism*, Bull. Amer. Math. Soc. (N.S.), 43 (2006), pp. 305–364.

[8] M. Golubitsky, I. N. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Vol. 2, Appl. Math. Sci. 69, Springer-Verlag, New York, 1988.

[9] M. Golubitsky, I. Stewart, and A. Török, *Patterns of synchrony in coupled cell networks with multiple arrows*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 78–100.

[10] I. Stewart, *The lattice of balanced equivalence relations of a coupled cell network*, Math. Proc. Cambridge Philos. Soc., 143 (2007), pp. 165–183.

[11] I. Stewart, M. Golubitsky, and M. Pivato, *Symmetry groupoids and patterns of synchrony in coupled cell networks*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 609–646.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 249–264
© 2008 Society for Industrial and Applied Mathematics

Surface Gap Solitons at a Nonlinearity Interface*

Tomáš Dohnal† and Dmitry Pelinovsky‡

Abstract. We demonstrate existence of waves localized at the interface of two nonlinear periodic media with different coefficients of the cubic nonlinearity via the one-dimensional Gross–Pitaevsky equation. We call these waves the surface gap solitons (SGSs). In the case of smooth symmetric periodic potentials, we study analytically bifurcations of SGSs from standard gap solitons and determine numerically the maximal jump of the nonlinearity coefficient allowing for SGS existence. We show that the maximal jump vanishes near the thresholds of bifurcations of gap solitons. In the case of continuous potentials with a jump in the first derivative at the interface, we develop a homotopy method of continuation of SGS families from the solution obtained via gluing of parts of the standard gap solitons and study existence of SGSs in the photonic band gaps. We explain the termination of the SGS families in the interior points of the band gaps from the bifurcation of linear bound states in continuous nonsmooth potentials.

Key words. surface gap solitons, Gross–Pitaevsky equation, nonlinear Schrödinger equation, nonlinearity interface

AMS subject classifications. 35Q60, 37J45, 78A40

DOI. 10.1137/060676751

1. Introduction.

We are concerned with localized waves at the interface of two periodic nonlinear media called surface gap solitons (SGSs). One of the first publications on optical solitons propagating along material interfaces is [17], where the interface of a linear and a focusing Kerr nonlinear medium is studied. In the last two years relevant publications in the context of nonlinear optics have dealt, for instance, with discrete surface solitons in nonlinear waveguide arrays [4, 10, 16], SGSs at the interface of a uniform and a periodic medium with the defocusing cubic nonlinearity [7], and surface vortex solitons at the interface of two periodic media with different mean values of the refractive index and with saturable nonlinearity [5, 6]. One of the typical models employed in the theory of gap solitons is the one-dimensional nonlinear Schrödinger (NLS) equation with cubic nonlinearity and periodic potential called the Gross–Pitaevsky equation.

We investigate here the existence of surface waves at the interface of two media with identical periodic linear parts of the refractive index and with different cubic nonlinearities. It is known that for most photonic materials a variation in the nonlinear part of the refractive index n_2 is necessarily accompanied by a larger change in the linear part n_0. Nevertheless,

*Received by the editors December 5, 2006; accepted for publication (in revised form) by B. Sandstede September 5, 2007; published electronically April 23, 2008.
http://www.siam.org/journals/siads/7-2/67675.html
†Seminar for Applied Mathematics, ETH Zürich, Zürich, Switzerland (dohnal@math.ethz.ch). This author is supported by an ETH Research Fellowship.
‡Department of Mathematics, McMaster University, Hamilton, Ontario, Canada, L8S 4K1 (dmpeli@math.mcmaster.ca). This author is supported by the Humboldt Research Fellowship hosted at Institut für Analysis, Dynamik und Modellierung, Fakultaet für Mathematik und Physik at Universität Stuttgart.

certain materials exhibit large variations in n_2 accompanied by small variations in n_0; see [1, 15]. Localized states have been studied theoretically in media with constant n_0 and spatially periodic n_2 in [3].

Each of the two periodic nonlinear media supports at least two families of standard gap solitons in every bounded nonempty frequency gap. One family is always unstable, while the other can be stable depending on the locations of spectral bands and bifurcations of eigenvalues from the band edges [13]. The potentially stable family looks like a single-humped envelope soliton with exponential decay and oscillations near the central peak. Multihumped envelope solitons may also exist in such periodic nonlinear media, but we shall focus herein on existence of a single-humped solution localized near the interface between the two periodic nonlinear media.

The paper is organized as follows. Section 2 reviews Floquet theory for the governing Gross–Pitaevsky equation and summarizes the results on existence of gap solitons. In section 3 we study the existence of SGSs for a smooth symmetric periodic potential function and find the maximal allowed jump in the nonlinearity coefficient between the two media for existence of SGSs. Section 4 discusses bifurcations and existence of SGSs for a continuous potential function with a derivative jump at the nonlinearity interface. Section 5 concludes the paper with conjectures on the stability of SGSs.

2. Background: Floquet theory and gap soliton existence. We consider the one-dimensional periodic cubic Schrödinger equation in the form

$$iu_t = -u_{xx} + V(x)u - \Gamma(x)|u|^2 u, \quad x \in \mathbb{R}, \quad t \geq 0, \tag{2.1}$$

where x and t are the spatial and temporal variables, respectively, $V(x)$ is a real, continuous, and d-periodic potential, and $\Gamma(x) = \Gamma_\pm$ for $\pm x > 0$ is a real nonlinearity coefficient with constants Γ_+ and Γ_-. The positive values of $\Gamma(x)$ correspond to the focusing nonlinearity and the negative values of $\Gamma(x)$ to the defocusing nonlinearity.

We are interested in the existence of stationary solutions of (2.1) localized near the interface at $x = 0$ and having the form

$$u(x,t) = e^{-i\omega t}\phi(x) \quad \text{s.t.} \quad \phi : \mathbb{R} \to \mathbb{R}, \quad \phi \to 0 \ \text{ as } \ |x| \to \infty. \tag{2.2}$$

The function $\phi(x)$ has to satisfy the second-order nonautonomous ODE

$$-\phi'' - \omega\phi + V(x)\phi - \Gamma(x)\phi^3 = 0, \tag{2.3}$$

which can be cast in the Hamiltonian form with the Hamiltonian function

$$H[\phi] = \frac{1}{2}\left[(\phi')^2 + \omega\phi^2 - V(x)\phi^2\right] + \frac{1}{4}\Gamma(x)\phi^4. \tag{2.4}$$

Since $\Gamma(x)$ is discontinuous at $x = 0$, $\phi(x)$ is a weak solution of the ODE (2.3) in $\phi \in C^2(\mathbb{R}_+ \cup \mathbb{R}_-)$, such that the second derivative $\phi''(x)$ may have a jump at $x = 0$. The continuously differentiable solution $\phi \in C^1(\mathbb{R})$ is a critical point of the energy functional

$$E_\omega[\phi] = \frac{1}{2}\int_{\mathbb{R}} \left[|\phi'|^2 + \omega|\phi|^2 - V(x)|\phi|^2\right] dx + \frac{1}{4}\int_{\mathbb{R}} \Gamma(x)|\phi|^4 dx,$$

such that the first variation $E'_\omega[\phi]$ recovers the ODE (2.3).

Replacing t by z in (2.1) and (2.2), the x-localized solution $u(x,z)$ can be viewed as a spatial soliton propagating along the direction z and localized in the transverse direction x. The parameter ω plays the role of the propagation constant. Other applications of the ODE (2.3) occur in the theory of stationary solutions of the nonlinear Maxwell and Klein–Gordon equations in one dimension.

As we show below, the localized solutions of the ODE (2.3) decay exponentially as $|x| \to \infty$ only if ω belongs to the frequency gaps in the continuous spectra of the operator $L := -\partial_{xx} + V(x)$ called the photonic band gaps. To do so, we recall the basic Floquet theory (see [2, 9]) for the Hill equation

$$L\psi(x) = -\psi''(x) + V(x)\psi(x) = \omega\psi(x), \qquad x \in \mathbb{R}. \tag{2.5}$$

The bounded solutions $\psi(x)$ of the Hill equation (2.5) are usually called Bloch functions. Given a real, continuous, and d-periodic potential $V(x)$, bounded solutions $\psi(x)$ exist for ω in a union of (possibly disjoint) spectral bands

$$\Sigma := [\omega_0, \omega_1] \cup [\omega_2, \omega_3] \cup [\omega_4, \omega_5] \cup \dots,$$

where $\omega_{2n-2} < \omega_{2n-1} \leq \omega_{2n}$, $n \in \mathbb{N}$, and $\omega_n \to \infty$ as $n \to \infty$. The set Σ represents the complete (purely continuous) spectrum of the operator L [2]. We shall *assume* for simplicity that all spectral bands are disjoint with $\omega_{2n-1} < \omega_{2n}$, $n \in \mathbb{N}$, such that all finite frequency gaps are nonempty.

For a fixed ω in the interior point of Σ, both fundamental solutions of the second-order ODE (2.5) are quasi-periodic in x and have the representation $\psi = p_\pm(x)e^{\pm ikx}$, where $p_\pm(x) = p_\pm(x+d)$ and $k \in \left[0, \frac{\pi}{d}\right]$. The parameter k parameterizes the frequency parameter ω, such that we shall use the notation $\omega = \omega_{2n,2n+1}(k)$ for the spectral band in $\omega \in [\omega_{2n}, \omega_{2n+1}]$. If the nth band is separated from the $(n+1)$th band (i.e., $\omega_{2n-1} < \omega_{2n}$ and $\omega_{2n+1} < \omega_{2n+2}$), then $\omega'_{2n,2n+1}(k) = 0$ and $\omega''_{2n,2n+1}(k) \neq 0$ at the endpoints $k = 0$ and $k = \frac{\pi}{d}$ [8].

When $\omega = \omega_n$, one of the solutions $\psi = \psi_n(x)$ is either d-periodic (corresponding to $k = 0$) or d-antiperiodic (corresponding to $k = \frac{\pi}{d}$), and the other fundamental solution $\psi(x)$ grows linearly in x. For a fixed $\omega \in \mathbb{R} \setminus \Sigma$ the two fundamental solutions of (2.5) grow exponentially in either x or $-x$ and have the representation $\psi = u_\pm(x)e^{\pm\kappa x}$, where $u_\pm(x)$ is either periodic or antiperiodic and $\kappa = \kappa(\omega) \in \mathbb{R}_+$. The functions $u_\pm(x)$ are periodic (antiperiodic) if the bounded solutions $\psi_n(x)$ are periodic (antiperiodic) at the band edges ω_{2n-1} and ω_{2n}, which surround the band gap.

Suppose that $\phi(x)$ is a localized solution of the ODE (2.3). It is then obvious from the linearized analysis that the solution $\phi(x)$ decays exponentially as $|x| \to \infty$ only if $\omega \in \mathbb{R} \setminus \Sigma$. It was shown under fairly general assumptions (see [13] and references therein) that the families of gap solitons of the ODE (2.3) with constant coefficient $\Gamma(x) = \Gamma_0$ undertake a local bifurcation from all points $\omega = \omega_{2m}$, $m \geq 0$, to the left if $\Gamma_0 > 0$ and from all points $\omega = \omega_{2m+1}$, $m \geq 0$, to the right if $\Gamma_0 < 0$ (the term *local bifurcation* means that $\|\phi\|_{L^\infty} \to 0$ as $\omega \to \omega_n$). This conjecture was rigorously proved in [11], where existence of exponentially decaying gap solitons in $H^1(\mathbb{R})$ was confirmed in every finite frequency gap $\omega \in (\omega_{2m-1}, \omega_{2m})$, $m \in \mathbb{N}$, and in the semi-infinite frequency gap $\omega < \omega_0$ for $\Gamma_0 > 0$. We use this result but

simplify our consideration by working with the class of symmetric potentials $V = V_0(x)$, where $V_0(-x) = V_0(x)$ on $x \in \mathbb{R}$. In particular, we shall perform numerical computations with

$$V_0(x) = \sin^2\left(\frac{\pi x}{d}\right), \quad d = 10, \tag{2.6}$$

which has a minimum at $x = 0$, i.e., at our interface location. The spectral bands and gaps of $V_0(x)$ are approximated numerically from the Hill equation (2.5). For instance, the first five band edges of the potential (2.6) are located as follows:

$$\omega_0 \approx 0.283, \quad \omega_1 \approx 0.291, \quad \omega_2 \approx 0.747, \quad \omega_3 \approx 0.843, \quad \omega_4 \approx 1.057.$$

As seen in Figure 1 of [13], the Bloch functions $\psi = \psi_n(x)$ at the band edges $\omega = \omega_n$, $n \geq 0$, have the following symmetry properties:

$$\begin{aligned} &\psi_n(-x) = \psi_n(x), \ n \in \{0, 1, 4, 5, 8, 9, \ldots\}, \\ &\psi_n(-x) = -\psi_n(x), \ n \in \{2, 3, 6, 7, \ldots\}. \end{aligned} \tag{2.7}$$

Symmetry properties (2.7) can be proved for any even potential $V_0(-x) = V_0(x)$. Indeed, since the Hill equation (2.5) is symmetric with respect to reflection $x \mapsto -x$ and admits only one linearly independent bounded eigenfunction $\psi = \psi_n(x)$ at $\omega = \omega_n$, the function $\psi_n(x)$ must be either even or odd in x. By the trace of the monodromy matrix [2], the periodic functions $\psi_n(x)$ correspond to the set $n \in S_+$ with $S_+ = \{0, 3, 4, 7, 8, \ldots\}$, and the antiperiodic functions $\psi_n(x)$ correspond to the set $n \in S_-$ with $S_- = \{1, 2, 5, 6, \ldots\}$. By Sturm's theorem [2], the periodic functions $\psi_n(x)$ with $n \in S_+$ have exactly $\text{ind}_{S_+}(n) - 1$ nodes on $x \in \left(-\frac{d}{2}, \frac{d}{2}\right)$, where $\text{ind}_{S_+}(n)$ is the order number of n in the set S_+. For instance, $\psi_0(x)$ has no nodes (positive definite), $\psi_3(x)$ has one node, $\psi_4(x)$ has two nodes, etc. Combining the symmetry with respect to reflections and the number of nodes, we conclude that the set of eigenfunctions $\{\psi_n(x)\}_{n \in S_+}$ alternates the symmetry in x, such that $\psi_0(x)$ is even, $\psi_3(x)$ is odd, $\psi_4(x)$ is even, etc. Similarly, the antiperiodic functions $\psi_n(x)$ with $n \in S_-$ have exactly $\text{ind}_{S_-}(n) - 1$ nodes on $x \in \left(-\frac{d}{2}, \frac{d}{2}\right)$. For instance, $\psi_1(x)$ has no nodes, $\psi_2(x)$ has one node, etc. We conclude again that the set of eigenfunctions $\{\psi_n(x)\}_{n \in S_-}$ alternates the symmetry in x, such that $\psi_1(x)$ is even, $\psi_2(x)$ is odd, etc.

Altogether, this set of facts is summarized in Table 1.

Table 1

Properties of the Bloch functions $\psi_n(x)$ and gap soliton bifurcations at the first eight band edges of an even potential $V_0(-x) = V_0(x)$.

n	0	1	2	3	4	5	6	7
symmetry	even	even	odd	odd	even	even	odd	odd
periodicity	S_+	S_-	S_-	S_+	S_+	S_-	S_-	S_+
# nodes on $\left(-\frac{d}{2}, \frac{d}{2}\right)$	0	0	1	1	2	2	3	3
sign of Γ_0 for local bifurcation	1	-1	1	-1	1	-1	1	-1

Let $\phi_0(x)$ be a single-humped solution of the ODE (2.3) with $\Gamma(x) = \Gamma_0$ and $V(x) = V_0(x)$ which bifurcates from the band edge $\omega = \omega_n$. By the local bifurcation theory [13], it inherits the symmetry properties (2.7) of the Bloch function $\psi_n(x)$. Therefore, $\phi_0(-x) = \phi_0(x)$ for branches of gap solitons to the left of ω_n with $n = \{0, 4, 8, \ldots\}$ (for $\Gamma_0 > 0$) and to the right

of ω_n with $n = \{1, 5, 9, \ldots\}$ (for $\Gamma_0 < 0$), while $\phi_0(-x) = -\phi_0(x)$ for branches of gap solitons to the left of ω_n with $n = \{2, 6, \ldots\}$ (for $\Gamma_0 > 0$) and to the right of ω_n with $n = \{3, 7, \ldots\}$ (for $\Gamma_0 < 0$). See Figures 2–3 in [13] for gap solitons $\phi_0(x)$ in the potential (2.6).

In this paper, we shall consider the existence of SGSs in the ODE (2.3) with piecewise constant coefficient $\Gamma(x) = \Gamma_\pm$ for $\pm x > 0$ and potential $V(x)$ of the following two classes:

$$\text{(2.8)} \qquad \text{(i) } V = V_0(x), \qquad \text{(ii) } V = V_0(x-\delta)\chi_{(-\infty,0)} + V_0(x+\delta)\chi_{[0,\infty)},$$

where $\chi_{[a,b]} = 1$ on $x \in [a, b]$ and zero otherwise, while $0 < \delta < d$. Here $V_0(x)$ is a smooth, even, d-periodic function on $x \in \mathbb{R}$. We note that $V(x)$ in (ii) is continuous and even on $x \in \mathbb{R}$ but smooth and periodic only on each $\pm x > 0$.

One can develop a general shooting method for numerical approximations of SGSs from the condition that a localized solution $\phi(x)$ of the second-order ODE (2.3) with $\omega \in (\omega_{2m-1}, \omega_{2m})$, $m \in \mathbb{N}$, decays to zero at infinity according to two fundamental solutions $p_\pm(x)e^{\mp\kappa x}$ as $x \to \pm\infty$, where $\kappa = \kappa(\omega)$ is a positive number. Solving the ODE (2.3) with $\Gamma(x) = \Gamma_+$ for a general initial value $\phi(0)$ and $\phi'(0)$ to $x > 0$ and the same ODE with $\Gamma(x) = \Gamma_-$ to $x < 0$, one can construct a continuously differentiable solution $\phi(x)$ on $x \in \mathbb{R}$ which decays to zero as $x \to \pm\infty$ if and only if the projections to the growing fundamental solutions $p_\pm(x)e^{\pm\kappa x}$ are zero at infinity. The system of two constraints for two initial values constitutes a well-posed problem of numerical analysis. This numerical approach was adopted in recent work [18]. Practical implementations of this algorithm are unclear as the shooting method may depend sensitively on starting approximations of the initial value and may require long computational time to search through all appropriate initial values. In addition, the ODE solvers of the shooting method may develop numerical instabilities in approximations of growing solutions.

Due to these reasons, we shall develop an alternative view on numerical approximations of SGSs, starting with local bifurcation analysis and using the homotopy continuation method to trace the solution families along parameters ω, $\Gamma_\pm$, and δ. Using these analytical and numerical results, we have obtained the following main results.

(1) We prove analytically that any gap soliton for $\Gamma_+ = \Gamma_-$ can be continued to the SGS for sufficiently small $|\Gamma_+ - \Gamma_-|$ under a nondegeneracy assumption.
(2) We prove analytically that the maximal difference $|\Gamma_+ - \Gamma_-|$ leading to SGS existence converges to 0 when ω approaches the band edge which features the local bifurcation of a gap soliton.
(3) SGSs are computed numerically when the potential $V(x)$ is given by (2.8)(i), and the maximal $|\Gamma_+ - \Gamma_-|$ allowing their existence is found. Our numerical results confirm the analytical results (1)–(2) above.
(4) Existence of SGSs for $V(x)$ in (2.8)(ii) with $\Gamma_+ > 0$ and $\Gamma_- < 0$ is studied. We numerically show that local bifurcations may occur from a countable set of points in the parameter domain $(\omega, \delta) \in (\omega_{2m-1}, \omega_{2m}) \times (0, d)$, $m \in \mathbb{N}$.
(5) We numerically compute the points of local bifurcation of SGSs for the potential (2.8)(ii) and use the homotopy continuation of the bifurcating solution. As a result, we show that the family of SGSs exists typically in a subset of the plane (ω, δ).
(6) We analytically show that the termination of families of SGSs for the potential (2.8)(ii) is related to existence of linear bound states for the nonsmooth potential.

Results (1)–(3) are reported in section 3, and results (4)–(6) are described in section 4.

3. Bifurcations of SGSs for smooth potentials. In this section we study continuation of SGSs from gap solitons existing for $\Gamma_+ = \Gamma_-$ in the case of a smooth potential function $V(x)$. A prototypical example of such potential is the symmetric function $V_0(x)$ in (2.8)(i).

3.1. Existence of bifurcations from gap solitons. Let $\gamma = (\Gamma_+ + \Gamma_-)/2$ and $\nu = (\Gamma_+ - \Gamma_-)/2$. Then, the ODE (2.3) can be rewritten in the form

$$F(\phi, \nu) = -\phi'' - \omega\phi + V(x)\phi - \gamma\phi^3 - \nu\,\mathrm{sign}(x)\phi^3 = 0, \tag{3.1}$$

where $F(\phi, \nu) : H^1(\mathbb{R}) \times \mathbb{R} \mapsto H^{-1}(\mathbb{R})$ is a nonlinear operator acting on a function $\phi(x)$ in space $H^1(\mathbb{R})$ and parameter $\nu \in \mathbb{R}$.

We assume that there exists a solution $\phi_0(x) \in H^1(\mathbb{R})$ for $\omega \in \mathbb{R} \setminus \Sigma$ and some γ and $V(x)$, such that $F(\phi_0, 0) = 0$. The Jacobian $D_\phi F(\phi_0, 0)$ is given by the Schrödinger operator $\mathcal{L} : H^2(\mathbb{R}) \mapsto L^2(\mathbb{R})$, where

$$\mathcal{L} = -\partial_x^2 - \omega + V(x) - 3\gamma\phi_0^2(x). \tag{3.2}$$

Since $\omega \in \mathbb{R}\setminus\Sigma$, we have $\phi_0^2(x) \to 0$ exponentially fast as $|x| \to \infty$, such that the term $-3\gamma\phi_0^2(x)$ is a relatively compact perturbation to the unbounded operator $L - \omega$, where $L = -\partial_x^2 + V(x)$. By a standard argument (see Corollary 2 in section XIII.4 in [14]), the essential spectrum of $\mathcal{L}$ and $(L - \omega)$ coincide. Since $\omega \in \mathbb{R} \setminus \Sigma$, the zero point is isolated from the essential spectrum of $\mathcal{L}$. If we further assume that $\mathcal{L}$ has the trivial kernel in $H^1(\mathbb{R})$, then $\mathcal{L}$ is invertible on $L^2(\mathbb{R})$. Since the translational invariance is broken if $V(x) \neq 0$, $\mathcal{L}$ generally has the trivial kernel, unless a bifurcation of branches of gap solitons occur. By the standard analysis based on the implicit function theorem, there exists a unique smooth continuation of $\phi_\nu(x)$ from $\phi_0(x)$ in $H^1(\mathbb{R})$ for sufficiently small ν, such that $F(\phi_\nu, \nu) = 0$ and $\phi_\nu(x) \to \phi_0(x)$ in $H^1(\mathbb{R})$ as $\nu \to 0$.

In other words, we have proved above that if a gap soliton exists for $\Gamma_+ = \Gamma_-$ and $\omega \in \mathbb{R} \setminus \Sigma$ and the linearized operator $\mathcal{L}$ is nondegenerate, then the gap soliton is uniquely continued into the SGS for small nonzero $|\Gamma_+ - \Gamma_-|$. We confirm this prediction via numerical analysis of the ODE (2.3) with $V(x)$ in (2.8)(i) for ω taken in the semi-infinite band gap and the first two finite gaps. Numerical approximations of $\phi_0(x)$ for $\Gamma_+ = \Gamma_-$ are obtained from the Newton–Raphson iterations and the homotopy continuation method. The initial guess for the Newton iteration is taken from an asymptotic expansion leading to the NLS approximation [13] when ω is close to the local bifurcation threshold ω_n. After a successful convergence for one such ω we use a standard homotopy continuation and generate a family of gap solitons $\phi_0(x)$ parameterized by ω. The discretization of the ODE (2.3) is based on a fourth-order central difference approximation of ∂_{xx} on a truncated domain with zero Dirichlet boundary conditions.

3.2. Numerical computations of SGSs. We now proceed to construct SGSs, i.e., solutions $\phi(x)$ of the second-order ODE (2.3) with $\Gamma_+ \neq \Gamma_-$. When $\phi_0(x)$ is obtained for a given value of ω, we can apply the numerical homotopy continuation of the solution by deviating Γ_- from Γ_+. At each step, the SGS $\phi(x)$ is thus found via Newton's iterations. The final value of Γ_-, up to which the iteration converges, is denoted by Γ_*.

Figure 1 shows the values of Γ_* for $\Gamma_+ = +1$ (a) and $\Gamma_+ = -1$ (b). The computational tolerance in Γ_* is 0.006 inside the band gaps and 0.002 near the band edges. In the case

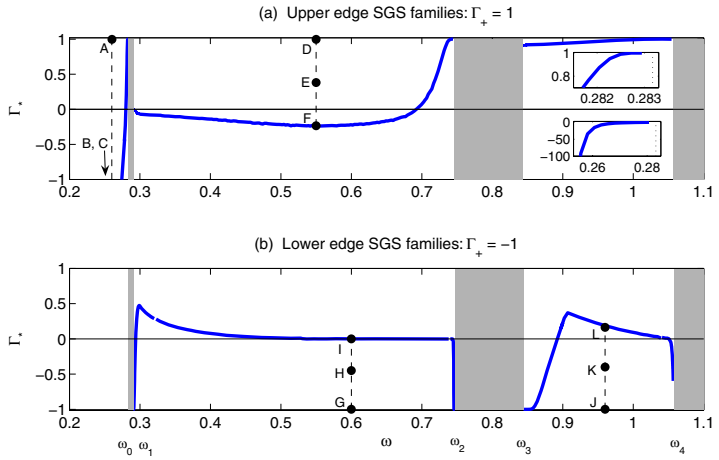


Figure 1. *The values of* Γ_* *for SGSs originating from symmetric GS families of the first three frequency gaps of* $V(x) = \sin^2(\pi x/10)$. *In* (a) *the upper inset zooms in and the lower inset zooms out on the graph in the semi-infinite gap. The points A–L are referenced in Figure 2.*

$\Gamma_+ = 1$, local bifurcations of small-amplitude gap solitons occur from the lower band edges. Figure 1(a) shows that the SGSs exist in the semi-infinite gap, as well as in the first two frequency gaps. In the case $\Gamma_+ = -1$, local bifurcations of gap solitons occur from the upper band edges. Figure 1(b) shows that the SGSs exist in the first and second frequency gaps. The two insets of Figure 1(a) show that Γ_* decreases quickly as ω moves away from the edge of the first band and that the convergence $\Gamma_* \uparrow 1$ as $\omega \uparrow \omega_0$ is smooth. We further see from Figure 1 that the interval of existence shrinks as ω approaches the value ω_n for any band edge, where gap solitons undertake a local bifurcation. In addition, the interval of existence is extremely large in the semi-infinite gap $(-\infty, \omega_0)$, but it becomes narrow in the finite gaps $(\omega_{2m-1}, \omega_{2m})$ for $m \geq 1$.

For comparison, the family of SGSs in the gap (ω_1, ω_2) exists for $-0.24 < \Gamma_* < 1$ in the case $\Gamma_+ = +1$ and $-1 < \Gamma_* < 0.47$ in the case $\Gamma_+ = -1$. The family of SGSs in the gap (ω_3, ω_4) exists in a very narrow region of $0.92 < \Gamma_* < 1$ in the case $\Gamma_+ = +1$ and in a bigger interval $-1 < \Gamma_* < 0.37$ in the case $\Gamma_+ = -1$ (similarly to that in the first gap).

Figure 2 shows profiles of SGSs which correspond to the twelve points labeled A–L in Figure 1. The solid lines correspond to the gap solitons from which the homotopy in Γ_- is started (i.e., points A, D, G, and J). Clearly, the total power and maximum amplitude of the SGSs increase as $|\Gamma_+ - \Gamma_-|$ increases. Also notice that the profiles become more concentrated on the half $x > 0$ in the case $\Gamma_+ = +1$ (see Figure 2 (a–b)) and on the half $x < 0$ in the case $\Gamma_+ = -1$ (see Figure 2 (c–d)) as $|\Gamma_+ - \Gamma_-|$ increases. This is in accord with the law of refraction: when $\Gamma_+ = +1$ and Γ_- decreases from 1, the half $x > 0$ becomes relatively more focusing and therefore attracts more energy of the soliton, while when $\Gamma_+ = -1$ and Γ_- increases from -1, the situation is the opposite.

3.3. Asymptotic analysis near gap soliton bifurcation points. Now we shall explain why the existence interval shrinks to zero when ω approaches the value ω_n where a local bifurcation

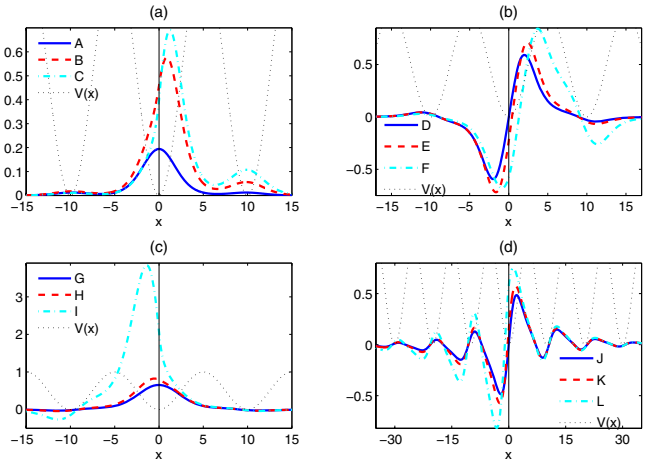


Figure 2. *The profiles of SGSs corresponding to the points A–L in Figure 1. Values of ω are A–C:* 0.26, *D–F:* 0.55, *G–I:* 0.6, *J–L:* 0.96. *Values of Γ_- are A:* 1, *B:* -3.9, *C:* -15.3, *D:* 1, *E:* 0.38, *F:* -0.235, *G:* -1, *H:* -0.45, *I:* 0.002, *J:* -1, *K:* -0.4, *L:* 0.164.

of gap solitons occurs. As $\omega \to \omega_n$, we have $\|\phi_0\|_{L^\infty} \to 0$ and $\mathcal{L} \to (L-\omega_n)$. Since the operator $(L-\omega_n)$ is not invertible, the implicit function theorem cannot be used and the solution $\phi_0(x)$ cannot be continued beyond $\nu = 0$. In order to give a more precise explanation of this phenomenon, we adopt the NLS approximation for local bifurcation of gap solitons from [13] (see also the review in [12]). In particular, we consider an asymptotic solution to the ODE (2.3),

$$\begin{aligned} &\omega = \omega_n + \varepsilon^2\,\Omega + \mathcal{O}(\varepsilon^4), \\ &\phi(x) = \varepsilon\,A(X)\psi_n(x) + \varepsilon^2\,A'(X)\tilde{\psi}_n(x) + \varepsilon^3\,\phi^{(3)}(x,X) + \mathcal{O}(\varepsilon^4), \end{aligned} \tag{3.3}$$

where $X = \varepsilon\,x$, $\varepsilon << 1$, the function $A(X)$ and parameter Ω are defined below, and ψ_n and $\tilde{\psi}_n$ are the d-periodic (or d-antiperiodic) Bloch functions and generalized Bloch functions, respectively, of the Hill equation (2.5) for $\omega = \omega_n$, such that

$$(L-\omega_n)\psi_n = 0, \qquad (L-\omega_n)\tilde{\psi}_n = 2\psi_n'. \tag{3.4}$$

The correction term $\phi^{(3)}(x,X)$ at $\mathcal{O}(\varepsilon^3)$ solves the nonhomogeneous problem

$$(L-\omega_n)\phi^{(3)} = \Omega A\psi_n + A''\psi_n + 2A''\tilde{\psi}_n' + \Gamma(X)A^3\psi_n^3. \tag{3.5}$$

To ensure boundedness of $\phi^{(3)}(x,X)$ with respect to the variable x, and, hence, legitimacy of the expansion (3.3), one has to apply the Fredholm alternative which imposes the orthogonality condition of the right-hand side of (3.5) with respect to $\psi_n(x)$ on $x \in [0,d]$. The orthogonality condition is written as

$$\Omega A + \mu A'' + \rho\Gamma(X)A^3 = 0, \tag{3.6}$$

where

$$\mu = 1 + 2\frac{(\tilde{\psi}_n', \psi_n)}{(\psi_n, \psi_n)}, \qquad \rho = \frac{(\psi_n^2, \psi_n^2)}{(\psi_n, \psi_n)},$$

and we have used the standard L^2 inner product $(\cdot, \cdot)$ over one period $x \in [0, d]$. It is shown in [13] that $\mu = \frac{1}{2}\omega''_{2n,2n+1}(k)$ with either $k = 0$ or $k = \frac{\pi}{d}$ at the point $\omega = \omega_n$, where $\omega_{2n,2n+1}(k)$ is the dispersion relation between $\omega \in [\omega_{2n}, \omega_{2n+1}]$ and $k \in [0, \frac{\pi}{d}]$.

Due to the nature of the nonlinearity interface, the function $\Gamma(X)$ is the same as $\Gamma(x)$, i.e., $\Gamma(X) = \Gamma_\pm$ for $\pm X > 0$. We shall prove that no localized solution of the ODE (3.6) exists under the condition $\Gamma_- \neq \Gamma_+$. Indeed, consider the Hamiltonian of the ODE (3.6):

$$H[A] = \frac{1}{2}\left[\mu (A')^2 + \Omega A^2\right] + \frac{1}{4}\rho\Gamma(X)A^4. \tag{3.7}$$

If $A(X)$ solves the ODE (3.6), then

$$\frac{d}{dX}H[A(X)] = \frac{1}{4}\rho\Gamma'(X)A^4(X) = \frac{1}{4}\rho(\Gamma_+ - \Gamma_-)\delta(X)A^4(X),$$

where $\delta(X)$ is the Dirac delta-function. If $A(X)$ is a localized solution on $X \in \mathbb{R}$, then the integration on $X \in \mathbb{R}$ gives the constraint

$$0 = \lim_{x\to+\infty} H[A(X)] - \lim_{x\to-\infty} H[A(X)] = \frac{1}{4}\rho(\Gamma_+ - \Gamma_-)A^4(0),$$

since $H[A(X)] \to 0$ if $A(X), A'(X) \to 0$ as $|X| \to \infty$. Therefore, $A(0) = 0$ if $\Gamma_+ \neq \Gamma_-$. Consider now $H[A(X)]$ on $X > 0$. It is clear from the decaying conditions as $X \to \infty$ that $H[A(X)] = \text{const} = 0$, which together with the fact that $A(0) = 0$ leads to $0 = \lim_{X\downarrow 0} H[A(X)] = \frac{1}{2}\mu|A'(0)|^2$, such that $A'(0) = 0$. The only solution of the ODE (3.6) with $A(0) = A'(0) = 0$ is the zero solution $A(X) \equiv 0$.

If $\Gamma_+ = \Gamma_- = \Gamma_0$ and $\text{sign}(\mu) = \text{sign}(\rho\Gamma_0) = -\text{sign}(\Omega)$, the ODE (3.6) has the standard sech-soliton decaying as $|X| \to \infty$. However, the result above shows that the sech-soliton with $\Gamma_+ = \Gamma_-$ cannot be homotopically continued to a decaying solution of (3.6) for $\Gamma_+ \neq \Gamma_-$. This proves that $\Gamma_* \to \Gamma_+$ as $\omega \to \omega_n$, where ω_n is a local bifurcation value.

4. Bifurcations of SGSs for nonsmooth potentials.

In this section, we study local bifurcations of solutions of the ODE (2.3) when $V(x)$ is a continuous function with the jump in the first derivative at the nonlinearity interface. The prototypical example of such potentials is given by (2.8)(ii), where $V_0(x)$ is an even potential (in our numerical computations we use V_0 from (2.6)). We shall consider the existence of SGSs under the normalization $\Gamma_+ = -\Gamma_- = +1$.

4.1. SGS numerical construction via gluing.

The point (δ_*, ω_*) in the parameter domain $\delta \in (0, d)$ and $\omega \in (\omega_{2m-1}, \omega_{2m})$, $m \in \mathbb{N}$, is defined to be a point of a local bifurcation of SGSs according to the following two-step algorithm.

(i) *Construction of continuous solutions.* Let $\phi_\pm(x; \omega)$ denote the family of single-humped gap solitons parameterized by $\omega \in (\omega_{2m-1}, \omega_{2m})$ and centered at $x = 0$ corresponding to (2.3) with $\Gamma(x) \equiv \Gamma_\pm$, respectively. These families bifurcate from the points $\omega = \omega_{2m}$ for $\Gamma_+ > 0$

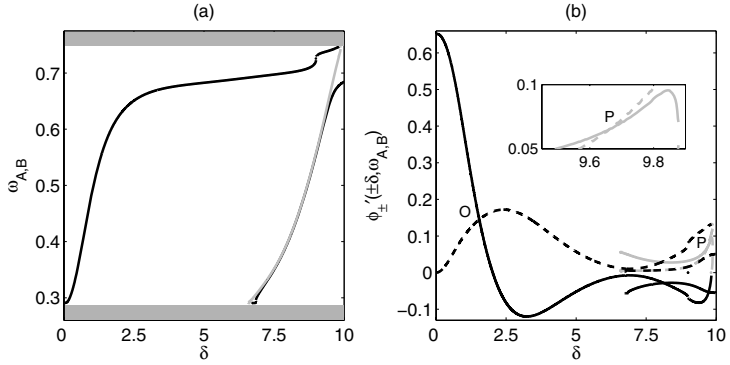


Figure 3. *Two-step search for (ω_*, δ_*) in the gap (ω_1, ω_2). (a) Result of step* (i)*—parametrization of the families of continuous solutions* (4.1)*: black line $\omega_A(\delta)$, gray line $\omega_B(\delta)$. (b) Step* (ii)*—search for δ_*: solid black $\phi_+'(\delta;\omega_A)$, dashed black $\phi_-'(-\delta;\omega_A)$, solid gray $-\phi_+'(\delta;\omega_B)$, and dashed gray $\phi_-'(-\delta;\omega_B)$. Labeled points correspond to SGSs.*

and $\omega = \omega_{2m-1}$ for $\Gamma_- < 0$. In order to find continuous solutions, we now study for each fixed $\delta \in (0, d)$ the two functions

$$f_A(\omega) = \phi_-(-\delta;\omega) - \phi_+(\delta;\omega), \quad f_B(\omega) = \phi_-(-\delta;\omega) + \phi_+(\delta;\omega)$$

and find their zeros denoted by $\omega_{A,B} = \omega_{A,B}(\delta)$, respectively. For each δ existence of zeros of either $f_A(\omega)$ or $f_B(\omega)$ is guaranteed by continuity of $\phi_\pm$ as functions of ω and by the fact that $\phi_-(-\delta;\omega_{2m-1}) = \phi_+(\delta;\omega_{2m}) = 0$ and $\phi_-(-\delta;\omega_{2m}) \neq 0$, $\phi_+(\delta;\omega_{2m-1}) \neq 0$. Moreover, several zeros of these functions may occur for the same δ.

When a zero $\omega_A(\delta)$ or $\omega_B(\delta)$ is found, a δ-parameterized family of continuous solutions $\phi_A(x;\delta)$ or $\phi_B(x;\delta)$, respectively, is constructed by gluing two individual gap solitons:

$$\begin{aligned} \phi_A(x;\delta) &= \phi_-(x-\delta;\omega_A)\chi_{(-\infty,0)} + \phi_+(x+\delta;\omega_A)\chi_{[0,\infty)}, \\ \phi_B(x;\delta) &= \phi_-(x-\delta;\omega_B)\chi_{(-\infty,0)} - \phi_+(x+\delta;\omega_B)\chi_{[0,\infty)}. \end{aligned} \tag{4.1}$$

The functions $\phi_{A,B}(x;\delta)$ decay as $|x| \to \infty$ and are smooth in x everywhere except at the nonlinearity interface $x = 0$, where they generally have a jump in the first derivative.

Note that it is important to consider both ϕ_A and ϕ_B due to the sign invariance of the ODE (2.3). Each sign produces a branch of continuous solutions of the ODE (2.3).

Figures 3 (a) and 4 (a) present the numerically computed $\omega_{A,B}(\delta)$ for $\delta \in (0, d)$ in the gaps (ω_1, ω_2) and (ω_3, ω_4), respectively. The lack of smoothness in the curves in these figures is due to an insufficient resolution in the search algorithm and can be corrected with a finer resolution. Note that when $\omega_{A,B}(\delta)$ is multiple-valued, as seen in Figures 3 (a) and 4 (a), we may have several decaying solutions $\phi_A(x)$ and/or $\phi_B(x)$ for the same δ.

(ii) *Construction of SGSs.* Next, we search for continuously differentiable solutions within the above family $\phi_{A,B}(x;\delta)$. To ensure the continuity of the first derivative of $\phi(x;\delta)$ at $x = 0$, we search for zeros of the two functions

$$g_A(\delta) = \phi_-'(-\delta;\omega_A) - \phi_+'(\delta;\omega_A), \quad g_B(\delta) = \phi_-'(-\delta;\omega_B) + \phi_+'(\delta;\omega_B).$$

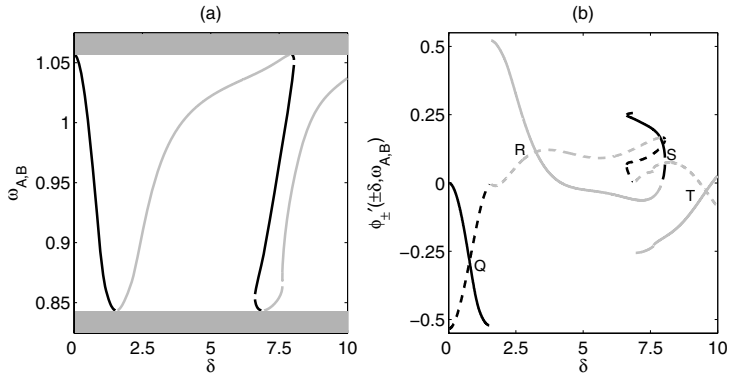


Figure 4. *Analogous to Figure 3 but for the gap* (ω_3, ω_4).

If a zero of either $g_A(\delta)$ or $g_B(\delta)$, denoted by δ_*, exists, then the function $\phi_A(x;\delta_*)$ or $\phi_B(x;\delta_*)$, respectively, in (4.1) has a continuous first derivative across the point $x = 0$. Figures 3 (b) and 4 (b) present the numerical results on computing δ_*. The labeled intersection points O, P, Q, R, S, and T correspond to zeros of $g_{A,B}(\delta)$. They are found as intersection points of solid and dashed curves of the same color. The solid black line shows the plot of $\phi'_+(\delta;\omega_A)$, and the dashed black line shows $\phi'_-(-\delta;\omega_A)$. Similarly, the solid gray line plots $-\phi'_+(\delta;\omega_B)$, and the dashed gray line plots $\phi'_-(-\delta;\omega_B)$. Therefore, an intersection of a solid black and a dashed black line (points O, Q, S) gives zeros δ_* of $g_A(\delta)$ and, thus, a solution $\phi_A(x;\delta_*)$. Similarly, an intersection of a solid gray and a dashed gray line (points P, R, T) gives zeros δ_* of $g_B(\delta)$ and, thus, a solution $\phi_B(x;\delta_*)$.

Table 2 shows the approximate computed values of δ_* and corresponding $\omega_* = \omega_{A,B}(\delta_*)$ at the points O–T for branches A, B of solutions given by (4.1). Note that additional points (δ_*, ω_*) can be obtained by generalizing the above functions $f_{A,B}$ and $g_{A,B}$ to

$$f_{j_A}(\omega) = \phi_-(-(jd+\delta);\omega) - \phi_+(jd+\delta;\omega), \quad f_{j_B}(\omega) = \phi_-(-(jd+\delta);\omega) + \phi_+(jd+\delta;\omega)$$

and

$$g_{j_A}(\delta) = \phi'_-(-(jd+\delta);\omega_A) - \phi'_+(jd+\delta;\omega_A), \quad g_{j_B}(\delta) = \phi'_-(-(jd+\delta);\omega_B) + \phi'_+(jd+\delta;\omega_B)$$

for $j \in \{1, 2, \ldots\}$ with V still defined as in (2.8)(ii). Nontrivial points (ω_*, δ_*) may exist for any such j. For example, we have found one such point for $j = 1$. The computed value is $(\omega_*, \delta_*) \approx (0.73, 7.33)$, and the resulting SGS corresponds to the point Z in Figure 5(a). Such additional solutions are SGSs of smaller amplitude compared to those for $j = 0$.

Table 2

Bifurcation points for SGSs in the domain $\omega \in (\omega_1, \omega_2) \cup (\omega_3, \omega_4)$ *and* $\delta \in (0, d)$.

Point	O	P	Q	R	S	T
Branch of solution	A	B	A	B	A	B
ω_*	0.58	0.70	0.94	0.97	1.03	1.03
δ_*	1.54	9.66	0.78	3.24	7.97	9.57

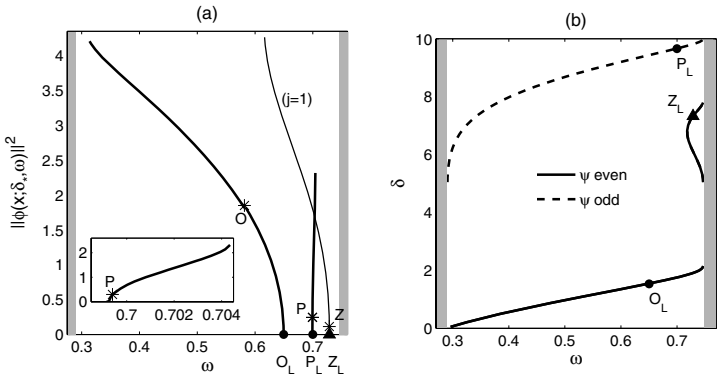


Figure 5. (a) *SGS continuation curves, total power versus frequency, in the gap* (ω_1, ω_2). *Labeled points* O, P *correspond to those in Figure* 3 (b). *Point* Z *is discussed in section* 4.1. *Points* O_L, P_L, *and* Z_L *are SGS termination points.* (b) *Point spectrum of the linear Schrödinger operator inside* (ω_1, ω_2) *for all* $\delta \in (0, d)$. *Solid/dashed lines: eigenvalues with even/odd eigenfunctions.*

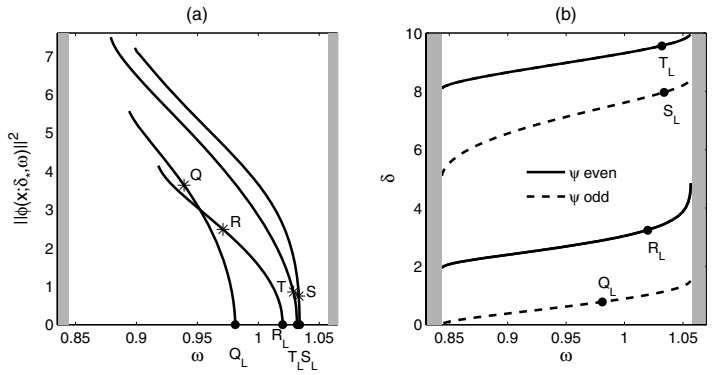


Figure 6. (a) *SGS continuation curves, total power versus frequency, in the gap* (ω_3, ω_4). *Labeled points* Q–T *correspond to those in Figure* 4 (b). *Points* Q_L–T_L *are SGS termination points.* (b) *Point spectrum of the linear Schrödinger operator inside* (ω_3, ω_4) *for all* $\delta \in (0, d)$. *Solid/dashed lines: eigenvalues with even/odd eigenfunctions.*

4.2. Numerical homotopy continuation of SGSs. Assuming the existence of a point (ω_*, δ_*), we have constructed the SGS of the ODE (2.3), where the potential function $V(x)$ is given by (2.8)(ii) and $(\omega, \delta) = (\omega_*, \delta_*)$. The SGS denoted as $\phi_*(x)$ is represented by one of the functions in (4.1) with $(\omega, \delta) = (\omega_*, \delta_*)$. Each of these solutions can be used as a starting point for a numerical homotopy continuation to generate a family of SGSs parameterized by $\omega \subset (\omega_{2m-1}, \omega_{2m})$ for a given value of $\delta = \delta_*$. Similarly, for a fixed $\omega = \omega_*$ a family parameterized by $\delta \subset (0, d)$ can be constructed. Under the same assumption that the operator $\mathcal{L} = -\partial_x^2 - \omega_* + V(x) - 3\Gamma(x)\phi_*^2(x)$ is invertible, the implicit function theorem implies that there exists a unique smooth continuation of the particular solution $\phi_*(x)$ to the family of solutions along parameters ω and δ.

We restrict our numerical studies to the continuation in ω. Numerical results of such continuation from the SGSs at points O–T are shown in Figures 5 (a) and 6 (a). The curves

plot the total soliton power $\|\phi\|^2_{L^2(\mathbb{R})}$ as a function of frequency ω for fixed $\delta = \delta_*$. Note that each curve corresponds to a different value of δ_* and hence a different potential $V(x)$. The values of δ_* can be read in Table 2. Termination of a continuation curve is defined when the total power of the soliton becomes zero or when Newton iteration convergence fails. As the figures show, the latter case is always accompanied by the slope of the continuation curve becoming infinite, suggesting a violation of the implicit function theorem assumptions. The former termination case is studied in the following subsection.

4.3. Analysis of termination points of SGSs. We shall now consider the termination points of the solution families plotted in Figures 5 (a) and 6 (a), where the soliton power becomes zero. The points are labeled O_L–T_L, and their corresponding values of δ and ω are given in Table 3.

Table 3
Termination points for the six SGS families in Figures 5 (a) and 6 (a).

Point	O_L	P_L	Q_L	R_L	S_L	T_L
δ	1.54	9.66	0.78	3.24	7.97	9.57
ω	0.65	0.70	0.98	1.02	1.03	1.03

The termination points are expected to be related to the existence of nontrivial bound states in the (point) spectrum of the Schrödinger operator for the same potential $V(x)$, i.e., with exponentially decaying solutions of the linear ODE

$$-\psi'' - \omega\psi + V(x)\psi = 0, \qquad \psi : \mathbb{R} \mapsto \mathbb{R}, \tag{4.2}$$

for $V(x)$ in (2.8)(ii) and $\omega \in \mathbb{R}\backslash\Sigma$. The point spectrum is nonempty due to the singularity of $V(x)$ at $x = 0$.

4.3.1. Numerical results. Results of numerical computations of the point spectrum contained in the first two finite gaps (ω_1, ω_2) and (ω_3, ω_4) are shown in Figures 5 (b) and 6 (b) for all values $\delta \in [0, d]$. The eigenfunctions ψ are either even (solid lines) or odd (dashed lines). For the six values of δ corresponding to the SGS families in Figures 5 (a) and 6 (a) the eigenvalues are marked by black dots and are in perfect agreement with the values of ω at the termination points O_L–T_L. The symmetry (even/odd) of the bound states at O_L–T_L also matches that of the eigenfunctions at the marked points in the point spectrum. The eigenvalue curve originating as well as ending at ω_2 in Figure 5 (b) corresponds to the termination point Z_L of the SGS family for $j = 1$ in Figure 5 (a). The termination point Z_L for the same value of δ is shown by a triangle.

4.3.2. Bifurcation analysis for $|\delta|$ small. In this subsection we consider bifurcations of point spectrum of the Schrödinger operator from the band edges for small values of $|\delta|$ (or, due to the d periodicity of V, equivalently for δ near 0 from above and near d from below). This analysis will prove the existence of the spectral curves near $\delta = 0$ and $\delta = 10$ in Figures 5 (b) and 6 (b), i.e., the existence of curves with points O_L and Q_L locally to $\delta = 0$ and the curves with points P_L and T_L locally to $\delta = 10$.

In order to construct solutions of the spectral problem (4.2), we first consider exponentially decaying solutions of the ODE on the half-line

$$-\psi_+'' - \omega\psi_+ + V_0(x+\delta)\psi_+ = 0, \qquad \psi_+ : \mathbb{R}_+ \mapsto \mathbb{R}.$$

By using the fundamental solution of the Hill equation (2.5), we can express $\psi_+(x)$ in the form $\psi_+ = e^{-\kappa x}u_-(x+\delta)$, where $u_-(x)$ are periodic or antiperiodic bounded solutions of the Hill equation (2.5) with $V(x) = V_0(x)$.

As $V(x)$ is even, the function $\psi_+(x)$ admits a symmetric (even) reflection about $x = 0$ if $\psi_+'(0) = 0$, which is equivalent to the condition

$$G_1(\delta,\kappa) = u_-'(\delta) - \kappa u_-(\delta) = 0,$$

and it admits an antisymmetric (odd) reflection about $x = 0$ if $\psi_+(0) = 0$, which is equivalent to the condition

$$G_2(\delta,\kappa) = u_-(\delta) = 0.$$

Since eigenvalues of the spectral problem (4.2) are simple and the eigenfunctions are either even or odd, all eigenvalues of the spectral problem (4.2) in the band gaps $\omega \in \mathbb{R}\backslash\Sigma$ are defined by zeros of the functions $G_1(\delta,\kappa)$ and $G_2(\delta,\kappa)$ in κ for a given value of δ, where $\kappa \geq 0$ and the values of κ are related to the values of ω in the band gaps. Both functions $G_{1,2}$ are analytic in $\delta \in \mathbb{R}$ and periodic with period d. Both functions admit analytic continuation in the parameter $\kappa \in \mathbb{R}_+$ [8].

If $\delta = 0$, the only zeros of $G_1(\delta,\kappa)$ and $G_2(\delta,\kappa)$ occur at $\kappa = 0$, i.e., at the band edges $\omega = \omega_n$. Indeed, if $G_1(0,\kappa) = 0$, then $\psi_+'(0) = 0$, such that $\psi_+(x) = \psi_n(x)$ is an even function on $x \in \mathbb{R}$. However, $\psi_+(x)$ decays exponentially as $x \to \infty$ and grows exponentially as $x \to -\infty$ if $\kappa > 0$. Therefore, $G_1(0,\kappa) = 0$ is equivalent to $\kappa = 0$. A similar argument works for $G_2(0,\kappa) = 0$.

(i) *Bifurcation of even eigenfunctions.* Let us first consider the zeros of $G_1(\delta,\kappa)$. Computing the derivatives of $G_1(\delta,\kappa)$ in δ and κ at $(\delta,\kappa) = (0,0)$, we obtain

$$\begin{aligned}
\partial_\delta G_1(0,0) &= u_-''(0) = \psi_n''(0) = (V_0(0) - \omega_n)\psi_n(0),\\
\partial_\kappa G_1(0,0) &= -\tilde\psi_n'(0) - \psi_n(0),
\end{aligned}$$

where $\tilde\psi_n$ is the generalized Bloch function; see (3.4). The fact that $\tilde\psi_n = -\frac{\partial u_-}{\partial\kappa}\big|_{\kappa=0}$ is clear from differentiation of (2.5) with respect to κ.

It is found in [12] that

$$D(x) = \psi_n(x)\tilde\psi_n'(x) - \psi_n'(x)\tilde\psi_n(x) + \psi_n^2(x)$$

is constant in x, i.e., $D(x) = D(0)$, and that

$$D(0) = \frac{1}{2}\omega_{2n-1,2n}''(k)(\psi_n,\psi_n), \tag{4.3}$$

where either $k = 0$ or $k = \frac{\pi}{d}$ at the bifurcation point $\omega = \omega_n$. Since $\psi_n'(0) = 0$, $D(0) = \psi_n(0)(\tilde\psi_n'(0)+\psi_n(0))$, and the leading-order approximation for the root of $G_1(\delta,\kappa)$ near $(\delta,\kappa) =$

$(0,0)$ is given by

$$\delta = \frac{\tilde{\psi}_n'(0) + \psi_n(0)}{\psi_n(0)(V_0(0) - \omega_n)}\kappa + \mathcal{O}(\kappa^2) = \frac{D(0)}{\psi_n^2(0)(V_0(0) - \omega_n)}\kappa + \mathcal{O}(\kappa^2),$$

where $\psi_n(0) \neq 0$ (which is met since $\psi_n'(0) = 0$). Using (4.3) and the facts that $\omega_n > 0$ and $V_0(0) = 0$ for the numerical example (2.6), we get

$$\delta = -\frac{\omega_{2n-1,2n}''(k)(\psi_n, \psi_n)}{2\psi_n^2(0)\omega_n}\kappa + \mathcal{O}(\kappa^2).$$

Therefore, the bifurcation occurs for $\delta > 0$ if $\omega_{2n-1,2n}''(k) < 0$ (e.g., for ω to the right of ω_1) and for $\delta < 0$ if $\omega_{2n-1,2n}''(k) > 0$ (e.g., for ω to the left of ω_0 and ω_4); see Table 1. Note that the negative values of δ correspond to the values of δ below the level $\delta = d$ due to periodicity of the function $G_1(\delta, \kappa)$ in δ. The above local existence analysis for even bound states is confirmed by the solid lines near $\delta = 0$ in Figure 5 (b) and near $\delta = d = 10$ in Figure 6 (b).

(ii) *Bifurcation of odd eigenfunctions.* Similarly to (i), we study the zeros of $G_2(\delta, \kappa)$. We compute the derivatives of $G_2(\delta, \kappa)$ in δ and κ at $(\delta, \kappa) = (0, 0)$,

$$\partial_\delta G_2(0,0) = u_-'(0) = \psi_n'(0),$$
$$\partial_\kappa G_2(0,0) = -\tilde{\psi}_n(0),$$

such that the leading-order approximation for the root of $G_2(\delta, \kappa)$ near $(\delta, \kappa) = (0, 0)$ is given by

$$\delta = \frac{\tilde{\psi}_n(0)}{\psi_n'(0)}\kappa + \mathcal{O}(\kappa^2) = -\frac{\omega_{2n-1,2n}''(k)(\psi_n, \psi_n)}{2(\psi_n'(0))^2}\kappa + \mathcal{O}(\kappa^2),$$

where $\psi_n'(0) \neq 0$ (which is met since $\psi_n(0) = 0$). From the expansion, we conclude that the bifurcation occurs for $\delta > 0$ if $\omega_{2n-1,2n}''(k) < 0$ (e.g., for ω to the right of ω_3) and for $\delta < 0$ if $\omega_{2n-1,2n}''(k) > 0$ (e.g., for ω to the left of ω_2). The dashed lines near $\delta = 0$ in Figure 6 (b) and near $\delta = d = 10$ in Figure 5 (b) confirm this analysis.

Note that there are curves in Figures 5 (b) and 6 (b) which do not bifurcate from $\delta = 0$ and $\delta = d = 10$ but still bifurcate from the band edge $\omega = \omega_n$. Bifurcations of these curves cannot be confirmed from the analytical theory above, unless the values of $G_{1,2}(\delta; 0)$ for $0 < \delta < d$ are approximated numerically.

5. Conclusion.

We have employed methods of bifurcation theory for the existence problem of SGSs supported by the nonlinearity interface and the periodic potential. Two bifurcation problems are considered numerically. The first bifurcation takes place from the standard gap solitons existing at the zero jump of the nonlinearity coefficient. The second bifurcation takes place from the bound state consisting of parts of two standard gap solitons glued together in a continuously differentiable SGS. Three asymptotic results are described in the article. We show that the standard gap solitons can be continued generally for small jumps in the nonlinearity coefficient. On the contrary, no SGSs for nonzero jump of the nonlinearity coefficient exist in the NLS approximation which is valid near the band edges. In addition, we analytically study bifurcations of eigenvalues of the Schrödinger operator with a nonsmooth potential from band edges of the Hill equation.

One can argue that the SGSs bifurcating from a standard gap soliton or a gluing combination of two gap solitons inherit stability properties of gap solitons in the neighborhood of the local bifurcation points. Stability of standard gap solitons was considered analytically and numerically in [13]. The stability properties can change far from the bifurcation points. Detailed computations of stability of the SGSs will be the subject of a forthcoming work.

Acknowledgment. Dmitry Pelinovsky thanks the people at ETH Zürich for hospitality during his visit.

REFERENCES

[1] D. Blömer, A. Szameit, F. Dreisow, T. Schreiber, S. Nolte, and A. Tünnermann, *Nonlinear refractive index of fs-laser-written waveguides in fused silica*, Opt. Express, 14 (2006), pp. 2151–2157.
[2] M. S. Eastham, *The Spectral Theory of Periodic Differential Equations*, Scottish Academic Press, Edinburgh, 1973.
[3] G. Fibich, Y. Sivan, and M. I. Weinstein, *Bound states of nonlinear Schrödinger equations with a periodic nonlinear microstructure*, Phys. D, 217 (2006), pp. 31–57.
[4] J. Hudock, S. Suntsov, D. Christodoulides, and G. Stegeman, *Vector discrete nonlinear surface waves*, Opt. Express, 13 (2005), pp. 7720–7725.
[5] Y. V. Kartashov, A. A. Egorov, V. A. Vysloukh, and L. Torner, *Surface vortex solitons*, Opt. Express, 14 (2006), pp. 4049–4057.
[6] Y. V. Kartashov and L. Torner, *Multipole-mode surface solitons*, Opt. Lett., 31 (2006), pp. 2172–2174.
[7] Y. V. Kartashov, V. A. Vysloukh, and L. Torner, *Surface gap solitons*, Phys. Rev. Lett., 96 (2006), 073901.
[8] W. Kohn, *Analytic properties of Bloch waves and Wannier functions*, Phys. Rev. (2), 115 (1959), pp. 809–821.
[9] W. Magnus and S. Winkler, *Hill's Equation*, Interscience Tracts in Pure and Applied Mathematics 20, John Wiley & Sons, New York, London, Sydney, 1966.
[10] K. G. Makris, S. Suntsov, D. N. Christodoulides, G. I. Stegeman, and A. Hache, *Discrete surface solitons*, Opt. Lett., 30 (2005), pp. 2466–2468.
[11] A. Pankov, *Periodic nonlinear Schrödinger equation with application to photonic crystals*, Milan J. Math., 73 (2005), pp. 259–287.
[12] D. Pelinovsky, *Asymptotic reductions of the Gross–Pitaevskii equation*, in Emergent Nonlinear Phenomena in Bose–Einstein Condensates: Theory and Experiment, P. Kevrekidis, D. Frantzeskakis, and R. Carretero, eds., Springer-Verlag, Heidelberg, 2007, pp. 377–398.
[13] D. E. Pelinovsky, A. A. Sukhorukov, and Y. Kivshar, *Bifurcations and stability of gap solitons in periodic structures*, Phys. Rev. E (3), 70 (2004), 036618.
[14] B. Simon and M. Reed, *Methods of Modern Mathematical Physics* IV: *Analysis of Operators*, Academic Press, New York, 1978.
[15] Y. Sivan, G. Fibich, and M. I. Weinstein, *Waves in nonlinear lattices—ultrashort optical pulses and Bose–Einstein condensates*, Phys. Rev. Lett., 97 (2006), 193902.
[16] S. Suntsov, K. G. Makris, D. N. Christodoulides, G. I. Stegeman, A. Hache, R. Morandotti, H. Yang, G. Salamo, and M. Sorel, *Observation of discrete surface solitons*, Phys. Rev. Lett., 96 (2006), 063901.
[17] W. J. Tomlinson, *Surface wave at a nonlinear interface*, Opt. Lett., 5 (1980), pp. 323–325.
[18] D. A. Zezyulin, G. L. Alfimov, V. V. Konotop, and V. M. Perez-Garcia, *Control of nonlinear modes by scattering-length management in Bose–Einstein condensates*, Phys. Rev. A (3), 76 (2007), 013621.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 265–283
© 2008 Society for Industrial and Applied Mathematics

Bounded Solutions of Nonlocal Complex Ginzburg–Landau Equations for a Subcritical Bifurcation*

V. A. Volpert†, A. A. Nepomnyashchy‡, L. G. Stanton†, and A. A. Golovin†

Abstract. Stable periodic solutions of a system of two nonlocal coupled complex Ginzburg–Landau (CGL) equations describing the dynamics of a subcritical Hopf bifurcation in a spatially extended system are found analytically in the limit of large dispersion coefficients. The domains in the parameter space where these solutions exist and are stable are determined. It is shown that the existence and stability depend on the sign of the coupling parameter and on the ratio of the dispersion coefficients. Numerical simulations of the system of nonlocal coupled CGL equations confirm the analytical results and exhibit other bounded dynamic regimes, such as standing waves and spatio-temporal chaos.

Key words. complex Ginzburg–Landau equation, nonlocal equations, Hopf bifurcation, subcritical instability

AMS subject classifications. 35Q99, 37G99

DOI. 10.1137/070687190

1. Introduction.

A Ginzburg–Landau equation,

$$A_t = \mu A + (\alpha_1 + i\alpha_2)A_{xx} + (\beta_1 + i\beta_2)A|A|^2, \tag{1.1}$$

for the amplitude of an unstable mode $A(x,t)$ as a function of slow temporal and spatial variables describes the behavior of a system near a Hopf bifurcation point. This has been a subject of continuing attention over the last three decades [1]. Most works address various types of solutions in the case of a supercritical bifurcation ($\mu = 1$, $\beta_1 < 0$). However, in applications, subcritical bifurcations ($\beta_1 > 0$) occur as often as supercritical ones. Indeed, subcritical Hopf bifurcations are found in convection in binary fluids [7, 13, 14, 15, 20, 24], in nonlinear optics and lasers [16, 19, 22], in directional solidification [5], in combustion [2, 17], and in many other applications including lesser known examples of explosive crystallization fronts [25] and frontal polymerization [4].

One might expect that unless higher order nonlinear saturation terms are accounted for in the equations, all the solutions for $\mu = 1$ in case of a subcritical bifurcation will blow up in a finite time. In this case, the amplitude equations would be of little use as a means to describe the behavior of the original system. However, this is not necessarily the case. Beginning with the work by Hocking and Stewartson [8], bounded solutions were found for

*Received by the editors April 2, 2007; accepted for publication (in revised form) by D. Barkley September 13, 2007; published electronically April 23, 2008. This work was supported by the National Science Foundation grant DMS-0505878.
http://www.siam.org/journals/siads/7-2/68719.html
†Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208-3125 (v-volpert@northwestern.edu, l-stanton@northwestern.edu, a-golovin@northwestern.edu).
‡Department of Mathematics and Minerva Center for Nonlinear Physics of Complex Systems, Technion – Israel Institute of Technology, Haifa 32000, Israel (nepom@techunix.technion.ac.il). This author acknowledges the support of the Israel Science Foundation grant 812/06.

the supercritical parameter value $\mu = 1$ in the case of a subcritical bifurcation $\beta_1 > 0$ under certain conditions on the magnitude of the coefficients of the equation. These results were apparently forgotten [1] and then rediscovered and extended by Bretherton and Spiegel [3] and Schöpf and Kramer [23]. Further extensions of these results can be found in [10, 11, 16, 21, 22].

The works cited above were performed for a single complex Ginzburg–Landau (CGL) equation (1.1). However, for a large class of systems with reflection symmetry, a more complete description of a Hopf bifurcation problem in the case of a short-wave instability involves a system of two *nonlocal coupled* Ginzburg–Landau equations for the amplitudes of two counterpropagating waves, $A^{\pm}(x,t)$ [12, 18]:

$$A_t^+ = \mu A^+ + (\alpha_1 + i\alpha_2)A_{xx}^+ + (\beta_1 + i\beta_2)A^+|A^+|^2 + (p_1 + ip_2)A^+\langle|A^-|^2\rangle, \tag{1.2a}$$

$$A_t^- = \mu A^- + (\alpha_1 + i\alpha_2)A_{xx}^- + (\beta_1 + i\beta_2)A^-|A^-|^2 + (p_1 + ip_2)A^-\langle|A^+|^2\rangle. \tag{1.2b}$$

Here the coupling term involves the averaged quantities defined as

$$\langle f\rangle = \lim_{L\to\infty}\frac{1}{2L}\int_{-L}^{L} f(x)dx,$$

where x is the spatial variable. The averaging procedure is defined for any positive, locally integrable function that is bounded for $-\infty < x < \infty$. For a periodic function $f(x)$, the average coincides with the average over a period. The averaged terms appear in the equations due to the fact that the effect of one wave on the other, traveling with a group velocity in the opposite direction on the intermediate time scale, enters only through its average on the slowest time scale.

We emphasize that it is these equations rather than equations with local cubic coupling that generically arise in the case of Hopf bifurcation in systems with right-left symmetries. Ginzburg–Landau equations with local coupling in the nonlinear terms cannot be derived from the original equations except in the nongeneric case of an asymptotically small group velocity of waves. Thus, in order to understand the dynamics of systems undergoing subcritical Hopf bifurcations, nonlocal coupled CGL equations must be considered, as is done in this paper.

As noted in [23], in the limit $\alpha_2 \to \infty$, $\beta_2 \to \infty$, (1.1) reduces to the nonlinear Schrödinger (NLS) equation, which has a continuum of soliton solutions that reduces to a discrete set when the equation is perturbed. The method used in [23] allows one to find not only soliton solutions but also spatially periodic solutions of (1.1) in the case of large α_2 and β_2. In this paper, we extend the analysis done in [23] to the system of equations (1.2) and show that the system (1.2) may have stable, bounded solutions for the supercritical parameter value $\mu = 1$ in the case of a subcritical bifurcation $\beta_1 > 0$ under the assumption that α_2 and β_2 are sufficiently large. We also perform numerical simulations that confirm our analytical results.

2. Periodic solutions.

Equations (1.2) can be rescaled as

$$A_t^+ = A^+ + (1 + i\alpha)A_{xx}^+ + (1 + i\beta)A^+|A^+|^2 + (p + ip_i)A^+\langle|A^-|^2\rangle, \tag{2.1a}$$

$$A_t^- = A^- + (1 + i\alpha)A_{xx}^- + (1 + i\beta)A^-|A^-|^2 + (p + ip_i)A^-\langle|A^+|^2\rangle \tag{2.1b}$$

(we retain the same notation for the amplitudes $A^{\pm}$). It is convenient to write these equations in the real form by substituting

$$A^{\pm}(x,t) = R^{\pm}(x,t)\exp\left(i\Theta^{\pm}(x,t)\right)$$

into (2.1) and separating real and imaginary parts to obtain

$$R_t^{\pm} = R^{\pm} + R_{xx}^{\pm} - R^{\pm}(\Theta_x^{\pm})^2 - \alpha(2R_x^{\pm}\Theta_x^{\pm} + R^{\pm}\Theta_{xx}^{\pm}) + (R^{\pm})^3 + pR^{\pm}\langle(R^{\mp})^2\rangle, \tag{2.2a}$$

$$R^{\pm}\Theta_t^{\pm} = R^{\pm}\Theta_{xx}^{\pm} + 2R_x^{\pm}\Theta_x^{\pm} + \alpha(R_{xx}^{\pm} - R^{\pm}(\Theta_x^{\pm})^2) + \beta(R^{\pm})^3 + p_i R^{\pm}\langle(R^{\mp})^2\rangle. \tag{2.2b}$$

The superscript $\pm$ means that each of the equations (2.2) includes two equations, one with the upper superscript and the other with the lower superscript. Note that the terms $\langle(R^{\mp})^2\rangle$ do not depend on x. Thus, the change of variables

$$\tilde{\Theta}^{\pm} = \Theta^{\pm} - p_i \int \langle(R^{\mp})^2\rangle\, dt$$

allows us to eliminate the term proportional to p_i in (2.2b). We retain the same notation, $\Theta^{\pm}$, for the phase but set $p_i = 0$ in (2.2b).

We consider the case when the coefficients α and β in the equations are large and set

$$\alpha = \frac{1}{\varepsilon}, \quad \beta = \frac{F}{\varepsilon}, \quad |\varepsilon| \ll 1, \quad F = O(1).$$

The equations then take the form

$$R_t^{\pm} = R^{\pm} + R_{xx}^{\pm} - R^{\pm}(\Theta_x^{\pm})^2 - \frac{1}{\varepsilon}\left(2R_x^{\pm}\Theta_x^{\pm} + R^{\pm}\Theta_{xx}^{\pm}\right) + (R^{\pm})^3 + pR^{\pm}\langle(R^{\mp})^2\rangle, \tag{2.3a}$$

$$R^{\pm}\Theta_t^{\pm} = R^{\pm}\Theta_{xx}^{\pm} + 2R_x^{\pm}\Theta_x^{\pm} + \frac{1}{\varepsilon}\left(R_{xx}^{\pm} - R^{\pm}(\Theta_x^{\pm})^2\right) + \frac{F}{\varepsilon}(R^{\pm})^3. \tag{2.3b}$$

Multiplying (2.3a) by ε^2 and (2.3b) by ε and taking the sum of the resulting equations yield

$$\varepsilon^2 R_t^{\pm} + \varepsilon R^{\pm}\Theta_t^{\pm} = \varepsilon^2 R^{\pm} + (1+\varepsilon^2)[R_{xx}^{\pm} - R^{\pm}(\Theta_x^{\pm})^2] + (F+\varepsilon^2)(R^{\pm})^3 + \varepsilon^2 pR^{\pm}\langle(R^{\mp})^2\rangle. \tag{2.4a}$$

Multiplying (2.3a) by ε and (2.3b) by ε^2 and taking the difference of the resulting equations yield

$$-\varepsilon R_t^{\pm} + \varepsilon^2 R^{\pm}\Theta_t^{\pm} = -\varepsilon R^{\pm} + (1+\varepsilon^2)[R^{\pm}\Theta_{xx}^{\pm} + 2R_x^{\pm}\Theta_x^{\pm}] + \varepsilon(F-1)(R^{\pm})^3 - \varepsilon pR^{\pm}\langle(R^{\mp})^2\rangle. \tag{2.4b}$$

We seek the solution of (2.4) as an expansion in powers of ε,

$$R^{\pm} = R_0^{\pm} + \varepsilon^2 R_2^{\pm} + \ldots, \quad \Theta^{\pm} = \frac{1}{\varepsilon}\Theta_{-1}^{\pm} + \varepsilon\Theta_1^{\pm} + \ldots.$$

Substituting the expansions into (2.4) and collecting like powers of ε, we obtain

$$O(\varepsilon^{-2}): \qquad R_0^{\pm}(\Theta_{-1}^{\pm})_x^2 = 0, \tag{2.5a}$$

$$O(\varepsilon^{-1}): \qquad R_0^{\pm}(\Theta_{-1}^{\pm})_{xx} + 2(R_0^{\pm})_x(\Theta_{-1}^{\pm})_x = 0, \tag{2.5b}$$

$$O(\varepsilon^{0}): \qquad R_0^{\pm}(\Theta_{-1}^{\pm})_t = (R_0^{\pm})_{xx} + F(R_0^{\pm})^3, \tag{2.5c}$$

$$O(\varepsilon): \qquad -(R_0^{\pm})_t + R_0^{\pm}(\Theta_{-1}^{\pm})_t = -R_0^{\pm} + R_0^{\pm}(\Theta_1^{\pm})_{xx} + 2(R_0^{\pm})_x(\Theta_1^{\pm})_x + (F-1)(R_0^{\pm})^3 - pR_0^{\pm}\langle(R_0^{\mp})^2\rangle. \tag{2.5d}$$

Equations (2.5a) and (2.5b) imply that $\Theta_{-1}^{\pm}$ does not depend on x and is a function of t only, $\Theta_{-1}^{\pm} = \Theta_{-1}^{\pm}(t)$. We assume that $F > 0$ and $\gamma_{\pm}(t) \equiv (\Theta_{-1}^{\pm})_t > 0$, in which case (2.5c) has a family of solutions

$$R_0^{\pm}(x,t) = \left[\frac{2\gamma_{\pm}(t)}{[2-m_{\pm}(t)]F}\right]^{1/2} \mathrm{dn}\left(\left[\frac{\gamma_{\pm}(t)}{[2-m_{\pm}(t)]}\right]^{1/2} x|m_{\pm}(t)\right). \tag{2.6}$$

Here $0 < m_{\pm} < 1$ is any function of t, and $\mathrm{dn}(u|m)$ is the Jacobi elliptic function that varies between $(1-m)^{1/2}$ and 1 with the period $2K(m)$, where $K(m)$ is the complete elliptic integral of the first kind. Thus, (2.6) represents a spatially periodic solution with the period

$$\lambda_{\pm} = 2K(m_{\pm}(t))\left[\frac{2-m_{\pm}(t)}{\gamma_{\pm}(t)}\right]^{1/2}. \tag{2.7}$$

In what follows we assume that the period $\lambda_{\pm}$ of the solution is constant, i.e., that there is a relation (2.7) between the time-dependent functions $m_{\pm}(t)$ and $\gamma_{\pm}(t)$. We remark that $m_{\pm} \to 1$ implies $K(m_{\pm}) \to \infty$ and the solution (2.6) degenerates into the pulse $(2\gamma_{\pm}/F)^{1/2}\,\mathrm{sech}(\gamma_{\pm}^{1/2}x)$, whereas $m_{\pm} \to 0$ corresponds to small harmonic oscillations.

Next we multiply (2.5d) by $R_0^{\pm}$ to rewrite it as

$$[(R_0^{\pm})^2(\Theta_1^{\pm})_x]_x = -\frac{1}{2}[(R_0^{\pm})^2]_t + (1+\gamma_{\pm})(R_0^{\pm})^2 - (F-1)(R_0^{\pm})^4 + p(R_0^{\pm})^2\langle(R_0^{\mp})^2\rangle. \tag{2.8}$$

The local wavenumber $(\Theta_1^{\pm})_x$ must be a bounded function, from which, taking into account that $R_0^{\pm}$ is periodic, we can draw a number of important conclusions. First, the integral of the right-hand side of (2.8) over this period must vanish. Indeed, suppose this is not the case and the integral over a period is equal to $a \neq 0$. Then integrating (2.8) over N periods of $R_0^{\pm}$, from x to $x + N\lambda_{\pm}$, yields

$$(R_0^{\pm})^2(\Theta_1^{\pm})_x\,\big|_{x+N\lambda_{\pm}} - (R_0^{\pm})^2(\Theta_1^{\pm})_x\big|_x = aN. \tag{2.9}$$

Since

$$(R_0^{\pm})^2\,\big|_{x+N\lambda_{\pm}} = (R_0^{\pm})^2\big|_x$$

due to periodicity of $R_0^{\pm}$, we obtain

$$(\Theta_1^{\pm})_x\,\big|_{x+N\lambda_{\pm}} = (\Theta_1^{\pm})_x\big|_x + \frac{aN}{(R_0^{\pm})^2|_x}. \tag{2.10}$$

That N can be arbitrarily large implies that $(\Theta_1^{\pm})_x$ is unbounded, contrary to our assumption about the local wavenumber. Thus, $a = 0$, and using (2.10) with $N = 1$ we arrive at another important conclusion that $(\Theta_1^{\pm})_x$ is periodic with period $\lambda_{\pm}$. Moreover, since $R_0^{\pm}$ is an even function, $(\Theta_1^{\pm})_x$ defined by (2.5d) is an odd one, and hence the phase $\Theta_1^{\pm}$ is even and periodic with the same period as $R_0^{\pm}$.

Upon some calculations using

$$\int_0^{K(m)} \mathrm{dn}^2\,u\,du = E(m), \quad \int_0^{K(m)} \mathrm{dn}^4\,u\,du = \frac{2}{3}(2-m)E(m) - \frac{1}{3}(1-m)K(m),$$

the condition that the integral of the right-hand side of (2.8) over the period vanishes yields

$$\frac{d}{dt}\left[\frac{K_\pm E_\pm}{\lambda_\pm^2 F}\right] = (1+\gamma_\pm)\frac{2K_\pm E_\pm}{\lambda_\pm^2 F} - \frac{16(K_\pm)^4(F-1)}{3\lambda_\pm^4 F^2}\left[2(2-m_\pm)\frac{E_\pm}{K_\pm} - (1-m_\pm)\right] + 16p\frac{K_\pm E_\pm}{\lambda_\pm^2 F}\frac{K_\mp E_\mp}{\lambda_\mp^2 F}. \tag{2.11}$$

Here $K_\pm = K(m_\pm)$, $E_\pm = E(m_\pm)$, and $E(m)$ is the complete elliptic integral of the second kind. Using (2.7) to eliminate $\lambda_\pm$, the above equation can be written as

$$D_\pm \frac{1}{\gamma_\pm}\frac{d\gamma_\pm}{dt} = 1 - B_\pm\gamma_\pm + pC_\mp\gamma_\mp, \tag{2.12}$$

where

$$D_\pm = D(m_\pm), \quad D(m) = \frac{1}{4E(m)}\frac{E^2(m) - (1-m)K^2(m)}{E(m) - 2(1-m)K(m)/(2-m)} > 0,$$

$$B_\pm = B(m_\pm), \quad B(m) = \frac{1}{3F}\left[F - 4 - 2(F-1)\frac{(1-m)K(m)}{(2-m)E(m)}\right],$$

$$C_\pm = C(m_\pm), \quad C(m) = \frac{2}{F}\frac{E(m)}{(2-m)K(m)} > 0.$$

The critical points $\gamma_\pm^s$ of (2.12), i.e., the solutions of

$$-B_+\gamma_+^s + pC_-\gamma_-^s = -1, \tag{2.13a}$$

$$-B_-\gamma_-^s + pC_+\gamma_+^s = -1, \tag{2.13b}$$

are given by

$$\gamma_+^s = \frac{1}{C_+}\frac{\Gamma_- + p}{\Gamma_-\Gamma_+ - p^2}, \qquad \gamma_-^s = \frac{1}{C_-}\frac{\Gamma_+ + p}{\Gamma_-\Gamma_+ - p^2}, \tag{2.14}$$

where

$$\Gamma_\pm = \Gamma(m_\pm), \quad \Gamma(m) = B(m)/C(m). \tag{2.15}$$

These critical points $\gamma_\pm^s$ are time-independent, as the corresponding $m_\pm$ are (see (2.7)). The behavior of the function $\Gamma(m)$ is important for subsequent calculations and is shown in Figure 1. One can show that $\Gamma(0) = -1$ and

$$\Gamma \sim (4\ln 2 - \ln(1-m))(F-4)/12 \quad \text{as } m \to 1.$$

For $F > 4$, $\Gamma(m)$ is a monotonically increasing function that has one zero at $m = m_0$, which depends on F. The function $m_0(F)$ is shown in Figure 2. It monotonically decreases from $m_0 = 1$ at $F = 4$ to zero as $F \to \infty$. For $F < 4$, $\Gamma(m) < 0$ for all $0 < m < 1$. It has a single maximum if $2 < F < 4$ and is a monotonically decreasing function if $0 < F < 2$. Thus

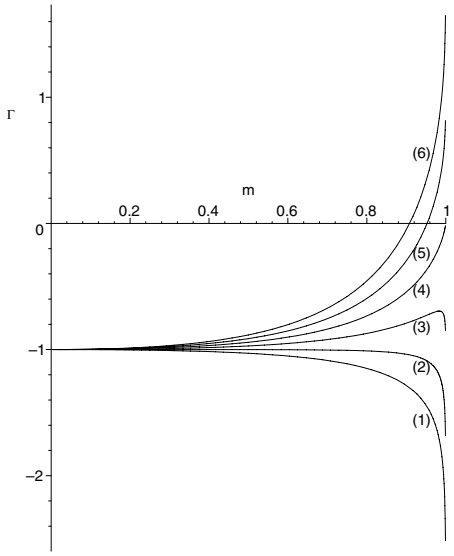


Figure 1. *The graph of the function* $\Gamma(m)$ *defined by* (2.15) *for various values of* F. *The curves are plotted for* $F = 1$ *(curve* (1)*),* $F = 2$ *(curve* (2)*),* $F = 3$ *(curve* (3)*),* $F = 4$ *(curve* (4)*),* $F = 5$ *(curve* (5)*), and* $F = 6$ *(curve* (6)*).*

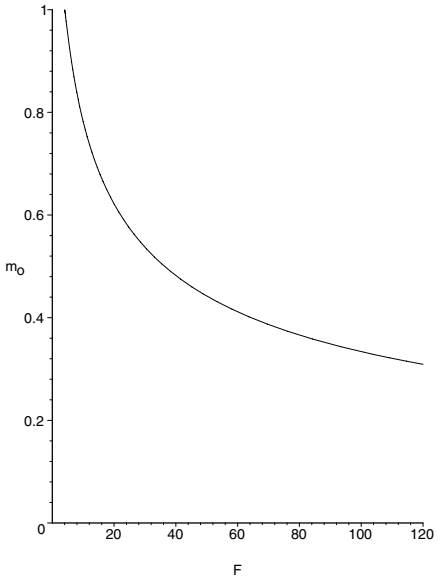
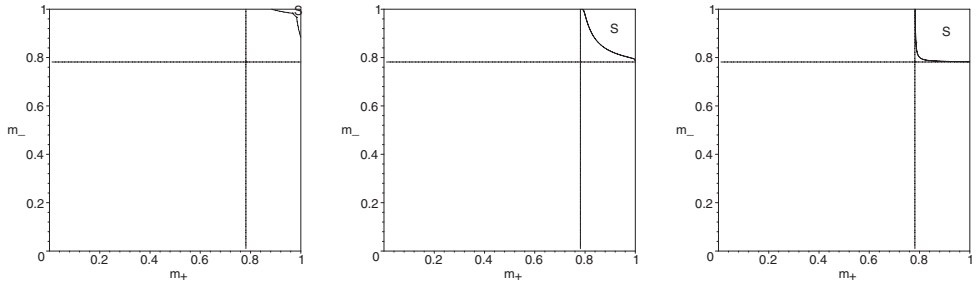


Figure 2. *The graph of the functions* $m_0(F)$.

we obtain a two-parameter (m_+, m_-) family of solutions (2.14) that determine the amplitude (2.6). The corresponding wavelengths are given by (2.7). Since $\gamma_\pm^s$ are required to be positive, there are certain conditions on the choice of the parameters m_+ and m_-. To describe these conditions, we consider the cases $p > 0$ and $p < 0$ separately. Consider first $p > 0$. Then,



(a) $p = 2.5$ (b) $p = 0.5$ (c) $p = 0.1$

Figure 3. *Existence regions for $F = 10$ and three different values of $p > 0$. See text.*

since $C_\pm > 0$, the stationary equations (2.13) imply that $m_\pm$ must be such that

$$\Gamma_\pm > 0, \quad p^2 < p_e^2(m_+, m_-) \equiv \Gamma_+ \Gamma_-, \tag{2.16}$$

which, in turn, requires $F > 4$ (otherwise, $\Gamma(m) < 0$ for $0 < m < 1$) and $m_\pm > m_0$. Thus, for $p > 0$, positive critical points $\gamma_\pm$ exist only if $F > 4$ and for parameter values $m_\pm$ that lie in the square $m_0 < m_+ < 1$, $m_0 < m_- < 1$ subject to the condition $p < p_e$. This result is illustrated by Figure 3, where the existence regions are shown in the (m_+, m_-)-plane for $F = 10$ and three different values of $p > 0$. The existence regions are marked by S, since all the existing solutions are stable, as shown in the next section. The existence region is bounded by the curve $p_e(m_+, m_-) = p$ and the line segments $m_+ = 1$, $m_0 < m_- < 1$ and $m_- = 1$, $m_0 < m_+ < 1$. Note that the ends of the curve $p_e(m_+, m_-) = p$ are the points $(m_0, 1)$, $(1, m_0)$, which is not necessarily seen on the scale of the figure. To better understand the form of the solution $R_0^\pm$, we consider a point (m_+, m_-) close to one of the existence boundaries. As (m_+, m_-) approaches the existence boundary $p_e(m_+, m_-) = p$, $\gamma_\pm^s \to \infty$. As a result, $\lambda_\pm \to 0$, and $R_0^\pm$ in (2.6) goes to infinity so that the solution (2.6) has the form of an array of closely spaced spikes. As (m_+, m_-) approaches other existence boundaries, the form of the solution is different. Consider the case $m_+ \to 1$, $m_0 < m_- < 1$ (the case $m_- \to 1$, $m_0 < m_+ < 1$ is similar). Then both γ_+^s and γ_-^s have finite limits. As a result, $\lambda_+ \to \infty$, and, as mentioned earlier, R_0^+ degenerates into a pulse

$$R_0^+ = \sqrt{\frac{2\gamma_+}{F}}\,\mathrm{sech}(x\sqrt{\gamma_+}).$$

Different types of solutions corresponding to different values of $m_\pm$ are shown in Figure 4. Figure 3 also shows that decreasing p increases the existence region. In the limit $p \to 0$, the existence region occupies the entire square $m_0 < m_\pm < 1$. The evolution of the existence region for other values of $F > 4$ as p varies is similar to that for $F = 10$. The main difference is the size of the square $m_0 < m_\pm < 1$, since m_0 depends on F (see Figure 2).

Next, consider the case $p < 0$. We distinguish between two subcases, $F > 4$ and $F < 4$. As discussed earlier, an important difference between them is in the behavior of the function $\Gamma(m)$. Consider first $F > 4$, so that Γ is a monotonically increasing function of m. Suppose m_1 is such that $\Gamma(m_1) = -p$. Then the critical points $\gamma_\pm^s$ are positive if either $m_1 < m_\pm < 1$

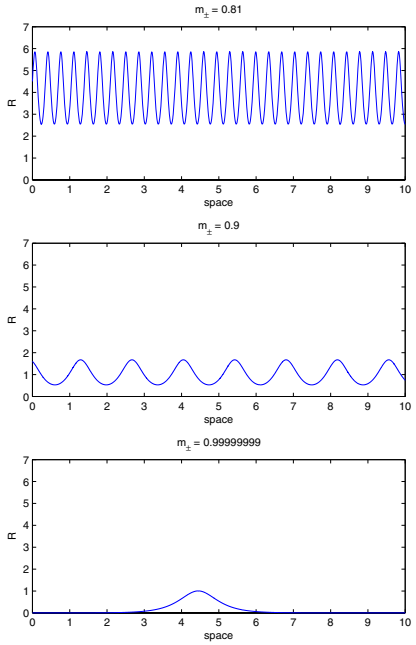
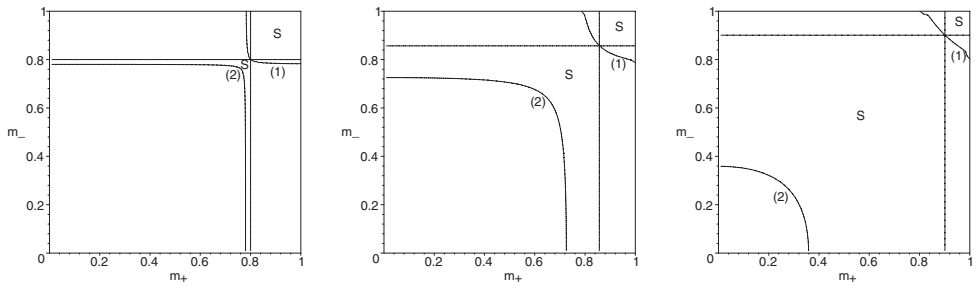


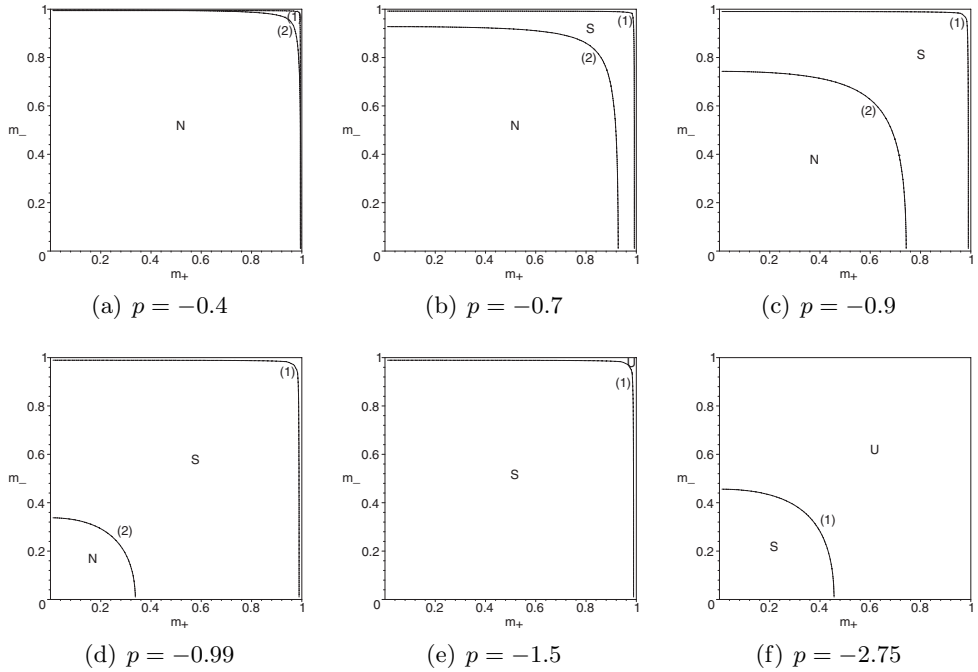
Figure 4. *Different types of solutions* (2.6), (2.14) *for* $F = 10$, $p = 0.1$.



(a) $p = -0.1$ (b) $p = -0.5$ (c) $p = -0.95$

Figure 5. *Existence regions for* $F = 10$ *and three different values of* $p < 0$. *See text.*

(in which case both the numerators and denominators in (2.14) are positive) or $0 < m_\pm < m_1$ and $p_e^2(m_+, m_-) < p^2$ (in which case both the numerators and denominators in (2.14) are negative). Note that the condition $p_e^2 < p^2$ gives a restriction on the range of $m_\pm$ only if $-1 < p < 0$. Otherwise, it is satisfied for all $0 < m_\pm < m_1$. Figure 5 illustrates these results for $F = 10$ and three different values of p. Here the vertical and horizontal lines inside the unit square are $m_+ = m_1$ and $m_- = m_1$, respectively. Curves (1) and (2) represent the boundaries $p_e^2(m_+, m_-) = p^2$. Existence regions are marked by S as, for these parameter values, all the existing solutions are stable. For other parameter values, the solutions may be unstable, as discussed in the next section. We observe that increasing $|p|$ results in the broadening of the existence region inside the $0 < m_\pm < m_1$ square. At $|p| = 1$, curve (2) disappears so that, for



(a) $p = -0.4$ (b) $p = -0.7$ (c) $p = -0.9$

(d) $p = -0.99$ (e) $p = -1.5$ (f) $p = -2.75$

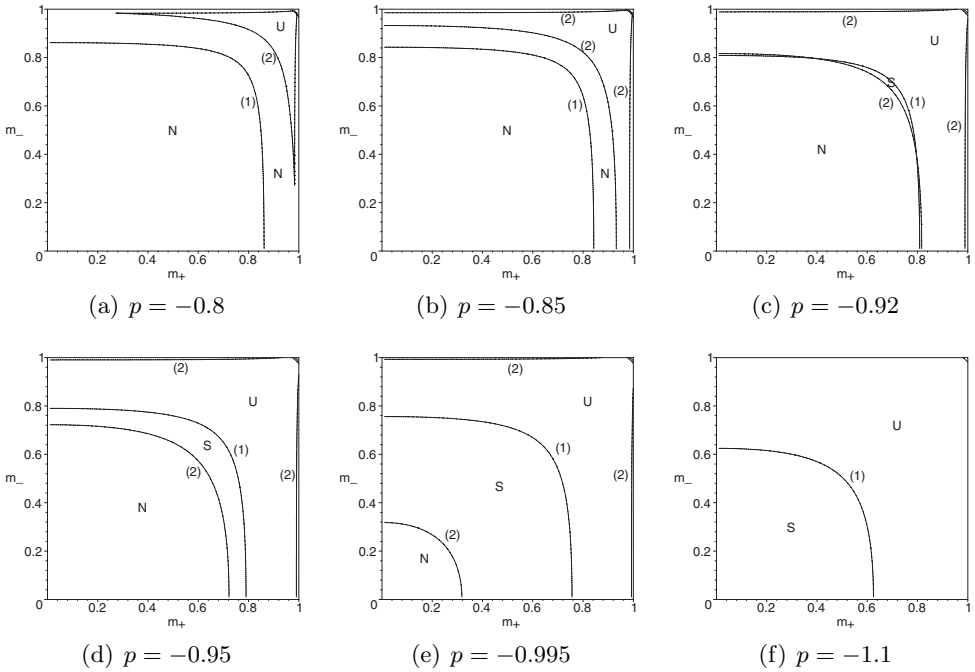
Figure 6. *Existence and stability regions for $F = 3.9$ and six different values of $p < 0$. Here (1) is the stability boundary $p_s^2 = p^2$; (2) is the existence boundary $p_e^2 = p^2$. N denotes the region where solution does not exist; S denotes the region where the solution exists and is stable; U denotes the region where the solution exists and is unstable. See text.*

$|p| > 1$, the existence regions are the two squares, $0 < m_\pm < m_1$ and $m_1 < m_\pm < 1$.

For $F < 4$, $\Gamma(m) < 0$, so that the numerators in (2.14) are negative, and for $\gamma_\pm^s > 0$, the denominator must be negative as well, i.e., $p_e^2(m_+, m_-) < p^2$. This condition implies that $|p| > \max_{0<m<1} \Gamma(m)$. It is convenient to consult the graph of $\Gamma(m)$ (see Figure 1) to understand the evolution of the existence regions as p varies. For $2 < F < 4$, $\Gamma(m)$ has a local maximum. Thus, if $|p|$ is less than this maximum value, there are no solutions. If $|p|$ slightly exceeds the maximum, an existence island appears. It broadens as $|p|$ increases. As p reaches $\Gamma(0) = -1$, the existence region occupies the entire unit square with the exception of a little region of sufficiently large $m_\pm$. These results are illustrated in Figures 6 and 7, where the evolution of the existence region as p varies is shown for $F = 3.9$ and $F = 3.1$, respectively. Here curve (2) is the existence boundary $p_e^2 = p^2$. The nonexistence region is marked by N. The existence region is marked by either S or U, depending on whether the solution is stable or unstable (stability is discussed in the next section). Note that the little region of nonexistence that occurs for sufficiently large $m_\pm$ is not seen on the scale of Figures 6 and 7.

Finally, as shown in the next section, for $0 < F < 2$ all the solutions are unstable, and therefore we do not discuss this case here.

We remark that the existence of bounded solutions and their stability (see section 3) are analyzed based on the first several terms of the expansions in powers of ϵ. We expect that



(a) $p = -0.8$ (b) $p = -0.85$ (c) $p = -0.92$

(d) $p = -0.95$ (e) $p = -0.995$ (f) $p = -1.1$

Figure 7. *Existence and stability regions for $F = 3.1$ and six different values of $p < 0$. Here (1) is the stability boundary $p_s^2 = p^2$; (2) is the existence boundary $p_e^2 = p^2$. N denotes the region where solution does not exist; S denotes the region where the solution exists and is stable; U denotes the region where the solution exists and is unstable. See text.*

the higher-order terms will not significantly change the existence and stability properties. Though we do not have a rigorous proof of this statement, we believe that, unlike solitary waves, which are described by nongeneric homoclinic/heteroclinic solutions in the x-space and therefore their very existence is sensitive to the equation perturbations (see, e.g., [9]), the cnoidal waves correspond to periodic solutions which are generic in systems with the $x \to -x$ symmetry, and thus we expect that the influence of the higher-order terms will not lead to drastic changes. Our expectation that the higher-order terms will not significantly change the existence and stability properties is somewhat backed up by our direct numerical simulations of the system (2.1) presented below. Though exhaustive numerical studies that would cover the entire parameter space have not been performed, the computations for selected parameter values agree with the analytical results that do not account for higher-order terms.

3. Stability.

In this section, we study stability of the critical points $\gamma_\pm^s$ of the system (2.12). Thus, we analyze the stability of the solutions (2.6) with respect to a specific class of perturbations. Specifically, we restrict our stability analysis to the perturbations that are periodic with constant periods $\lambda_\pm$ and keep the shape of the cnoidal wave (2.6). Therefore, the results of the stability analysis presented in this section give the necessary conditions for stability of solutions (2.6) rather than the sufficient conditions. In general, one could impose perturbations of an incommensurate period and get a modulational instability of the periodic wave. This is beyond the scope of the present paper.

Instabilities in the CGL equations in the NLS limit can be thought of as being of two kinds. One is the instability inherited from the NLS equation. The other is the instability due to the presence of nonconservative terms in the CGL equations. As a solution of the NLS equation, the cnoidal wave (2.6) is stable within the class of solutions with the same period (and the same Floquet exponent; see [6] and the references therein). Thus, all the instabilities that we obtain in this section for solutions of the CGL equations, with respect to the disturbances with the same period, have nothing to do with the NLS equation; they are specifically due to the presence of nonconservative terms in the equations.

In the case of general perturbations that do not have to be of the same period as the solution, the cnoidal solutions of the NLS equation are unstable in the focusing case $F > 0$ that is considered in this paper (again, see [6] and the references therein). This instability is inherited by the Ginzburg–Landau equations. We observe it in our numerical computations that are described below. However, we do not perform any stability analyses in this paper to discover these instabilities.

It is convenient to rewrite (2.11) as an equation for $m_\pm$. Using (2.7) to eliminate $\gamma_\pm$, we reduce (2.11) to

$$\widehat{D}_\pm \frac{dm_\pm}{dt} = 1 - \frac{8a_\pm}{\lambda_\pm^2 F} + \frac{8pb_\mp}{\lambda_\mp^2 F}, \tag{3.1}$$

where

$$\widehat{D}_\pm = \widehat{D}(m_\pm), \quad a_\pm = a(m_\pm), \quad b_\pm = b(m_\pm),$$

$$\widehat{D}(m) = \frac{E^2(m) - (1-m)K^2(m)}{4m(1-m)K(m)E(m)}, \quad a(m) = E(m)K(m)\Gamma(m), \quad b(m) = E(m)K(m).$$

We linearize (3.1) about its critical point $m_\pm^s$,

$$m_\pm = m_\pm^s + e^{\sigma t}\widetilde{m}_\pm,$$

and assume that the periods $\lambda_\pm$ are fixed and time-independent to obtain the dispersion relation

$$\sigma^2 \widehat{D}_+\widehat{D}_- + \sigma\left[\widehat{D}_- \frac{8}{F\lambda_+^2}\frac{da_+}{dm_+} + \widehat{D}_+ \frac{8}{F\lambda_-^2}\frac{da_-}{dm_-}\right] + \frac{64}{(F\lambda_+\lambda_-)^2}\left[\frac{da_+}{dm_+}\frac{da_-}{dm_-} - p^2\frac{db_+}{dm_+}\frac{db_-}{dm_-}\right] = 0. \tag{3.2}$$

We remark that $\widehat{D} > 0$, and for $F > 4$, the function $a(m)$ is an increasing function of m, $0 < m < 1$. Thus, the coefficient of σ to the first power in the quadratic equation (3.2) is positive provided $F > 4$. The stability boundary in this case is determined by

$$p^2 = \frac{da_+}{dm_+}\frac{da_-}{dm_-}\left[\frac{db_+}{dm_+}\frac{db_-}{dm_-}\right]^{-1} \equiv p_s^2, \quad p_s > 0. \tag{3.3}$$

Finally, if $F > 4$, the critical point is stable if $|p| < p_s$ and unstable otherwise.

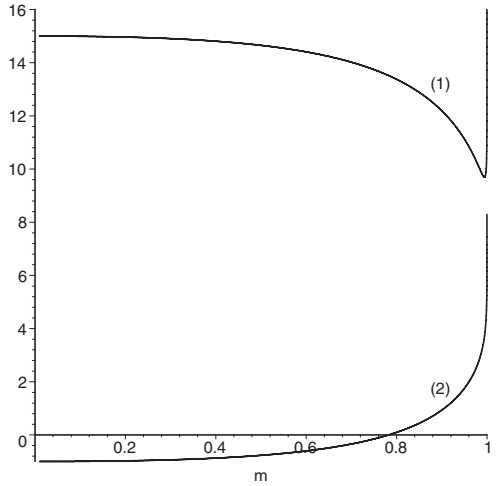


Figure 8. *The graphs of the functions $f(m)$ (curve (1)) and $\Gamma(m)$ (curve (2)). Here $F = 10$.*

In order to determine how the stability of the critical point depends on parameter values, we turn to various cases considered in the previous section. Let us first consider the case $p > 0$. Then, as discussed in the previous section, $F > 4$ in order for the critical points to be positive. Thus, the stability condition is $|p| < p_s$. It is easy to see that

$$f(m) \equiv \frac{da(m)}{dm} \Big/ \frac{db(m)}{dm} = \Gamma + \frac{d\Gamma}{dm} b(m) \left[\frac{db(m)}{dm}\right]^{-1} > \Gamma.$$

Thus, $p_e < p_s$ so that all the existing solutions in the case $p > 0$ are stable.

Consider next the case $p < 0$ and $F > 4$. All the solutions $m_\pm$ such that $m_0 < m_\pm < 1$ are stable. Indeed,

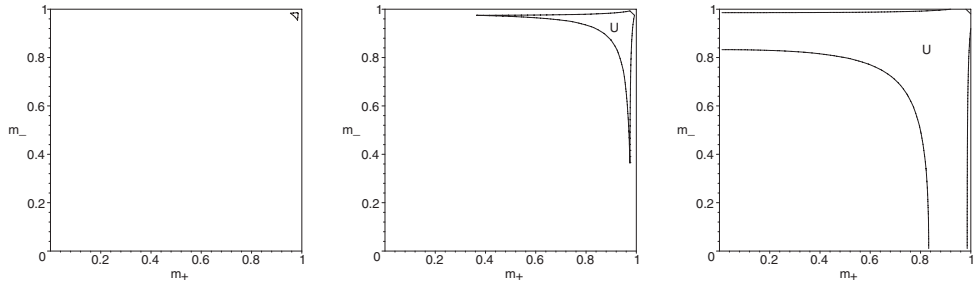
$$p_s^2 = f(m_+) f(m_-) > \Gamma_+ \Gamma_- > \Gamma^2(m_1) = p^2.$$

Stability of solutions in the other existence region, $0 < m_\pm < m_1$, $p_e^2 < p^2$, depends on the parameter values. It is convenient to analyze the stability by comparing the graphs of $f(m)$ and $\Gamma(m)$ (Figure 8). We remark that the function $f(m)$ is positive for $0 < m < 1$, has one minimum, and goes to infinity as m approaches 1:

$$f(m) \sim (4 \ln 2 - \ln(1 - m))(F - 4)/6, \quad m \to 1.$$

If $|p|$ is less than the minimum of the function $f(m)$, then any existing solution is stable, because $p^2 < f(m_-) f(m_+) = p_s^2$ in this case. The onset of instability occurs at $|p| = \min f(m)$. For $|p|$ slightly above this value, there is a little island of unstable solutions, which, as $|p|$ increases, gradually occupies the entire existence region $0 < m_\pm < m_1$. For $|p| > f(0) = 2F - 6$, only a small corner near the point $m_\pm = m_1$ is stable. The minimum of the function $f(m)$ can be accurately approximated as

$$\min_{0<m<1} f(m) \approx \frac{3}{2}(F - 4).$$



(a) $p = -10.5$ (b) $p = -12.4$ (c) $p = -14$

Figure 9. *Evolution of the instability region (marked by U) for $F > 4$ as p varies. Here $F = 10$. See text for a detailed discussion.*

For $F = 10$, the approximation yields the value of 9, while the numerical value (see Figure 8) is 9.6. This explains why the solutions depicted in Figure 5 are all stable: the values of $|p|$ there are less than the minimum of f. The stability results are illustrated in Figure 9 for $F = 10$ and larger values of $|p|$. As follows from the above discussion, the instability first occurs at $|p| \approx 9.6$, and for $|p| = 10.5$ the unstable solutions occupy a noticeable region in the upper right corner of the $0 < m_\pm < m_1$ square; see Figure 9(a). As $|p|$ increases, the instability region expands. Note that in this figure, we do not see the $m_1 < m_\pm < 1$ square that was shown in Figure 5. Indeed, for $|p|$ large, the value of m_1 is so close to 1 that this square is not seen on the scale of the figure (one can use the asymptotic expression in (2.16) to see that for $|p| = 10$, m_1 differs from 1 by less than 10^{-7}).

Next, consider the case $p < 0$ and $0 < F < 5/2$. In this case $da/dm < 0$. Indeed, the condition $da/dm < 0$ can be written as $F < F_*(m)$, where $F_*(m)$ can be explicitly found. The function $F_*(m)$ is a monotonically increasing function of m, $0 < m < 1$, which varies from $F_*(0) = 5/2$ to $F_*(1) = 4$, so that $da/dm < 0$ for $0 < F < 5/2$. Since $da/dm < 0$, the coefficient of the linear term in σ in the quadratic equation (3.2) is negative, so that there is a positive solution σ of (3.2), and all the existing solutions $m_\pm$ are unstable.

It turns out that all the solutions in the case $p < 0$ and $5/2 \le F < 3$ are also unstable. This follows from the fact that in this case $p_s^2 < p_e^2$. To better understand it, consider again Figure 7. Here we see that in order for the solutions to become stable, the relative positions of curves (1) and (2) have to change. The above inequality, $p_s^2 < p_e^2$, means that the curves will never change their relative positions. This inequality is equivalent to

$$\frac{f(m_+)}{\Gamma(m_+)}\frac{f(m_-)}{\Gamma(m_-)} < 1,$$

which is true if

$$\frac{f(m)}{|\Gamma(m)|} < 1,$$

i.e., if

$$\hat{f} \equiv f(m) + \Gamma(m) < 0.$$

One can see that $\hat{f}$ is a monotonically decreasing function of m with $\hat{f}(0) = 2F - 6$. Thus, $\hat{f} < 0$ for all $0 < m < 1$ if $F \le 3$.

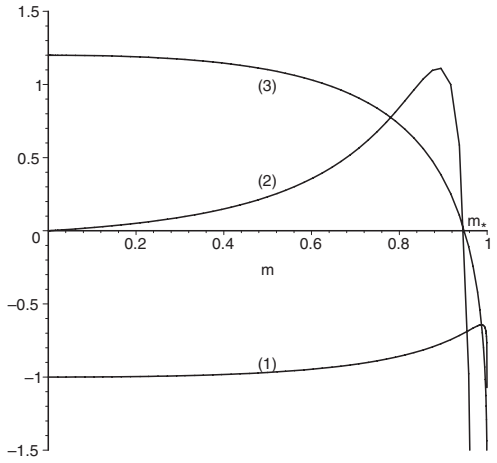


Figure 10. *The graphs of the functions* $\Gamma(m)$ *(curve* (1)*),* da/dm *(curve* (2)*), and* $f(m)$ *(curve* (3)*). Here* $F = 3.1$*. Note that* da/dm *and* $f(m)$ *become equal to zero at the same point* $m = m_*$*.*

Finally, consider the case $p < 0$ and $3 < F < 4$. In order for the solution (m_+, m_-) to be stable, we need

$$f(m_+)f(m_-) > p^2 \tag{3.4}$$

and the coefficient of the linear in σ term in (3.2) to be positive, i.e.,

$$\frac{\widehat{D}_+}{\lambda_-^2}\frac{da_-}{dm_-} + \frac{\widehat{D}_-}{\lambda_+^2}\frac{da_+}{dm_+} > 0. \tag{3.5}$$

Figure 10 illustrates the following result: the solution (m_+, m_-) is stable if $m_\pm < m_*$ and (3.4) is satisfied; otherwise, it is unstable. Indeed, if $m_\pm > m_*$, then $da_\pm/dm_\pm < 0$, and condition (3.5) is not satisfied; if one of the numbers m_+, m_- is greater than m_* and the other one is less than m_*, then $f(m_+)$ and $f(m_-)$ have different signs, and condition (3.4) is not satisfied. If $m_\pm < m_*$, then $da_\pm/dm_\pm > 0$, so that (3.5) is satisfied, which proves the above statement.

The main conclusion that follows from this stability analysis is that increasing $|p|$ decreases the stability region. At $|p| = f(0) = 2F - 6$ the stability region shrinks to zero. No stable solutions exist for $|p| > 2F - 6$.

The evolution of the stability boundary as p varies is shown in Figures 6 and 7 for $F = 3.9$ and $F = 3.1$, respectively.

4. Numerical simulations. In this section, we verify the stability and existence of our analytical solution through numerical simulations. We have used a pseudospectral code with periodic boundary conditions. A semi-implicit scheme has been used with a Crank–Nicolson method to handle the linear terms and with an Adams–Bashforth method for the nonlinear terms.

We first examine the case $F > 4$. Using (2.6) perturbed with small-amplitude noise as an initial condition, the solutions were observed to be stable for chosen values of $m_\pm$ in the stable

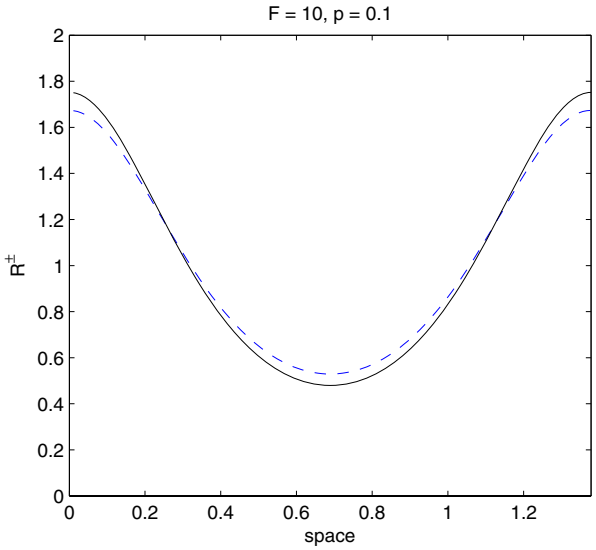


Figure 11. *Plots of the analytic solution (dashed) and the stable numerical solution (solid); here* $\epsilon = 0.1$ *and* $m_{\pm} = 0.9$.

regions shown in Figures 3, 5, and 9. A plot of the stable solution together with the analytical solution is shown in Figure 11. If the initial condition is taken to be small-amplitude noise, the system eventually evolves to a standing wave regime whose period is equal to the domain length (see Figure 12). The standing wave regime manifests itself as a periodic alternation of relatively long time intervals of nearly homogeneous spatial distributions of variables and short intervals of strongly inhomogeneous distributions characterized by the formation of wavenumber kinks and by phase-slip events. The oscillations in both waves, A^+ and A^-, are synchronized in time but not in space, due to the nonlocal nature of the coupling. For $p < -\min f(m)$, there exists a region of unstable solutions, as seen in Figure 9. In this region, the solution destabilizes to the standing waves similar to those shown in Figure 12. One can see that the spatial structure of this standing wave contains a wavenumber kink corresponding to a phase slip. We note that the formation of the standing waves shown in Figure 12 is the finite-domain effect: with the increase of the computational domain length L the solution never stabilizes to the standing waves but stays chaotic, as shown in Figure 13. This chaotic behavior consists of wiggling localized pulses accompanied by phase slips.

We next examine the case $F < 4$. In this case, solutions do not exist for $p > 0$, and numerical simulations using small-amplitude noise as an initial condition blow up in finite time. For $p < 0$, the solutions (2.6) were observed to be stable for values of $m_{\pm}$ in the stable regions seen in Figures 6 and 7. For values of $m_{\pm}$ in the unstable regions, the solutions blow up in finite time. This is also the case for all solutions with $F < 3$.

It should be noted that in all cases, the solutions destabilize as $m_{\pm}$ approaches existence boundaries. As $m_{\pm} \to 1$, the solution (2.6) becomes an array of pulse-like peaks with $\lambda_{\pm} \to \infty$. The tails of these peaks then approach 0 and destabilize, since the trivial solution ($A^{\pm} = 0$) is unstable. Eventually the solution evolves to standing waves. Conversely, as $m_{\pm}$ approaches

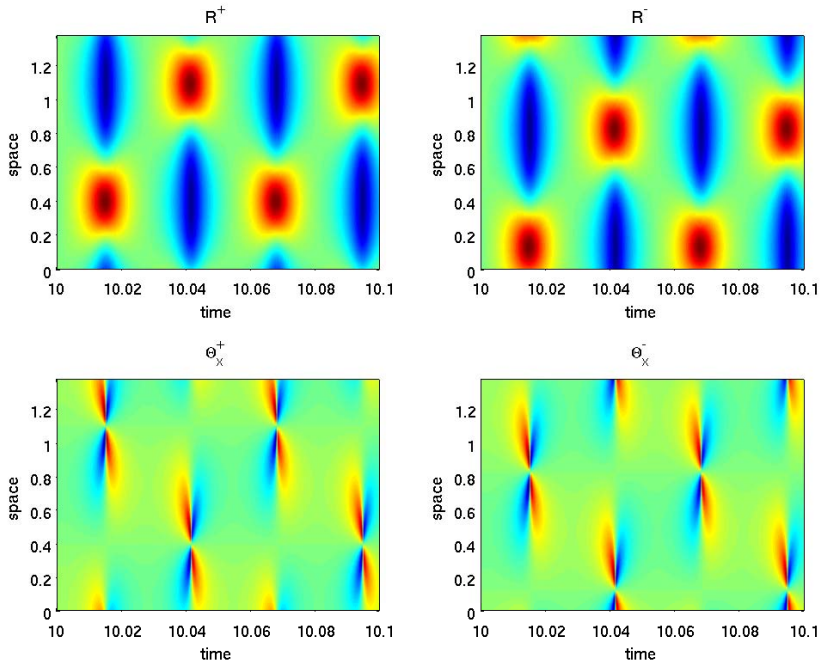


Figure 12. *Spatio-temporal plots of* $R^{\pm}$ *and* $\Theta_x^{\pm}$ *for* $F = 10$, $p = 0.1$, $m_{\pm} = 0.9$, $L = \lambda_{\pm} = 1.3779$ *showing alternating standing waves.*

the other existence boundary, $\lambda_{\pm} \to 0$, and the solution amplitude tends to infinity. Since the amplitude was assumed to be an $O(1)$ quantity, our analysis is no longer valid in this case. In the corresponding numerical simulations the solution blows up in finite time. It should also be noted that stable solutions were observed only for initial conditions with one period in the domain. In the case of starting with multiple periods, the solution then evolves to standing waves in a small domain or to chaotic behavior in large domains.

5. Conclusions. We have investigated the spatio-temporal dynamics of a system with subcritical Hopf bifurcation with reflection symmetry described near the instability threshold by a system of two nonlocally coupled CGL equations (2.1). We have shown that in the limit of large dispersion coefficients α and β, the system (2.1) may have bounded solutions corresponding to periodically modulated waves in the original physical system. These solutions are generalizations of the solutions of a single subcritical CGL equation found in [23]. In the parameter space, corresponding to the periods of the modulation waves and the coupling coefficient, we have found the regions where these solutions exist. We have shown that the existence of the solutions essentially depends on the sign of the coupling parameter and the ratio of the dispersion coefficients. We have performed a linear stability analysis and have found the necessary conditions for the stability of these solutions. We have also performed numerical simulations of the system (2.1). We have shown that for the parameters corresponding to the stability domain and the initial conditions close enough to the analytical solution (2.6), the

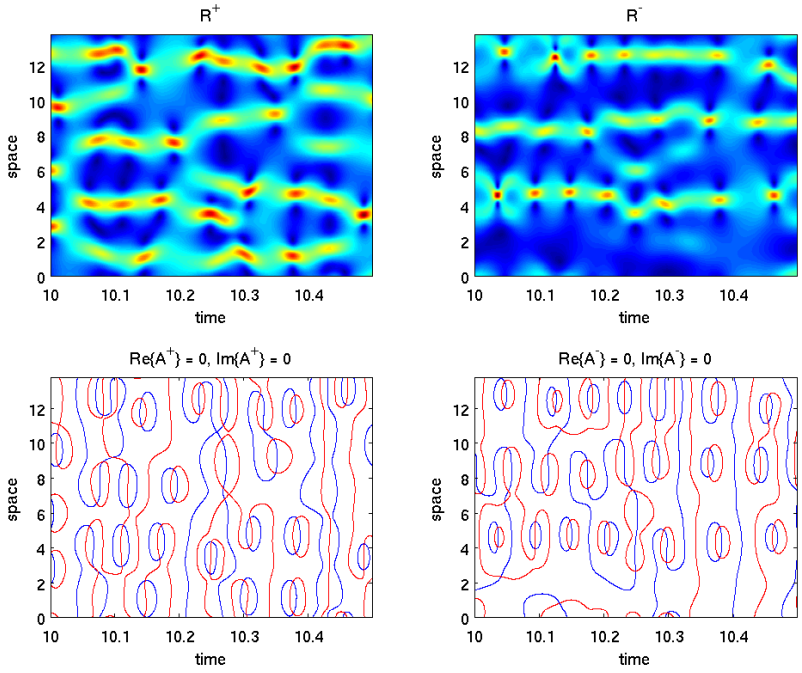


Figure 13. *Spatio-temporal plot of* $R^{\pm}$ *and zero level sets of* $\mathrm{Re}(A^{\pm})$ *(blue) and* $\mathrm{Im}(A^{\pm})$ *(red) for* $L = 13.779$*,* $F = 10$*,* $p = 0.1$*,* $\alpha = 20$*,* $\epsilon = 1/\alpha = 0.05$ *showing chaotic behavior.*

system indeed evolves toward this solution. However, for different initial conditions, as well as for the parameter values for which (2.6) is unstable, we have found that the system evolves toward standing waves, spatio-temporal chaos, or a finite-time blow-up, depending on the size of the computational domain.

We would also like to discuss two possible extensions of this work that were suggested by one of the anonymous referees of the paper. These extensions are beyond the scope of the present paper and may be a subject of a separate study.

The first extension addresses a supercritical Hopf bifurcation; i.e., we consider $\mu = -1$ instead of $\mu = 1$ in (1.2). In this case, solutions of the same type as (2.6) are expected to exist. Indeed, their characteristics and stability are governed by the same equation (2.11) with the only difference being that the factor $(F - 1)$ in the second term of the right-hand side of the equation is replaced by $(F + 1)$. Of course, the existence and stability conditions for the supercritical solutions will be different from those for subcritical solutions. However, if F is large, the relative difference between the factors $(F + 1)$ and $(F - 1)$ is small, and all the characteristics of the sub- and supercritical solutions will be very similar. This brings up an interesting question of why the criticality, which is typically a determining parameter in Hopf bifurcations, is not as crucial in the NLS limit. We believe that the reason for this is that the cnoidal waves we study are actually not created just by the Hopf bifurcation. They are basically the solutions of the NLS equation and, in fact, form a family of neutrally stable solutions. The Ginzburg–Landau nonconservative terms do not create these solutions,

as usually the Hopf bifurcation does, but rather create a slow dynamics inside this already existing family. Since what we study is exactly this dynamics, all the nonconservative terms are as important as the criticality, and the final results depend equally strongly on both the criticality and such parameters as F and p rather than the criticality itself.

The other extension addresses (2.1) in the case when the coefficients α and β have different signs, i.e., the parameter $F = \beta/\alpha < 0$. This case can be called defocusing by analogy with the terminology for the NLS equation, while the case $F > 0$ that has been considered in this paper is the focusing case. In the defocusing case, solutions of the dn type (see (2.6)) do not exist. However, one can find bounded solutions of the sn type in this case, both for sub- and supercritical bifurcations. All the characteristics and stability of these solutions can be studied similarly to how the dn-type solutions are studied in this paper.

Acknowledgment. The authors would like to thank an anonymous referee for useful comments that are reflected in this version of the paper.

REFERENCES

[1] I. S. Aranson and L. Kramer, *The world of the complex Ginzburg-Landau equation*, Rev. Modern Phys., 74 (2002), pp. 99–143.
[2] M. R. Booty, S. B. Margolis, and B. J. Matkowsky, *Interaction of pulsating and spinning waves in condensed phase combustion*, SIAM J. Appl. Math., 46 (1986), pp. 801–843.
[3] C. S. Bretherton and E. A. Spiegel, *Intermittency through modulational instability*, Phys. Lett. A, 96 (1983), pp. 152–156.
[4] D. M. G. Comissiong, L. K. Gross, and V. A. Volpert, *Nonlinear dynamics of frontal polymerization with autoacceleration*, J. Engrg. Math., 53 (2005), pp. 59–78.
[5] S. H. Davis, *Theory of Solidification*, Cambridge University Press, Cambridge, UK, 2001.
[6] Th. Gallay and M. Haragus, *Stability of small periodic waves for the nonlinear Schrödinger equation*, J. Differential Equations, 234 (2007), pp. 544–581.
[7] R. Heinrichs, G. Ahlers, and D. S. Cannell, *Traveling waves and spatial variation in the convection of a binary mixture*, Phys. Rev. A (3), 35 (1987), pp. 2761–2764.
[8] L. M. Hocking and K. Stewartson, *On the nonlinear response of a marginally unstable plane parallel flow to a two-dimensional disturbance*, Proc. Roy. Soc. London A, 326 (1972), pp. 289–313.
[9] T. Kapitula, N. Kutz, and B. Sandstede, *Stability of pulses in the master-modelocking equation*, J. Opt. Soc. Amer. B Opt. Phys., 19 (2002), pp. 740–746.
[10] E. Kaplan, E. Kuznetsov, and V. Steinberg, *Burst and collapse in traveling-wave convection of a binary-fluid*, Phys. Rev. E (3), 50 (1994), pp. 3712–3722.
[11] E. Kaplan, E. Kuznetsov, and V. Steinberg, *Phase gradient mechanism of self-focusing and collapse in nonlinear dispersive traveling waves*, Europhys. Lett., 28 (1994), pp. 237–243.
[12] E. Knobloch and J. Deluca, *Amplitude equations for traveling-wave convection*, Nonlinearity, 3 (1990), pp. 975–980.
[13] P. Kolodner, D. Bensimon, and C. M. Surko, *Traveling-wave convection in an annulus*, Phys. Rev. Lett., 60 (1988), pp. 1723–1726.
[14] P. Kolodner, G. Flatgen, and I. G. Kevrekidis, *Controlling dispersive chaos in binary-fluid convection*, Phys. Rev. Lett., 83 (1999), pp. 730–733.
[15] P. Kolodner, S. Slimani, N. Aubry, and R. Lima, *Characterization of dispersive chaos and related states of binary-fluid convection*, Phys. D, 85 (1995), pp. 165–224.
[16] L. Kramer, S. Popp, E. A. Kuznetsov, and S. K. Turitsyn, *Optical pulse collapse in defocusing active medium*, JETP Lett., 61 (1995), pp. 904–910.
[17] S. B. Margolis, H. G. Kaper, G. K. Leaf, and B. J. Matkowsky, *Bifurcation of pulsating and spinning reaction fronts in condensed two-phase combustion*, Combust. Sci. Technol., 43 (1985), pp. 127–165.

[18] B. J. Matkowsky and V. Volpert, *Coupled nonlocal complex Ginzburg-Landau equations in gasless combustion*, Phys. D, 54 (1992), pp. 203–219.
[19] J. D. Moores, *On the Ginzburg-Landau laser mode-locking model with 5th-order saturable absorber term*, Opt. Commun., 96 (1993), pp. 65–70.
[20] E. Moses, J. Fineberg, and V. Steinberg, *Multistability and confined traveling-wave patterns in a convecting binary mixture*, Phys. Rev. A (3), 35 (1987), pp. 2757–2760.
[21] S. Popp, O. Stiller, E. Kuznetsov, and L. Kramer, *The cubic complex Ginzburg-Landau equation for a backward bifurcation*, Phys. D, 114 (1998), pp. 81–107.
[22] J. A. Powell and P. K. Jakobsen, *Localized states in fluid convection and multiphoton lasers*, Phys. D, 64 (1993), pp. 132–152.
[23] W. Schöpf and L. Kramer, *Small-amplitude periodic and chaotic solutions of the complex Ginzburg-Landau equation for a subcritical bifurcation*, Phys. Rev. Lett., 66 (1991), pp. 2316–2319.
[24] W. Schöpf and W. Zimmermann, *Convection in binary fluids—amplitude equations, codimension-2 bifurcation, and thermal fluctuations*, Phys. Rev. E (3), 47 (1993), pp. 1739–1764.
[25] I. Smagin, *Stability of an Explosive Crystallization Front*, Master's thesis, Technion – Israel Institute of Technology, Haifa, Israel, 2007.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 284–310
© 2008 Society for Industrial and Applied Mathematics

Asynchronous and Synchronous Dispersals in Spatially Discrete Population Models*

Abdul-Aziz Yakubu†

Abstract. This study is on the role of synchronous and asynchronous dispersals in a discrete-time single-species population model with dispersal between two patches, where predispersal dynamics are compensatory or overcompensatory and dispersal is synchronous or asynchronous or mixed synchronous and asynchronous. It is known that single-species dispersal-linked population models behave as single-species single-patch models whenever all predispersal local dynamics are compensatory and dispersal is synchronous. However, the dynamics of the corresponding model connected by asynchronous and mixed synchronous-asynchronous dispersals depend on the dispersal rates, intrinsic growth rates, and the parameter that models the possible modes of dispersal. The species becomes extinct on at least one patch when the asynchronous dispersal rates are high, while it persists when the rates are low. In mixed synchronous-asynchronous systems, depending on the model parameters, the pioneer species either becomes extinct on all patches or persists on all patches. Overcompensatory predispersal dynamics with synchronous dispersal can lead to multiple attractors with fractal basin boundaries. However, the associated models with either asynchronous or mixed synchronous and asynchronous dispersals exhibit multiple attractors with fewer numbers of distinct attractors. That is, the long-term dynamics of synchronous dispersal-linked systems can be more sensitive to initial population sizes than that of the corresponding asynchronous and mixed synchronous-asynchronous systems. Also, synchronous, asynchronous, and mixed synchronous-asynchronous dispersals can "stabilize" the local patch dynamics from overcompensatory to compensatory dynamics. In our mixed synchronous-asynchronous model, the dominant mode of dispersal usually drives the dynamics of the full system.

Key words. asynchronous dispersal, compensatory dynamics, mixed synchronous-asynchronous dispersals, multiple attractors, overcompensatory dynamics, synchronous dispersal

AMS subject classifications. 37E05, 39A11, 54H20, 92B05, 92D25, 92D40

DOI. 10.1137/070688122

1. Introduction.

In host-parasite systems, the timing of density effects and parasitism can have a profound impact on the population dynamics [35]. Doebeli made a similar observation in a two-patch, single-species, dispersal-linked model of coupled Smith–Slatkin difference equations. He showed that differences in the timing of reproduction and dispersal enhance the stabilizing effect of dispersal [7]. Hastings [22], Gyllenberg, Söderbacka, and Ericsson [17], Doebeli [7, 8], Gonzalez-Andújar and Perry [15], and Castillo-Chavez and Yakubu [5, 44] have studied single-species discrete-time dispersal-linked models that implicitly assume no difference in the timing of reproduction and dispersal *(dispersal synchrony)*. Their work showed that

*Received by the editors April 12, 2007; accepted for publication (in revised form) by C. Castillo-Chavez October 31, 2007; published electronically April 23, 2008. This research was partially supported by grants from the National Science Foundation, the National Security Agency, and North East Fisheries Science Center (Woods Hole, MA).
http://www.siam.org/journals/siads/7-2/68812.html
†Department of Mathematics, Howard University, Washington, DC 20059 (ayakubu@howard.edu).

the interaction between local dynamics and symmetric synchronous dispersal can lead to the replacement of chaotic local dynamics by periodic dynamics for some initial population sizes.

In this paper, we introduce a single-species two-patch dispersal-linked model where predispersal dynamics are *compensatory* (equilibrium dynamics) or *overcompensatory* (oscillatory dynamics) and dispersal is synchronous or asynchronous or mixed synchronous and asynchronous [2, 3, 39, 44, 45]. The novelty of our model is in the embedding of synchronous and asynchronous models into a single framework. Depending on a single continuous parameter, the model is capable of exhibiting synchronous dispersal, asynchronous dispersal, and *mixed* synchronous and asynchronous dispersals. Under dispersal synchrony (that is, where there is no asynchronous dispersal), our model reduces to that of Hastings [22, 23], Gyllenberg, Söderbacka, and Ericsson [17], Doebeli [7, 8], and Castillo-Chavez and Yakubu [5, 44], whereas it reduces to a model of Doebeli when dispersal is asynchronous (that is, where there is no synchronous dispersal) [7, 8].

A large number of researchers have carried out extensive studies on the interplay between local dynamics and dispersal in dispersal-linked models. Early work on this was done by Cohen and Levin [6], Gadgil [14], Hastings [23], Levin [27, 28], Levin and Paine [29], and Levins [30, 31], and later work was done by Allen [1], Doebeli [7], Doebeli and Ruxton [8], Earn, Levin, and Rohani [9], Gonzalez-Andújar and Perry [15], Gyllenberg, Söderbacka, and Ericsson [17], Hanski [18], Hanski and Gilpin [19], Hastings [22], and Castillo-Chavez and Yakubu [4, 5, 44]. In this paper, we focus on the impact of synchronous and asynchronous modes of dispersals on local populations with discrete nonoverlapping generations [7, 8, 9, 15, 17, 22, 44]. In particular, we extend Doebeli's idea that the detailed timing of dispersal can affect the global dynamics of dispersal-linked systems [7, 8].

We review, in section 2, the impact of compensatory and overcompensatory dynamics on "unstructured" single-species, single-patch discrete-time models. The Beverton–Holt [2, 3, 4, 5, 11, 12, 13, 20, 21, 38], bobwhite quail "hump-with-tail" [10, 44], Ricker [4, 7, 8, 22, 24, 25, 32, 33, 34, 35, 36, 37, 40, 44, 45], and Smith–Slatkin [20, 36, 41, 45] models are used to describe either compensatory or overcompensatory dynamics. Only *pioneer* species are considered (pioneer species are species that persist at very small population sizes when left in isolation with no outside interference) [11, 12, 13].

In section 3, three basic single-species dispersal-linked models consisting of two subpopulations (with nonoverlapping generations) connected by one of the three modes of dispersals (synchronous, asynchronous, and mixed synchronous-asynchronous dispersals) are introduced. To understand the behavior of the mixed synchronous-asynchronous model, in section 4, we review prior work on the model with dispersal synchrony. Single-species dispersal-linked population models under the *same* qualitative local compensatory dynamics are known to behave as single-patch systems whenever dispersal is synchronous [5, 44]. When predispersal local dynamics are overcompensatory, dispersal synchrony can fracture the basins of attraction through its support of multiple attractors. We highlight, in section 4, the possible structures of the coexisting attractors where local populations (in the absence of dispersal) live on either a preselected *n-cycle* attractor or a chaotic attractor (overcompensatory dynamics). Hastings and others have observed similar multiple attractors in synchronous models [4, 5, 22, 44].

The model under dispersal asynchrony is studied in sections 5 and 6. We show, in section 5, that the dynamics of the full system depend on the asynchronous dispersal rates. The species

becomes extinct on at least one patch when asynchronous dispersal rates are high, while it persists when the rates are low. In sharp contrast to dispersal synchrony, dispersal asynchrony impacts compensatory local dynamics [7, 23, 28, 29].

The difference in the timing of reproduction and dispersal enlarges the asynchrony of interactions, and Doebeli predicted the "likelihood" of simple system dynamics due to asynchronous dispersal [7]. In general, dispersal can give rise to multiple attractors with interesting basin structures, whenever the local patch dynamics are overcompensatory [4, 5, 7, 9, 16, 19, 22, 26, 44]. In section 6, several examples are introduced to show that dispersal-linked models with "unstructured" overcompensatory predispersal patch dynamics connected by asynchronous symmetric or asymmetric dispersal support multiple attractors with a smaller number of distinct attractors than the corresponding model under dispersal synchrony. We use MATLAB and the Dynamics software of Nusse and Yorke to study the differences among the structures of the attractors and the differences between the synchronous and asynchronous cases [39]. Our results show that asynchronous dispersal can stabilize or shift the predispersal local dynamics from an attracting period four to a period two or to a fixed point or to a limit cycle attractor. That is, both synchronous and asynchronous dispersals can generate period-doubling reversals in dispersal-linked models under overcompensatory dynamics.

Models under mixed synchronous-asynchronous dispersals are studied in sections 7, 8, and 9. As in synchronous models, in mixed models, the pioneer species either persists on all patches or becomes extinct on all patches. In section 7, we derive conditions for the extinction (respectively, persistence) of the species on all patches. Mixed synchronous-asynchronous systems under compensatory and overcompensatory local dynamics are studied in sections 8 and 9, respectively. When the local dynamics are overcompensatory, mixed models exhibit multiple attractors with a smaller number of distinct attractors than the corresponding model under dispersal asynchrony. Section 10 discusses some possible implications of the results of this paper, and relevant mathematical details of all technical terms are collected in the appendix.

2. Predispersal local patch dynamics. In this section, we review single-species discrete-time population models without dispersal. As in [7, 22, 44], the equation for the local dynamics in each Patch $i \in \{1, 2\}$ at generation t *after* reproduction but *before* dispersal is modeled by

$$x_i(t+1) = x_i(t)g_i(x_i(t)) \qquad (i = 1, 2), \tag{1}$$

where $x_i(t)$ denotes the population size and the per capita growth functions, $g_i : [0, \infty) \to (0, \infty)$ are assumed to be strictly decreasing, positive, and twice differentiable (C^2 on $[0, \infty)$), where $g_i(0) > 1$ and $\lim_{x_i \to \infty} g_i(x_i) < 1$. System (1) is a discrete-time, single-species, population model with two (uncoupled) patches. It describes the population dynamics of pioneer species [4, 5, 11, 12, 13, 44].

Predispersal Patch i local reproduction function $f_i(x_i) = x_i g_i(x_i)$ describes the local dynamics of the species, where x_i is the measure of the size of the population in the patch. Each f_i has a unique positive fixed point denoted by X_i. Since g_i is a strictly decreasing continuous function, $f_i(x_i) > x_i$ whenever $0 < x_i < X_i$ and $f_i(x_i) < x_i$ whenever $x_i > X_i$. Consequently, $I_i \equiv f_i([0, X_i])$ is a global attractor. That is, every initial population eventually reaches a limit in I_i.

We focus on two types of local dynamics—*compensatory* and *overcompensatory* dynamics.

Definition 1. *Patch i predispersal local dynamics are compensatory whenever all positive population sizes approach the positive equilibrium at X_i monotonically under f_i iterations* [4, 38, 44].

Definition 2. *Patch i predispersal local dynamics are overcompensatory whenever some positive population sizes "overshoot" the positive equilibrium at X_i under f_i iterations (that is, $f_i'(X_i) < 0$)* [4, 38, 44].

If f_i increases monotonically from zero with the rate of increase slowing down as x_i gets large, then all population sizes "undershoot" the globally attracting positive equilibrium, and by Definition 1 Patch i local dynamics are compensatory. The Beverton–Holt stock recruitment model, $f_i(x_i) = \frac{a_i x_i}{1+b_i x_i}$, portrays compensatory dynamics in Patch i whenever $a_i > 1$ and $b_i > 0$ [4, 44, 45]. If f_i is an orientation-reversing one-hump map with a stable positive fixed point (respectively, an unstable positive fixed point), then the return to the stable fixed point takes the form of damped oscillations (respectively, the local behavior near the unstable fixed point takes the form of divergent oscillations), and by Definition 2 Patch i dynamics are overcompensatory. Whenever $r_i > 1$ and f_i is Ricker's model, $f_i(x_i) = x_i \exp(r_i - x_i)$, then the dynamics in Patch i are overcompensatory [2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 20, 21, 22, 32, 33, 34, 35, 36, 37, 38, 39, 40, 44, 45]. In general, f_i supports either an *n-cycle* (nonchaotic) attractor with $n > 1$ or a chaotic (interval) attractor whenever Patch i dynamics are overcompensatory and the positive fixed point is unstable.

A detailed description of functions under compensatory or overcompensatory dynamics requires the introduction of the concept of an *α-monotone concave map*.

Definition 3. *f_i is an α-monotone concave map if $f_i'(x_i) > 0$ and $f_i''(x_i) < 0$ for each $x_i \in [0, \alpha)$* [4, 38, 44].

Patch i population is under compensatory dynamics at population sizes in the interval $[0, \alpha)$ whenever f_i is an *α-monotone concave map* with a unique positive fixed point in the open interval $(0, \alpha)$ (see [44, Definition 3]). The bobwhite quail "hump-with-tail" model $f_i(x_i) = x_i(k_i + \frac{K_i}{1+x_i^{n_i}})$, the Ricker model $f_i(x_i) = x_i \exp(r_i - x_i)$, and the Smith–Slatkin model $f_i(x_i) = \frac{a_i x_i}{1+(b_i x_i)^{l_i}}$ describe overcompensatory and compensatory dynamics (depending on parameter values). If $l_i = 1$ and $a_i > 1$, the Smith–Slatkin model reduces to the Beverton–Holt model, an *∞-monotone concave map* (compensatory dynamics [2, 36, 42, 44]).

3. Synchronous and asynchronous dispersal-linked two-patch model. Hastings [22], Gyllenberg, Söderbacka, and Ericsson [17], Doebeli [7, 8], Yakubu and Castillo-Chavez [44], and others have studied discrete-time single-species dispersal-linked population models that implicitly assume that the timing of reproduction and dispersal *do not* differ from patch to patch. A two-patch version of these models with *dispersal synchrony* is given by the following system of coupled nonlinear difference equations:

$$\left.\begin{aligned} x_1(t+1) &= (1-d_1)f_1(x_1(t)) + d_2 f_2(x_2(t)), \\ x_2(t+1) &= d_1 f_1(x_1(t)) + (1-d_2)f_2(x_2(t)). \end{aligned}\right\} \tag{2}$$

In system (2), reproduction occurs prior to dispersal within each generation and in each patch. After reproduction, the constant fraction $d_1 \in (0,1)$ of the population disperses from Patch 1 to Patch 2 while the constant fraction $d_2 \in (0,1)$ disperses from Patch 2 to Patch 1.

Doebeli, in 1995, studied a simple two-patch discrete-time model of coupled Smith–Slatkin single-species ecological models where the timing of reproduction and dispersal *differs* from patch to patch. In Doebeli's two-patch model, in each generation reproduction occurs in Patch 1 first, followed by the dispersal, from Patch 1 to Patch 2, of the fraction d_1 of the population. As a result, Patch 2 population experiences the effects of its own density as well as that of the newly dispersed individuals from Patch 1 to Patch 2. In Patch 2, the fraction d_2 of the population disperses from Patch 2 to Patch 1 after reproduction [7]. The dynamics of the two-patch system under *asynchronous* dispersal are then described by the following system of coupled nonlinear difference equations:

$$\left.\begin{array}{rcl} x_1(t+1) &=& (1-d_1)f_1(x_1(t)) + d_2x_2(t)g_2(x_2(t)+d_1f_1(x_1(t))), \\ x_2(t+1) &=& (1-d_2)x_2(t)g_2(x_2(t)+d_1f_1(x_1(t))), \end{array}\right\} \tag{3}$$

where $0 < d_1, d_2 < 1$ and $f_1(x_1) = x_1g_1(x_1)$.

In Doebeli's simple model with dispersal asynchrony, at the next generation, the population size in Patch 1 is increased by the dispersal from Patch 2. However, unlike the Patch 1 population size, the population size in Patch 2 at the next generation is *not* increased by the dispersal from Patch 1. By their own nature, such simple models do not incorporate many of the important biological factors. However, they often provide useful insights to help our understanding of complex processes.

To embed synchronous and asynchronous dispersals into a single framework, we let the constant parameter $\gamma \in [0,1]$ span the range of possible modes of dispersal, where $\gamma = 0$ implies synchronous dispersal, $\gamma = 1$ implies asynchronous dispersal, and $\gamma \in (0,1)$ implies *mixed* synchronous and asynchronous dispersal. This leads to the following equations describing the dispersal phase:

$$\left.\begin{array}{rcccl} x_1(t+1) &=& F_1(x_1(t),x_2(t)) &=& (1-d_1)f_1(x_1(t)) + d_2x_2(t)g_2(\overline{x}(t)), \\ x_2(t+1) &=& F_2(x_1(t),x_2(t)) &=& (1-\gamma)d_1f_1(x_1(t)) + (1-d_2)x_2(t)g_2(\overline{x}(t)), \end{array}\right\} \tag{4}$$

where

$$\overline{x}(t) = x_2(t) + \gamma d_1 f_1(x_1(t)).$$

Unlike Doebeli's model, in models (2) and (4), at the next generation the population size in Patch 1 is increased by the dispersal from Patch 2, while that of Patch 2 is increased by the dispersal from Patch 1. In each generation, reproductions in Patch 2 of models (3) and (4) experience crowding from the dispersal from Patch 1.

The vector of population densities $x(t) = (x_1(t), x_2(t))$ is written as $x = (x_1, x_2)$ so that the *dispersal-linked* function is

$$F : \mathbb{R}^2_+ \to \mathbb{R}^2_+,$$

where

$$F(x_1, x_2) = (F_1(x_1, x_2), F_2(x_1, x_2)).$$

Then F^t is the dispersal-linked function composed with itself t times. $F^t_i(x)$ is the ith component of F^t evaluated at the point x in $\mathbb{R}^2_+$. In system (4), F^t gives the population densities in generation t.

When $\gamma = 0$ (respectively, $\gamma = 1$), dispersal is synchronous (respectively, asynchronous), and system (4) reduces to system (2) (respectively, system (3)). Dispersal is *symmetric* when $d_1 = d_2$, while it is *asymmetric* when $d_1 \neq d_2$. When there are no dispersals, $d_1 = d_2 = 0$ and system (4) reduces to the uncoupled system (1). The predispersal basic demographic reproductive number in each patch ($d_1 = d_2 = 0$) is

$$\Re_d^i = g_i(0).$$

$\Re_d^i > 1$ guarantees the successful invasion and survival of the discretely reproducing population in Patch i, while $\Re_d^i < 1$ guarantees the extinction of the initial population in the patch (no dispersal). We assume throughout that the species is a pioneer in each Patch $i \in \{1, 2\}$. That is, $\Re_d^i > 1$.

In system (4), there is no population explosion.

Lemma 1. *In system (4), the positive cone is positively invariant and no point has an unbounded orbit.*

4. Dispersal synchrony in two-patch models. In this section, we consider system (4) with *only* dispersal synchrony (that is, system (2) or system (4) with $\gamma = 0$). Others have studied synchronous dispersal models, and in this section we review some of these prior works. When all local dynamics are compensatory, Yakubu and Castillo-Chavez proved that system (2) supports a positive equilibrium that attracts all positive initial population sizes [44]. That is, when all local dynamics are compensatory, the qualitative dynamics of system (2) with symmetric or asymmetric synchronous dispersal between patches is qualitatively equivalent to those of each of the local single patches before dispersal. With synchronous symmetric dispersal and symmetric initial population sizes, system (2) behaves as a single patch system whenever the local reproduction functions are identical ($f_1 = f_2$) and the predispersal local dynamics are either compensatory or overcompensatory.

In 1993, Hastings [22] and Gyllenberg, Söderbacka, and Ericsson [17] used two identical logistic difference equations in system (2) with parameters in the chaotic regime to illustrate that synchronous dispersal-linked population models are capable of supporting multiple attractors with complicated attraction-basin boundaries. In a recent paper, Yakubu and Castillo-Chavez studied the role of synchronous dispersal in generating multiple attractors where local dynamics are overcompensatory [44]. They focused on situations where the local populations (in the absence of dispersal) live on either a preselected *n-cycle* attractor or a chaotic attractor. Yakubu and Castillo-Chavez supported the results of Hastings and obtained that synchronous dispersal can force the preselected (chaotic or nonchaotic) attractor to coexist with one or more "new" attractors (multiple attractors). Example 1 illustrates multiple attractors in system (2) with synchronous symmetric dispersal, where the local dynamics are governed by the Ricker model. In Example 1, we choose the values of the parameters so that the predispersal local dynamics and the full system dynamics under synchronous symmetric dispersal are as listed in Table 1.

Example 1. *Consider system (2) with the Ricker model*

$$f_i(x_i) = x_i \exp(r_i - x_i)$$

for each $i \in \{1, 2\}$. Set the following parameter values:

$$r = r_1 = r_2 \in (2, 2.52) \quad \textit{and} \quad d_1 = d_2 = 0.03.$$

Table 1

Predispersal local dynamics versus postdispersal synchronous dynamics.

r values	Predispersal attractors	Synchronous attractors
1. $(2, 2.53)$	2-cycle	two 2-cycles (see Fig. 1)
2. $(2.53, 2.59)$	4-cycle	4- and 2-cycles
3. $(2.66, 2.68)$	8-cycle	8-, 4-, and 2-cycles
4. $(2.69, 2.6901)$	16-cycle	16-, 8-, 4-, and 2-cycles
5. $(2.695, 2.701)$	chaotic attractor	four attractors (see Fig. 2)

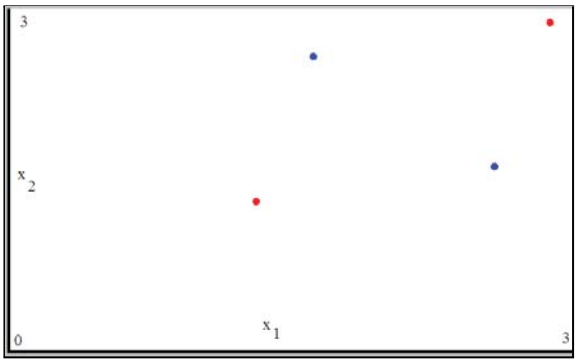


Figure 1. *Two coexisting attractors: Symmetric* 2*-cycle (red dots) and an asymmetric* 2*-cycle (blue dots) in the* (x_1, x_2)*-plane, where* $r_1 = r_2 = 2.1$ *and* $d_1 = d_2 = 0.03$*. Figure* 1 *is plotted over* 3000 *time steps.*

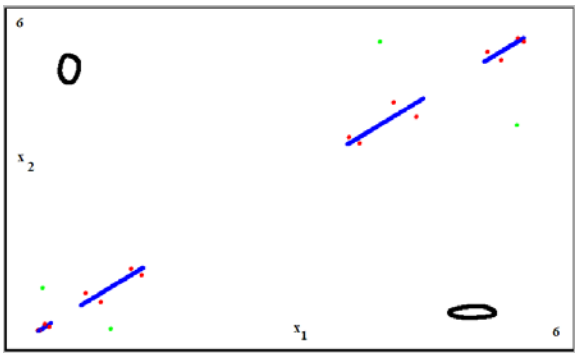


Figure 2. *Four attractors: Symmetric* 4*-piece chaotic attractor (blue region), asymmetric* 4*-cycle (green dots), asymmetric* 16*-cycle (red dots), and a period-*2 *limit cycle (black region), where* $r_1 = r_2 = 2.7$ *and* $d_1 = d_2 = 0.03$*. Figure* 2 *is plotted over* 5000 *time steps.*

The predispersal local dynamics in Example 1, $f_i(x_i) = x_i \exp(r_i - x_i)$, have a stable positive fixed point at $X_i = r_i$ whenever $0 < r_i < 2$ [44]. As r_i is increased past 2, the fixed point X_i undergoes a period-doubling bifurcation route to chaos [33, 34, 35, 39]. In Example 1, the predispersal identical local patches are on a 2-cycle attractor (overcompensatory dynamics), and the full system with symmetric dispersal supports multiple attractors—a symmetric 2-cycle attractor coexisting with an asymmetric 2-cycle attractor (see Figure 1 and Table 1). To study the impact of increasing the identical intrinsic growth rates, $r = r_1 = r_2$, the symmetric dispersal rates are kept fixed at $d_1 = d_2 = 0.03$ while r is increased past 2.52.

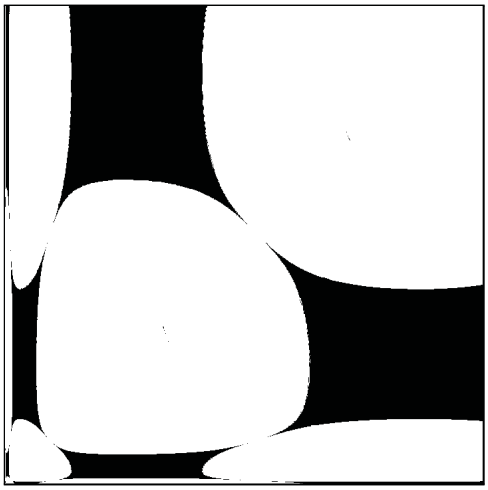


Figure 3. *Basins of attraction of two coexisting* 2*-cycle attractors, where the parameters are exactly as in Figure* 1*. The white region is the basin of attraction of the black "specks" (*2*-cycle) in the figure, and the black region is the basin of attraction of the other coexisting attractor. On the horizontal axis* $0 \leq x_1 \leq 4$*, and on the vertical axis* $0 \leq x_2 \leq 4$*. Figure* 3 *is plotted over* 5000 *time steps.*

When $r \in (2.53, 2.59)$, the predispersal identical local patches are on a 4-cycle attractor, and the full system with symmetric dispersal supports two coexisting attractors consisting of a symmetric 4-cycle attractor and an asymmetric 2-cycle attractor (see Table 1). At $r \in (2.6, 2.65)$, the predispersal identical local patches are on a 4-cycle attractor, while the full system with symmetric dispersal supports three coexisting attractors consisting of a symmetric 4-cycle attractor, an asymmetric 4-cycle attractor, and an asymmetric 2-cycle attractor (see Table 1). For values of $r \in (2.66, 2.68)$, the predispersal identical local patches are on an 8-cycle attractor, and the full system with symmetric dispersal supports three coexisting attractors consisting of a symmetric 8-cycle attractor, an asymmetric 4-cycle attractor, and an asymmetric 2-cycle attractor (see Table 1). At $r \in (2.69, 2.6901)$, the predispersal identical local patches are on a 16-cycle attractor, while the full system with symmetric dispersal supports four coexisting attractors consisting of a symmetric 16-cycle attractor, an asymmetric 8-cycle attractor, an asymmetric 4-cycle attractor, and an asymmetric 2-cycle attractor (see Table 1). When $r \in (2.695, 2.701)$, the predispersal identical local patches are on a chaotic attractor, and the full system with symmetric dispersal supports four coexisting attractors consisting of a symmetric chaotic attractor, a period-2 limit cycle attractor, an asymmetric 4-cycle attractor, and an asymmetric 16-cycle attractor (see Figure 2 and Table 1).

The qualitative structure and number of the attractors in dispersal-linked population models are the result of a complex interaction between the dispersal rate and predispersal local patch dynamics. The *basins of attraction*, the set of all population sizes that eventually settle into an attractor under iteration, may provide critical information on a variety of issues including the final attractor observed. In Example 1, the *Dynamics* software of Nusse and Yorke is used to study the nature of the basins of attraction of the multiple attractors in Figures 1 and 2 [39]. As in [44], Figures 3 and 4 highlight that the basins of attraction

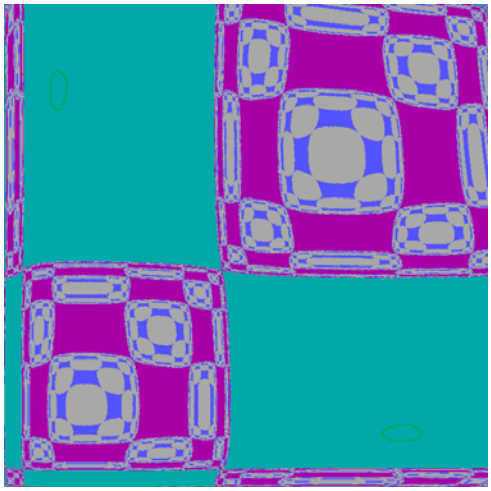


Figure 4. *Basins of attraction of the* 4 *attractors, where the parameters are exactly as in Figure* 2*. On the horizontal axis* $0 \leq x_1 \leq 6$*, and on the vertical axis* $0 \leq x_2 \leq 6$*. Figure* 4 *is plotted over* 5000 *time steps.*

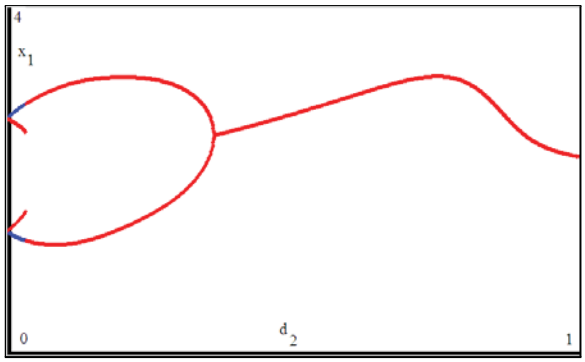


Figure 5. *The full system shifts from two* 2*-cycle attractors (red* 2*-cycle and blue* 2*-cycle) to a single (red)* 2*-cycle attractor or to a single (red) fixed point attractor with synchronous asymmetric dispersal, where* $r_1 = r_2 = 2.1$ *and* $d_1 = 0.03$ *while* d_2 *is varied continuously between* 0.03 *and* 1*.*

become thinner and their boundaries exhibit increasing fractal structures as the number of attractors increases or as the period of the attractors increases.

4.1. Asymmetric dispersal synchrony. In Figures 1, 2, 3, and 4, only synchronous *symmetric* dispersal is assumed. To illustrate the impact of synchronous *asymmetric* dispersal on Figure 1, we keep the parameters fixed at $r = 2.1$ and $d_1 = 0.03$, while d_2 is varied continuously between 0 and 1 (see Figure 5). The full system stabilizes or shifts from the two coexisting 2-cycle attractors to a single 2-cycle attractor (saddle-node bifurcation reversal) or to a single fixed point attractor (period-doubling reversal). Similarly, when the predispersal local dynamics are chaotic, asymmetric dispersal synchrony can change the number and nature of the coexisting attractors. For example, when $r = 2.7$, $d_1 = 0.03$, and $d_2 = 0.7$, the full system appears to support a single attractor, an attracting fixed point at $(4.308, 1.513)$ where the predispersal local dynamics are chaotic (see Figure 6).

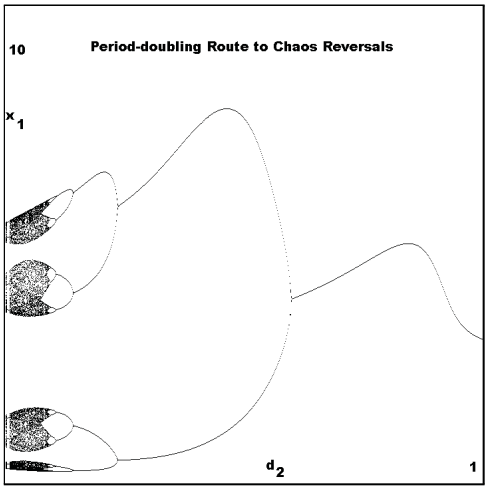


Figure 6. *The full system shifts from four coexisting attractors to a single fixed point attractor with synchronous asymmetric dispersal, where* $r_1 = r_2 = 2.7$ *and* $d_1 = 0.03$ *while* d_2 *is varied continuously between* 0.03 *and* 1.

As in [44], Figures 5 and 6 highlight that for most values of asymmetric dispersal rates ($d_1 \neq d_2$) simple dynamics are supported in the full system, in sharp contrast to Figures 1, 2, 3, and 4, where dispersal is symmetric and multiple attractors with complicated basins of attraction are supported. That is, asymmetry enhances the stabilizing effect of dispersal in system (2), where dispersal is completely synchronous.

5. Dispersal asynchrony in two-patch models. The work of Doebeli shows the dependence of the dynamics of dispersal-linked models on asynchronous dispersal rates where the predispersal local dynamics are chaotic (overcompensatory dynamics) and are governed by the Smith–Slatkin model [7]. In this section, we consider system (4) with *only* dispersal asynchrony (that is, system (3) or system (4) with $\gamma = 1$), where the predispersal dynamics are noncyclic (compensatory), cyclic, and chaotic (overcompensatory). Next, we show that the species becomes extinct (respectively, persists) on at least one patch when the asynchronous dispersal rates are high (respectively, low). We collect these in the following result.

Theorem 1. *In system* (3)*, we have the following:*

(i) $(1-d_2)\Re_d^2 < 1$ *implies that the* ω*-limit set of every positive population vector is a subset of* $[0,\infty)\times\{0\}$*. Hence, the species becomes extinct in Patch* 2*.*

(ii) $(1-d_1)\Re_d^1 > 1$ *and* $(1-d_2)\Re_d^2 > 1$ *imply that* $(0,0)$ *is unstable and there is no catastrophic extinction of the species in both Patches* 1 *and* 2*.*

(iii) $\Re_d = \max\{(1-d_1)\Re_d^1, (1-d_2)\Re_d^2\} < 1$ *implies that* $(0,0)$ *is globally asymptotically stable. Hence, the species becomes extinct in both Patches* 1 *and* 2*.*

(iv) $(1-d_1)\Re_d^1 > 1$ *and* $(1-d_2)\Re_d^2 < 1$ *imply that* $(g_1^{-1}(\frac{1}{1-d_1}), 0)$ *is a globally stable fixed point in* $(0,\infty)\times[0,\infty)$ *whenever the Patch* 1 *dynamics are compensatory. Hence, the species persists in Patch* 1*, while it becomes extinct in Patch* 2*.*

The proof of Theorem 1 is in the appendix.

The four cases in Theorem 1 are not exhaustive. For example, conditions for the persistence of the species in both Patches 1 and 2 via a stable positive equilibrium population

vector can be obtained if one assumes the persistence of the species in Patch 2 where the asynchronous dispersal rates are low and the local dynamics are compensatory. To illustrate this in the simplest setting, we assume that the predispersal Patch 1 local population has reached the positive equilibrium X_1, and we let $f_1(x_1) \equiv X_1$ [43, 46]. Then system (3) reduces to the system

$$\left.\begin{array}{rcl} x_1(t+1) &=& (1-d_1)X_1 + d_2 x_2(t) g_2(x_2(t)+d_1X_1), \\ x_2(t+1) &=& (1-d_2)x_2(t)g_2(x_2(t)+d_1X_1). \end{array}\right\} \tag{5}$$

If the dispersal rate from Patch 2 to Patch 1 is low, then system (5) supports a globally stable positive equilibrium whenever the predispersal local patch dynamics are compensatory and the Patch 1 carrying capacity is small. That is, dispersal asynchrony like dispersal synchrony is capable of supporting the persistence of the pioneer species in all patches. We summarize these in the following result.

Theorem 2. *In system (5), let each local patch dynamics be modeled by f_i, an α-monotone concave map, with the positive fixed point $X_i \in (0,\alpha)$. If $(1-d_2)\Re_d^2 > 1$, then the positive equilibrium population vector,*

$$\left((1-d_1)X_1 + \frac{d_2}{1-d_2}\left(g_2^{-1}\left(\frac{1}{1-d_2}\right) - d_1X_1\right),\ g_2^{-1}\left(\frac{1}{1-d_2}\right) - d_1X_1\right),$$

is globally attracting whenever $X_1 < \frac{g_2^{-1}(\frac{1}{1-d_2})}{d_1}$. That is, the dispersal-linked system supports a globally stable positive fixed point whenever the predispersal local patch dynamics are compensatory.

The proof of Theorem 2 is in the appendix.

In Example 2, we use compensatory local dynamics via the Beverton–Holt model to illustrate the dependence of the dynamics of system (3) on the asynchronous dispersal rates.

Example 2. *Consider system (3) with*

$$f_i(x_i) = \frac{a_i x_i}{1+b_i x_i}$$

for each $i \in \{1,2\}$. Set the following parameter values:

$$a_1 = 2,\ a_2 = 2.1,\ b_1 = b_2 = 1,\ \text{and}\ d_1 = d_2 = 0.01.$$

In Example 2, the local dynamics in both patches are compensatory, where $g_1(0) = a_1$, $g_2(0) = a_2$, $X_1 = 0.01$, $X_2 = 0.1$, $(1-d_2)\Re_d^2 > 1$, and $X_1 < \frac{g_2^{-1}(\frac{1}{1-d_2})}{d_1}$. Consequently, the resulting system with symmetric dispersal asynchrony supports a globally stable positive equilibrium population vector at $(1.002, 1.069)$ (Theorem 2). We study the impact of increasing the dispersal parameters on Example 2. In Figures 7, 8, and 9 the parameters a_1, a_2, b_1, and b_2 are kept fixed at their current values.

In Figure 7, symmetric dispersal is assumed and $d_1 = d_2$ is varied continuously between 0 and 1. The population in each patch decreases to zero monotonically with increasing symmetric dispersal rates (see Figure 7). That is, when the symmetric dispersal rate is high and

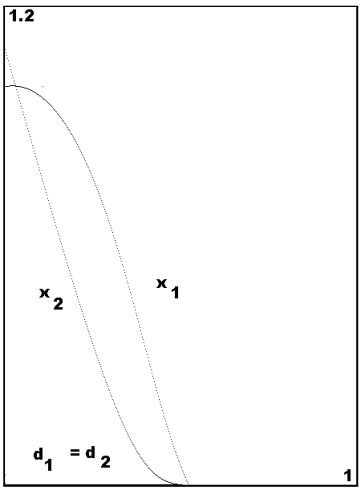


Figure 7. *When the symmetric dispersal rate $d_1 = d_2$ is low the species persists in both Patches* 1 *and* 2. *However, when $d_1 = d_2 > 0.53$ it becomes extinct in both Patches* 1 *and* 2.

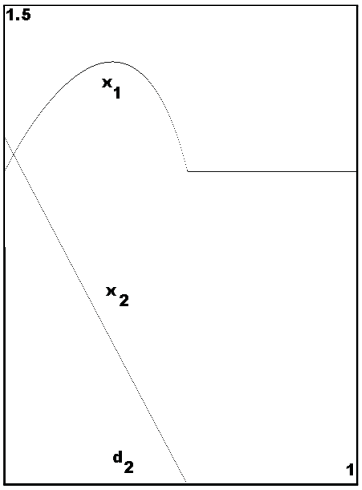


Figure 8. *Persistence in both Patches* 1 *and* 2 *with asymmetric dispersal asynchrony.*

$\Re_d < 1$, the species becomes extinct in both Patches 1 and 2 (Theorem 1), and when the symmetric dispersal rate is low and $\Re_d > 1$, the species persists in both Patches 1 and 2.

Asymmetric dispersal is assumed in Figures 8 and 9. In Figure 8, d_1 is fixed at $d_1 = 0.01$ while d_2 is varied continuously between 0 and 1. As the values of d_2 increase, the Patch 2 population decreases monotonically to zero while the Patch 1 population first increases to a maximum value before decreasing monotonically to the carrying capacity in Patch 1. As in Figure 7, the species persists in both Patches 1 and 2 when the asymmetric dispersal rate is low. In contrast to Figure 7, when the dispersal rate from Patch 2 to Patch 1 is high and $(1 - d_1)\Re_d^1 > 1$, the species persists in Patch 1, while it is extinct in Patch 2 (Theorem 1 and Figure 8). In Figure 8, this explains the sudden leveling of the graphs of x_1 and x_2 at high levels of d_2. In Figure 9, d_2 is fixed at $d_2 = 0.01$ while d_1 is varied continuously between

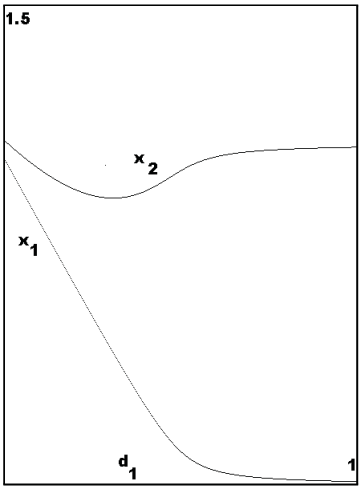


Figure 9. *Persistence in both Patches* 1 *and* 2 *with asymmetric dispersal asynchrony.*

0 and 1. As the values of d_1 increase, the population in Patch 1 decreases monotonically to a very small positive value, while the Patch 2 population first decreases to a positive minimum value before increasing monotonically to a value close to the carrying capacity in Patch 2 (see Figure 9). In Figure 9, for all values of the asymmetric dispersal rates, the species persists in both Patch 1 and Patch 2.

Figures 7, 8, and 9 show that dispersal asynchrony is capable of shifting the local dynamics from persistence of the pioneer species to its extinction on at least one patch. Thus, dispersal asynchrony impacts local patch dynamics. Clearly, these new results have highlighted only a few possibilities with the selected examples.

6. Multiple attractors: Asynchronous versus synchronous symmetric dispersal. Population models with "unstructured" overcompensatory predispersal local patch dynamics connected by either asynchronous or synchronous dispersals are capable of supporting multiple attractors. However, asynchronous symmetric dispersal-linked models are more likely to support multiple attractors with smaller numbers of distinct attractors than the corresponding models under dispersal synchrony. In this section, we use examples to highlight the differences among the attractors and the differences between the asynchronous and synchronous cases.

6.1. Symmetric dispersal asynchrony. In this section, we consider the asynchronous dispersal model, system (3), with symmetric dispersal (that is, $d_1 = d_2 = d$), where the per capita growth rates are identical (that is, $g_1 = g_2 = g$). Doebeli, in 1994, used two identical Smith–Slatkin difference equations with parameters in a chaotic regime to describe the predispersal local dynamics, where the asynchronous dispersal is symmetric [7].

In section 3, we generated multiple attractors in synchronous systems under symmetric dispersal ($d_1 = d_2$), where the identical predispersal local reproduction function is the Ricker model [10, 37, 44]. To study the corresponding asynchronous symmetric dispersal case, we repeat those results using system (3) and the identical Ricker model as the predispersal local dynamics, where asynchronous dispersal is symmetric. As in Example 1, in Example 3 we choose the values of the parameters so that the predispersal local dynamics and full system

Table 2

Predispersal local dynamics versus postdispersal asynchronous dynamics.

r values	Predispersal attractors	Asynchronous attractors
1. $(2, 2.06)$	2-cycle	a fixed point
2. $(2.07, 2.09)$	2-cycle	a limit cycle
3. $(2.098, 2.2)$	2-cycle	a 2-cycle (Fig. 10)
4. $(2.6, 2.65)$	4-cycle	two 4-cycles
5. $(2.66, 2.68)$	8-cycle	two 4-cycles
6. $(2.69, 2.6901)$	16-cycle	8-cycle and 4-cycle
7. $(2.695, 2.701)$	chaotic	8-cycle and 4-cycle (Fig. 11)
8. 2.8	chaotic	two chaotic attractors (Fig. 12)

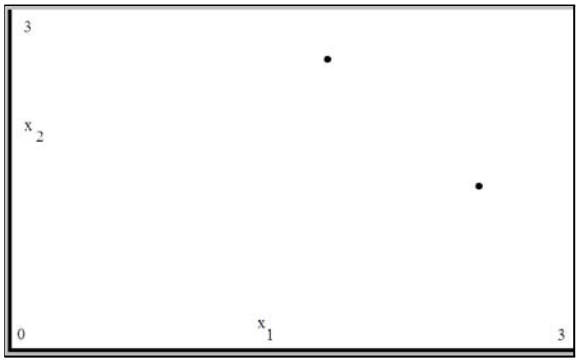


Figure 10. *A single* 2*-cycle (black dots) attractor in system* (3) *with* $g(x_i) = \exp(r - x_i)$ *and symmetric asynchronous dispersal, where the parameters are exactly as in Figure* 1*. Figure* 10 *is plotted over* 3000 *time steps.*

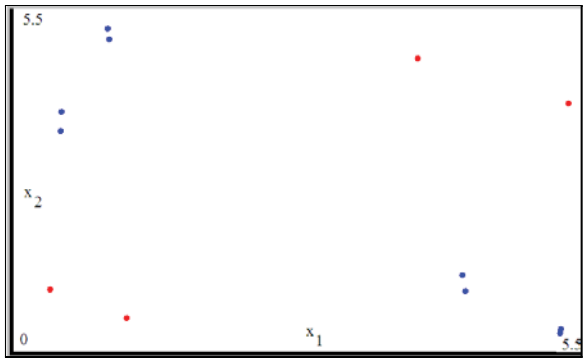


Figure 11. *Multiple attractors in system* (3) *with* $g(x_i) = \exp(r - x_i)$ *and symmetric asynchronous dispersal: A* 4*-cycle (red dots) attractor coexisting with an* 8*-cycle (blue dots) attractor, where the parameters are exactly as in Figure* 2*. Figure* 11 *is plotted over* 3000 *time steps.*

under asynchronous symmetric dispersal are as listed in Table 2.

Example 3. *Consider system* (3) *with the Ricker predispersal identical local dynamics*

$$g(x_i) = \exp(r - x_i).$$

Set the following parameter values:

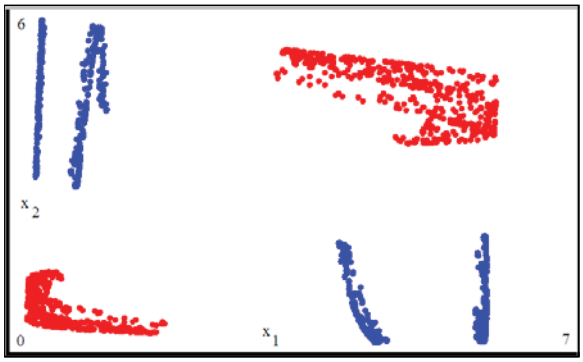


Figure 12. *Multiple chaotic attractors in system* (3)*, with* $g(x_i) = \exp(r - x_i)$ *and symmetric asynchronous dispersal. A* 2*-piece chaotic attractor along the diagonal (red region) coexisting with a* 4*-piece chaotic attractor off the diagonal (blue region), where* $r = 2.8$ *and* $d = 0.03$ *in Example* 3*. Figure* 12 *is plotted over* 5000 *time steps.*

$$r \in (2, 2.52) \text{ and } d = d_1 = d_2 = 0.03.$$

For values of the parameter $r \in (2, 2.52)$, the predispersal identical local patches are on a 2-cycle attractor, and the full system with symmetric dispersal synchrony supports two 2-cycle attractors (see Table 1). However, the dynamics of the corresponding system under symmetric dispersal asynchrony depends on the value of r. It supports a single fixed point attractor when $r \in (2, 2.06)$, a single limit cycle attractor when $r \in (2.07, 2.09)$, a single 2-cycle attractor when $r \in (2.098, 2.2)$ (no multiple attractors; see Figure 10), and two 2-cycle attractors when $r \in (2.3, 2.5)$. At $r \in (2.6, 2.65)$, the predispersal identical local patches are on a 4-cycle attractor, and the full system with symmetric dispersal asynchrony supports two 4-cycle attractors, where the corresponding synchronous model supports a 4-cycle attractor coexisting with a 2-cycle attractor. When $r \in (2.66, 2.68)$, the predispersal identical local patches are on an 8-cycle attractor, and the full system with symmetric dispersal synchrony supports three 4-cycle attractors, where the corresponding asynchronous model supports two 4-cycle attractors. For values of $r \in (2.69, 2.6901)$, the predispersal identical local patches are on a 16-cycle attractor, and the full system with symmetric dispersal synchrony supports four attractors, where the corresponding asynchronous model supports an 8-cycle attractor coexisting with a 4-cycle attractor. At $r \in (2.695, 2.701)$, the predispersal identical local patches are on a chaotic attractor, and the full system with symmetric dispersal synchrony supports four attractors, where the corresponding asynchronous model supports an 8-cycle attractor coexisting with a 4-cycle attractor (see Table 2 and Figure 11). Figure 12 demonstrates that population models under dispersal asynchrony are capable of supporting coexisting chaotic attractors.

As in Figures 3 and 4, the *Dynamics* software of Nusse and Yorke is used to study the nature of the basins of attraction of the multiple attractors in Figure 11 (see Figure 13) [39].

Figures 1, 2, 3, 4, 10, 11, 12, and 13 together with Tables 1 and 2 illustrate that asynchronous symmetric dispersal-linked models support multiple attractors with simpler basins of attraction than the corresponding synchronous symmetric ones. However, in both dispersal-linked models, our results show that the boundary between the initial population sizes leading

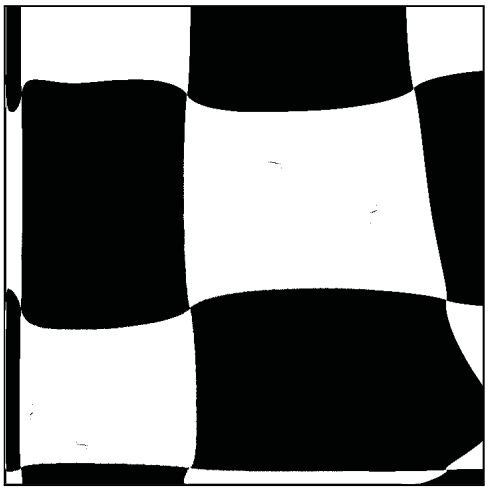


Figure 13. *Basins of attraction of the two coexisting* 4*-cycle and* 8*-cycle attractors in Figure* 11*, where the parameters are exactly as in Figure* 2*. The black region is the basin of attraction of the* 4*-cycle (four black "specks" in the white region) and the white region is that of the* 8*-cycle. On the horizontal axis* $0 \le x_1 \le 4$*, and on the vertical axis* $0 \le x_2 \le 4$*. Figure* 13 *is plotted over* 5000 *time steps.*

to each of the coexisting attractors is a fractal that fills up the entire set of initial population sizes. Consequently, both deterministic dispersal-linked models exhibit sensitive dependence of the long-term dynamical behavior on initial population sizes. Fractal basin boundaries have been studied in synchronous dispersal-linked models [5, 22, 44], epidemic models [4], as well as in physics [16, 26, 39].

6.2. Asymmetric dispersal asynchrony. In Example 3, asynchronous *symmetric* dispersal is assumed. To illustrate the impact of asynchronous *asymmetric* dispersal we now assume asymmetric dispersal in system (3), where the per capita growth rates are identical (that is, $g_1 = g_2 = g$ and $d_1 \neq d_2$).

As in Examples 1 and 3, we use the Ricker model as predispersal local population dynamics. To study the impact of asynchronous *asymmetric* dispersal on Figure 10, we keep the parameters in Example 3 fixed at $r = 2.1$ and $d_1 = 0.03$ while d_2 is varied continuously between 0 and 1 (see Figures 14 and 15). The full system stabilizes or shifts from the single 2-cycle attractor to a single limit cycle attractor (discrete Hopf bifurcation) or to a single fixed point attractor.

Similarly, when the predispersal local dynamics are chaotic, asymmetric dispersal asynchrony can change the number and nature of the coexisting attractors (see Figure 16). Figures 14, 15, and 16 highlight that the stabilizing effect of dispersal is much larger with asymmetry. Thus, asynchronous or synchronous asymmetric dispersals can stabilize or shift the local dynamics from a stable cycle or to a stable fixed point or to a stable limit cycle. However, high asynchronous dispersal rates can lead to the extinction of the species on at least one patch (Theorem 1).

7. Mixed synchronous-asynchronous dispersals in two-patch models. In this section and the next two sections, we consider system (4) with *mixed* synchronous and asynchronous

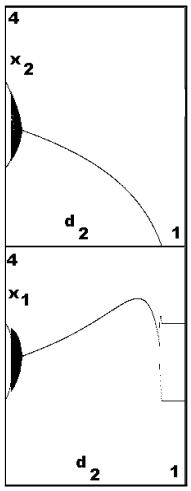


Figure 14. *After period-doubling reversals and Hopf bifurcation, Patch* 2 *population decreases monotonically to zero, while Patch* 1 *population increases monotonically to a maximum value before decreasing to the Patch* 1 *predispersal* 2*-cycle dynamics, where* $r_1 = 2.1$*,* $d_1 = 0.03$*, and* d_2 *is varied continuously between* 0 *and* 1*.*

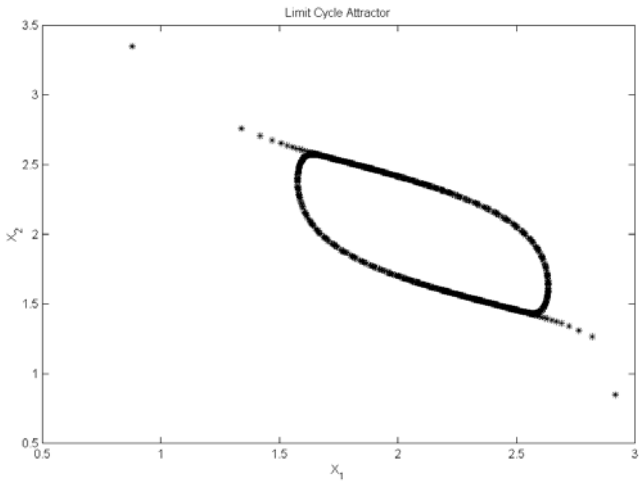


Figure 15. *A limit cycle attractor with asymmetric dispersal asynchrony, where* $r_1 = 2.1$*,* $d_1 = 0.03$*, and* $d_2 = 0.032$*.*

dispersals (that is, system (4) with $0 < \gamma < 1$). Recall that there is no population explosion in system (4). Consequently, by regular perturbation analysis at the endpoints $\gamma = 0$ and $\gamma = 1$, one obtains that when γ is sufficiently small (respectively, large) the qualitative dynamics of the dispersal-linked system under mixed synchronous and asynchronous dispersals are similar to that of the corresponding system under only synchronous (respectively, asynchronous) dispersal.

If $\gamma = 0$ and dispersal is synchronous ($0 < d_1, d_2 < 1$), it is known that the species always persists in both patches, where $g_1(0), g_2(0) > 1$ [44]. However, if $\gamma = 1$ and dispersal is asynchronous, then the species does not always persist in both patches (Theorem 3). If $\gamma \neq 1$ in

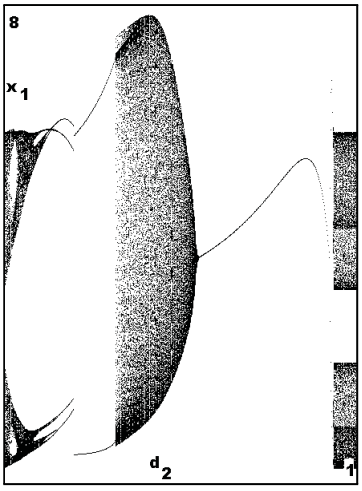


Figure 16. *Period-doubling reversals and Hopf bifurcation, where $r_1 = 2.8$, $d_1 = 0.03$, and d_2 is varied continuously between 0 and 1 (predispersal local dynamics is chaotic; see Figure 12).*

system (4), then $x_1 > 0$ or $x_2 > 0$ implies that $F_1(x) > 0$ and $F_2(x) > 0$. Consequently, in the mixed synchronous-asynchronous dispersal model with $\gamma \neq 1$, the single species either persists in both patches or becomes extinct in both patches. As in the model with asynchronous dispersal, in this section we obtain conditions that guarantee the extinction of the species in both Patches 1 and 2 of the mixed synchronous-asynchronous dispersal model.

Theorem 3. *In system (4), if $\gamma \in (0,1)$, then*

$$(1-\gamma)d_1d_2\Re_d^1\Re_d^2 < \left(1-(1-d_1)\,\Re_d^1\right)\left(1-(1-d_2)\Re_d^2\right)$$

implies that $(0,0)$ is globally asymptotically stable, where $(1-d_i)\,\Re_d^i < 1$ for each $i \in \{1,2\}$. Hence, the species becomes extinct in both Patches 1 and 2.

The proof of Theorem 3 is in the appendix.

By Theorem 3, when the product of the dispersal rates (d_1 and d_2), the intrinsic growth rates ($\Re_d^1$ and $\Re_d^2$), and $(1-\gamma)$ is smaller than the product of $\left(1-(1-d_1)\Re_d^1\right)$ and $\left(1-(1-d_2)\Re_d^2\right)$, then the species becomes extinct in all patches. In the next section, we use the Beverton–Holt model to provide examples of species extinction and persistence in mixed synchronous-asynchronous models.

8. Mixed synchronous-asynchronous dispersal models and compensatory dynamics. Mixed synchronous-asynchronous systems can exhibit species persistence in both Patches 1 and 2. To illustrate this with a simple example, we proceed as in system (3) and assume that the predispersal Patch 1 local population has reached the positive equilibrium X_1, and we let $f_1(x_1) \equiv X_1$ [43, 46]. Then the mixed model, system (4), reduces to

$$\left.\begin{aligned} x_1(t+1) &= (1-d_1)X_1 + d_2x_2(t)g_2(x_2(t)+\gamma d_1X_1), \\ x_2(t+1) &= (1-\gamma)d_1X_1 + (1-d_2)x_2(t)g_2(x_2(t)+\gamma d_1X_1). \end{aligned}\right\} \tag{6}$$

System (6) supports a globally stable positive equilibrium whenever the predispersal local patch dynamics are compensatory. We summarize these in the following result.

Theorem 4. *In system (6), let each local patch dynamics be modeled by f_i, an α-monotone concave map, with the positive fixed point $X_i \in (0, \alpha)$. Then the positive equilibrium population vector,*

$$\left((1-d_1)X_1 + \frac{d_2}{1-d_2}(\widehat{X}_2 - (1-\gamma)d_1X_1),\ \widehat{X}_2\right),$$

is globally attracting, where $\widehat{X}_2$ is the unique positive solution of the equation

$$(1-\gamma)d_1X_1 + (1-d_2)x_2g_2(x_2 + \gamma d_1X_1) = x_2.$$

That is, the dispersal-linked mixed system supports a globally stable positive fixed point whenever the predispersal local patch dynamics are compensatory.

The proof of Theorem 4 is similar to that of Theorem 2 and is omitted.

In Example 4, we use compensatory local dynamics and the Beverton–Holt model to study the dependence of the dynamics of system (4) on the mixed synchronous-asynchronous dispersal rates.

Example 4. *Consider system (4) with*

$$f_i(x_i) = \frac{a_ix_i}{1+b_ix_i}$$

for each $i \in \{1, 2\}$. Set the following parameter values:

$$a_1 = 2,\ a_2 = 2.1,\ b_1 = b_2 = 1,\ \gamma = 0.95,\ \text{and}\ d_1 = d_2 = 0.01.$$

As in Example 2, in Example 4 the local dynamics in both patches are compensatory and the system exhibits a globally stable positive equilibrium population vector at $(1.002, 1.070)$ (Theorem 4). When symmetric dispersal is assumed and $d_1 = d_2$ is varied continuously between 0 and 1, as in the asynchronous dispersal model the population in each patch decreases to zero monotonically with increasing values of the symmetric dispersal coefficients (see Figure 7). In particular, when $d_1 = d_2 > 0.7$,

$$(1-\gamma)d_1d_2\Re_d^1\Re_d^2 < \left(1-(1-d_1)\,\Re_d^1\right)\left(1-(1-d_2)\Re_d^2\right)$$

and the species becomes extinct in both patches (Theorem 3).

When $\gamma = 1$, $d_1 = d_2 > 0.53$, and all the other parameters remain at their current values in Example 4, dispersal is asynchronous and the species becomes extinct in both patches (Figure 7 and Theorem 1). To illustrate species persistence in mixed synchronous-asynchronous dispersal models, where the species is extinct in the associated asynchronous dispersal model, we keep all parameters fixed at their current values and let $\gamma = 0.75$ in Example 4. With this choice of parameters, the species persists for all values of the symmetric dispersal coefficients $d_1 = d_2 \in (0, 1)$ (see Figure 17).

As in asynchronous dispersal models, our numerical explorations show that asymmetric mixed synchronous-asynchronous dispersals are capable of shifting the population dynamics from persistence to extinction in both patches. Furthermore, our results show that in mixed systems the parameter that spans the range of possible modes of dispersal is also capable of forcing a similar shift from extinction to persistence of the species in all patches.

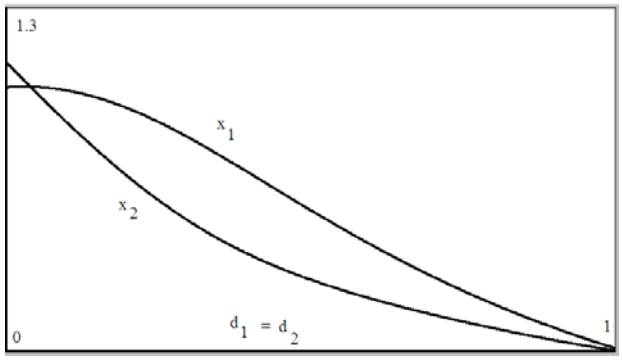


Figure 17. *Persistence in both Patches* 1 *and* 2 *with symmetric mixed synchronous-asynchronous dispersal, where* $\gamma = 0.75$, $d_1 = d_2 \in (0, 1)$, *and all other parameters remain fixed at their current values in Example* 4.

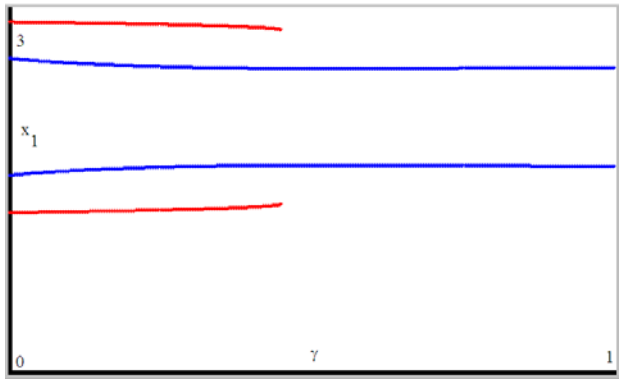


Figure 18. *Example* 5 *has two coexisting* 2*-cycle attractors for values of* $\gamma \in [0.0.45)$ *and only a single* 2*-cycle attractor for values of* $\gamma \in (0.45, 1)$.

9. Mixed synchronous-asynchronous dispersal models and overcompensatory dynamics. The qualitative dynamics of mixed synchronous-asynchronous dispersal systems are similar to those of the associated synchronous (respectively, asynchronous) dispersal systems when γ, the parameter that spans the range of possible modes of dispersal, is close to 0 (respectively, close to 1). In this section, we highlight the possible behaviors of mixed synchronous-asynchronous systems, where the local dynamics are overcompensatory and are governed by the Ricker model.

Example 5. *Consider system* (4) *with the Ricker model*

$$f_i(x_i) = x_i \exp(r_i - x_i)$$

for each $i \in \{1, 2\}$. *Set the following parameter values:*

$$r = r_1 = r_2 = 2.1, \ \gamma \in [0, 1], \ \text{and} \ d = d_1 = d_2 = 0.03.$$

With our choice of parameters, Example 4 exhibits two coexisting 2-cycle attractors (see Example 1 and Figures 1 and 18) when $\gamma \in [0, 0.45)$. That is, for these values of the parameters, the qualitative dynamics of the system with mixed synchronous-asynchronous dispersals

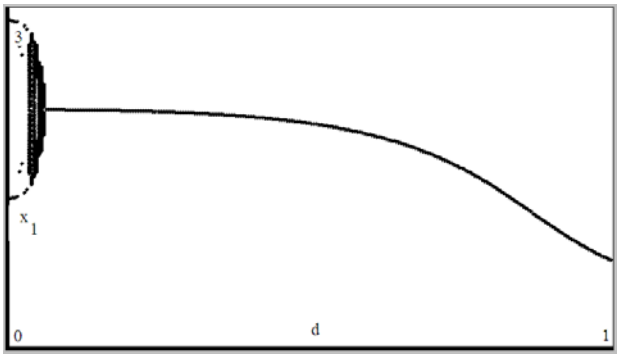


Figure 19. *The full mixed system shifts from a* 2*-cycle attractor to a limit cycle attractor, and then to a fixed point attractor, where* $r_1 = r_2 = 2.1$ *and* $\gamma = 0.4$ *while* $d_1 = d_2$ *is varied continuously between* 0 *and* 1*.*

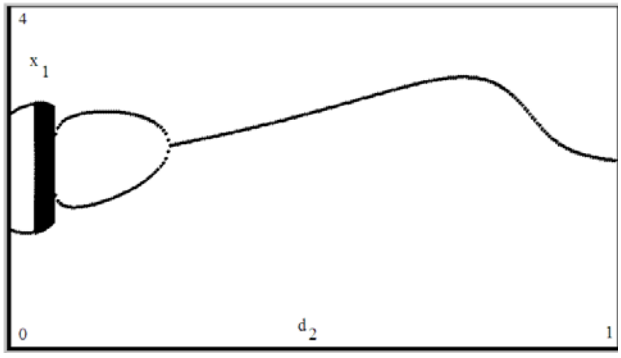


Figure 20. *The full mixed system shifts from a* 2*-cycle attractor to a limit cycle attractor back to a* 2*-cycle attractor, and then to a fixed point attractor, where* $r_1 = r_2 = 2.1$*,* $\gamma = 0.4$*, and* $d_1 = 0.03$ *while* d_2 *is varied continuously between* 0 *and* 1*.*

are the same as those of the corresponding system with only dispersal synchrony. Furthermore, our simulations show that the full system shifts from the two (multiple) 2-cycle attractors to a single 2-cycle attractor (see Example 3 and Figures 10 and 18) when $\gamma \in [0.45, 1]$. In this case, the qualitative dynamics of the mixed synchronous-asynchronous dispersal model are the same as those of the corresponding system with only asynchronous dispersal.

To study the impact of symmetric mixed synchronous-asynchronous dispersals on Example 5, we let $\gamma = 0.4$ and vary d continuously between 0 and 1 while r is kept fixed at 2.1. Figure 19 shows that with increasing values of d, the mixed system shifts from a 2-cycle attractor to a limit cycle attractor and then to a fixed point attractor.

Mixed models under asymmetric dispersals can exhibit qualitative dynamics that are different from those of the associated mixed models under symmetric dispersal. To demonstrate this difference, we let $r_1 = r_2 = 2.1$, $\gamma = 0.4$, and $d_1 = 0.03$, and we vary d_2 continuously between 0 and 1 (see Figures 19 and 20). Unlike Figure 19, Figure 20 shows that with increasing values of the asymmetric dispersal coefficient d_2, the mixed system shifts from a 2-cycle attractor to a limit cycle attractor and then returns to a 2-cycle attractor. This return to a 2-cycle attractor after a limit cycle attractor is different from the bifurcations in Figure 19

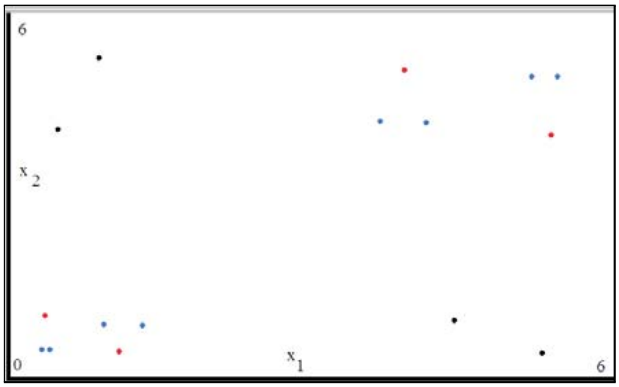


Figure 21. *Three coexisting attractors in Example 6: Two 4-cycle (red and black dots) attractors coexisting with an 8-cycle (blue dots) attractor. Figure 21 is plotted over 3000 time steps.*

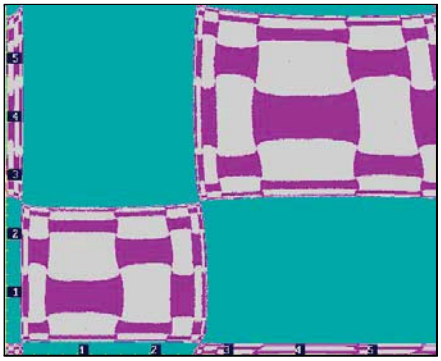


Figure 22. *Basins of attraction of the three coexisting attractors in Figure 21. Figure 22 is plotted over 5000 time steps.*

with the symmetric mixed dispersal. However, with increasing values of the asymmetric dispersal coefficient, as in Figure 19, the 2-cycle attractor undergoes a period-doubling reversal bifurcation (see Figure 20).

Recall that Figures 1, 2, 3, 4, 10, 11, 12, and 13 and Tables 1 and 2 demonstrate that synchronous dispersal-linked systems can be more sensitive to initial population sizes than the corresponding asynchronous ones. Next, we use an example with an intermediate value of the parameter γ to illustrate that mixed dispersal-linked systems can be more (respectively, less) sensitive to initial population sizes than the corresponding asynchronous (respectively, synchronous) systems.

Example 6. *In Example 5, set the following parameter values:*

$$r_1 = r_2 = 2.7,\ \gamma = 0.5,\ \textit{and}\ d_1 = d_2 = 0.03.$$

Figure 21 exhibits the three coexisting attractors (two 4-cycle attractors and an 8-cycle attractor) of Example 6, and Figure 22 shows their basins of attraction. However, the associated synchronous model of Example 6 ($\gamma = 0$) has four coexisting attractors (see Figures 2 and 4), while the associated asynchronous model ($\gamma = 1$) has two coexisting attractors (see

Figures 11 and 13). These examples illustrate that when predispersal dynamics are cyclic, then synchronous-asynchronous dispersal-linked mixed systems become more sensitive to initial population sizes as the values of γ decrease.

10. Conclusions. In this paper, we generalize the classical single species, discrete-time, two-patch synchronous dispersal linked-model to a mixed synchronous-asynchronous dispersal model which includes a model of Doebeli for asynchronous dispersal. We extend an idea of Doebeli and show how the detailed timing of dispersal can affect the global dynamics of dispersal-linked systems [7, 8]. The dynamics of discrete-time population models connected by either asynchronous or mixed synchronous-asynchronous dispersals depend on the dispersal rates and the predispersal local patch dynamics. In asynchronous models, the species becomes extinct on at least one patch when the dispersal rates are high, while it persists when the dispersal rates are low. However, in mixed synchronous-asynchronous systems, depending on the dispersal rates, the range of possible modes of dispersals, and the intrinsic growth rates in the two patches, the pioneer species either persists in all patches or becomes extinct in all patches.

The results of Hastings and Levin on continuous-time metapopulation models and those of Yakubu and Castillo-Chavez on discrete-time models connected by synchronous dispersal predict that an equilibrium population is stable only if it corresponds to a stable equilibrium within each patch [22, 29, 44]. Our results support this prediction for two-patch discrete-time models under asynchronous dispersal rates as long as the dispersal rate from Patch 2 to Patch 1 is low and the predispersal local patch dynamics are compensatory.

Local unstructured populations under compensatory or overcompensatory dynamics tend to support single attractors; that is, the population has a single outcome. However, in dispersal-linked population models with overcompensatory local dynamics, both synchronous and asynchronous dispersals can fracture the basins of attraction through their support of multiple attractors. Hastings [22] and Yakubu and Castillo-Chavez [44] showed that both symmetric and asymmetric synchronous dispersals are capable of generating multiple attractors where the predispersal local patch dynamics are overcompensatory. Our results show that asynchronous dispersal-linked systems support multiple attractors with a smaller number of distinct attractors than the corresponding synchronous systems. The interactions via dispersal of various forms of intraspecific competition has not only led to the generation of a dynamical landscape capable of supporting multiple attractors but also has aided our understanding of the role that initial population sizes play in the ultimate fate (life-history) of dispersal-linked systems. As the complexity of the local dynamics increases, dispersal-linked deterministic systems exhibit sensitive dependence of the long-term behavior on the initial population sizes. The smaller number of distinct attractors makes synchronous dispersal-linked systems more sensitive to initial population sizes than the corresponding asynchronous systems [5, 16, 22, 26, 44]. Complex overcompensatory local dynamics give rise to sensitive dependence of mixed dispersal-linked dynamics on initial population sizes when the dominant dispersal mode is synchronous. However, our results show that mixed dynamics are less sensitive to initial population sizes when the dominant dispersal mode is asynchronous.

Asymmetric dispersal is capable of stabilizing or shifting the predispersal local dynamics from overcompensatory to compensatory dynamics. Thus, asymmetry enhances the stabilizing role of dispersal in synchronous, asynchronous, and mixed synchronous-asynchronous models.

Our results show that it is possible for the long-term qualitative dynamics of models with mixed synchronous-asynchronous dispersals to be identical to that of the corresponding model with either only dispersal synchrony or only dispersal asynchrony. That is, in mixed synchronous-asynchronous models, the dominant mode of dispersal is capable of driving the population dynamics of the dispersal-linked systems.

Appendix.

Proof of Lemma 1. Recall that $F_1(x) = (1-d_1)f_1(x_1) + d_2x_2g_2(x_2+\gamma d_1f_1(x_1))$ and $F_2(x) = (1-\gamma)d_1f_1(x_1) + (1-d_2)x_2g_2(x_2+\gamma d_1f_1(x_1))$, where $x=(x_1,x_2)\in\mathbb{R}^2_+$. Since each $g_i>0$, $0\le\gamma\le 1$, and $0<d_i<1$, we have $F_1(x)>0$ and $F_2(x)>0$ whenever $x_1,x_2>0$. That is, the positive cone is positively invariant.

Next, we show that for each $i\in\{1,2\}$ the sequence $\{F_i^t(x)\}_{t\ge0}$ is bounded. By the monotonicity condition on g_2 and the fact that $0<d_1,d_2<1$, we obtain that

$$F_1(x)+F_2(x)\le f_1(x_1)+f_2(x_2).$$

If $x_i\le\max I_i$, then $f_i(x_i)=x_ig_i(x_i)\le\max I_i$, but if $x_i>\max I_i$, then $f_i(x_i)=x_ig_i(x_i)<x_i$, where $I_i\equiv f_i([0,X_i])$. As a result,

$$F_1(x)+F_2(x)\le\begin{cases}\max I_1+\max I_2 & \text{if } x_1\le\max I_1 \text{ and } x_2\le\max I_2,\\ x_1+\max I_2 & \text{if } x_1>\max I_1 \text{ and } x_2\le\max I_2,\\ \max I_1+x_2 & \text{if } x_1\le\max I_1 \text{ and } x_2>\max I_2,\\ x_1+x_2 & \text{if } x_1>\max I_1 \text{ and } x_2>\max I_2.\end{cases}$$

Hence, each sequence $\{F_i^t(x)\}_{t\ge0}$ is bounded. Consequently, no point in system (4) has an unbounded orbit.

Proof of Theorem 1. By Lemma 1, the ω-limit set of every point in $[0,\infty)\times[0,\infty)$ is nonempty. As a result, we consider an arbitrary point $y=(y_1,y_2)\in[0,\infty)\times[0,\infty)$. Let $x=(x_1,x_2)\in\omega(y)$. By definition, there exists $n_i\to+\infty$ such that $F^{n_i}(y)\to x$ as $n_i\to+\infty$.

To prove (i), we need to show that $F_2^{n_i}(y)\to0$ as $n_i\to+\infty$. Define the function $V:\mathbb{R}^2_+\to\mathbb{R}_+$ by $V(y_1,y_2)=y_2$. Next, we show that V is a Lyapunov function for system (3). Hence, it decreases to a limit point with second coordinate zero.

If $y_2>0$, then $V(F(y))<(1-d_2)\Re_d^2y_2$ and $V(F(y))<V(y)$ whenever $(1-d_2)\Re_d^2<1$. Therefore, for all points $y=(y_1,y_2)$ satisfying $y_2>0$ we know that $V(F(y))<V(y)$. If $x_2>0$, then $V(F(x))<V(x)$. However, this is impossible for an ω-limit point. This proves (i).

To prove (ii), notice that the eigenvalues of the Jacobian matrix $DF(0,0)$ are $(1-d_1)\Re_d^1$ and $(1-d_2)\Re_d^2$. Consequently, $(0,0)$ is unstable whenever $(1-d_1)\Re_d^1>1$ and $(1-d_2)\Re_d^2>1$. Furthermore, $F_1(x)\ge(1-d_1)f_1(x_1)$ and $(1-d_1)\Re_d^1>1$ imply that $\{0\}$ is an unstable fixed point of the one-dimensional map $\widehat{f}_1(x_1)=(1-d_1)f_1(x_1)$. Also, $(1-d_1)\Re_d^1>1$ implies that $\widehat{f}_1$ has a unique positive fixed point, denoted by $\widehat{X}_1$. Moreover, $\widehat{f}_1(x_1)>x_1$ whenever $0<x_1<\widehat{X}_1$ and $\widehat{f}_1(x_1)<x_1$ whenever $x_1>\widehat{X}_1$. Consequently, $\widehat{I}_1\equiv\widehat{f}_1([0,\widehat{X}_1])$ is a global attractor; that is, every initial population eventually reaches a limit in $\widehat{I}_1$. Let $\widehat{J}_1\equiv\widehat{f}_1\left([\widehat{X}_1,\max\widehat{I}_1]\right)$, where $\max\widehat{I}_1$ is the largest endpoint of $\widehat{I}_1$. Notice that the smallest endpoint of $\widehat{J}_1$, $\min\widehat{J}_1$, is positive. That is, $\min\widehat{J}_1>0$. Furthermore, $\lim_{t\to\infty}F_1^t(x)\ge\lim_{t\to\infty}\widehat{f}_1^t(x_1)\ge\min\widehat{J}_1$ for any $x_1>0$. Hence, the species persists in Patch 1.

To prove (iii) and (iv), notice that when Patch 2 is empty (that is, when $(1-d_2)\Re_d^2 < 1$), the Patch 1 population is governed by the "limiting equation" $F_1(x) = (1-d_1)f_1(x_1)$. That Patch 1 dynamics are compensatory implies that $(0,0)$ is globally stable when $\Re_d < 1$, while $(0,0)$ is unstable and $(g_1^{-1}(\frac{1}{1-d_1}),0)$ is globally stable when $(1-d_1)\Re_d^1 > 1$ and $(1-d_2)\Re_d^2 < 1$. This completes the proof.

The following result is useful in the proof of Theorem 2.

Lemma 2. *Let $\widehat{f}_2(x_2) = (1-d_2)x_2 g_2(x_2 + d_1X_1)$. If Patch 2 dynamics are compensatory and $(1-d_2)\Re_d^2 > 1$, then $\widehat{f}_2$ is orientation-preserving and all positive densities approach the positive equilibrium at $\widehat{X}_2 = g_2^{-1}(\frac{1}{1-d_2}) - d_1X_1$ monotonically under $\widehat{f}_2$ iterations.*

Proof of Lemma 2. First, notice that $(1-d_2)\Re_d^2 > 1$ implies that $\widehat{f}_2$ has a unique positive fixed point at $\widehat{X}_2$. To prove the result, we need to show that $\widehat{X}_2$ is globally stable in $(0,\infty)$. If we know that $\widehat{f}_2$ cannot support 2-*cycles*, then Sharkovskii's theorem implies that $\widehat{f}_2$ cannot have cycles except for a fixed point. Using the monotonicity condition on g_2 and the fact that $\widehat{X}_2 > 0$, we obtain that zero and infinity are repellors under $\widehat{f}_2$ iterations. Since no point overshoots X_2, we obtain that the unique positive fixed point of $\widehat{f}_2$, $\widehat{X}_2$ is globally stable in $(0,\infty)$ and no point overshoots it under $\widehat{f}_2$ iterations.

Now, we prove that $\widehat{f}_2$ cannot support 2-*cycles*. Suppose that $\widehat{f}_2$ has a 2-*cycle*. Since Patch 2 predispersal local dynamics are compensatory and $(1-d_2)\Re_d^2 > 1$, we have that the fixed point $\widehat{X}_2$ must be unstable and $(\widehat{f}_2)'(\widehat{X}_2) < -1$. That is, $(\widehat{f}_2)'(\widehat{X}_2) = (1-d_1)(g_2(X_2) + \widehat{X}_2 g_2'(X_2)) < -1$. Since $g_i(X_i) = 1$ and $\widehat{X}_2 = X_2 - d_1X_1$, we have

$$\begin{aligned}(1-d_1)f_2'(X_2) &= (1-d_1)(1 + X_2 g_2'(X_2))\\ \leq (1-d_1)(1 + (\widehat{X}_2 + d_1X_1)g_2'(X_2)) &< -1 + (1-d_1)d_1X_1 g_2'(X_2) < 0.\end{aligned}$$

This contradicts the fact that $(1-d_1)f_2'(X_2) > 0$ (compensatory dynamics). As a result, $\widehat{f}_2$ cannot support 2-*cycles*. This establishes Lemma 2.

Proof of Theorem 2. Notice that system (4) is essentially a one-dimensional system. By Lemma 2, $\lim_{t\to\infty} F_2^t(x) = (g_2^{-1}(\frac{1}{1-d_2}) - d_1X_1)$ for each point $x = (x_1, x_2) \in (0,\infty)\times(0,\infty)$. Consequently, $\lim_{t\to\infty} F_1^t(x) = (1-d_1)X_1 + \frac{d_2}{1-d_2}(g_2^{-1}(\frac{1}{1-d_2}) - d_1X_1)$, and the positive fixed point is globally asymptotically stable in $(0,\infty)\times(0,\infty)$.

Proof of Theorem 3. To prove Theorem 3, we proceed as in the proof of Theorem 1 and define the function $V : \mathbb{R}_+^2 \to \mathbb{R}_+$ by

$$V(y_1, y_2) = y_1 + by_2,$$

where

$$0 < \frac{d_2\Re_d^2}{1-(1-d_2)\Re_d^2} < b < \frac{1-(1-d_1)\,\Re_d^1}{(1-\gamma)d_1\Re_d^1}.$$

Next, we show that V is a Lyapunov function for system (4), where $\gamma \neq 0, 1$. Hence, it decreases to a limit point with both coordinates equal to zero.

$V(F(y)) = ((1-d_1) + (1-\gamma)bd_1)\, f_1(y_1) + (d_2 + b(1-d_2))y_2 g_2(y_2 + \gamma d_1 f_1(y_1))$. If $y = (y_1, y_2) \neq (0,0)$, then

$$V(F(y)) < ((1-d_1) + (1-\gamma)bd_1)\,\Re_d^1 y_1 + (d_2 + b(1-d_2))\Re_d^2 y_2.$$

With our choice of the positive constant b, we have

$$((1-d_1)+(1-\gamma)bd_1)\,\Re_d^1 < 1$$

and

$$(d_2+b(1-d_2))\Re_d^2 < b.$$

Hence $V(F(y)) < V(y)$ whenever $0 < \frac{d_2\Re_d^2}{1-(1-d_2)\Re_d^2} < \frac{1-(1-d_1)\Re_d^1}{(1-\gamma)d_1\Re_d^1}$. Therefore, for all points $y=(y_1,y_2)\neq(0,0)$ we know that $V(F(y)) < V(y)$. Now, proceed exactly as in the proof of Theorem 1 to complete the proof.

Acknowledgment. The author thanks the referees for their useful suggestions.

REFERENCES

[1] L. J. Allen, *Persistence, extinction, and critical patch number for island populations*, J. Math. Biol., 24 (1987), pp. 617–625.

[2] M. Begon, J. L. Harper, and C. R. Townsend, *Ecology: Individuals, Populations and Communities*, Blackwell Science, Malden, MA, 1996.

[3] F. Brauer and C. Castillo-Chavez, *Mathematical Models in Population Biology and Epidemiology*, Texts Appl. Math. 40, Springer-Verlag, New York, 2001.

[4] C. Castillo-Chavez and A. A. Yakubu, *Dispersal, disease and life-history evolution*, Math. Biosci., 173 (2001), pp. 35–53.

[5] C. Castillo-Chavez and A. A. Yakubu, *Intraspecific competition, dispersal and disease dynamics in discrete-time patchy environments*, in Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction to Models, Methods and Theory, C. Castillo-Chavez, S. Blower, P. van den Driessche, D. Kirschner, and A.-A. Yakubu, eds., Springer-Verlag, New York, 2001, pp. 165–181.

[6] D. Cohen and S. A. Levin, *The interaction between dispersal and dormancy strategies in varying and heterogeneous environments*, in Mathematical Topics in Population Biology, Morphogenesis and Neurosciences, E. Teramoto and M. Yamaguti, eds., Springer-Verlag, Berlin, 1987, pp. 110–122.

[7] M. Doebeli, *Dispersal and dynamics*, Theoret. Population Biol., 47 (1995), pp. 82–106.

[8] M. Doebeli and G. D. Ruxton, *Evolution of dispersal rates in metapopulation model: Branching and cyclic dynamics in phenotype space*, Evolution, 5 (1997), pp. 1730–1741.

[9] D. J. Earn, S. A. Levin, and P. Rohani, *Coherence and conservation*, Science, 290 (2000), pp. 1360–1364.

[10] P. L. Errington, *Some contributions of a fifteen year local study of the northern bobwhite to a knowledge of population phenomena*, Ecol. Monogr., 15 (1945), pp. 1–34.

[11] J. E. Franke and A.-A. Yakubu, *Extinction and persistence of species in discrete competitive systems with a safe refuge*, J. Math. Anal. Appl., 23 (1996), pp. 746–761.

[12] J. E. Franke and A.-A. Yakubu, *Geometry of exclusion principles in discrete systems*, J. Math. Anal. Appl., 168 (1992), pp. 385–400.

[13] J. E. Franke and A.-A. Yakubu, *Mutual exclusion versus coexistence for discrete competitive systems*, J. Math. Biol., 30 (1991), pp. 161–168.

[14] M. Gadgil, *Dispersal: Population consequences and evolution*, Ecology, 52 (1971), pp. 253–261.

[15] J. L. Gonzalez-Andújar and J. N. Perry, *Chaos, metapopulations and dispersal*, Ecol. Model., 65 (1993), pp. 255–263.

[16] C. Grebogi, E. Ott, and J. A. Yorke, *Chaos, strange attractors, and fractal basin boundaries in nonlinear dynamics*, Science, 228 (1987), pp. 632–638.

[17] M. Gyllenberg, G. Söderbacka, and S. Ericsson, *Does migration stabilize local population dynamics? Analysis of a discrete metapopulation model*, Math. Biosci., 118 (1993), pp. 25–49.

[18] I. Hanski, *Single-species metapopulation dynamics: Concepts, models and observations*, Biol. J. Linn. Soc., 42 (1991), pp. 17–38.

[19] I. A. Hanski and M. E. Gilpin, *Metapopulation Biology: Ecology, Genetics, and Evolution*, Academic Press, San Diego, 1997.
[20] M. P. Hassell, *The Dynamics of Competition and Predation*, Studies in Biol. 72, Edward Arnold, London, 1976.
[21] M. P. Hassell, J. H. Lawton, and R. M. May, *Patterns of dynamical behavior in single species populations*, J. Anim. Ecol., 45 (1976), pp. 471–486.
[22] A. Hastings, *Complex interactions between dispersal and dynamics: Lessons from coupled logistic equations*, Ecology, 75 (1993), pp. 1362–1372.
[23] A. Hastings, *Dynamics of a single species in a spatially varying environment: The stabilizing role of high dispersal rates*, J. Math. Biol., 16 (1982), pp. 49–55.
[24] S. M. Henson, R. F. Constantino, J. M. Cushing, B. Dennis, and R. A. Desharnais, *Multiple attractors, saddles, and population dynamics in periodic habitats*, Bull. Math. Biol., 61 (1999), pp. 1121–1149.
[25] S. M. Henson and J. M. Cushing, *Hierarchical models of intraspecific competition: Scramble versus contest*, J. Math. Biol., 34 (1996), pp. 755–772.
[26] I. Kan, *Open sets of diffeomorphisms having two attractors, each with an everywhere dense basin*, Bull. Amer. Math. Soc. (N.S.), 31 (1994), pp. 68–74.
[27] S. A. Levin, *Dispersion and population interactions*, Amer. Naturalist, 108 (1974), pp. 207–228.
[28] S. A. Levin, *The problem of pattern and scale in ecology*, Ecology, 73 (1992), pp. 1943–1967.
[29] S. Levin and R. T. Paine, *Disturbance, patch formation, and community structure*, Proc. Nat. Acad. Sci. USA, 68 (1974), pp. 2744–2747.
[30] R. Levins, *Some demographic and genetic consequences of environmental heterogeneity for biological control*, Bull. Entomol. Soc. Amer., 15 (1969), pp. 237–240.
[31] R. Levins, *The effect of random variation of different types on population growth*, Proc. Nat. Acad. Sci. USA, 62 (1969), pp. 1061–1062.
[32] R. M. May and G. F. Oster, *Bifurcations and dynamic complexity in simple ecological models*, Amer. Naturalist, 110 (1976), pp. 573–579.
[33] R. M. May, *Simple mathematical models with very complicated dynamics*, Nature, 261 (1977), pp. 459–469.
[34] R. M. May, *Stability and Complexity in Model Ecosystems*, Princeton University Press, Princeton, NJ, 1974.
[35] R. M. May, M. P. Hassell, R. M. Anderson, and D. W. Tonkyn, *Density dependence in host-parasitoid models*, J. Anim. Ecol., 50 (1981), pp. 855–865.
[36] J. Maynard Smith and M. Slatkin, *The stability of predator-prey systems*, Ecology, 54 (1973), pp. 384–391.
[37] J. G. Milton and J. Bélair, *Chaos, noise, and extinction in models of population growth*, Theoret. Population Biol., 37 (1990), pp. 273–290.
[38] A. J. Nicholson, *Compensatory reactions of populations to stresses, and their evolutionary significance*, Aust. J. Zool., 2 (1954), pp. 1–65.
[39] H. E. Nusse and J. A. Yorke, *Dynamics: Numerical Explorations*, Springer-Verlag, New York, 1997.
[40] W. E. Ricker, *Stock and recruitment*, J. Fisheries Research Board of Canada, 11 (1954), pp. 559–623.
[41] T. Royama, *Analytical Population Dynamics*, Pop & Comm. Biol. Series 10, Chapman & Hall, Boca Raton, FL, 1992.
[42] H. L. Smith, *Cooperative systems of differential equations with concave nonlinearities*, Nonlinear Anal., 10 (1986), pp. 1037–1052.
[43] H. R. Thieme, *Convergence results and a Poincaré-Bendixson trichotomy for asymptotically autonomous differential equations*, J. Math. Biol., 30 (1992), pp. 755–763.
[44] A. Yakubu and C. Castillo-Chavez, *Interplay between local dynamics and dispersal in discrete-time metapopulation models*, J. Theoret. Biol., 218 (2002), pp. 273–288.
[45] P. Yodzis, *Introduction to Theoretical Ecology*, Harper & Row, New York, 1989.
[46] X.-Q. Zhao, *Asymptotic behavior for asymptotically periodic semiflows with applications*, Comm. Appl. Nonlinear Anal., 3 (1996), pp. 43–66.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 311–340
© 2008 Society for Industrial and Applied Mathematics

Periodic Solutions in Hamiltonian Systems, Averaging, and the Lunar Problem*

Patricia Yanguas†, Jesús F. Palacián†, Kenneth R. Meyer‡, and H. Scott Dumas‡

Abstract. We investigate the existence, characteristic multipliers, and stability of periodic solutions to a Hamiltonian vector field which is a small perturbation of a vector field tangent to the fibers of a circle bundle. Our primary examples are the planar lunar and spatial lunar problems of celestial mechanics, i.e., the restricted three-body problem where the infinitesimal is close to one of the primaries. By averaging the perturbation over the fibers of the circle bundle one obtains a Hamiltonian system on the reduced (orbit) space of the circle bundle. Our goal in the first part of the paper is to state and prove results which have hypotheses on the reduced system and have conclusions about the full system. Starting with the classical work of Reeb, we give a summary of lemmas, corollaries, and theorems about the existence, characteristic multipliers, and stability of periodic solutions to Hamiltonian systems which are perturbations of circle bundle flows. By reformulating the classical results in modern language and giving alternative proofs in place of the original proofs, we are able to infer new consequences of these classical results. The second part of the paper is devoted to applications of the general results. We apply these general results to the planar and spatial lunar problem. After scaling, the lunar problem is a perturbation of the Kepler problem, which after regularization is a circle bundle flow. We find the classical near-circular periodic solutions and the near-rectilinear periodic solutions. Then we compute their approximate multipliers and show that there is a "twist." However, the twist is too degenerate to apply the classical KAM theorem on invariant tori. We also find symmetric periodic solutions which are continuations of elliptic solutions of the Kepler problem.

Key words. averaging, normalization, reduced space, N-body problem, periodic solutions, twist condition

AMS subject classifications. 34C20, 34C25, 37J40, 70F10, 70K65

DOI. 10.1137/070696453

1. Introduction.

For us the lunar problem is the circular restricted three-body problem where the infinitesimal is close to one of the primaries. After scaling the restricted problem, the lunar problem is a perturbation of the Kepler problem, and Moser [39] has shown that the Kepler problem after regularization is a circle bundle flow. Thus, the lunar problem is a prototype for Hamiltonian systems that arise as perturbations of circle bundle flows.

By averaging the perturbation over the fibers of the circle bundle, Reeb [43] and Moser [39] obtained a Hamiltonian vector field on the base (or reduced) space; see also [32, 33]. They were able to give sufficient conditions for the existence of periodic solutions by looking at the system on the base alone.

*Received by the editors July 6, 2007; accepted for publication (in revised form) by J. Marsden December 1, 2007; published electronically April 23, 2008. The authors appear in reverse alphabetical order.
http://www.siam.org/journals/siads/7-2/69645.html

†Departamento de Ingeniería Matemática e Informática, Universidad Pública de Navarra, 31006 Pamplona, Spain (yanguas@unavarra.es, palacian@unavarra.es). The work of these authors was partially supported by Project MTM2005-08595 of Ministerio de Educación y Ciencia (Spain) and Project Resolución 18/2005 of Departamento de Educación y Cultura, Gobierno de Navarra (Spain).

‡Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025 (ken.meyer@uc.edu, scott.dumas@uc.edu).

Since then a number of papers have appeared which analyze systems by looking at the reduced system only; see [7, 16, 21, 27, 28, 29, 41] and the many references therein. One starts with a small parameter which is a measure of the perturbation of an integrable system where all solutions are periodic. Then one normalizes (or averages) the perturbation term-by-term in the small parameter. After a finite number of terms have been normalized, the higher-order perturbations are truncated, giving an approximation of the full system. This approximation is well defined on the lower-dimensional reduced space. Being lower-dimensional, sometimes just two-dimensional, the system on the reduced space is easier to understand. But not all the features of the full system are accurately reflected by the reduced system; it typically does not display the breakdown of invariant tori, ergodic regions, solenoids, etc.

Our goal in the first part of the paper is to state, prove, and apply results which have hypotheses on the reduced system and have conclusions about the full system. Starting with the work of Reeb, we give in section 2 a summary of lemmas, corollaries, and theorems about the existence, characteristic multipliers, and parametric stability of periodic solutions for Hamiltonian systems which are perturbations of circle bundle flows. Some of the results are old, some are just extensions, and a few are new. Lemma 2.1 is the key to an original direct proof of Reeb's theorems using symplectic geometry arguments. Corollary 2.3 constitutes a new application of Krein–Gel'fand theory [47] about the stability of linear Hamiltonian vector fields with periodic coefficients. Theorem 2.5 connects, through Lemma 2.1, the theory of reduction in Hamiltonian systems with the existence of KAM tori in a simple way. Theorem 2.6, about the existence of symmetric periodic solutions, is also new.

The second part of the paper is devoted to applying these general results to the lunar problem. In section 3 we apply these general results to the planar lunar problem and in section 4 to the spatial lunar problem. In the planar problem we find Hill's classical near-circular periodic solutions, compute their approximate multipliers, and then show that there is a "twist" term. The twist is of too high an order in the perturbation parameter to apply the classical KAM theorem. We also find symmetric periodic solutions which are continuations of elliptic solutions of the Kepler problem.

For the spatial problem we again find the classical Hill periodic solutions, but also the near-rectilinear periodic solutions, and we compute their approximate multipliers. These solutions are shown to be parametrically stable and elliptic. Again we compute a twist term for all these periodic solutions. We pay particular attention to the near-rectilinear periodic solutions and show that they are not collision orbits.

2. Averaging theorems.

Here we summarize some general results from the classic paper by Reeb [43] on averaging Hamiltonian systems on manifolds, along with some obvious corollaries. We also extend these results to systems with discrete symmetries.

Let (M, Ω) be a symplectic manifold of dimension $2n$, $\mathcal{H}_0 : M \to \mathbb{R}$ a smooth Hamiltonian which defines a Hamiltonian vector field $Y_0 = (d\mathcal{H}_0)^{\#}$ with symplectic flow ϕ_0^t (see [1]). Let $\mathbb{I} \subset \mathbb{R}$ be an interval such that each $h \in \mathbb{I}$ is a regular value of $\mathcal{H}_0$ and $\mathcal{N}_0(h) = \mathcal{H}_0^{-1}(h)$ is a compact connected circle bundle over a base space $B(h)$ with projection $\pi : \mathcal{N}_0(h) \to B(h)$. Assume the vector field Y_0 is everywhere tangent to the fibers of $\mathcal{N}_0(h)$; i.e., assume that all the solutions of Y_0 in $\mathcal{N}_0(h)$ are periodic. There is no loss of generality [22] in assuming that all these periodic solutions have periods smoothly depending only on the value of the

Hamiltonian; i.e., the period is a smooth function $T = T(h)$ (sometimes the dependence on h will be omitted in the notation).

For example, consider a pair of harmonic oscillators

$$\ddot{x} + x = 0, \qquad \ddot{y} + y = 0,$$

which may be written as the Hamiltonian system

$$\dot{x} = \frac{\partial H}{\partial u} = u, \qquad \dot{u} = -\frac{\partial H}{\partial x} = -x, \qquad \dot{y} = \frac{\partial H}{\partial v} = v, \qquad \dot{v} = -\frac{\partial H}{\partial y} = -y,$$

with Hamiltonian

$$H = \frac{1}{2}(x^2 + u^2) + \frac{1}{2}(y^2 + v^2).$$

In polar coordinates

$$r^2 = x^2 + u^2, \qquad \theta = \tan^{-1} u/x, \qquad \rho^2 = y^2 + v^2, \qquad \phi = \tan^{-1} v/y,$$

the equations become

$$\dot{r} = 0, \qquad \dot{\theta} = -1, \qquad \dot{\rho} = 0, \qquad \dot{\phi} = -1,$$

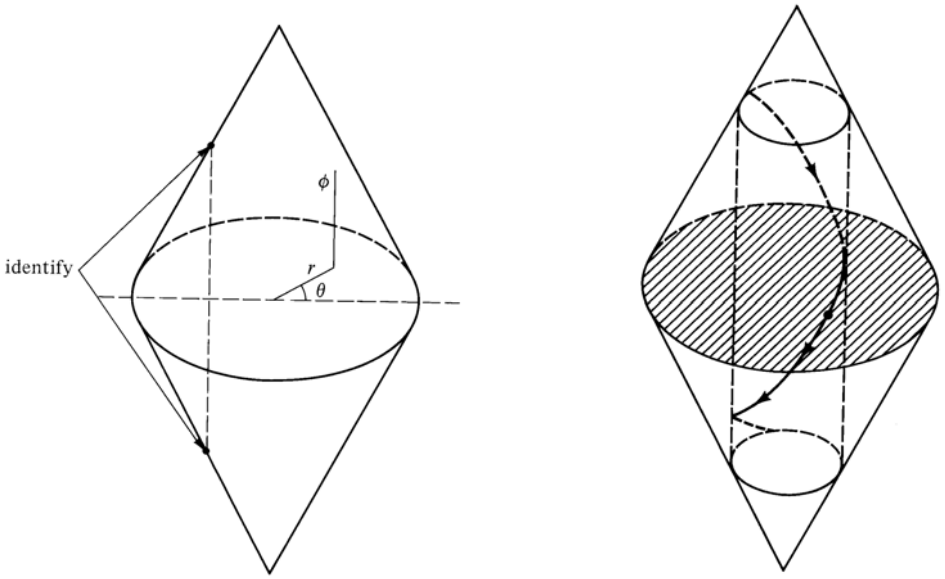
and they admit the two integrals r and ρ.

The energy level $E = H^{-1}(\frac{1}{2})$ is a 3-sphere and is invariant under the flow. All the solutions are 2π-periodic, and so the orbits are circles. Thus the 3-sphere is a union of circles. We can use polar coordinates to coordinatize the sphere provided we are careful to observe the proper conventions.

Starting with the polar coordinates r, θ, ρ, ϕ for $\mathbb{R}^4$, we note that on the 3-sphere, $E = r^2 + \rho^2 = 1$; so we may discard ρ and take $0 \le r \le 1$. We will use r, θ, ϕ as coordinates on S^3. Now r, θ with $0 \le r \le 1$ are just polar coordinates for the closed unit disk. For each point of the open disk, there is a circle with coordinate ϕ (defined mod 2π), but when $r = 1$, $\rho = 0$; so the circle collapses to a point over the boundary of the disk. The geometric model of S^3 is two solid cones with points on the boundary cones identified, as shown in Figure 1a. Through each point in the open unit disk with coordinates r, θ there is a line segment (the dashed line) perpendicular to the disk. The angular coordinate ϕ is measured on this segment, $\phi = 0$ is the disk, $\phi = \pi$ is the upper boundary cone, and $\phi = -\pi$ is the lower boundary cone. Each point on the upper boundary cone with coordinates r, θ, $\phi = \pi$ is identified with the point on the lower boundary cone with coordinates r, θ, $\phi = -\pi$.

In this model there are two special orbits where $r = 0$ and $\rho = 0$. Other than these two special circles, on each orbit, as θ increases by 2π, so does ϕ. Thus, each such orbit meets the open disk where $\phi = 0$ (the shaded disk in Figure 1b) in one point. We can identify each such orbit with the unique point where it intersects the disk. One special orbit meets the disk at the center, and so we can identify it with the center. The other is the outer boundary circle, which is a single orbit. When we identify this circle with a point, the closed disk with its outer circle identified with a point becomes a 2-sphere.

Thus, the 3-sphere S^3 is the union of circles. The quotient space obtained by identifying a circle with a point is a 2-sphere (the Hopf fibration of S^3).



(a) A model of S^3 (b) An orbit on S^3

Figure 1. *S^3 as a circle bundle over S^2.*

Let D be the open disk $\phi = 0$ (the shaded disk in Figure 1b). The union of all the orbits which meet D is a product of a circle and a 2-disk, so each point not on the special circle $r = 1$ lies in an open set that is the product of a 2-disk and a circle. By reversing r and ρ in the discussion above, the circle where $r = 1$ has a similar neighborhood. So locally the 3-sphere is the product of a disk and a circle, but the sphere is not the product of a 2-manifold and a circle (the sphere has a trivial fundamental group, but such a product would not).

In higher dimensions, consider n harmonic oscillators all with frequency 1; i.e., let $M = \mathbb{R}^{2n}$, $\mathcal{H}_0 = \frac{1}{2}\sum_1^n (x_i^2 + y_i^2)$, and $\mathcal{N} = \mathcal{H}_0^{-1}(h) = S^{2n-1}$ (the sphere of radius $\sqrt{2h}$). Then all solutions are 2π-periodic and B is the complex projective $(n-1)$-space, $\mathbb{CP}^{n-1}$. $\mathbb{CP}^1$ is homeomorphic to the 2-sphere, so when $n = 2$ the reduced space is $B = S^2$ as illustrated above.

Another example is the geodesic flow on the n-sphere S^n; i.e., $M = TS^n$ (the tangent bundle of the sphere), $\mathcal{H}_0 : M \to \mathbb{R} : v_p \mapsto |v_p|$ ($\mathcal{H}_0(v_p)$ is the length of the vector $v_p \in T_pM$), $\mathcal{N} = \{v_p \in TS^n : |v_p| = h\}$ (the h-sphere bundle), and B is $G_{2,n+1}$, the Grassmannian manifold of oriented 2-planes in $\mathbb{R}^{n+1}$ (see, for instance, [38]). If $n = 2$, then B is S^2, whereas it is $S^2 \times S^2$ when $n = 3$.

2.1. Reeb's theorems. Here we state and prove two of Reeb's theorems in more modern terminology. Our proof gives more of the Hamiltonian structure and therefore leads to further applications.

Theorem 2.1. *The base space B inherits a symplectic structure ω from (M, Ω); i.e., (B, ω) is a symplectic manifold.*

This is the original reduction theorem. Now let us look at a perturbation of this situation.

Let ε be a small parameter, $\mathcal{H}_1 : M \to \mathbb{R}$ be smooth, $\mathcal{H}_\varepsilon = \mathcal{H}_0 + \varepsilon\mathcal{H}_1$, $Y_\varepsilon = Y_0 + \varepsilon Y_1 = d\mathcal{H}_\varepsilon^{\#}$, $\mathcal{N}_\varepsilon(h) = \mathcal{H}_\varepsilon^{-1}(h)$, and ϕ_ε^t be the flow defined by Y_ε.

Let the average of $\mathcal{H}_1$ be

$$\bar{\mathcal{H}} = \frac{1}{T}\int_0^T \mathcal{H}_1(\phi_0^t)dt,$$

which is a smooth function on $B(h)$, and let $\bar{\phi}^t$ be the flow on $B(h)$ defined by $\bar{Y} = d\bar{\mathcal{H}}^{\#}$.

A critical point of $\bar{\mathcal{H}}$ is *nondegenerate* if the Hessian at the critical point is nonsingular, and the function $\bar{\mathcal{H}}$ is a *Morse function* if all its critical points are nondegenerate. The *index* of a nondegenerate critical point p of $\bar{\mathcal{H}}$ is the dimension of the maximal linear subspace where the Hessian of $\bar{\mathcal{H}}$ at p is negative definite.

Theorem 2.2. *If $\bar{\mathcal{H}}$ has a nondegenerate critical point at $\pi(p) = \bar{p} \in B$ with $p \in \mathcal{N}_0$, then there are smooth functions $p(\varepsilon)$ and $T(\varepsilon)$ for ε small with $p(0) = p$, $T(0) = T$, and $p(\varepsilon) \in \mathcal{N}_\varepsilon$, and the solution of Y_ε through $p(\varepsilon)$ is $T(\varepsilon)$-periodic.*

If $\bar{\mathcal{H}}$ is a Morse function, then Y_ε has at least $\chi(B)$ periodic solutions, where $\chi(B)$ is the Euler–Poincaré characteristic of B.

The proof of Theorem 2.2 yields additional corollaries. The essence of the proof of the local part of Theorem 2.2 is the existence of symplectic coordinates for a tubular neighborhood of the orbit through p. Here we give the proof of the existence of these coordinates.

Lemma 2.1. *Let $p \in \mathcal{N}_0(h)$, with $h \in \mathbb{I}$ fixed. Then there are symplectic coordinates (I, θ, y), valid in a tubular neighborhood of the periodic solution $\phi_0^t(p)$ of $Y_0(h)$, where (I, θ) are action-angle coordinates and $y \in \mathbb{N}$, where $\mathbb{N}$ is an open neighborhood of the origin in $\mathbb{R}^{2n-2}$. The point p corresponds to $(I, \theta, y) = (0, 0, 0)$.*

In these coordinates $\mathcal{H}_0$ is a function of I only; i.e., $\mathcal{H}_0 = \mathcal{H}_0(I)$. A local cross section is $\theta = \alpha$, and a local cross section in an energy level is $\theta = \alpha, I = \beta$, where α, β are constants. In addition, $y \in \mathbb{N}$ are coordinates in the cross section in the energy level.

The Hamiltonian is

$$\mathcal{H}_\varepsilon(I, \theta, y) = \mathcal{H}_0(I) + \varepsilon\mathcal{H}_1(I, \theta, y) = \mathcal{H}_0(I) + \varepsilon\bar{\mathcal{H}}(I, y) + O(\varepsilon^2). \tag{1}$$

Proof. By the Hamiltonian flow box theorem [36, 40] there are local symplectic coordinates $u = (u_1, \dots, u_{2n})$ for M in a neighborhood W of p such that p corresponds to $u_j = 0$, $j = 1, \dots, 2n$ (note that we locate p at the origin). The Hamiltonian is $\mathcal{H}_0 = u_{n+1}$ (so we take $h = 0$), and Y_0 is the differential equation

$$\dot{u}_1 = \partial\mathcal{H}_0/\partial u_{n+1} = 1, \qquad \dot{u}_j = 0, \quad j = 2, \dots, 2n.$$

A local cross section to the flow of Y_0 is $\Sigma = \{u : u_1 = 0\} \cap W$, and a local cross section in an energy level $\mathcal{H}^{-1}(0)$ is $\sigma = \{u : u_1 = u_{n+1} = 0\} \cap W$.

The validity of these coordinates can be extended to a tubular neighborhood $U = \{\phi_0^t(q) : q \in \Sigma, t \in \mathbb{R}\}$ of the Y_0-orbit through p. Let $Z = \Sigma \times \mathbb{R}$, and let $\eta : Z \to U : (q, t) \to \phi_0^t(q)$ be a symplectic map. The vector field Y_0 on U lifts to $\dot{u}_1 = 1$, $\dot{u}_j = 0$, $j = 2, \dots, 2n$, on Z.

Recall that we assume that the period T depends smoothly on the value of $\mathcal{H}_0$, which in these coordinates means that the period depends smoothly on u_{n+1}, i.e., that $T(u_{n+1})$ is

smooth. Let $F(w)$ satisfy $dF/dw = -2\pi/T(w)$, and let $f = F^{-1}$. Change variables on Z from $\{u_1, u_{n+1}\}$ to $\{I, \theta\}$ by

$$I = f(u_{n+1}), \qquad \theta = -\frac{u_1}{f'(u_{n+1})}.$$

One checks that $dI \wedge d\theta = du_1 \wedge du_{n+1}$, so this is a symplectic change of variables. Since $I = f(u_{n+1})$, we have $\mathcal{H}_0 = u_{n+1} = F(I)$ and $\dot\theta = -F'(I) = 2\pi/T(I)$. Thus, when t increases by $T(I)$, the variable θ increases by 2π and so can be considered as an angular variable. Therefore, $(I, u_2, \dots, u_n, \theta, u_{n+2}, \dots, u_{2n})$ is a full set of symplectic coordinates for Z and, via η, a full set of symplectic coordinates for U. Let $v = (u_2, \dots, u_n, u_{n+2}, \dots, u_{2n})$. So the Hamiltonian is

$$\mathcal{H}_\varepsilon(I, \theta, v) = \mathcal{H}_0(I) + \varepsilon \mathcal{H}_1(I, \theta, v). \tag{2}$$

We use the method of Lie transforms to effect the average. Let $W_1(I, \theta, v)$ be the solution of

$$\mathcal{H}_1 + \{\mathcal{H}_0, W_1\} = \mathcal{H}_1 + \frac{\partial \mathcal{H}_0}{\partial I}\frac{\partial W_1}{\partial \phi} = \bar{\mathcal{H}},$$

that is,

$$W_1 = \int^\theta \frac{\partial \mathcal{H}_0}{\partial I}^{-1} \{\mathcal{H}_1 - \bar{\mathcal{H}}\} d\theta.$$

Since $\bar{\mathcal{H}}$ is the mean value of $\mathcal{H}_1$, the function W_1 is 2π-periodic in θ. Let us change variables by $v = V(I, \theta, y, \varepsilon)$, where $V(I, \theta, y, \varepsilon)$ is the solution of

$$\frac{dv}{d\varepsilon} = J\nabla_y W_1(I, \theta, y), \qquad v(0) = y,$$

where J denotes the skew-symmetric matrix

$$J = \begin{bmatrix} 0 & E \\ -E & 0 \end{bmatrix}$$

and E stands for the identity matrix. Since V is the solution of a Hamiltonian equation, the change of variables is symplectic, and since W_1 is 2π-periodic in θ, so is V. The resulting Hamiltonian becomes

$$\mathcal{H}_\varepsilon(I, \theta, y) = \mathcal{H}_0(I) + \varepsilon \bar{\mathcal{H}}(I, y) + O(\varepsilon^2)$$

by the theory of Lie transforms (cf. [36, p. 168ff]). ■

The proofs of Theorems 2.1 and 2.2 follow from this lemma.

Proof of Reeb's theorems. First, each orbit of $\mathcal{H}_0$ in the level set $\mathcal{H}_0^{-1}(0)$ intersects σ once, so σ can be considered a coordinate patch on the base space B, and y provides symplectic coordinates for σ. Thus, B has an atlas of symplectic charts, and therefore B is a symplectic manifold. This proves Theorem 2.1.

Up to terms of order ε the equations are

$$\dot I = O(\varepsilon^2), \quad \dot\theta = 2\pi/T(I) + O(\varepsilon^2), \quad \dot y = \varepsilon J \nabla_y \bar{\mathcal{H}}(I, y) + O(\varepsilon^2).$$

The return time for θ to increase from 0 to 2π is $T+O(\varepsilon^2)$, and the section map in an energy level ($I=0$) is $P:\sigma\to\sigma: y\mapsto P(y)$, where $P(y)=y+\varepsilon TJ\nabla_y\bar{\mathcal{H}}(0,y)+O(\varepsilon^2)$. A fixed point of P gives rise to a periodic solution, and so we must solve $P(y)=y$ or, equivalently, $TJ\nabla_y\bar{\mathcal{H}}(0,y)+O(\varepsilon)=0$. By hypothesis $y=0$ is a nondegenerate critical point of $\bar{\mathcal{H}}$ when $I=0$ or $\nabla_y\bar{\mathcal{H}}(0,0)=0$ and $\partial^2\bar{\mathcal{H}}/\partial y^2(0,0)$ is nonsingular. Thus by the implicit function theorem there is a function $\bar{y}(\varepsilon)=O(\varepsilon)$ such that $P(\bar{y}(\varepsilon))=\bar{y}(\varepsilon)$. This fixed point of P is the initial condition for the periodic solution asserted in Theorem 2.2. ■

2.2. Corollaries. Only the last sentence in Theorem 2.2 gives a truly global result. Those conversant with Morse theory [13] will see that there is a sharper global result.

Corollary 2.1. *Let $\bar{\mathcal{H}}$ be a Morse function, let β_j be the jth Betti number of B, and let C_j be the number of critical points of index j. Then $C_j\geq\beta_j$ or, better yet,*

$$
\begin{aligned}
C_0 &\geq \beta_0,\\
C_1-C_0 &\geq \beta_1-\beta_0,\\
C_2-C_1+C_0 &\geq \beta_2-\beta_1+\beta_0,\\
&\cdots\\
C_k-C_{k-1}+C_{k+2}-\cdots\pm C_0 &\geq \beta_k-\beta_{k-1}+\beta_{k+2}-\cdots\pm\beta_0 \qquad (k<2n-2),\\
C_{2n-2}-C_{2n-3}+C_{2n-4}-\cdots+C_0 &= \beta_{2n-2}-\beta_{2n-3}+\beta_{2n-4}-\cdots+\beta_0=\chi(B).
\end{aligned}
\tag{3}
$$

For these better inequalities on a Morse function, see [37]. The lower estimate on the number of periodic solutions in Theorem 2.2 is $\chi(B)$, the alternating sum of the Betti numbers which could be 0 or negative, whereas the Morse inequalities $C_j\geq\beta_j$ give a lower estimate which is the sum of the Betti numbers. Moreover, the estimates give some information on the number of critical points of various indices. For example, Milnor [37] remarks that if $C_{j+1}=C_{j-1}=0$, then $C_j=\beta_j$.

The nontrivial characteristic multipliers of the periodic solution given in Theorem 2.2 are the eigenvalues of

$$\mathcal{P}=\frac{\partial P}{\partial y}(\bar{y}(\varepsilon))=E+\varepsilon TJ\frac{\partial^2\bar{\mathcal{H}}}{\partial y^2}(0,0)+O(\varepsilon^2),$$

where E is the identity matrix. The eigenvalues of the Hamiltonian matrix

$$A=J\frac{\partial^2\bar{\mathcal{H}}}{\partial y^2}(0,0) \tag{4}$$

are the characteristic exponents of the critical point of $\bar{Y}$ at $\bar{p}$ on B. Thus, the lemma also yields the following corollary.

Corollary 2.2. *Let p be as in Theorem 2.2 and let the characteristic exponents of $\bar{Y}(\bar{p})$ be $\lambda_1,\lambda_2,\ldots,\lambda_{2n-2}$. Then the characteristic multipliers of the periodic solution through $p(\varepsilon)$ are*

$$1,\ 1,\ 1+\varepsilon\lambda_1T+O(\varepsilon^2),\ 1+\varepsilon\lambda_2T+O(\varepsilon^2),\ldots,1+\varepsilon\lambda_{2n-2}T+O(\varepsilon^2).$$

This result was used in [33]. We shall say that a periodic solution is *elliptic* or *linearly stable* if the monodromy matrix is diagonalizable and all its eigenvalues have unit modulus.

One must be careful in applying this corollary, because it gives only an approximation of the characteristic multipliers. Consider the case $2\bar{\mathcal{H}}=(u_1^2+v_1^2)-(u_2^2+v_2^2)$, where $y=$

(u_1, u_2, v_1, v_2), so the eigenvalues are $i, i, -i, -i$. When $T = 1$, Corollary 2.2 says that the multipliers are 1, 1, $1+\varepsilon i+O(\varepsilon^2)$, $1+\varepsilon i+O(\varepsilon^2)$, $1-\varepsilon i+O(\varepsilon^2)$, and $1-\varepsilon i+O(\varepsilon^2)$, which looks like an elliptic periodic solution. But higher-order terms can change the stability. Consider now a perturbation of this example, namely, $2\bar{\mathcal{H}} = (u_1^2 + v_1^2) - (u_2^2 + v_2^2) + 2\varepsilon v_1 v_2$. Now the estimates of the multipliers would be 1, 1, $1 + \varepsilon i + \frac{1}{2}\varepsilon^2 + O(\varepsilon^3)$, $1 + \varepsilon i - \frac{1}{2}\varepsilon^2 + O(\varepsilon^3)$, $1 - \varepsilon i + \frac{1}{2}\varepsilon^2 + O(\varepsilon^3)$, and $1 - \varepsilon i - \frac{1}{2}\varepsilon^2 + O(\varepsilon^3)$, which gives an unstable periodic solution. The solution of this problem lies in the Krein–Gel'fand concept of parametric stability [47], which we briefly summarize below.

For the moment consider the linear constant coefficient Hamiltonian system

$$\dot{y} = Cy = J\nabla H(y), \qquad H = \frac{1}{2} y^T S y, \tag{5}$$

where S is a symmetric matrix and $C = JS$ is a Hamiltonian matrix. System (5) (or the Hamiltonian matrix C) is *stable* if all its solutions are bounded for all t, and it is said to be *parametrically stable* or *strongly stable* if it and all sufficiently small linear constant coefficient Hamiltonian perturbations of it are stable. If system (5) is parametrically stable, then it is stable, and it is stable if and only if C is diagonalizable and has only purely imaginary eigenvalues.

Let $\pm\alpha_1 i, \pm\alpha_2 i, \ldots, \pm\alpha_s i$ be the eigenvalues of the stable matrix C, and V_j, $j = 1, \ldots, s$, be the maximal real linear subspace where C has eigenvalues $\pm\alpha_j i$. So V_j is a C-invariant symplectic subspace, C restricted to V_j has eigenvalues $\pm\alpha_j i$, and $\mathbb{R}^{2n} = V_1 \oplus V_2 \oplus \cdots \oplus V_s$. Let H_j be the restriction of H to V_j.

Theorem 2.3 (see [47]). *System* (5) *is parametrically stable if and only if*

- *all the eigenvalues of C are purely imaginary,*
- *C is nonsingular,*
- *C is diagonalizable over the complex numbers, and*
- *the Hamiltonian H_j is positive or negative definite for each j.*

Thus, $2H = (u_1^2+v_1^2)+(u_2^2+v_2^2)$ is parametrically stable, as the corresponding eigenvalues are $\pm i$ (double); hence $H_1 = H$ is positive definite. The Hamiltonian $2H = (u_1^2+v_1^2)-4(u_2^2+v_2^2)$ has eigenvalues $\pm i$ and $\pm 2i$, so $2H_1 = u_1^2 + v_1^2$ is positive definite and $2H_2 = -4(u_2^2 + v_2^2)$ is negative definite; therefore, H is parametrically stable. However, $2H = (u_1^2 + v_1^2) - (u_2^2 + v_2^2)$ has eigenvalues $\pm i$ (double), and, as $H_1 = H$ is not positive or negative definite, it cannot be parametrically stable.

Now consider the linear T-periodic Hamiltonian system

$$\dot{y} = D(t)y = J\nabla H(y), \qquad H = \frac{1}{2} y^T R(t) y, \tag{6}$$

where $R(t) = R(t+T)$ is symmetric and $D(t) = JR(t)$ is Hamiltonian. The periodic system (6) is *stable* if all its solutions are bounded for all t, and it is said to be *parametrically stable* or *strongly stable* if it and all sufficiently small linear T-periodic Hamiltonian perturbations of it are stable. The monodromy matrix is $M = Z(T)$, where $Z(t)$ is a fundamental matrix solution of (6). If the system is parametrically stable, then it is stable, and (6) is stable if and only if its monodromy matrix is diagonalizable and has only eigenvalues (multipliers) of unit modulus.

Let $\beta_1^{\pm 1}, \beta_2^{\pm 1}, \ldots, \beta_s^{\pm 1}$ be the eigenvalues of M and V_j, $j = 1, \ldots, s$, be the maximal real linear subspace where M has eigenvalues $\beta_j^{\pm 1}$. So V_j is an M-invariant symplectic subspace, M restricted to V_j (denoted by M_j) is symplectic and has eigenvalues $\beta_j^{\pm 1}$, and $\mathbb{R}^{2n} = V_1 \oplus V_2 \oplus \cdots \oplus V_s$.

For periodic systems we need to define the analogue of the quadratic form H_j. There are at least three ways to do this: (1) define a bilinear form on the eigenvectors corresponding to $\beta_j^{\pm 1}$ [47], (2) use Floquet theory and take logs of M, or (3) use a Cayley transformation. All three ways yield the same result, and we choose the latter because of its simplicity. The particular Möbius transformation

$$\Psi : z \mapsto w = (z-1)(z+1)^{-1}, \qquad \Psi^{-1} : w \mapsto z = (1+w)(1-w)^{-1}$$

is known as the *Cayley transformation*. One checks that $\Psi(1) = 0$, $\Psi(i) = i$, and $\Psi(-1) = \infty$, and so Ψ takes the unit circle in the z-plane to the imaginary axis in the w-plane, the interior of the unit circle in the z-plane to the left half w-plane, etc. Ψ can be applied to any matrix B which does not have -1 as an eigenvalue, and if λ is an eigenvalue of B, then $\Psi(\lambda)$ is an eigenvalue of $\Psi(B)$.

Lemma 2.2. *If M is a symplectic matrix which does not have eigenvalue -1, then $C = \Psi(M)$ is a Hamiltonian matrix. Moreover, if M has only eigenvalues of unit modulus and is diagonalizable, then $C = \Psi(M)$ has only purely imaginary eigenvalues and is diagonalizable.*

Proof. Simply check. ∎

M_j is the restriction of M to V_j and is a symplectic matrix, so $C_j = \Psi(M_j)$ is a Hamiltonian matrix and $S_j = JC_j$ is a symmetric matrix.

Theorem 2.4 (see [47]). *System (6) is parametrically stable if and only if*

- *all the eigenvalues of M have unit modulus,*
- *M does not have eigenvalue $+1$ or -1,*
- *M is diagonalizable over the complex numbers, and*
- *the symmetric matrix S_j is positive or negative definite for each j.*

Corollary 2.3. *If one or more of the λ_j of Corollary 2.2 is real or has nonzero real part, then the periodic solution through $p(\varepsilon)$ is unstable.*

If the matrix A in (4) is the coefficient matrix of a parametrically stable system, then the periodic solution through $p(\varepsilon)$ is elliptic. In particular, if $\bar{p}$ is a nondegenerate maximum or minimum of $\bar{\mathcal{H}}$, then the periodic solution through $p(\varepsilon)$ is elliptic. If $\bar{\mathcal{H}}$ is a Morse function, then there are at least two elliptic periodic solutions, since $\bar{\mathcal{H}}$ must have a nondegenerate maximum and minimum.

The authors believe this application of Krein–Gel'fand theory to be new.

Proof. The first sentence is obvious. Recall that the nontrivial multipliers are the eigenvalues of the symplectic matrix $\mathcal{P} = E + \varepsilon TA + O(\varepsilon^2)$. Applying Cayley's transformation to $\mathcal{P}$ yields the Hamiltonian matrix

$$\mathcal{A} = \Psi(\mathcal{P}) = \Psi(E + \varepsilon TA + O(\varepsilon^2)) = \frac{1}{2}\varepsilon TA + O(\varepsilon^2) = \frac{1}{2}\varepsilon T(A + O(\varepsilon)).$$

If A is the matrix of a parametrically stable system, the matrix $A + O(\varepsilon)$ is stable for all small ε, and hence so is $\mathcal{A}$. Thus all eigenvalues of $\mathcal{P} = \Psi^{-1}(\mathcal{A})$ have unit modulus. ∎

2.3. KAM tori. One can also detect invariant tori using KAM theory.

Theorem 2.5. *Let p be as in Theorem 2.2 and suppose there are symplectic action-angle variables $(I_1,\dots,I_{n-1},\theta_1,\dots\theta_{n-1})$ at $\bar{p}$ in B such that*

$$\bar{\mathcal{H}} = \sum_{k=1}^{n-1} \omega_k I_k + \frac{1}{2}\sum_{k=1}^{n-1}\sum_{j=1}^{n-1} C_{kj} I_k I_j + \mathcal{H}^{\#}, \tag{7}$$

where the ω_k are nonzero, $C_{kj} = C_{jk}$, and $\mathcal{H}^{\#}(I_1,\dots,I_{n-1},\theta_1,\dots\theta_{n-1})$ is at least cubic in $I_1,\dots,I_{n-1}$.

Assume that $\det C_{kj} \neq 0$. *That is, assume the system has been put into Birkhoff normal form and the "twist" condition is satisfied. Furthermore, assume $dT/dh \neq 0$; i.e., assume the period varies with $\mathcal{H}_0$ in a nontrivial way.*

Then near the periodic solutions given in Theorem 2.2 there are invariant KAM tori of dimension n. In particular, when $n = 2$, the periodic solution of Theorem 2.2 is orbitally stable.

Proof. In the tubular neighborhood constructed in Lemma 2.1, a full set of symplectic coordinates is $(I, I_1,\dots,I_{n-1},\theta,\theta_1,\dots\theta_{n-1})$ and the Hamiltonian is

$$\mathcal{H} = \mathcal{H}_0(I) + \varepsilon\left\{\sum_{k=1}^{n-1}\omega_k I_k + \frac{1}{2}\sum_{k=1}^{n-1}\sum_{j=1}^{n-1} C_{kj} I_k I_j\right\} + \cdots,$$

and the theorem follows by Theorem 14 on page 185 of [5]. ∎

This is just one of many KAM theorems. We place it here because of its simplicity. We will return to KAM-type results in subsequent papers.

2.4. Symmetric periodic solutions. In some cases the problem admits a discrete symmetry. Let $R : M \to M$ be an antisymplectic involution; i.e., $R^*\Omega = -\Omega$ and R^2 is the identity map of M. Then $F = \{p \in M : R(p) = p\}$ is a Lagrangian submanifold of M. The system defined by $\mathcal{H}_0$ (or $\mathcal{H}_\varepsilon$) is *reversible* or *admits R as a symmetry* if $\mathcal{H}_0 \circ R = \mathcal{H}_0$ (or $\mathcal{H}_\varepsilon \circ R = \mathcal{H}_\varepsilon$).

Now R maps an orbit of Y_0 into itself and so is well defined on B. Let $\bar{R}$ be R on B, so $\bar{R} : B \to B$, $\bar{R}^*\omega = -\omega$, $\bar{R}^2$ is the identity map on B, and $\bar{\mathcal{H}} \circ \bar{R} = \bar{\mathcal{H}}$. Let $\bar{F} = \{p \in B : \bar{R}(p) = p\}$.

A classical result [8] is the following lemma.

Lemma 2.3. *If a solution of Y_0 (or Y_ε) starts on F at time $t = 0$ and returns to F after time $t = T$, then the solution is $2T$-periodic, and its orbit is mapped onto itself by R.*

Similarly, if a solution of $\bar{Y}$ starts on $\bar{F}$ at time $t = 0$ and returns to $\bar{F}$ after time $t = T$, then the solution is $2T$-periodic and its orbit is mapped onto itself by $\bar{R}$.

These statements follow from the general identities

$$\phi_\varepsilon^t \circ R = R \circ \phi_\varepsilon^{-t}, \qquad \bar{\phi}^t \circ \bar{R} = \bar{R} \circ \bar{\phi}^{-t}. \tag{8}$$

Such periodic solutions are called *symmetric periodic solutions*. Let $\bar{\phi}^t(\bar{p})$ be a symmetric 2τ-periodic solution of $\bar{Y}$ and $\bar{q} = \bar{\phi}^\tau(\bar{p})$; then there are symplectic coordinate systems (ξ,ζ)

and (X, Z) for $\bar{B}$ at $\bar{p}$ and $\bar{q}$ with $(\xi(\bar{p}), \zeta(\bar{p})) = (0,0)$ and $(X(\bar{q}), Z(\bar{q})) = (0,0)$ such that

$$\bar{R}(\xi, \zeta) = (\xi, -\zeta) \qquad \text{and} \qquad \bar{R}(X, Z) = (X, -Z)$$

(see [34]). Locally $\bar{F}$ is given by $\zeta = 0$ near $\bar{p}$ and by $Z = 0$ near $\bar{q}$. Let $\bar{\phi}^t(\xi, \zeta) = (X(t, \xi, \zeta), Z(t, \xi, \zeta))$. In these coordinates the solution is a symmetric periodic solution if $Z(\tau, 0, 0) = 0$. Such a periodic solution is called a *nondegenerate symmetric periodic solution* if

$$\det \frac{\partial Z}{\partial \xi}(\tau, 0, 0) \neq 0.$$

In general, a nondegenerate symmetric periodic solution persists under small symmetric perturbations. However, in our case the problem is somewhat degenerate, requiring the use of the method and implicit function theorem of Arenstorf [2, 3, 4]. But first we present another lemma. For simplicity let $n = 2$.

Lemma 2.4. *Let $\mathcal{H}_0$ admit R as a symmetry, and let p and (I, θ, y) be as in Lemma 2.1. If $p \in F$, then $R(I, \theta, y) = (I, -\theta, \bar{R}(y))$ and $R(I, \pi + \theta, y) = (I, \pi - \theta, \bar{R}(y))$.*

Let $n = 2$ and $\bar{\phi}^t(\bar{p})$ be a nondegenerate symmetric 2τ-periodic solution of $\bar{Y}$. There exists a set of symplectic action-angle variables (I_1, θ_1) for $\bar{B}$, valid in a neighborhood of $\{\bar{\phi}^t(\bar{p}) : 0 \leq t \leq \tau\}$, such that $\bar{\mathcal{H}}$ is independent of θ_1. Thus $\bar{\mathcal{H}} = \bar{\mathcal{H}}(I, I_1)$, and in these coordinates $\bar{R}(I_1, \theta_1) = (I_1, -\theta_1)$ and $\bar{R}(I_1, \pi + \theta_1) = (I_1, \pi - \theta_1)$, so $\bar{F} = \{(I_1, \theta_1) : \theta_1 \equiv 0 \mod \pi\}$.

If the periodic solution corresponds to $I = I_1 = 0$, then the solution is nondegenerate if

$$\frac{\partial^2 \bar{\mathcal{H}}}{\partial I_1^2}(0,0) \neq 0.$$

Since the reduced space B depends on $\mathcal{H}$ or I, we have coordinates such that

$$\mathcal{H}_\varepsilon = \mathcal{H}_0(I) + \varepsilon \bar{\mathcal{H}}(I, I_1) + O(\varepsilon^2). \tag{9}$$

Proof. By Lemma 2.3 we have $\phi_0^t(p) = R\phi_0^{-t}(p)$ since $R(p) = p$. By construction, θ is t measured from p, so $R : \theta \mapsto -\theta$. Also, $R(I, \pi + \theta, y) = (I, -\pi - \theta + 2\pi, \bar{R}(y)) = (I, \pi - \theta, \bar{R}(y))$.

The proof of the existence of the action-angle variables (I_1, θ_1) for B follows the proof of Lemma 2.1 and the paragraph above. ∎

Theorem 2.6. *Let $\partial \mathcal{H}_0 / \partial I$ be nonzero, let $n = 2$, and let $\bar{p} \in \bar{B}$ with $\bar{R}(\bar{p}) = \bar{p}$ be an initial point for a nondegenerate τ-periodic solution of $\bar{Y}$. Let $p \in M$ with $R(p) = p$ be a point on the orbit which projects to $\bar{p}$. Let α, β be positive integers with α fixed, β large, and ε small.*

Then near the initial condition p, the flow of Y_ε has a symmetric periodic solution where $T(\varepsilon) = \alpha\tau + O(\varepsilon) = \beta T + O(\varepsilon)$.

Arenstorf's method of establishing the existence of symmetric periodic solutions has been around for a long time and has been applied to several problems [3, 4, 14, 25]. The authors believe that the above theorem is new at this level of generality.

Proof. Choose coordinates $(I, I_1, \theta, \theta_1)$ by Lemma 2.4 so that

$$\mathcal{H}_\varepsilon = \mathcal{H}_0(I) + \varepsilon \bar{\mathcal{H}}(I, I_1) + O(\varepsilon^2),$$

and the equations of motion are

$$\dot{I} = O(\varepsilon^2), \qquad \dot{\theta} = -\frac{\partial \mathcal{H}_0}{\partial I}(I) + O(\varepsilon),$$
$$\dot{I}_1 = O(\varepsilon^2), \qquad \dot{\theta}_1 = -\varepsilon \frac{\partial \bar{\mathcal{H}}}{\partial I_1}(I, I_1) + O(\varepsilon^2).$$

Since these equations are autonomous, we may take the fast angle θ as the independent variable so that the equations become

$$\frac{\partial I}{\partial \theta} = 0, \qquad \frac{\partial I_1}{\partial \theta} = 0, \qquad \frac{\partial \theta_1}{\partial \theta} = \varepsilon G(I, I_1) = \varepsilon \left\{ \frac{\partial \mathcal{H}_0}{\partial I}(I) \right\}^{-1} \frac{\partial \bar{\mathcal{H}}}{\partial I_1}(I, I_1), \tag{10}$$

plus $O(\varepsilon^2)$ terms. For the moment, ignore the $O(\varepsilon^2)$ terms and seek a symmetric periodic solution of the approximate equations. Let α and β be relatively prime integers, $\nu = G(0,0)^{-1}$, and set $\varepsilon = \nu\alpha/\beta$. Start with initial conditions $I = I_1 = \theta_1 = 0$ and integrate the approximate equations on θ from 0 to $\beta\pi$ to obtain the approximate solution

$$I = 0, \qquad I_1 = 0, \qquad \theta_1 = \alpha\pi.$$

This approximate solution satisfies the symmetry conditions and so to this level of approximation is a symmetric periodic solution.

Fixing α and taking β large, the parameter ε becomes small, and so we might expect that this approximate solution could be continued into the full problem. However, the problem is complicated by the fact that taking β large corresponds to integrating the equations over a large variation of θ. As Arenstorf has observed, the usual implicit function theorem cannot be applied since one cannot set $\varepsilon = 0$ to find an approximate solution. Thus, we must follow Arenstorf and make careful estimates.

First, we fix the integer α and the initial condition $I = 0$ once and for all. Let the superscript f denote the full solution of (10) including $O(\varepsilon^2)$ terms, the superscript a the approximate solution, and the superscript e the error term. Integrate the full equations with initial condition $I_1 = K$, and integrate from $\theta = 0$ to $\theta = \beta\pi$ to obtain

$$\theta_1^f(\beta\pi, \varepsilon, K) = \theta_1^a(\beta\pi, \varepsilon, K) + \theta_1^e(\beta\pi, \varepsilon, K), \tag{11}$$

where $\theta_1^a(\beta\pi, \varepsilon, K) = \varepsilon\beta\pi\, G(0, K)$.

The error term θ_1^e is due to the $O(\varepsilon^2)$ terms appended to (10). Bounding these $O(\varepsilon^2)$ terms by $C\varepsilon^2$, and taking the $O(\varepsilon)$-Lipschitz constant of the θ_1-flow to be $L\varepsilon$, we apply to θ_1^e a standard Gronwall estimate of the form $\{u(0) = 0 \text{ and } d|u|/d\theta \le L\varepsilon|u| + C\varepsilon^2\} \Rightarrow |u(\theta)| \le \varepsilon C(e^{\varepsilon L\theta} - 1)/L$ (see, e.g., Hartman [23]) to conclude that

$$|\theta_1^e| \le \varepsilon C(e^{\varepsilon L\beta\pi} - 1)/L. \tag{12}$$

A similar estimate holds for the first partial derivatives of θ_1^e.

The approximate equation has solution $\theta_1^a(\beta\pi, \varepsilon, K) = \alpha\pi$ by taking $\varepsilon = \nu\alpha/\beta$ and $K = 0$. Also by assumption $\partial\theta_1^a/\partial K$ is nonzero. From the estimate (12) the error term can be made arbitrarily small by taking β large with $\varepsilon = \nu\alpha/\beta$, since in this case the estimate (12) reads

$|\theta_1^e| \le C\nu\alpha(e^{L\nu\alpha\pi} - 1)/(L\beta)$. Similarly, the derivatives of θ_1^e can be made small by taking β large. These estimates ensure that we remain in a compact neighborhood of the approximate solution. Thus, the implicit function theorem of Arenstorf [2, 3, 4] applies, and there exists β_0 such that if $\beta > \beta_0$, then there is a solution $K_s(\beta)$ such that

$$\theta_1^f(\beta\pi, \nu\alpha/\beta, K_s(\beta)) = \alpha\pi.$$

This gives the initial conditions for a symmetric periodic solution. ■

2.5. Weinstein's theorem. For completeness we add this much deeper global result on the existence of periodic solutions which is not a corollary of Reeb's theorems. Let X be a topological space; then the category of X in the sense of Lusternik–Schnirelmann, $\mathrm{cat}(X)$, is the smallest number of open sets that are contractible in X and that cover X [26, 31]. One of the main uses of this concept is in the theorem that says that every smooth function on a compact manifold M has at least $\mathrm{cat}(M)$ critical points. Weinstein extended the connection between critical points of functions and periodic solutions of Hamiltonian systems to prove the following theorem.

Theorem 2.7. *Assume B is compact and simply connected in the sense that $H^1(B, \mathbb{R}) = 0$, where $H^1(B, \mathbb{R})$ is the one-dimensional cohomology group of B over the real numbers, and let $\ell = cat(B)$ be the Lusternik–Schnirelmann category of B. Then for small ε the flow of Y_ε has at least ℓ periodic solutions with periods near T (there is no nondegeneracy assumption)* [45, 46].

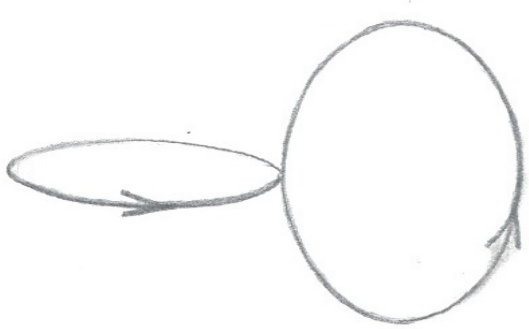
The n-sphere S^n has category 2 and all other compact manifolds have category greater than 2. For $X = S^n \times S^n$, we produce three contractible open sets that cover X, so $\mathrm{cat}(S^n \times S^n) = 3$. (We illustrate this in Figure 2 for the case $n = 1$.)

It is helpful to think of X as a cellular complex. Starting with two n-cells, identify one point in one cell with one point in the other cell to form the wedge product (or wedge sum) of two spheres $S^n \vee S^n$. (The wedge product of two circles is thus a figure eight, as in Figure 2a; for a precise definition of $S^n \vee S^n$, see [24, p. 10].) Now attach a $2n$-cell to $S^n \vee S^n$ to form X (Figure 2b). Take the $2n$-cell to be the first contractible set. For the second set, delete one point from each of the two spheres in $S^n \vee S^n$. This set is $D^n \vee D^n$ and can easily be "fattened up" to an open set in $S^n \times S^n$. These two sets cover all but the two points deleted from $S^n \times S^n$. For the third set, choose any contractible open set in X that covers these two points. (Figure 2c shows sets that could be fattened up to form the second and third contractible open sets for $n = 1$.)

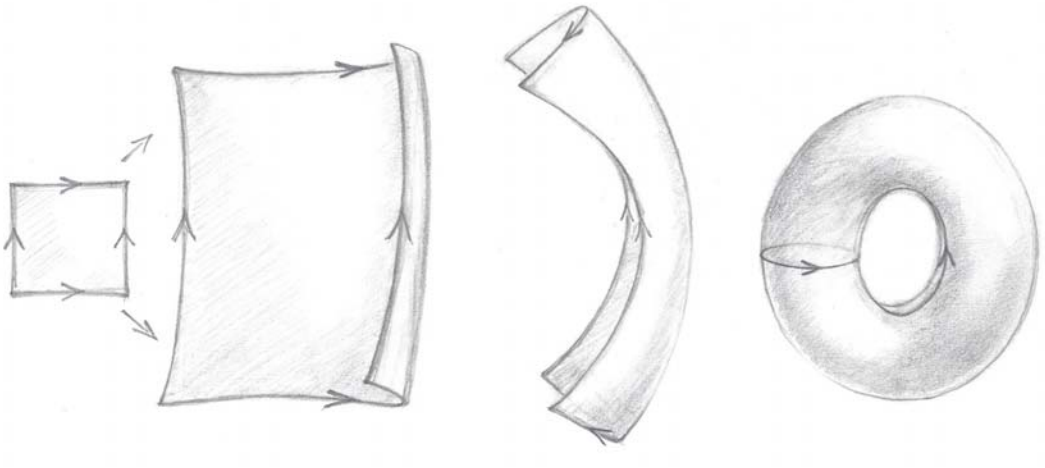
3. The planar lunar problem.

3.1. The Hamiltonians. For us the lunar problem is the restricted three-body problem where the infinitesimal particle is close to one of the primaries [35, 36]. Note that, in this context, the terminology "lunar problem" means that the zero mass point can move about either primary, which is more general than the way it is historically defined, where the infinitesimal mass point moves only about the smaller primary (or secondary).

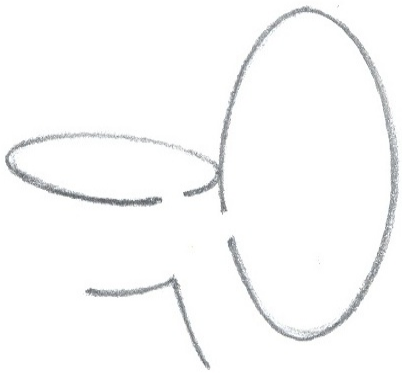
Here we summarize the normalization and reduction as given in [42] and then apply the general theorems from section 2. Figure 3 is a sketch of the planar lunar problem in the rotating frame—the projection on the $x_1\,x_2$-plane. The primary bodies are point particles



(a) $S^1 \vee S^1$



(b) Attaching the 2-cell to $S^1 \vee S^1$ to form $S^1 \times S^1$



(c) The second and third contractible sets (before "fattening")

Figure 2. *The Lusternik–Schnirelmann category of* $S^1 \times S^1$.

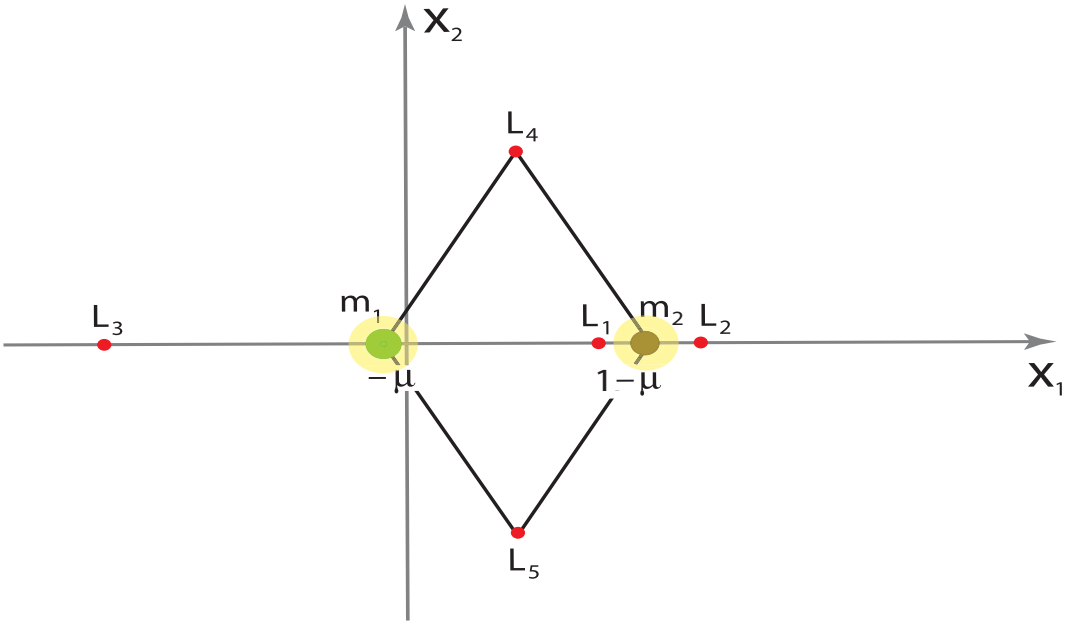


Figure 3. *The lunar problem.*

with masses m_1 and m_2 and are located at the points $(-\mu, 0)$ and $(1-\mu, 0)$, respectively. Parameter $\mu = m_1/(m_1+m_2)$ (it is assumed that $m_1 \geq m_2$). The motion of the infinitesimal particle is confined to either one of the yellow regions around the primaries. The points $L_1, \ldots, L_5$ are the equilibria of the restricted three-body problem in the rotating frame. The infinitesimal particle touches neither L_1 nor L_2.

We start with the Hamiltonian of the planar circular restricted three-body problem in rotating coordinates given by

$$\mathcal{H} = \frac{1}{2}(y_1^2 + y_2^2) - (x_1 y_2 - x_2 y_1) - \frac{\mu}{\sqrt{(x_1 - 1 + \mu)^2 + x_2^2}} - \frac{1-\mu}{\sqrt{(x_1 + \mu)^2 + x_2^2}}. \tag{13}$$

We now change coordinates in order to bring $\mathcal{H}$ into suitable form. First we perform the linear change from y_2 and x_1 to $y_2 - \mu$ and $x_1 - \mu$, respectively, to bring one primary to the origin. Then, we introduce a small parameter ε by replacing $y = (y_1, y_2)$ by $\varepsilon^{-1}(1-\mu)^{1/3} y$ and $x = (x_1, x_2)$ by $\varepsilon^2 (1-\mu)^{1/3} x$. By doing so we restrict $\mathcal{H}$ to a particular case where the infinitesimal particle is moving around one of the primaries. This change is symplectic with multiplier $\varepsilon^{-1}(1-\mu)^{-2/3}$; thus $\mathcal{H}$ must be replaced by $\varepsilon^{-1}(1-\mu)^{-2/3}\mathcal{H}$.

In the next step, we scale time by dividing t by ε^3 and multiplying $\mathcal{H}$ by ε^3. Then we expand the resulting Hamiltonian in powers of ε to get

$$\mathcal{H}_\varepsilon = \frac{1}{2}(y_1^2 + y_2^2) - \frac{1}{\sqrt{x_1^2 + x_2^2}} - \varepsilon^3 (x_1 y_2 - x_2 y_1) + \frac{1}{2}\varepsilon^6 \mu(-2x_1^2 + x_2^2) + \cdots. \tag{14}$$

The zeroth-order term is the Hamiltonian of the Kepler problem and the $O(\varepsilon^3)$ term is due to the rotating coordinates. It is not until $O(\varepsilon^6)$ that the second primary influences the motion.

Moser has shown [39] that the n-dimensional Kepler problem can be regularized and the regularized flow is equivalent to the geodesic flow on S^n. Let us be more specific for our case. Let $\mathcal{K} = \mathcal{H}_0$ be the Hamiltonian of the planar Kepler problem defined on $(\mathbb{R}^2 \setminus \{0\}) \times \mathbb{R}^2$, $K_0 = \{(x, y) \in (\mathbb{R}^2 \setminus \{0\}) \times \mathbb{R}^2 : \mathcal{K}(x, y) = -\frac{1}{2}\}$. Let S^2 be the unit sphere, $\hat{S}^2$ the unit sphere punctured at the north pole, TS^2 ($T\hat{S}^2$) the tangent bundle of the (punctured) 2-sphere, and $T_0S^2 = \{v \in TS^2 : \|v\| = 1\}$ ($T_0\hat{S}^2 = \{v \in T\hat{S}^2 : \|v\| = 1\}$) the unit (punctured) sphere bundle.

The elliptic domain $\mathcal{E}$ is the set of points in K_0 which gives rise to elliptic orbits. All the solutions of the Kepler problem in $\mathcal{E}$ are periodic with the same period. Thus $\mathcal{E}$ is a circle bundle but is not compact. The base is two punctured disks, as we show below.

Moser constructs a symplectic diffeomorphism from $(\mathbb{R}^2 \setminus \{0\}) \times \mathbb{R}^2$ onto $T\hat{S}^2$ which, when restricted to K_0, maps onto $T_0\hat{S}^2$. After changing the time variable for the Kepler problem, the diffeomorphism takes Kepler flow on K_0 to the geodesic flow on $T_0\hat{S}^2$. The geodesic flow on $T_0\hat{S}^2$ obviously extends to all of T_0S^2 and is considered the regularized Kepler problem.

All the geodesics on T_0S^2 are periodic, so T_0S^2 is a circle bundle with base S^2. Moser shows that a small perturbation of the Kepler problem can be carried over as a small perturbation of the geodesic problem. He then shows that the average of the perturbation over the geodesic flow defines a smooth flow on the base. We next proceed to construct this flow on the base.

Now express (14) in mixed polar and Delaunay coordinates (see, for instance, [9, 18]) so that the Hamiltonian becomes

$$\mathcal{H}_\varepsilon = -\frac{1}{2\,L^2} - \varepsilon^3 G - \frac{1}{4}\varepsilon^6 \mu r^2 \big(1 + 3\cos(2\vartheta)\big) + \cdots .$$

Here, (ℓ, g, L, G) are the usual Delaunay variables, ℓ the mean anomaly, g the argument of the pericenter, and L the square of the semimajor axis. G is the third component of the angular momentum vector $\mathbf{G} = (0, 0, G)$; thus $0 \le |G| \le L$ and G can be positive, negative, or zero. This is a coordinate system on $\mathcal{E}$. Finally, (r, ϑ) are the usual polar coordinates.

We eliminate the mean anomaly ℓ to a certain order by means of a special Lie transformation well suited for perturbed Kepler problems, the so-called normalization of Delaunay [18, 19]. We arrive at

$$\mathcal{H}_\varepsilon = -\frac{1}{2\,L^2} - \varepsilon^3 G + \frac{1}{8}\varepsilon^6\,\mu\,L^2\left(3\,G^2 - 5\,L^2 + 15\,(G^2 - L^2)\cos(2g)\right) + \cdots . \tag{15}$$

Here only finitely many terms have been put into normal form. This normalization is effectively the average of the perturbations over the periodic orbits of the Kepler problem in $\mathcal{E}$.

The base space (or reduced space or orbit space) for the regularized Kepler problem is a 2-sphere S^2 [39]. Figure 4 may be helpful in visualizing this base space. The flow is given by the circles around the poles on the sphere. A coordinate system for the reduced space is $\mathbf{a} = \mathbf{G} + LA$, where A is the Laplace–Runge–Lenz vector. One has $A = e\,(\cos g, \sin g, 0)$ and then $a_1 = e\cos g$, $a_2 = e\sin g$, and $a_3 = G$ on $\mathcal{E}$, where $e = \sqrt{1 - G^2/L^2}$ is the eccentricity. One can check that $|\mathbf{a}| = L$ and the vector $\mathbf{a}$ uniquely determines an orbit of the Kepler problem on the energy level $h = -1/(2L^2)$. Each point on the sphere $a_1^2 + a_2^2 + a_3^2 = L^2$ corresponds to a bounded orbit of the Kepler problem. Points $(0, 0, \pm L)$ correspond to the circular orbits, the circle $a_3 = 0$ (the equator, or the green circle in Figure 4) corresponds to collision orbits, and the other points on the sphere correspond to elliptic orbits. The

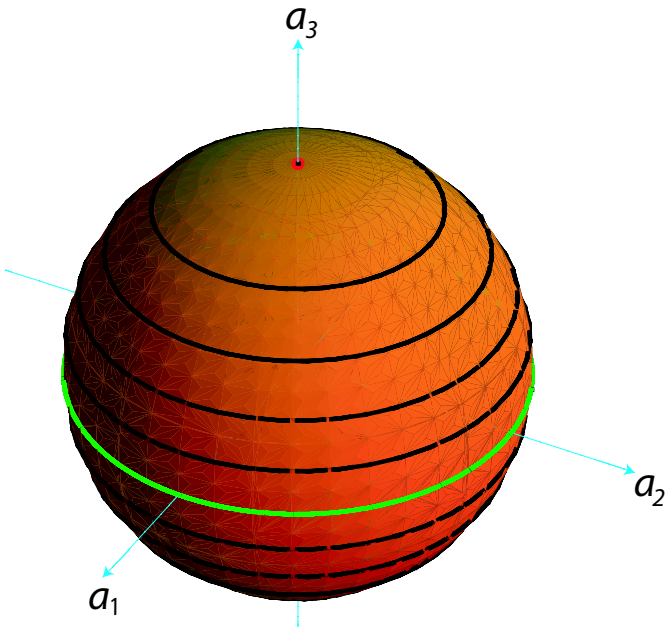


Figure 4. *The base space for the regularized Kepler problem.*

complement of $(0, 0, \pm L) \cup \{a_3 = 0\}$ is the reduced space of the elliptic domain $\mathcal{E}$.

Now compute from (15) the Hamiltonian on the reduced space of $\mathcal{E}$. Use $\cos(2g) = (a_1^2 - a_2^2)/(L^2 - G^2)$, so

$$\mathcal{H}_\varepsilon = -\frac{1}{2L^2} - \varepsilon^3 a_3 + \frac{1}{8}\varepsilon^6 \mu L^2 \left(3a_3^2 - 5L^2 - 15(a_1^2 - a_2^2)\right) + \cdots .$$

We first drop the higher-order nonnormalized terms and then use $a_3^2 = L^2 - a_1^2 - a_2^2$, dropping additive constants and dividing by ε^3 to get the Hamiltonian

$$\bar{\mathcal{H}} = -a_3 - \frac{3}{4}\varepsilon^3 \mu L^2 \left(3a_1^2 - 2a_2^2\right) + \cdots . \tag{16}$$

We note that this Hamiltonian is well defined and smooth on the exceptional set $(0, 0, \pm L) \cup \{a_3 = 0\}$. Since Moser proved that the averaged (normalized) Hamiltonian of the perturbation is defined and smooth on all of S^2, (16) is the Hamiltonian on the full reduced space S^2.

To obtain the equations of motion, note that $\{a_1, a_2\} = a_3$, $\{a_2, a_3\} = a_1$, $\{a_3, a_1\} = a_2$, and $\dot{a}_j = \sum_l \{a_j, a_l\} \partial \bar{\mathcal{H}} / \partial a_l$. So the equations of motion become

$$\begin{aligned} \dot{a}_1 &= a_2 + 3\varepsilon^3 \mu L^2 a_2 a_3 + \cdots , \\ \dot{a}_2 &= -a_1 + \tfrac{9}{2} \varepsilon^3 \mu L^2 a_3 a_1 + \cdots , \\ \dot{a}_3 &= -\tfrac{15}{2} \varepsilon^3 \mu L^2 a_1 a_2 + \cdots . \end{aligned} \tag{17}$$

3.2. Analysis of equilibria. We now apply the results of section 2 to the Hamiltonians for the planar lunar problem. Just from the facts that $B = S^2 = \{a_1^2 + a_2^2 + a_3^2 = L^2\}$, $H^1(S^2) = 0$,

and the Lusternik–Schnirelmann category of S^2 is 2, by Weinstein's theorem, Theorem 2.7, we conclude that there are at least two periodic solutions of the corresponding flow defined by Y_ε with period near $T = 2\pi L^3$. This applies to any (small) perturbation of the planar Kepler problem.

Looking at the Hamiltonian on B yields more information about these periodic solutions. The Hamiltonian (16) has two nondegenerate critical points, a maximum at $\mathbf{a} = (0, 0, -L)$ and a minimum at $\mathbf{a} = (0, 0, L)$, which by Reeb's theorem, Theorem 2.2, and Corollaries 2.2 and 2.3 correspond to elliptic periodic solutions of the planar restricted three-body problem of period $T(\varepsilon) = T + O(\varepsilon^3)$. (Note that $(0, 0, -L)$ and $(0, 0, L)$ are parametrically stable points according to Corollary 2.3, as they are respectively a minimum and a maximum.) These are the classical Hill's orbits of the restricted problem, which are the continuation of the circular solutions of the Kepler problem (see [12, 36] and the references therein). The maximum gives the prograde orbit, which is located at the north pole of the sphere in Figure 4 (it is represented by a red point), and the minimum provides the retrograde orbit (the south pole in Figure 4). The index of $(0, 0, -L)$ is 2, whereas the index of $(0, 0, L)$ is 0. Hence $C_0 = C_2 = 1$ and $C_j = 0$ for $j \notin \{0, 2\}$. The Betti numbers of S^2 are $\beta_0 = \beta_2 = 1$, and the others are zero. Moreover, the Euler–Poincaré characteristic of S^2 is 2, which is consistent with the Betti and C_j numbers. Thus, for all j, $C_j = \beta_j$, and the Morse inequalities (given in Corollary 2.1) become equalities. Note that in this case the Lusternik–Schnirelmann category and the Euler–Poincaré characteristic of S^2 yield the same estimate, which coincides with the number of critical points of the Hamiltonian $\bar{\mathcal{H}}$.

Since $\bar{\mathcal{H}} = -a_3 + \cdots$ the linearized equations about $(0, 0, \pm L)$ are

$$\dot{a}_1 = a_2, \qquad \dot{a}_2 = -a_1,$$

and so the characteristic exponents at these critical points are $\pm i$ (see Figure 4).

Thus, these near-circular periodic solutions are elliptic with characteristic multipliers 1, 1, $1 + \varepsilon^3 T i + O(\varepsilon^6)$, *and* $1 - \varepsilon^3 T i + O(\varepsilon^6)$.

As a last step, we have to undo the initial scalings and the shift to return to the Hamiltonian $\mathcal{H}$. Taking into account that the periodic solutions are near-circular, they have approximate radii $|x| \approx L^2$ and periods near $2\pi L^3$. Hence, because of the scalings, we conclude that the periodic solutions of $\mathcal{H}$ have radii $|x| \approx \varepsilon^2 L^2$ and periods $T(\varepsilon) \approx 2\pi\varepsilon^3 L^3$.

3.3. A twist condition. To see if Theorem 2.5 applies at $(0, 0, \pm L)$ we need several changes of variables. We start by moving the equilibria $(0, 0, \pm L)$ to the origin of a coordinate system. Therefore, we define

$$\bar{a}_1 = a_1, \quad \bar{a}_2 = a_2, \quad \bar{a}_3 = a_3 \mp L,$$

and then we introduce (local) symplectic coordinates Q and P as

$$Q = \sqrt{2}\frac{L\,\bar{a}_1}{\sqrt{2\,L \pm \bar{a}_3}} = \sqrt{2}\sqrt{L \mp G}\cos g,$$
$$P = \pm\sqrt{2}\frac{L\,\bar{a}_2}{\sqrt{2\,L \pm \bar{a}_3}} = \pm\sqrt{2}\sqrt{L \mp G}\sin g.$$

By recalling that (ℓ, g, L, G) are symplectic variables, it is almost straightforward to check that $\{Q, P\} = 1$; thus Q has the role of a coordinate, whereas P corresponds to its conjugate

momentum. These coordinates are valid in the hemispheres $\pm a_3 > 0$ (i.e., $\pm G < L$).

Now, to write $\bar{\mathcal{H}}$ in these coordinates, first note that

$$\frac{1}{2}(Q^2 + P^2) = L \mp G = L \mp a_3,$$

and also

$$a_1^2 = \frac{Q^2}{2L^2}(L \pm a_3), \qquad a_2^2 = \frac{P^2}{2L^2}(L \pm a_3).$$

Making this change of variables and dropping additive constants gives

$$\bar{\mathcal{H}} = \pm\frac{1}{2}(Q^2 + P^2) - \frac{3}{16}\varepsilon^3\mu(2P^2 - 3Q^2)(P^2 + Q^2 - 4L) + \cdots.$$

Change to action-angle variables by

$$Q = \sqrt{2I_1}\cos\theta_1, \qquad P = \sqrt{2I_1}\sin\theta_1$$

(note that $dQ \wedge dP = dI_1 \wedge d\theta_1$) to get

$$\bar{\mathcal{H}} = \pm I_1 - \frac{3}{4}\varepsilon^3\mu\, I_1\,(2L - I_1)\,(-2 + 5\cos^2\theta_1) + \cdots,$$

and then average over θ_1 to get

$$\bar{\mathcal{H}} = \pm I_1 - \frac{3}{8}\varepsilon^3\mu\, I_1\,(2L - I_1) + \cdots.$$

Note that the second derivative of $\bar{\mathcal{H}}$ with respect to I_1 is

$$\frac{\partial^2\bar{\mathcal{H}}}{\partial I_1^2} = \frac{3}{4}\varepsilon^3\,\mu, \tag{18}$$

and it does not vanish. Thus there is a twist term, but the hypothesis of Theorem 2.5 does not hold, as there is an additional ε^3 in front of the twist term.

This suggests, but does not prove, that these near-circular periodic solutions are stable and enclosed by invariant KAM tori.

We push the normalization up to order ε^8 in order to prove that the periodic solutions associated with the equilibria $(0, 0, \pm L)$ are not circular but have a small eccentricity. The terms factorized by ε^8 are

$$\tfrac{5}{32}\varepsilon^8\,\mu\,(1-\mu)^{1/3}\,e\,L^4\cos g\left(13\,G^2 - 7\,L^2 - 35\,(G^2 - L^2)\cos(2\,g)\right). \tag{19}$$

Now, after incorporating these terms into the Hamiltonian $\mathcal{H}_\varepsilon$ given by (15), the equilibria $(0, 0, \pm L)$ are transformed to

$$\left(\mp\frac{15}{16}\,\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6,\quad 0,\quad \pm\frac{\sqrt{256L^2 - 225\varepsilon^{10}\mu^2(1-\mu)^{2/3}L^{12}}}{16}\right).$$

Now the above equilibria do not correspond to circular solutions because their eccentricity is given by $e = \frac{15}{16}\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^5+\cdots$. The magnitude of their angular momentum vector is $G = \pm L \mp \frac{225}{512}\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{11}+\cdots$. This implies that the periodic solutions associated with $(0,0,\pm L)$ are indeed elliptic periodic solutions whose projections onto configuration space yield elliptic orbits with eccentricity close to zero. The inclination is zero for the periodic solution related to $(0,0,L)$, while it is π for the periodic solution related to $(0,0,-L)$. This proves that up to terms of order ε^8 the periodic solutions are near-circular periodic solutions.

Thus, these equilibria correspond to near-circular elliptic periodic orbits.

3.4. Continuation of elliptic orbits. The planar restricted three-body problem is symmetric in the line of syzygy; i.e., $R : (x_1, x_2, y_1, y_2) \to (x_1, -x_2, -y_1, y_2)$ is an antisymplectic involution that leaves the Hamiltonian (13) invariant. The Lagrangian subspace $F = \{(x_1, 0, 0, y_2)\}$ corresponds to orthogonal crossings of the line of syzygy. In Delaunay variables $R : (\ell, g, L, G) \to (-\ell, -g, L, G)$ and $F = \{(0, 0, L, G)\}$.

On the reduced space $\bar{R} : (a_1, a_2, a_3) \to (a_1, -a_2, a_3)$ or $\bar{R} : (Q, P) \to (Q, -P)$. The Lagrangian subspace $\bar{F}$ is the meridian circle $\{(a_1, 0, a_3)\}$ or $\{(Q, 0)\}$. A point on $\bar{F}$ corresponds to a symmetric elliptic orbit of the Kepler problem and the periodic solution on $\bar{\mathcal{H}} =$ constant corresponds to a family of precessing Keplerian ellipses which start and end at a symmetric ellipse.

By Theorem 2.6 the existence of symmetric periodic solutions which are the continuation of this family of precessing Keplerian ellipses is ensured because, according to (18), the condition $\frac{\partial^2 \bar{\mathcal{H}}}{\partial I_1^2} \neq 0$ holds. These are the periodic solutions obtained by Arenstorf in [3].

4. The spatial lunar problem.

4.1. The Hamiltonians. The Hamiltonian of the spatial problem is given in the rotating frame by

$$\mathcal{H} = \frac{1}{2}(y_1^2 + y_2^2 + y_3^2) - (x_1 y_2 - x_2 y_1) - \frac{\mu}{\sqrt{(x_1 - 1 + \mu)^2 + x_2^2 + x_3^2}} - \frac{1-\mu}{\sqrt{(x_1 + \mu)^2 + x_2^2 + x_3^2}}.$$

We change variables, scale time, and scale the Hamiltonian in the same way as in the planar case in order to arrive to the lunar case of the spatial restricted circular three-body problem (see [25]). After expanding in powers of the small parameter, we end up with the system

$$\mathcal{H}_\varepsilon = \frac{1}{2}(y_1^2 + y_2^2 + y_3^2) - \frac{1}{\sqrt{x_1^2 + x_2^2 + x_3^2}} - \varepsilon^3\,(x_1 y_2 - x_2 y_1) + \frac{1}{2}\,\varepsilon^6\,\mu\,(-2x_1^2 + x_2^2 + x_3^2) + \cdots.$$

Now we have a perturbation of the spatial Kepler problem. Moser has shown that the three-dimensional Kepler problem can be regularized and the regularized flow is equivalent to the geodesic flow on S^3. We proceed just as in the planar problem to find and analyze the averaged equations on the reduced space.

The following step consists in expressing $\mathcal{H}_\varepsilon$ in such a way that we can perform Lie transformations conveniently (see [17]). We use polar-nodal coordinates $(r, \vartheta, \nu, R, G, N)$ and Delaunay coordinates (ℓ, g, ν, L, G, N). The angle ϑ is the argument of the latitude, and ν is the argument of the node. The coordinate R is the momentum conjugate to the radial

variable r, $G = |\mathbf{G}|$ is the magnitude of angular momentum, and N is the third component of the angular momentum $\mathbf{G}$, so $0 \le |N| \le G \le L$. Expressing $\mathcal{H}_\varepsilon$ in these variables, we get

$$\mathcal{H}_\varepsilon = -\frac{1}{2\,L^2} - \varepsilon^3\,N + \frac{1}{8}\,\varepsilon^6\,\mu\,r^2\,\Big(1 - 3\,c^2 - 3\,(1-c^2)\cos(2\,\vartheta) \\ - 3\,\big(1 - c^2 + (1+c^2)\cos(2\,\vartheta)\big)\cos(2\,\nu) + 6\,c\sin(2\,\nu)\sin(2\,\vartheta)\Big) + \cdots,$$

where $c = N/G$. After performing the normalization of Delaunay to a fixed finite order, we arrive at the Hamiltonian

$$\mathcal{H}_\varepsilon = -\frac{1}{2\,L^2} - \varepsilon^3 N + \frac{1}{16}\varepsilon^6\,\mu\,L^4\,\Big((2+3\,e^2)\,\big(1 - 3\,c^2 - 3\,(1-c^2)\cos(2\,\nu)\big) \\ - 15\,e^2\cos(2\,g)\big(1 - c^2 + (1+c^2)\cos(2\,\nu)\big) + 30\,c\,e^2\sin(2\,g)\sin(2\,\nu)\Big) + \cdots, \tag{20}$$

where $e = \sqrt{1 - G^2/L^2}$. This normal form Hamiltonian was calculated previously in [42]. The transformed Hamiltonian, after truncating higher-order terms, depends on the two angles g and ν and their associated momenta G and N, respectively, whereas L is an integral of motion. Applying reduction theory, once higher-order terms have been dropped, $\mathcal{H}_\varepsilon$ is defined on the orbit space, or base space, which is the four-dimensional space $S^2 \times S^2$ [39].

We can use the set of variables given by $\mathbf{a} = (a_1, a_2, a_3)$ and $\mathbf{b} = (b_1, b_2, b_3)$ with the constraints $a_1^2 + a_2^2 + a_3^2 = L^2$ and $b_1^2 + b_2^2 + b_3^2 = L^2$ to parameterize $S^2 \times S^2$, where $\mathbf{a} = \mathbf{G} + LA$ and $\mathbf{b} = \mathbf{G} - LA$. We recall that $\mathbf{G}$ is the angular momentum vector and A is the Laplace–Runge–Lenz vector; moreover, $|\mathbf{a}| = |\mathbf{b}| = L$. Notice that the a_i and b_i belong to the interval $[-L, L]$. The explicit expressions for $\mathbf{a}$ and $\mathbf{b}$ in terms of Delaunay variables are found in Coffey, Deprit, and Miller [11] and in Cushman [15].

In particular, $2G = ((a_1 + b_1)^2 + (a_2 + b_2)^2 + (a_3 + b_3)^2)^{1/2}$, so $G = 0$ in $S^2 \times S^2$ if and only if $a_1 + b_1 = a_2 + b_2 = a_3 + b_3 \equiv 0$, $a_1^2 + a_2^2 + a_3^2 = L^2$, and $b_1^2 + b_2^2 + b_3^2 = L^2$. Thus, the subset of $S^2 \times S^2$ given by $\mathcal{R} = \{(\mathbf{a}, -\mathbf{a}) \in \mathbb{R}^6 \mid a_1^2 + a_2^2 + a_3^2 = L^2\}$ is a two-dimensional set homeomorphic to S^2 consisting of the rectilinear trajectories. In Delaunay elements the circular orbits satisfy the condition $G = L$, and in terms of $\mathbf{a}$ and $\mathbf{b}$ this implies that $a_1 = b_1$, $a_2 = b_2$, and $a_3 = b_3$. So the circular orbits define the two-dimensional set homeomorphic to S^2 given by $\mathcal{C} = \{(\mathbf{a}, \mathbf{a}) \in \mathbb{R}^6 \mid a_1^2 + a_2^2 + a_3^2 = L^2\}$. Similarly, equatorial trajectories satisfy $G = |N|$ and are given by the two-dimensional set $\mathcal{E} = \{(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^6 \mid a_1^2 + a_2^2 + a_3^2 = L^2$, $b_1 = -a_1$, $b_2 = -a_2$, $b_3 = a_3\}$, which is again homeomorphic to S^2. Just as in the planar case, the introduction of these invariants extends the use of the Delaunay variables, as we can include equatorial, circular, and rectilinear solutions [41]. The other points on $S^2 \times S^2$ correspond to elliptic orbits of the Kepler problem.

After several simplifications and manipulations over $\mathcal{H}_\varepsilon$, including the dropping of the constant term $-1/(2L^2)$ and division by ε^3, we arrive at

$$\bar{\mathcal{H}} = -\frac{1}{2}\,(a_3 + b_3) - \frac{1}{8}\,\varepsilon^3\,\mu\,L^2\,\big(3\,a_1^2 - 3\,a_2^2 - 3\,a_3^2 - 12\,a_1\,b_1 + 3\,b_1^2 \\ + 6\,a_2\,b_2 - 3\,b_2^2 + 6\,a_3\,b_3 - 3\,b_3^2\big) + \cdots. \tag{21}$$

The Poisson structure on $S^2 \times S^2$ in these coordinates is

$$\{a_1, a_2\} = 2\,a_3, \qquad \{a_2, a_3\} = 2\,a_1, \qquad \{a_3, a_1\} = 2\,a_2,$$
$$\{b_1, b_2\} = 2\,b_3, \qquad \{b_2, b_3\} = 2\,b_1, \qquad \{b_3, b_1\} = 2\,b_2, \qquad \{a_i, b_j\} = 0.$$

The corresponding equations of motion are

$$\begin{aligned}
\dot{a_1} &= a_2 - \tfrac{3}{2}\,\varepsilon^3\,\mu\,L^2\,(a_3\,b_2 - a_2\,b_3) + \cdots,\\
\dot{a_2} &= -a_1 + \tfrac{3}{2}\,\varepsilon^3\,\mu\,L^2\,(2a_1\,a_3 - 2\,a_3\,b_1 - a_1\,b_3) + \cdots,\\
\dot{a_3} &= -\tfrac{3}{2}\,\varepsilon^3\,\mu\,L^2\,(2\,a_1\,a_2 - 2\,a_2\,b_1 - a_1\,b_2) + \cdots,\\
\dot{b_1} &= b_2 + \tfrac{3}{2}\,\varepsilon^3\,\mu\,L^2\,(a_3\,b_2 - a_2\,b_3) + \cdots,\\
\dot{b_2} &= -b_1 - \tfrac{3}{2}\,\varepsilon^3\,\mu\,L^2\,(a_3\,b_1 + 2\,a_1\,b_3 - 2\,b_1\,b_3) + \cdots,\\
\dot{b_3} &= \tfrac{3}{2}\,\varepsilon^3\,\mu\,L^2\,(a_2\,b_1 + 2\,a_1\,b_2 - 2\,b_1\,b_2) + \cdots.
\end{aligned} \tag{22}$$

We stress that the equations of motion are global in the whole base space B. Including terms of order ε^3 is enough to determine the relative equilibria of $\bar{\mathcal{H}}$.

4.2. Analysis of equilibria. Let us now turn to the application of the results of section 2 to the spatial lunar problem. Just from the facts that $B = S^2 \times S^2 = \{a_1^2 + a_2^2 + a_3^2 = L^2, b_1^2 + b_2^2 + b_3^2 = L^2\}$, $H^1(S^2 \times S^2) = 0$, and the Lusternik–Schnirelmann category of $S^2 \times S^2$ is 3, by Weinstein's theorem, Theorem 2.7, we can conclude that there are at least three periodic solutions of the corresponding flow defined by Y_ε with period near $T = 2\pi L^3$. This holds for any perturbation of the spatial Kepler problem.

Looking at the Hamiltonian on B yields more information about these periodic solutions. The Hamiltonian (21) starts as $\bar{\mathcal{H}} = -\frac{1}{2}(a_3 + b_3) + \cdots$, so it has a nondegenerate maximum at $(\mathbf{a}, \mathbf{b}) = (0, 0, -L, 0, 0, -L)$ and a nondegenerate minimum at $(\mathbf{a}, \mathbf{b}) = (0, 0, L, 0, 0, L)$, which by Reeb's theorem, Theorem 2.2, and Corollary 2.2 correspond to elliptic periodic solutions of the spatial restricted three-body problem of period $T(\varepsilon) = T + O(\varepsilon^3)$. These are the circular equatorial motions already encountered in the planar case. It also has two nondegenerate critical points of index 2 at $(\mathbf{a}, \mathbf{b}) = (0, 0, \pm L, 0, 0, \mp L)$ which correspond to rectilinear motions whose projection in the coordinate space leads to periodic orbits in the vertical axis x_3. They correspond to the rectilinear trajectories found by Belbruno [6] for small μ.

The Betti numbers of $S^2 \times S^2$ are $\beta_0 = \beta_4 = 1$, $\beta_2 = 2$, and all the others are zero. As we have seen, $\bar{\mathcal{H}}$ is a Morse function and has the minimum number of critical points consistent with the Morse inequalities found in Corollary 2.1.

Near the critical points we can use (a_1, a_2, b_1, b_2) as coordinates on $B = S^2 \times S^2$. From the equations (22) one sees that the characteristic exponents of all four critical points of Y_ε at the four equilibria are $\pm i$ (double). Thus, by Corollary 2.2, the characteristic multipliers of the corresponding periodic solutions are 1, 1, $1 + \varepsilon^3\,T\,i$, $1 + \varepsilon^3\,T\,i$, $1 - \varepsilon^3\,T\,i$, and $1 - \varepsilon^3\,T\,i$ plus terms of order ε^6. As we have said, the maxima and minima at $(0, 0, \pm L, 0, 0, \pm L)$ give rise to elliptic periodic solutions, but since the minimax critical points at $(0, 0, \pm L, 0, 0, \mp L)$ have not been shown to be parametrically stable, we cannot conclude at this point that they give rise to elliptic periodic solutions. The deeper analysis of the next subsection is needed to decide the stability of those periodic solutions arising from the minimax critical points.

4.3. Linear stability and the twist condition. The aim of this section is the analysis of the linear stability of the families of periodic solutions established before, using the methods given in section 2. We also check that the twist condition needed for the possible existence of invariant tori is too degenerate. Finally, we also deal with the nonlinear stability of the four critical points of $S^2 \times S^2$. We start with the points related to the periodic near-rectilinear solutions.

4.3.1. Points $(0, 0, \pm L, 0, 0, \mp L)$. After moving the origin to the point of interest through

$$a_1 = \bar{a}_1, \quad a_2 = \bar{a}_2, \quad a_3 = \bar{a}_3 \pm L, \quad b_1 = \bar{b}_1, \quad b_2 = \bar{b}_2, \quad b_3 = \bar{b}_3 \mp L,$$

we introduce the local transformation

$$Q_1 = \frac{\bar{a}_2}{\sqrt{\pm\bar{a}_3 + 2L}}, \qquad Q_2 = \frac{\bar{b}_2}{\sqrt{\mp\bar{b}_3 + 2L}},$$
$$P_1 = \mp\frac{\bar{a}_1}{\sqrt{\pm\bar{a}_3 + 2L}}, \qquad P_2 = \pm\frac{\bar{b}_1}{\sqrt{\mp\bar{b}_3 + 2L}},$$

with inverse

$$\bar{a}_1 = \mp P_1\sqrt{2L - P_1^2 - Q_1^2}, \qquad \bar{a}_2 = Q_1\sqrt{2L - P_1^2 - Q_1^2}, \qquad \bar{a}_3 = \mp(P_1^2 + Q_1^2),$$
$$\bar{b}_1 = \pm P_2\sqrt{2L - P_2^2 - Q_2^2}, \qquad \bar{b}_2 = Q_2\sqrt{2L - P_2^2 - Q_2^2}, \qquad \bar{b}_3 = \pm(P_2^2 + Q_2^2).$$

The variables (Q_1, Q_2, P_1, P_2) are a canonical set for which Q_1, Q_2 can be interpreted as coordinates, whereas P_1 and P_2 represent their associated momenta, respectively.

The resulting Hamiltonian is obtained after putting $\bar{\mathcal{H}}$ in terms of Q_i and P_i and dropping constant terms. We get

$$\begin{aligned}\bar{\mathcal{H}} \;=\;& \pm\tfrac{1}{2}(P_1^2 + Q_1^2) \mp \tfrac{1}{2}(P_2^2 + Q_2^2) - \tfrac{3}{4}\varepsilon^3\,\mu\,L^2\left(3L\,(P_1^2 + P_2^2) + L\,(Q_1^2 + Q_2^2)\right.\\ &+ (2P_1\,P_2 + Q_1\,Q_2)\,\sqrt{2\,L - P_1^2 - Q_1^2}\,\sqrt{2\,L - P_2^2 - Q_2^2}\\ &\left.- (P_2^2 + Q_1^2)\,(P_2^2 + Q_2^2) - P_1^2\,(P_1^2 + P_2^2 + Q_1^2 + Q_2^2)\right) + \cdots.\end{aligned}$$

The Hamiltonian $\bar{\mathcal{H}}$ is valid in a neighborhood of the points $(0, 0, \pm L, 0, 0, \mp L)$.

Next we scale variables through the change $\bar{Q}_j = \varepsilon^{-3/2}\,Q_j$ and $\bar{P}_j = \varepsilon^{-3/2}\,P_j$ for $j \in \{1, 2\}$. To make the change canonical we must divide $\bar{\mathcal{H}}$ by ε^3. Expanding this Hamiltonian in powers of ε (and keeping the same name for it), we arrive at the Hamiltonian

$$\begin{aligned}\bar{\mathcal{H}} \;=\;& \pm\tfrac{1}{2}(\bar{P}_1^2 + \bar{Q}_1^2) \mp \tfrac{1}{2}(\bar{P}_2^2 + \bar{Q}_2^2) - \tfrac{3}{4}\varepsilon^3\,\mu\,L^3\left(3(\bar{P}_1^2 + \bar{P}_2^2) + 4\bar{P}_1\,\bar{P}_2 + \bar{Q}_1^2 + \bar{Q}_2^2 + 2\bar{Q}_1\,\bar{Q}_2\right)\\ &+ \tfrac{3}{8}\varepsilon^6\,\mu\,L^2\left(2(\bar{P}_1^4 + \bar{P}_1^3\,\bar{P}_2 + \bar{P}_1^2\,\bar{P}_2^2 + \bar{P}_1\,\bar{P}_2^3 + \bar{P}_2^4) + 2\bar{P}_2\,(\bar{P}_1 + \bar{P}_2)\,\bar{Q}_1^2\right.\\ &\left.+ (\bar{P}_1^2 + \bar{P}_2^2)\,\bar{Q}_1\,\bar{Q}_2 + 2(\bar{P}_1^2 + \bar{P}_1\,\bar{P}_2 + \bar{P}_2^2)\,\bar{Q}_2^2 + \bar{Q}_1\,\bar{Q}_2\,(\bar{Q}_1 + \bar{Q}_2)^2\right) + \cdots.\end{aligned}$$

The eigenvalues associated with the linear differential equation given through the quadratic part of $\bar{\mathcal{H}}$ are the expressions

$$(23)\quad \pm\sqrt{1 + 20\bar{\varepsilon}^2 + 2\sqrt{5}\bar{\varepsilon}\sqrt{3 + 20\bar{\varepsilon}^2}}\,i = \pm\omega_1\,i, \quad \pm\sqrt{1 + 20\bar{\varepsilon}^2 - 2\sqrt{5}\bar{\varepsilon}\sqrt{3 + 20\bar{\varepsilon}^2}}\,i = \pm\omega_2\,i,$$

where $\bar{\varepsilon}$ stands for $\frac{3}{4}\varepsilon^3\,\mu\,L^3$ and $\omega_1 > 1 > \omega_2 > 0$. Note that $\omega_1 = \omega_2 = 1$ when $\varepsilon = 0$, and the quadratic part of $\bar{\mathcal{H}}$ is in 1-1 resonance. However, we now see that when $\varepsilon \neq 0$ the eigenvalues are distinct.

These equilibria are parametrically stable and correspond to elliptic periodic solutions.

We keep ε small but positive so that we may perform further normalization. By doing so, both ω_1 and ω_2 remain close to 1 but different from it. As the corresponding set of eigenvectors forms a basis of $\mathbb{R}^4$, the quadratic part of $\bar{\mathcal{H}}$ may be brought into normal form through a canonical change of variables. This linear change has to be applied to $\bar{\mathcal{H}}$. The columns of the transformation matrix are the eigenvectors related to $\pm\omega_1 i$ and $\pm\omega_2 i$ multiplied by scale constants chosen to make the change symplectic. We do not give the explicit expression for this change because it is lengthy and the procedure is standard; see, for instance, [10, 30]. Defining the new variables by (q_1, q_2, p_1, p_2) and using the same name for the Hamiltonian, its quadratic part becomes

$$\pm\omega_1\, i\, q_1\, p_1 \mp \omega_2\, i\, q_2\, p_2.$$

Next we introduce action-angle variables $(I_1, I_2, \varphi_1, \varphi_2)$ by means of

$$\begin{aligned} q_1 &= \sqrt{I_1/\omega_1}\,(\cos\varphi_1 - i\sin\varphi_1), & q_2 &= \sqrt{I_2/\omega_2}\,(\cos\varphi_2 - i\sin\varphi_2),\\ p_1 &= \sqrt{\omega_1\, I_1}\,(\sin\varphi_1 - i\cos\varphi_1), & p_2 &= \sqrt{\omega_2\, I_2}\,(\sin\varphi_2 - i\cos\varphi_2). \end{aligned}$$

It is easy to check that $dq_1 \wedge dp_1 + dq_2 \wedge dp_2 = dI_1 \wedge d\varphi_1 + dI_2 \wedge d\varphi_2$. This transformation brings the quadratic terms of $\bar{\mathcal{H}}$ to $\pm\omega_1\, I_1 \mp \omega_2\, I_2$, while its quartic terms are converted into a finite Fourier series in φ_1 and φ_2 whose coefficients are homogeneous quadratic polynomials in I_1 and I_2. We do not give the Hamiltonian because it is enormous.

Now we average $\bar{\mathcal{H}}$ over φ_1 and φ_2, arriving in both cases at

$$\begin{aligned} \bar{\mathcal{H}} \;=\; & \pm\omega_1\, I_1 \mp \omega_2\, I_2 + \frac{(7\omega_1^6 + 13\omega_1^4 + 13\omega_1^2 + 3)(\omega_1^2 - 1)^2}{30\mu\, L^4\, \omega_1^2\, (\omega_1^2 + 2)^2\, (2\omega_1^2 + 1)}\, I_1^2 \\ & + \frac{2(\omega_1^2 - 1)^2\, (\omega_1^4 - 14\omega_1^2 - 5)\, (2\omega_2^2 + 1)}{135\mu\, L^4\, \omega_1\, (\omega_1^2 + 2)^2\, \omega_2}\, I_1\, I_2 \\ & + \frac{(7\omega_2^6 + 13\omega_2^4 + 13\omega_2^2 + 3)(\omega_2^2 - 1)^2}{30\mu\, L^4\, \omega_2^2\, (\omega_2^2 + 2)^2\, (2\omega_2^2 + 1)}\, I_2^2 + \cdots . \end{aligned}$$

The coefficients of I_1^2, I_2^2 and I_1, I_2 may be expressed in terms of $\bar{\varepsilon}$, and, after expanding them in powers of $\bar{\varepsilon}$ about 0, one obtains a formula starting in $\bar{\varepsilon}^2$. The generating function responsible for this averaging step is too big to be reproduced here, but it is a finite Fourier series in the angles φ_1 and φ_2.

Now we can compute the determinant of the Hessian associated with $\bar{\mathcal{H}}$. Using the constraint which relates ω_1 and ω_2 through (23) given by

$$\omega_2 = \sqrt{\frac{4 - \omega_1^2}{2\omega_1^2 + 1}},$$

we get

$$\det\begin{bmatrix} \dfrac{\partial^2\bar{\mathcal{H}}}{\partial I_1^2} & \dfrac{\partial^2\bar{\mathcal{H}}}{\partial I_1\partial I_2} \\ \dfrac{\partial^2\bar{\mathcal{H}}}{\partial I_2\partial I_1} & \dfrac{\partial^2\bar{\mathcal{H}}}{\partial I_2^2} \end{bmatrix} = \frac{(\omega_1^2-1)^6\,(7\omega_1^8-28\omega_1^6-534\omega_1^4-604\omega_1^2-137)}{225\mu^2\,L^8\,\omega_1^2\,(\omega_1^2-4)\,(\omega_1^2+2)^4\,(2\omega_1^2+1)^2}+\cdots,$$

which does not vanish since the (positive) real roots of the determinant occur for $\omega_1 = 1$ or $\omega_1 = 3.37369\ldots$, but as ε does not vanish, ω_1 remains greater than 1. Unfortunately, Theorem 2.5 does not apply since the twist condition is at a higher order in ε.

This suggests, but does not prove, that there are families of invariant 3*-tori around these periodic solutions.*

We leave the question of the existence of invariant KAM tori about these periodic solutions to a future paper. However, we can say something about the stability of the equilibria $(0,0,\pm L,0,0,\mp L)$ of the reduced system on the base space.

For the analysis of the stability of these equilibria we use Arnold's theorem [44]. We fix ε small and positive for this analysis. We need to find $\bar{\mathcal{H}}_4$, i.e., the quartic terms of $\bar{\mathcal{H}}$, and then compute

$$\begin{aligned}\bar{\mathcal{H}}_4(-\omega_2,\omega_1) &= \det\begin{bmatrix} \dfrac{\partial^2\bar{\mathcal{H}}_4}{\partial I_1^2} & \dfrac{\partial^2\bar{\mathcal{H}}_4}{\partial I_1\partial I_2} & \omega_1 \\ \dfrac{\partial^2\bar{\mathcal{H}}_4}{\partial I_2\partial I_1} & \dfrac{\partial^2\bar{\mathcal{H}}_4}{\partial I_2^2} & \omega_2 \\ \omega_1 & \omega_2 & 0 \end{bmatrix} \\ &= \frac{(\omega_1^2-1)^2\,(\omega_1^{12}-16\,\omega_1^{10}+66\,\omega_1^8-268\,\omega_1^6-275\,\omega_1^4-132\,\omega_1^2-24)}{15\,\mu\,L^4\,\omega_1^2\,(\omega_1^2-4)\,(2\,\omega_1^4+5\,\omega_1^2+2)^2}.\end{aligned}$$

Since this term does not vanish for ω_1 close to (but larger than) 1, Arnold's theorem applies, and so the following statement holds.

The equilibrium points $(0,0,\pm L,0,0,\mp L)$ *are stable on the reduced space* $S^2\times S^2$.

Now, if higher-order terms are included in the Hamiltonian (20), we see that the equilibria $(0,0,\pm L,0,0,\mp L)$ are distorted a bit. Specifically, the terms factorized by ε^8 are

$$\begin{aligned}&\tfrac{5}{64}\,\varepsilon^8\,\mu\,(1-\mu)^{1/3}\,e^2\,L^6\,(\cos g\cos h - c\sin g\sin h)\\ &\quad\times\Big(-18-31\,e^2+5\,c^2\,(6+e^2)+5\,(1-c^2)\,(6+e^2)\cos(2\,h)\\ &\qquad+35\,e^2\cos(2\,g)\,(1-c^2+(1+c^2)\cos(2\,h))-70\,c\,e^2\sin(2\,g)\sin(2\,h)\Big).\end{aligned}\tag{24}$$

Thus, after incorporating terms of order ε^8, the equilibria are transformed to

$$\begin{aligned}&\Big(\pm\tfrac{105}{16}\,\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6,\quad 0,\quad \pm\tfrac{\sqrt{256\,L^2-11025\,\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{12}}}{16},\\ &\ \pm\tfrac{105}{16}\,\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6,\quad 0,\quad \mp\tfrac{\sqrt{256\,L^2-11025\,\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{12}}}{16}\Big).\end{aligned}$$

Hence, it is not difficult to deduce that these equilibria correspond to near-rectilinear solutions whose eccentricity is given by $e = 1-\frac{11025}{512}\,\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{10}+\cdots$. The magnitude of their

angular momentum vector is $G = \frac{105}{16}\,\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6 + \cdots$, and its third component is $N = -\frac{33075}{512}\,\varepsilon^{13}\,\mu^3\,(1-\mu)^{2/3}\,L^{14} + \cdots$. This implies that the periodic solutions associated with $(0,0,\pm L,0,0,\mp L)$ are indeed elliptic periodic solutions such that their projections in configuration space yield elliptic orbits with eccentricity close to 1 and inclination angles given by $\pm\cos^{-1}(-\frac{315}{32}\,\varepsilon^8\,\mu^2\,(1-\mu)^{1/3}\,L^8 + \cdots)$.

Thus, these equilibria correspond to elliptic periodic orbits close to rectilinear orbits.

Moreover, it can be proved that the projection of the periodic orbits onto configuration space lies in the plane defined by x_2 and x_3. More precisely, in (the averaged) Cartesian variables $x_1, x_2, x_3, y_1, y_2, y_3$, the coordinates of these periodic orbits up to terms of order ε^{10} are

$$\left(0, \pm\frac{105}{8y_3}\,\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6, \mp\frac{2}{y_3^2} \pm \frac{33075}{512}\,\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{12}, 0, -\frac{105}{32}\,\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6\,y_3^2, y_3\right).$$

We remark that y_3 acts as the parameter of the periodic solution.

Next, the points related to the periodic near-circular equatorial solutions are analyzed.

4.3.2. Points $(0,0,\pm L,0,0,\pm L)$. We first move the Hamiltonian to the origin by

$$a_1 = \bar{a}_1, \quad a_2 = \bar{a}_2, \quad a_3 = \bar{a}_3 \pm L, \quad b_1 = \bar{b}_1, \quad b_2 = \bar{b}_2, \quad b_3 = \bar{b}_3 \pm L;$$

then we change variables by

$$Q_1 = \frac{\bar{a}_2}{\sqrt{\pm\bar{a}_3 + 2L}}, \qquad Q_2 = \frac{\bar{b}_2}{\sqrt{\pm\bar{b}_3 + 2L}},$$
$$P_1 = \mp\frac{\bar{a}_1}{\sqrt{\pm\bar{a}_3 + 2L}}, \qquad P_2 = \mp\frac{\bar{b}_1}{\sqrt{\pm\bar{b}_3 + 2L}},$$

with inverse

$$\bar{a}_1 = \mp P_1\sqrt{2L - P_1^2 - Q_1^2}, \qquad \bar{a}_2 = Q_1\sqrt{2L - P_1^2 - Q_1^2}, \qquad \bar{a}_3 = \mp(P_1^2 + Q_1^2),$$
$$\bar{b}_1 = \mp P_2\sqrt{2L - P_2^2 - Q_2^2}, \qquad \bar{b}_2 = Q_2\sqrt{2L - P_2^2 - Q_2^2}, \qquad \bar{b}_3 = \mp(P_2^2 + Q_2^2).$$

The change of variables is canonical, with Q_1 and Q_2 as coordinates and P_1 and P_2 as their associated momenta.

The resulting Hamiltonian is obtained after writing $\bar{\mathcal{H}}$ in terms of Q_i and P_i and dropping constant terms, so

$$\begin{aligned}\bar{\mathcal{H}} \;=\;& \pm\tfrac{1}{2}(P_1^2 + Q_1^2) \pm \tfrac{1}{2}(P_2^2 + Q_2^2) - \tfrac{3}{4}\varepsilon^3\,\mu\,L^2\,\big(L\,(P_1^2 + P_2^2) - L(Q_1^2 + Q_2^2)\\ &- (2P_1\,P_2 - Q_1\,Q_2)\,\sqrt{2\,L - P_1^2 - Q_1^2}\,\sqrt{2\,L - P_2^2 - Q_2^2}\\ &- (P_2^2 - Q_1^2)\,(P_2^2 + Q_2^2) - P_1^2\,(P_1^2 - P_2^2 + Q_1^2 - Q_2^2)\big) + \cdots.\end{aligned}$$

The Hamiltonian $\bar{\mathcal{H}}$ is valid in a neighborhood of $(0,0,\pm L,0,0,\mp L)$.

Now we scale by $\bar{Q}_j = \varepsilon^{-3/2}\, Q_j$ and $\bar{P}_j = \varepsilon^{-3/2}\, P_j$ for $j \in \{1,2\}$. The canonical structure is preserved by dividing $\bar{\mathcal{H}}$ by ε^3. After expansion of this Hamiltonian in powers of ε we obtain

$$\begin{aligned}\bar{\mathcal{H}} \;=\;& \pm\tfrac{1}{2}(\bar{P}_1^2+\bar{Q}_1^2) \pm \tfrac{1}{2}(\bar{P}_2^2+\bar{Q}_2^2) - \tfrac{3}{4}\varepsilon^3\,\mu\,L^3\left(\bar{P}_1^2+\bar{P}_2^2-4\bar{P}_1\,\bar{P}_2-\bar{Q}_1^2-\bar{Q}_2^2+2\bar{Q}_1\,\bar{Q}_2\right)\\ &+\tfrac{3}{8}\varepsilon^6\,\mu\,L^2\left(2(\bar{P}_1^4-\bar{P}_1^3\,\bar{P}_2-\bar{P}_1^2\,\bar{P}_2^2-\bar{P}_1\,\bar{P}_2^3+\bar{P}_2^4)+(\bar{P}_1^2+\bar{P}_2^2)\,\bar{Q}_1\,\bar{Q}_2\right.\\ &\left.+\,2(\bar{P}_1^2-\bar{P}_1\,\bar{P}_2+\bar{P}_2^2)\,(\bar{Q}_1^2-\bar{Q}_2^2)+\bar{Q}_1\,\bar{Q}_2\,(\bar{Q}_1-\bar{Q}_2)^2\right)+\cdots.\end{aligned}$$

For $(0,0,L,0,0,L)$ the eigenvalues associated with the linear differential equation given through the quadratic part of $\bar{\mathcal{H}}$ are

$$\pm\sqrt{1+2\bar{\varepsilon}}\,i = \pm\omega_1\,i, \quad \pm\sqrt{1-2\bar{\varepsilon}-24\bar{\varepsilon}^2}\,i = \pm\omega_2\,i \tag{25}$$

with $\bar{\varepsilon} = \frac{3}{4}\varepsilon^3\,\mu\,L^3$ and $\omega_1 > 1 > \omega_2 > 0$. For the point $(0,0,-L,0,0,-L)$ the eigenvalues are

$$\pm\sqrt{1-2\bar{\varepsilon}}\,i = \pm\omega_1\,i, \quad \pm\sqrt{1+2\bar{\varepsilon}-24\bar{\varepsilon}^2}\,i = \pm\omega_2\,i. \tag{26}$$

In this case $\omega_2 > 1 > \omega_1 > 0$. We remark that if $\varepsilon = 0$, then $\omega_1 = \omega_2 = 1$; thus the quadratic part of $\bar{\mathcal{H}}$ is in 1-1 resonance. So we keep ε small but positive so that we can apply KAM theory. As a consequence, ω_1 and ω_2 are close to 1 but different from it.

The eigenvectors related to ω_1 and ω_2 form a basis of $\mathbb{R}^4$; thus the quadratic part of $\bar{\mathcal{H}}$ is brought into normal form through a canonical change of variables. This linear change must be applied to $\bar{\mathcal{H}}$. The columns of the matrix are the eigenvectors scaled so that the matrix is symplectic. After defining the new variables by (q_1,q_2,p_1,p_2), the quadratic part of $\bar{\mathcal{H}}$ becomes

$$\pm\omega_1\,i\,q_1\,p_1 \pm \omega_2\,i\,q_2\,p_2.$$

The values of the frequencies ω_1 and ω_2 are given in (25) if the quadratic part is $\omega_1\,i\,q_1\,p_1 + \omega_2\,i\,q_2\,p_2$, whereas if the quadratic part is $-\omega_1\,i\,q_1\,p_1 - \omega_2\,i\,q_2\,p_2$, we take the frequencies from (26). From now on when we refer to $(0,0,L,0,0,L)$ we assume that ω_1 and ω_2 are as in (25), and when we study the point $(0,0,-L,0,0,-L)$ we take the frequencies from (26).

We have

$$\begin{aligned} q_1 &= \sqrt{I_1/\omega_1}\,(\cos\varphi_1 - i\sin\varphi_1), & q_2 &= \sqrt{I_2/\omega_2}\,(\cos\varphi_2 - i\sin\varphi_2),\\ p_1 &= \sqrt{\omega_1\,I_1}\,(\sin\varphi_1 - i\cos\varphi_1), & p_2 &= \sqrt{\omega_2\,I_2}\,(\sin\varphi_2 - i\cos\varphi_2),\end{aligned}$$

and the change satisfies $dq_1\wedge dp_1 + dq_2\wedge dp_2 = dI_1\wedge d\varphi_1 + dI_2\wedge d\varphi_2$. This transforms the quadratic terms of $\bar{\mathcal{H}}$ into $\pm\omega_1\,I_1 \pm \omega_2\,I_2$, while the quartic terms are converted into a finite Fourier series in φ_1 and φ_2 whose coefficients are homogeneous quadratic polynomials in I_1 and I_2.

Now we average $\bar{\mathcal{H}}$ over φ_1 and φ_2. For the two equilibria we obtain

$$\begin{aligned}\bar{\mathcal{H}} \;=\;& \omega_1\,I_1 + \omega_2\,I_2 - \frac{(\omega_1^2-1)^2\,(\omega_1^2+3)}{24\mu\,L^4\,\omega_1^2}\,I_1^2 - \frac{(\omega_1^2-1)^2\,(21\omega_1^2-13)}{6\mu\,L^4\,\omega_1\,\omega_2}\,I_1\,I_2\\ &- \frac{(6\omega_1^2-5)^2\,(48\omega_1^4+62\omega_1^2-93)}{1728\mu\,L^4\,\omega_2^2}\,I_2^2+\cdots.\end{aligned}$$

In both cases the coefficients of I_1^2, I_2^2 and I_1, I_2 may be expressed in terms of $\bar{\varepsilon}$, and expanding them in powers of $\bar{\varepsilon}$ around 0 yields expressions starting in $\bar{\varepsilon}^2$. The generating functions computed in the averaging process in the two cases are enormous, but they are finite Fourier series in the angles φ_1 and φ_2.

At this point we can compute the determinants of the Hessian associated with $\bar{\mathcal{H}}$. First we calculate the constraint relating ω_1 to ω_2 through $\bar{\varepsilon}$ using (25) or (26), obtaining in both situations

$$\omega_2 = \sqrt{(2\omega_1^2 - 1)\,(-3\omega_1^2 + 4)}.$$

We end up with the same expression for the points $(0,0,L,0,0,L)$ and $(0,0,-L,0,0,-L)$, which is

$$\det \begin{bmatrix} \dfrac{\partial^2 \bar{\mathcal{H}}}{\partial I_1^2} & \dfrac{\partial^2 \bar{\mathcal{H}}}{\partial I_1 \partial I_2} \\ \dfrac{\partial^2 \bar{\mathcal{H}}}{\partial I_2 \partial I_1} & \dfrac{\partial^2 \bar{\mathcal{H}}}{\partial I_2^2} \end{bmatrix} = \frac{(\omega_1^2 - 1)^4\,(24\omega_1^6 - 1811\omega_1^4 + 1918\omega_1^2 - 403)}{144\mu^2\,L^8\,\omega_1^2\,\omega_2^2} + \cdots.$$

The determinant vanishes when $\omega_1 \in \{0.536925\ldots, 0.88488\ldots, 1, 8.62479\ldots\}$. However, ω_1 is near 1 (either above or below, but it never reaches this value as ε cannot be zero). Again Theorem 2.5 does not apply since the twist occurs at too high an order in ε.

This suggests, but does not prove, that there are families of invariant 3*-tori around these periodic solutions.*

Again, we leave the question of the existence of invariant KAM tori about these periodic solutions to a future paper. However, we can easily say something about the stability of the equilibria $(0,0,\pm L,0,0,\pm L)$ of the reduced system on the base space. Since the Hamiltonian $\bar{\mathcal{H}}$ is positive or negative definite at these points, the classical theorem already known to Dirichlet [20, 36] applies.

The equilibrium points $(0,0,\pm L,0,0,\pm L)$ *are stable on the reduced space* $S^2 \times S^2$.

Finally, we prove that the near-circular equatorial periodic solutions are indeed equatorial but not circular periodic solutions. We start by taking into account the terms of the averaged Hamiltonian given through (24). Hence, the equilibria $(0,0,\pm L,0,0,\pm L)$ are refined, yielding

$$\Big(\mp\tfrac{15}{16}\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6,\quad 0,\quad \pm\tfrac{\sqrt{256\,L^2 - 225\,\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{12}}}{16},$$
$$\pm\tfrac{15}{16}\,\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^6,\quad 0,\quad \pm\tfrac{\sqrt{256\,L^2 - 225\,\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{12}}}{16}\Big).$$

As a consequence of the above, the given equilibria no longer correspond to circular solutions, because their eccentricity is $e = \frac{15}{16}\varepsilon^5\,\mu\,(1-\mu)^{1/3}\,L^5 + \cdots$. The magnitude of their angular momentum vector is $G = L - \frac{225}{512}\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{11} + \cdots$, and the third component of angular momentum is $N = \pm L \mp \frac{225}{512}\varepsilon^{10}\,\mu^2\,(1-\mu)^{2/3}\,L^{11} + \cdots$. This means that the periodic solutions associated with $(0,0,\pm L,0,0,\pm L)$ are indeed elliptic periodic solutions whose projections in configuration space yield elliptic orbits with eccentricity close to zero; the inclination for the solution related to $(0,0,L,0,0,L)$ is zero, whereas it is π for the periodic solution related to $(0,0,-L,0,0,-L)$. This proves that up to terms of order ε^8 the periodic solutions are near-circular periodic solutions of equatorial type.

Thus, these equilibria correspond to elliptic periodic orbits remaining in the same plane as the two primaries.

Acknowledgments. We thank Professor Chris McCord (Northern Illinois University) for his remarks about the Lusternik–Schnirelmann category. We also thank Professor Sebastián Ferrer (Universidad de Murcia) for his fruitful comments on the relative equilibria in $S^2 \times S^2$.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, Advanced Book Program, Benjamin/Cummings, Reading, MA, 1978.
[2] R. F. Arenstorf, *A new method of perturbation theory and its application to the satellite problem of celestial mechanics*, J. Reine Angew. Math., 221 (1966), pp. 113–145.
[3] R. F. Arenstorf, *New periodic solutions of the plane three-body problem corresponding to elliptic motion in the lunar theory*, J. Differential Equations, 4 (1968), pp. 202–256.
[4] R. F. Arenstorf, *Periodic solutions of circular-elliptic type in the planar N-body problem*, Celestial Mech., 17 (1978), pp. 331–355.
[5] V. I. Arnold, V. V. Kozlov, and A. I. Neuishtadt, eds., *Dynamical Systems* III, Encyclopedia of Mathematical Sciences 3, Springer-Verlag, Berlin, 1988.
[6] E. A. Belbruno, *A new family of periodic orbits for the restricted problem*, Celestial Mech., 25 (1981), pp. 195–217.
[7] P. Benevieri, A. Gavioli, and M. Villarini, *Existence of periodic orbits for vector fields via Fuller index and the averaging method*, Electron. J. Differential Equations, 128 (2004).
[8] G. D. Birkhoff, *Dynamical Systems*, 2nd ed., Amer. Math. Soc. Colloq. Publ. 9, AMS, Providence, RI, 1966.
[9] D. Brouwer and G. M. Clemence, *Methods of Celestial Mechanics*, Academic Press, New York, 1961.
[10] R. C. Churchill and M. Kummer, *A unified approach to linear and nonlinear normal forms for Hamiltonian systems*, J. Symbolic Comput., 27 (1999), pp. 49–131.
[11] S. L. Coffey, A. Deprit, and B. R. Miller, *The critical inclination in artificial satellite theory*, Celestial Mech., 39 (1986), pp. 365–406.
[12] C. Conley, *On some new long periodic solutions of the plane restricted three-body problem*, Comm. Pure Appl. Math., 16 (1963), pp. 449–467.
[13] C. Conley, personal communication, 1968.
[14] J. M. Cors, C. Pinyol, and J. Soler, *Analytic continuation in the case of non-regular dependency on a small parameter with an application to celestial mechanics*, J. Differential Equations, 219 (2005), pp. 1–19.
[15] R. Cushman, *Reduction, Brouwer's Hamiltonian, and the critical inclination*, Celestial Mech., 31 (1983), pp. 401–429.
[16] R. Cushman and D. L. Rod, *Reduction of the semisimple* 1:1 *resonance*, Phys. D, 6 (1982), pp. 105–112.
[17] A. Deprit, *Canonical transformations depending on a small parameter*, Celestial Mech., 1 (1969), pp. 12–30.
[18] A. Deprit, *The elimination of the parallax in satellite theory*, Celestial Mech., 24 (1981), pp. 111–153.
[19] A. Deprit, *Delaunay normalizations*, Celestial Mech., 26 (1982), pp. 9–21.
[20] G. L. Dirichlet, *Über die stabilität des gleichgewichts*, J. Reine Angew. Math., 32 (1846), pp. 85–88.
[21] A. Elipe and S. Ferrer, *Bifurcations in the generalized van der Waals interaction: The polar case ($m = 0$)*, in Hamiltonian Dynamical Systems: History, Theory, and Applications, H. Dumas, K. Meyer, and D. Schmidt, eds., Springer-Verlag, New York, 1995, pp. 137–145.
[22] W. B. Gordon, *On the relation between period and energy in periodic dynamical systems*, J. Math. Mech., 19 (1969), pp. 111–114.
[23] P. Hartman, *Ordinary Differential Equations*, John Wiley & Sons, New York, 1964.
[24] A. Hatcher, *Algebraic Topology*, Cambridge University Press, Cambridge, UK, 2002.
[25] R. C. Howison and K. R. Meyer, *Doubly symmetric periodic solutions of the spatial restricted three-body problem*, J. Differential Equations, 163 (2000), pp. 174–197.

[26] I. M. James, *On category, in the sense of Lusternik-Schnirelmann*, Topology, 17 (1978), pp. 331–348.
[27] M. Kummer, *On the three-dimensional lunar problem and other perturbation problems of the Kepler problem*, J. Math. Anal. Appl., 93 (1983), pp. 142–194.
[28] M. Kummer, *Reduction in the rotating Kepler problem and related topics*, in Hamiltonian Dynamics and Celestial Mechanics, Contemp. Math. 198, AMS, Providence, RI, 1996, pp. 155–179.
[29] V. Lanchares and A. Elipe, *Bifurcations in biparametric quadratic potentials*, Chaos, 5 (1995), pp. 367–373.
[30] A. J. Laub and K. R. Meyer, *Canonical forms for symplectic and Hamiltonian matrices*, Celestial Mech., 9 (1974), pp. 213–238.
[31] L. Lusternik and L. Schnirelmann, *Méthodes Topologiques dans les Problèmes Variationnels*, Hermann, Paris, 1934.
[32] J. Marsden and A. Weinstein, *Reduction of symplectic manifolds with symmetry*, Rep. Math. Phys., 5 (1974), pp. 121–130.
[33] K. R. Meyer, *Symmetries and integrals in mechanics*, in Dynamical Systems, M. M. Peixoto, ed., Academic Press, New York, 1973, pp. 259–272.
[34] K. R. Meyer, *Hamiltonian systems with a discrete symmetry*, J. Differential Equations, 41 (1981), pp. 228–238.
[35] K. R. Meyer, *Periodic Solutions of the N-Body Problem*, Lecture Notes in Math. 1719, Springer-Verlag, New York, 1999.
[36] K. R. Meyer and G. R. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Appl. Math. Sci. 90, Springer-Verlag, New York, 1992.
[37] J. Milnor, *Morse Theory*, Ann. of Math. Stud. 51, Princeton University Press, Princeton, NJ, 1963.
[38] J. Milnor and J. D. Stasheff, *Characteristic Classes*, Ann. of Math. Stud. 76, Princeton University Press, Princeton, NJ, 1974.
[39] J. Moser, *Regularization of Kepler's problem and the averaging method on a manifold*, Comm. Pure Appl. Math., 23 (1970), pp. 609–636.
[40] J. Moser and E. J. Zehnder, *Notes on Dynamical Systems*, Courant Lecture Notes in Mathematics 12, New York University, Courant Institute of Mathematical Sciences, New York; AMS, Providence, RI, 2005.
[41] J. F. Palacián, *Normal forms for perturbed Keplerian systems*, J. Differential Equations, 180 (2002), pp. 471–519.
[42] J. F. Palacián and P. Yanguas, *Invariant manifolds of spatial restricted three-body problems: The lunar case*, in New Advances in Celestial Mechanics and Hamiltonian Systems, J. Delgado, E. A. Lacomba, J. Llibre, and E. Pérez-Chavela, eds., Kluwer/Plenum, New York, 2004.
[43] G. Reeb, *Sur certaines propriétés topologiques des trajectoires des systèmes dynamiques*, Acad. Roy. Belgique. Cl. Sci. Mém. Coll. in 8°, 27 (1952).
[44] C. L. Siegel and J. K. Moser, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, 1971.
[45] A. Weinstein, *Symplectic V-manifolds, periodic orbits of Hamiltonian systems, and the volume of certain Riemannian manifolds*, Comm. Pure Appl. Math., 30 (1977), pp. 265–271.
[46] A. Weinstein, *Bifurcations and Hamilton's principle*, Math. Z., 159 (1978), pp. 235–248.
[47] V. A. Yakubovich and V. M. Starzhinskii, *Linear Differential Equations with Periodic Coefficients* 1, 2, John Wiley & Sons, New York, 1975.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 341–360
© 2008 Society for Industrial and Applied Mathematics

Basis Markov Partitions and Transition Matrices for Stochastic Systems*

Erik Bollt†, Paweł Góra‡, Andrzej Ostruszka§, and Karol Życzkowski¶

Abstract. We analyze discrete-time dynamical systems subjected to an additive noise and their deterministic limit. In this work, we will introduce a notion by which a discrete-time stochastic system has something like a Markov partition for deterministic systems. For a chosen class of noise profiles, the Frobenius–Perron (FP) operator associated to the noisy system is exactly represented by a stochastic transition matrix of a finite size K. This feature allows us to introduce for these stochastic systems a basis Markov partition, defined herein, irrespectively of whether the deterministic system possesses a Markov partition or not. We show that in the deterministic limit, corresponding to $K \to \infty$, the sequence of invariant measures of the noisy systems tends, in the weak sense, to the invariant measure of the deterministic system. Thus, by introducing a small additive noise one may approximate transition matrices and invariant measures of deterministic dynamical systems.

Key words. stochastic dynamics, Markov partition, Frobenius–Perron operator, transition matrix

AMS subject classifications. 37H10, 37A05, 37M99, 37A30, 37A60, 15A99, 93E03

DOI. 10.1137/070686111

1. Introduction.

Markov partitions for deterministic dynamical systems serve a central role for determining their symbolic dynamics [3, 4, 5] whose grammar is described by a finite sized transition matrix that generates a so-called sofic shift [6, 14]. The conditions for such a projection were defined by Bowen for Anosov hyperbolic systems [3, 4] and stated succinctly for interval maps as a partition whose elements are each a homeomorphism onto a finite union of its elements [3, 5]. We remark here that a defining property in both cases is that the set of characteristic functions defined over the elements of the Markov partition project the transfer operator exactly onto an operator of finite type; that is, a matrix results, whereas an infinite matrix would be expected for a non-Markov system. We argue here that this should be the defining property of any generalization of Markov partitions, that is, a set of basis functions which project the Frobenius–Perron (FP) operator exactly onto a finite-rank matrix with no

*Received by the editors March 22, 2007; accepted for publication (in revised form) by W. Beyn January 4, 2008; published electronically April 23, 2008.
http://www.siam.org/journals/siads/7-2/68611.html

†Department of Mathematics & Computer Science and Department of Physics, Clarkson University, Potsdam, NY 13699-5805 (bolltem@clarkson.edu). This author was supported by the National Science Foundation under grant DMS-0404778.

‡Department of Mathematics and Statistics, Concordia University, 1455 de Maisonneuve Boulevard West, Montreal, Quebec H3G 1M8, Canada (pgora@mathstat.concordia.ca). This author was supported by a Canadian NSERC grant.

§Institute of Physics, Jagiellonian University, ul. Reymonta 4, 30–059 Kraków, Poland (ostruszk@if.uj.edu.pl).

¶Institute of Physics, Jagiellonian University, ul. Reymonta 4, 30–059 Kraków, Poland, and Center for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/44, 02–668 Warszawa, Poland (karol@tatry.if.uj.edu.pl). This author acknowledges the partial support of grant 1 P03B 042 26 of the Polish Ministry of Science and Information Technology and the Marie Curie Actions Transfer of Knowledge project COCOS (contract MTKD-CT-2004-517186).

residual. We present results here explicitly for random dynamical systems of the interval $[0,1]$ and generalizations to the L-torus $[0,1]^L$.

In physical literature on dynamical systems one often distinguishes a "natural" invariant measure of a hyperbolic system, which is stable with respect to an external noise [2, 7]. In mathematics this measure is known as the Sinai–Ruelle–Bowen (SRB) measure, and under certain assumptions one may rigorously prove its uniqueness [10]. Although the overall idea that adding the noise improves the convergence to the SRB measure is well known in the physics community, this work attempts to provide a more solid mathematical framework for this statement. In particular, for a certain class of the noise profiles, we are in position to characterize this convergence quantitatively.

First we recall the FP operator for a deterministic transformation. Associated with a discrete dynamical system acting on initial conditions, $\mathbf{x} \in M$ (say, a manifold $M \subset \Re^n$),

$$\begin{aligned} \tau &: M \to M, \\ x &\mapsto \tau(x), \end{aligned} \tag{1}$$

is another dynamical system over $L^1(M)$, the space of densities of ensembles of initial conditions

$$\begin{aligned} P_\tau &: L^1(M) \to L^1(M), \\ \rho(x) &\mapsto [P_\tau \rho](x). \end{aligned} \tag{2}$$

This FP operator (P_τ) is defined through a continuity equation [16],

$$\int_{\tau^{-1}(B)} \rho(x)dx = \int_B [P_\tau \rho](x)dx, \tag{3}$$

where B is a measurable subset of M, while PDF $\rho(x)$ belongs to $L^1(M)$. Equation (3) may be formally rewritten using the Dirac delta function:

$$[P_\tau \rho](x) = \int_M \delta(x - \tau(y))\rho(y)dy. \tag{4}$$

This heuristic form is particularly suitable for further investigation of dynamical systems with additive noise; see (6).

Now consider the stochastically perturbed dynamical system

$$\begin{aligned} \tau_\nu &: M \to M, \\ x &\mapsto \tau(x) + \xi, \end{aligned} \tag{5}$$

where ξ is a random variable with PDF ν, which is applied once per each iteration. We assume that the realizations ξ_n of ξ added to subsequent iterations form an identical independently distributed (i.i.d.) sequence. The random part ξ is assumed to be independent of state x which we tacitly assume to be relatively small, so that the deterministic part τ has primary influence. The "stochastic FP operator" has a form similar to that of the deterministic case [16],

$$[P_{\tau_\nu} \rho](x) = \int_M \nu(x - \tau(y))\rho(y)dy, \tag{6}$$

where the deterministic kernel, the delta function in (2), now becomes a stochastic kernel describing the PDF of the noise perturbation. We will denote the stochastic FP operator as $P_{\mathcal{P}}$ below. In the case that the random map (5) arises from the usual continuous Langevin process, the infinitesimal generator of the FP operator for normal ν corresponds to a general solution of a Fokker–Planck equation [16]. The FP operator formalism is particularly convenient in that it allows for an arbitrary noise distribution ν to be incorporated in a direct and simple way. Within the formalism, we can also study multiplicative noise ($x \to \eta\tau(x)$, modeling parametric noise). The kernel-type integral transfer operator is $\mathcal{K}(x,y) = \nu(x/\tau(y))/\tau(y)$ for $x \in \Re^+$, which can then also be finitely approximated as described in the next section and usefully reordered to a canonical block reduced form. In more generality, the theory of random dynamical systems [1] clearly classifies those random systems which give rise to explicit transfer operators with corresponding infinitesimal generators, and there are well-defined connections between the theories of random dynamical systems and of stochastic differential equations.

The main aim of this work is to investigate a class of stochastically perturbed dynamical systems for which the FP operator is represented by a finite stochastic transition matrix of size N. Such dynamical systems will be called *basis Markov* in analogy to deterministic dynamical systems possessing a Markov partition [15, 25], for which the associated FP operator is finite. The deterministic limit of the stochastic system corresponds to the divergence of the matrix size. In this limit, $N \to \infty$, the sequence of invariant measures of the stochastic systems acting in the N-dimensional Hilbert space converges, in the weak sense, to the invariant measure of the corresponding deterministic system.

The paper is organized as follows. The Ulam–Galerkin method of approximating the infinite dimensional FP operator and the concept of the Markov partition for a deterministic system are reviewed in sections 2 and 3, respectively. In section 4 we introduce the notion of basis Markov stochastic systems, while in section 5 we analyze a particular example of random systems perturbed by an additive noise with cosine profile. In section 6 we construct a fairly general example of the transition densities satisfying our assumptions (20). The key result on convergence of the invariant measures for stochastic and deterministic systems is proved in section 7. A discussion of isospectral matrices used to describe the FP operator is relegated to the appendix.

2. Ulam–Galerkin's method: Approximating the infinite dimensional operator. A Galerkin method may be used to approximate the FP operator by a Markov operator of finite rank. Formally, projection of the infinite dimensional linear space $L^1(M)$ results from discretely indexed basis functions $\{\phi_i(x)\}_{i=1}^{\infty} \subset L^1(M)$ onto a finite dimensional linear subspace generated by a subset of the basis functions,

$$\Delta_N = \mathrm{Span}(\{\phi_i(x)\}_{i=1}^{N}). \tag{7}$$

This projection,

$$p : L^1(M) \to \Delta_N, \tag{8}$$

is realized optimally by the Galerkin method in terms of the inner product, which we choose to be integration:

$$(f,g) \equiv \int_M f(x)g(x)dx \quad \forall f,g \in L^2(M). \tag{9}$$

Specifically, the infinite dimensional "matrix" is approximated by the $N \times N$ matrix,

$$A_{i,j} = ([P_\tau \phi_i], \phi_j) = \int_M [P_\tau \phi_i](x)\phi_j(x)dx, \quad 1 \le i,j \le N. \tag{10}$$

One approximates $\rho(x)$ through a finite linear combination of basis functions:

$$\rho(x) \simeq \sum_{i=1}^{N} d_i \phi_i(x). \tag{11}$$

The historically famous Ulam method [17, 27] for deterministic dynamical systems is equivalent to the interpretation for finding the fraction of the box B_i which maps by τ to B_j; the Ulam matrix is equivalent to the Galerkin matrix by using (10) and choosing the basis functions to be the family of characteristic functions

$$\phi_i(x) = \mathbf{1}_{B_i}(x) = \begin{cases} 1 & \text{if } x \in B_i, \\ 0 & \text{else.} \end{cases} \tag{12}$$

Specifically, we choose the ordered set of basis functions to be in terms of a nested refinement of boxes $\{B_i\}$ covering M. Though Galerkin's and Ulam's methods are formally equivalent in the deterministic case, we are of the opinion that the Galerkin description is a more natural description in the stochastic setting.

3. Markov partitions of deterministic systems, and exact projection. In this section, we discuss that a Markov partition is special for the FP operator of a deterministic dynamical system in that characteristic functions supported over those partition elements lead to an exact projection of the FP operator onto an operator of finite rank, a matrix.

For a one-dimensional transformation of the interval, the definition of a Markov partition [24] (see also [15, 25]) can be found in more recent references [3, 10, 18].

Definition. *A map of the interval $\tau : [a,b] \to [a,b]$ is* Markov *if there is a finite partition $\{I_j\}_{j=1}^N$ such that*

1. $\cup_{j=1}^N I_j = [a,b]$ *(covering property),*
2. $\mathrm{int}(I_j) \cap \mathrm{int}(I_k) = \emptyset$ *if $j \neq k$ (no overlap property),*
3. $\tau(I_j) = \cup_{i=1}^{q(j)} I_{k_i^{(j)}}$ *for some $k_i^{(j)} \in \{1,2,\ldots,N\}$, $i = 1,2,\ldots,q(j)$ (a grid interval maps completely across a union of intervals without "dangling ends" property).*

It is not hard to show that the set of characteristic functions forms a finite basis set of functions

$$\{\phi_j(x)\} = \{\mathbf{1}_{I_j}(x)\}_{j=1}^N, \tag{13}$$

such that Galerkin projection (10) is exactly onto an operator of finite rank or a matrix $A_{i,j}$. That is, (10) simplifies to

$$\begin{aligned} A_{i,j} = ([P_\tau \phi_i], \phi_j) &= \int_M [P_\tau \phi_i](x)\phi_j(x)dx, \\ &= \int_M \int_M \delta(x - \tau(y))\phi_i(y)\phi_j(x)dydx \\ &= \int_{I_j} \int_{I_i} \delta(x - \tau(y))dydx, \quad 1 \le i,j \le N. \end{aligned} \tag{14}$$

If the map τ is in addition piecewise linear on its Markov partition, then $P_\tau[\phi_i(x)]$ is a linear combination of $\phi_j(x)$.

Similarly, there is a well-defined notion of an Anosov diffeomorphism with a Markov partition [3, 4, 9, 23], and so for such systems, it can be shown that characteristic functions supported over the corresponding Markov partition create a basis set such that (10) results in an operator of finite rank.

We take these observations as motivation for the following definition, which is meant to generalize the notion of a Markov partition to stochastic systems.

Definition. *Let $\{M, \mathcal{B}, \mu\}$ be a measure space and a transformation $\tau : M \to M$ be measurable. Then the transformation τ is* "basis Markov" *if there exists a finite set of basis functions $\{\phi_i(x)\}_{i=1}^N : M \to [0,1] \in L^1(M)$ such that the FP operator is operationally closed within Δ_N, where $\Delta_N = \mathrm{Span}(\{\phi_i(x)\}_{i=1}^N)$. That is, for any density $\rho \in \Delta_N$, its image $[P_\tau \rho](x)$ belongs to Δ_N.*

Remark 1. If a transformation τ is basis Markov, then, if we perform Galerkin's method, $A_{i,j} = (P_\tau[\phi_i], \phi_j)_{N\times N}$, with that basis set, it allows that, for any initial density which can be written as a linear combination of these basis functions,

$$\rho_0(x) = \sum_{i=1}^{N} c_i \phi_i(x), \tag{15}$$

or stated simply,

$$\rho_0(x) \in \Delta_N. \tag{16}$$

The action of the FP operator on such initial densities, $\rho_1(x) = P_\tau[\rho_0(x)]$, has the matrix presentation,

$$\mathbf{c}' = A \cdot \mathbf{c}, \text{ where } \rho_1(x) = \sum_{i=1}^{N} c_i' \phi_i(x), \tag{17}$$

and is well known as a linear operator from an N-dimensional vector space into itself. This emphasizes that the FP operator projects exactly to an operator of finite rank—a matrix.

Note that for a general finite set of functions, if we take a general linear combination of those functions and then apply the FP operator, we do not expect that the resulting density can be written as a (finite) linear combination of basis functions.

The following is a direct consequence of our definition of basis Markov in relationship to the usual definition of a Markov map, stating the sense in which basis Markov is a generalization.

Remark 2. Equation (14) implies that any piecewise linear Markov map, together with the characteristic functions supported over the partition elements, is basis Markov.

4. Basis Markov stochastic systems: A general case due to separable noise. We analyze a dynamical system τ defined on an interval $M = [0,1]$ with both ends identified and subjected to a specific form of the additive noise,

$$x' = \tau(x) + \xi. \tag{18}$$

To specify the special case of the stochastic dynamical system written in (5), the stochastic perturbation will be characterized by the transition density $\mathcal{P}(x,y)$ of a transition from point x to y induced by noise. Describing the dynamics in terms of a probability density $\rho(x)$ its one-step evolution is governed by the stochastic FP operator,

$$\rho'(x) = [P_{\mathcal{P}}\rho](x) \;=\; \int_M \mathcal{P}\big(\tau(y),x\big)\,\rho(y)dy. \tag{19}$$

We will denote this stochastic FP operator by the symbol $P_{\mathcal{P}}$, referring to (6) in all that follows. The operator $P_{\mathcal{P}}$ acts on every probability density defined on M, and, in general, it cannot be represented by a finite matrix. However, in what follows we shall analyze a certain class of noise profiles for which such a representation is possible.

Definition. *The stochastic system of equations* (19) *is called* basis Markov *if there exists a finite set of basis functions* $\{\phi_i(x)\}_{i=1}^N : M \to [0,1] \in L^1(M)$ *such that the FP operator* $P_{\mathcal{P}}$ *is operationally closed within* Δ_N*, where* $\Delta_N = \mathrm{Span}(\{\phi_i(x)\}_{i=1}^N)$.

We assume that the transition probability, $\mathcal{P}(x,y) \geq 0$, satisfies the following properties [21, 22]:

$$\begin{aligned} &\text{(a)} \quad \mathcal{P}(x,y) \equiv \mathcal{P}(x-y) = \mathcal{P}(\xi),\\ &\text{(b)} \quad \mathcal{P}(x,y) \equiv \mathcal{P}(x \bmod 1, y \bmod 1),\\ &\text{(c)} \quad \mathcal{P}(x,y) = \sum_{m,n=0}^{N} A_{mn}\,u_n(x)\,v_m(y) \end{aligned} \tag{20}$$

for $x,y \in \mathbb{R}$ and some finite N. Property (a) ensures that the distribution of the random variable ξ does not depend on the position x, while the periodicity condition is provided in (b). A noise profile fulfilling property (c) is called *separable* (decomposable), and it allows us to represent the dynamics of an arbitrary system with such a noise in a finite dimensional Hilbert space. Here $A = (A_{mn})_{m,n=0,\ldots,N}$ is a yet undetermined real matrix of expansion coefficients. Note that A characterizes the noise and does not depend on the deterministic dynamics τ. We assume that the functions u_n, $n = 0,\ldots,N$, and v_m, $m = 0,\ldots,N$, are continuous in $M = [0,1)$ and linearly independent, and we can express $f \equiv 1$ as their linear combinations. Both sets of functions span bases in an $N+1$ Hilbert space. Their orthogonality is not required.

This term *separable noise* is concocted in analogy to *separable states* in quantum mechanics and *separable* probability distributions, since such a property was called $N+1$-separability by Tucci [26]. Making use of this crucial feature of the noise profile we may expand the kernel of the FP operator (19):

$$\rho'(y) = [P_{\mathcal{P}}\rho](y) = \int_0^1 \sum_{m,n=0}^{N} A_{mn}u_n(\tau(x))v_m(y)\rho(x)dx \tag{21}$$

$$\begin{aligned} &= \sum_{m,n=0}^{N} A_{mn}\left[\int_0^1 u_n(\tau(x))\rho(x)dx\right] v_m(y)\\ &= \sum_{n=0}^{N}\left[\int_0^1 u_n(\tau(x))\rho(x)dx\right]\tilde{v}_n(y) \end{aligned} \tag{22}$$

for $y \in M$, where

$$\tilde{v}_n = \sum_{m=0}^{N} A_{mn} v_m. \tag{23}$$

Thus, any initial density is projected by the FP operator $P_{\mathcal{P}}$ into the vector space spanned by the functions $\tilde{v}_m$, $m = 0, \ldots, N$.

Assuming that a given density $\rho(x)$ belongs to this space, we can expand it in this basis,

$$\rho(x) = \sum_{m=0}^{N} q_m \, \tilde{v}_m(x). \tag{24}$$

Expanding ρ' in an analogous way, we will describe it by the vector $\vec{q}' = \{q'_0, \ldots, q'_N\}$.

Let B denote a matrix of integrals,

$$B_{nm} = \int_0^1 u_n(\tau(x)) v_m(x) dx, \tag{25}$$

where $n, m = 0, \ldots, N$. Observe that B depends directly on the system τ and on the noise via the basis functions u and v. Making use of this matrix, the one-step dynamics (23) may be rewritten in a matrix form

$$q'_n = \sum_{m=0}^{N} D_{nm} \, q_m, \quad \text{where } D = BA \tag{26}$$

and A is implied by (20). In this way we have arrived at a representation of the FP operator $P_{\mathcal{P}}$ by a matrix D of size $(N+1) \times (N+1)$, the elements of which read

$$D_{nm} = \int_0^1 u_n(\tau(x)) \tilde{v}_m(x) dx, \quad n, m = 0, \ldots, N. \tag{27}$$

With (26), we now see that random dynamical systems with noise satisfying condition (20) allow a finite dimensional subspace which is preserved.

Although the probability is conserved under the action of $P_{\mathcal{P}}$, the matrix D need not be stochastic. This is due to the fact that the functions $\{\tilde{v}_m(x)\}$ forming the expansion basis in (24) were not normalized. We shall then compute their norms,

$$s_m = \int_0^1 \tilde{v}_m(y) dy = \sum_{n=0}^{N} A_{mn} b_n, \tag{28}$$

where

$$b_n = \int_0^1 v_n(y) dy. \tag{29}$$

Let $K \leq N+1$ denote the number of nonzero components of the vector $\vec{s}$, and let $k = 1, \ldots, K$ runs over all indexes $n \in 0, \ldots, N+1$, for which $s_k \neq 0$. Then the rescaled vectors,

$$V_k(y) := \tilde{v}_k(y)/s_k, \tag{30}$$

are normalized:

$$\int_0^1 V_k(y)dy = 1. \tag{31}$$

The normalization condition $\int_0^1 \rho(x)dx = 1$ implies

$$\int_0^1 \sum_{m=0}^{N} q_l \tilde{v}_m(x)dx = \sum_{m=0}^{N} q_m s_m = \sum_{k=1}^{K} q_k s_k = 1. \tag{32}$$

The same is true for the transformed density,

$$\sum_k q'_k s_k = 1. \tag{33}$$

Hence this scalar product is preserved during the time evolution. Making use of the rescaled coefficients

$$c_k := q_k s_k, \tag{34}$$

the dynamics (26) reads

$$c'_k = q'_k s_k = \sum_j D_{kj}\, q_j s_k = \sum_j D_{kj}\, \frac{s_k}{s_j}\, q_j s_j =: \sum_j T_{kj}\, c_j. \tag{35}$$

By construction the coefficients c_k sum to unity. Since some of them can in general be negative, the transition matrix

$$T_{kj} \;\equiv\; D_{kj}\, \frac{s_k}{s_j} = \sum_{ii'} D_{kj}\, \frac{A_{ki} s_i}{A_{ji'} s_{i'}} \tag{36}$$

need not be stochastic. In the above equation, all indices run from 1 to K and the coefficients s_k are nonzero by construction.

It is interesting to distinguish a special class of noises for which all functions corresponding to nonzero values of the components s_k are nonnegative: $\tilde{v}_k(x) \geq 0$ for $x \in [0,1]$ and $k = 1, \ldots, K$. This implies that for any probability density ρ its expansion coefficients q_k in (24) are not negative. Furthermore, the normalization constants of $\tilde{v}_k$ are nonnegative, $s_k > 0$, $k = 1, \ldots K$, and so are the coefficients c_k and c'_k given in (34), (35). Hence vectors c and c' form normalized K-point probability distributions, and so in this case the transition matrix T of size K is stochastic. The dimensionality $K \leq N + 1$ is determined by the parameter N and the choice of the basis functions $\{v_l(x)\}$ entering (20).

5. A special case: Cosine noise. We will now discuss a particularly simple case of the separable noise described above and introduced in [21]. Let

$$\mathcal{P}_N(\xi) = \mathrm{C_N} \cos^N(\pi\xi), \tag{37}$$

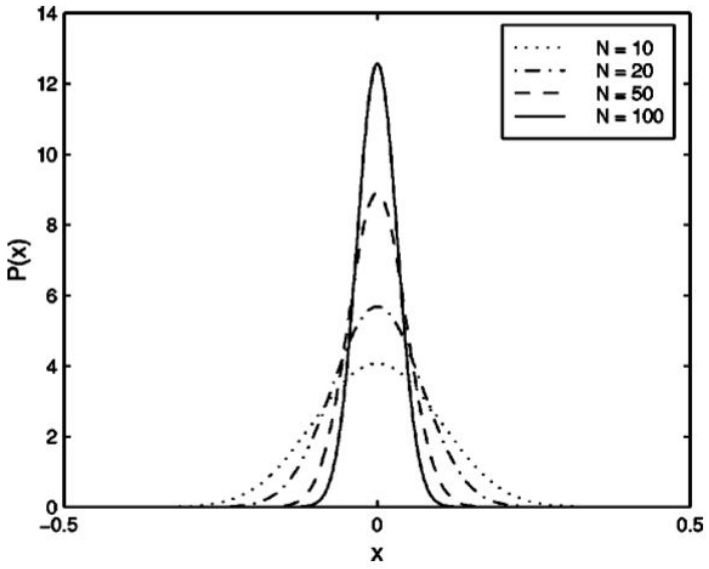


Figure 1. *The cosine noise of (37) closely resembles a normal noise profile, but with finite support. Several values of N are shown, with decreasing standard deviation with increasing N.*

where N is even ($N = 0, 2, \ldots$), and with the normalization constant

$$\mathrm{C_N} = \sqrt{\pi}\,\frac{\Gamma[N/2+1]}{\Gamma[(N+1)/2]}. \tag{38}$$

See Figure 1, in which we can see the decreasing standard deviation with respect to increasing N. This type of noise reminds us of a normal distribution, but of compact support.

The parameter N controls the strength of the noise measured by its variance

$$\sigma^2 = \frac{1}{2\pi^2}\Psi'\left(\frac{N}{2}+1\right) = \frac{1}{12} - \frac{1}{2\pi^2}\sum_{m=1}^{N/1}\frac{1}{m^2}, \tag{39}$$

where Ψ' stands for the derivative of the digamma function.

For the expansion (20) we use basis functions

$$\begin{aligned} u_m(x) &= \cos^m(\pi x)\sin^{N-m}(\pi x),\\ v_n(y) &= \cos^n(\pi y)\sin^{N-n}(\pi y), \end{aligned} \tag{40}$$

where $x \in M$ and $m, n = 0, \ldots, N$. Expanding the cosine as a sum to the Nth power in (37), we find that the $(N+1)\times(N+1)$ matrix A defined by (20) is diagonal:

$$A_{mn} = a_m\delta_{mn}, \text{ with } a_m = \mathrm{C_N}\binom{N}{m}. \tag{41}$$

Integrating trigonometric functions, we find the coefficients

$$b_m = \int_0^1 \cos^m(\pi x)\sin^{N-m}(\pi x)dx = \frac{2}{\pi N}\frac{\Gamma[(m+1)/2]\,\Gamma[(N-m+1)/2]}{\Gamma(N/2)} \tag{42}$$

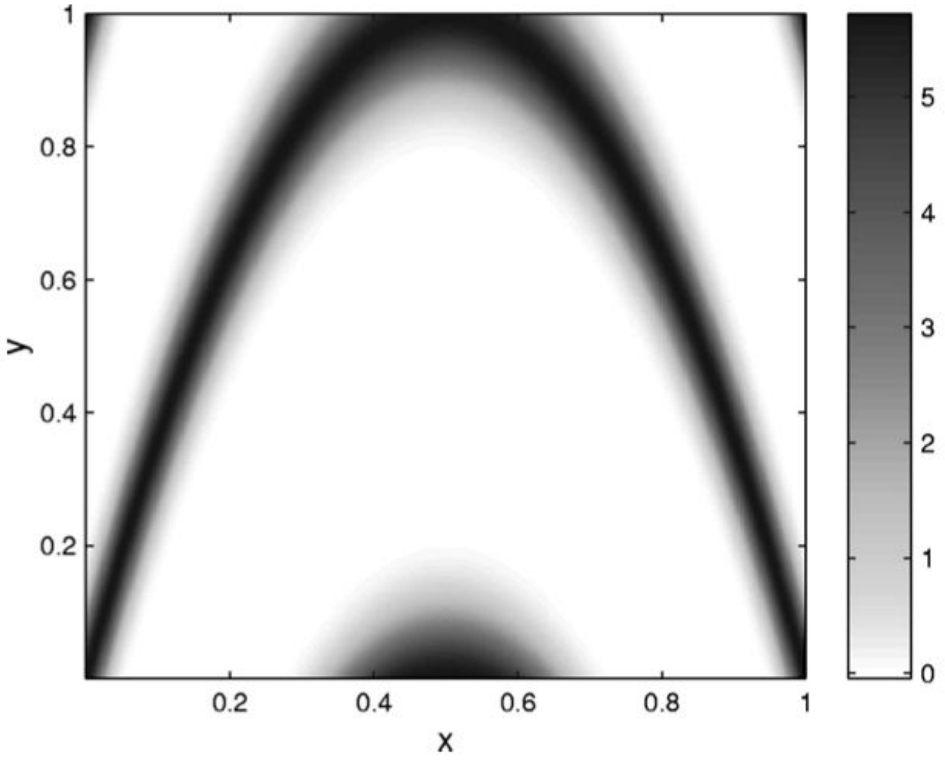


Figure 2. *The transition kernel $\mathcal{P}_N(f(x), y)$ for the logistic map $\tau(x) = 4x(1-x)$, with $N = 20$ and with cosine noise due to $N = 20$; compare to Figure 1. Note the periodicity of x of period 1.*

and

$$s_m = a_m b_m, \tag{43}$$

which are nonzero only for even values of m. Hence the size $K \times K$ of the transition matrix reads

$$K = N/2 + 1, \tag{44}$$

and the expression (36) takes the form

$$T_{kj} = D_{mn} \frac{a_m b_m}{a_n b_n}, \quad \text{where } k, j = 1, \ldots, K,\ m = 2(k-1),\ n = 2(j-1). \tag{45}$$

For the noise (37) discussed here all functions $\tilde{v}_m$ for even m, which contribute to the matrix T, are nonnegative; hence, as discussed in the previous section, the transition matrix T is stochastic. We find in this case that the transition kernel reminds us of a fuzzy but periodically repeated version of the map. See Figure 2. However, the FP operator embeds to a transition matrix T, which "appears" roughly as a different form of the original map; see Figure 3.

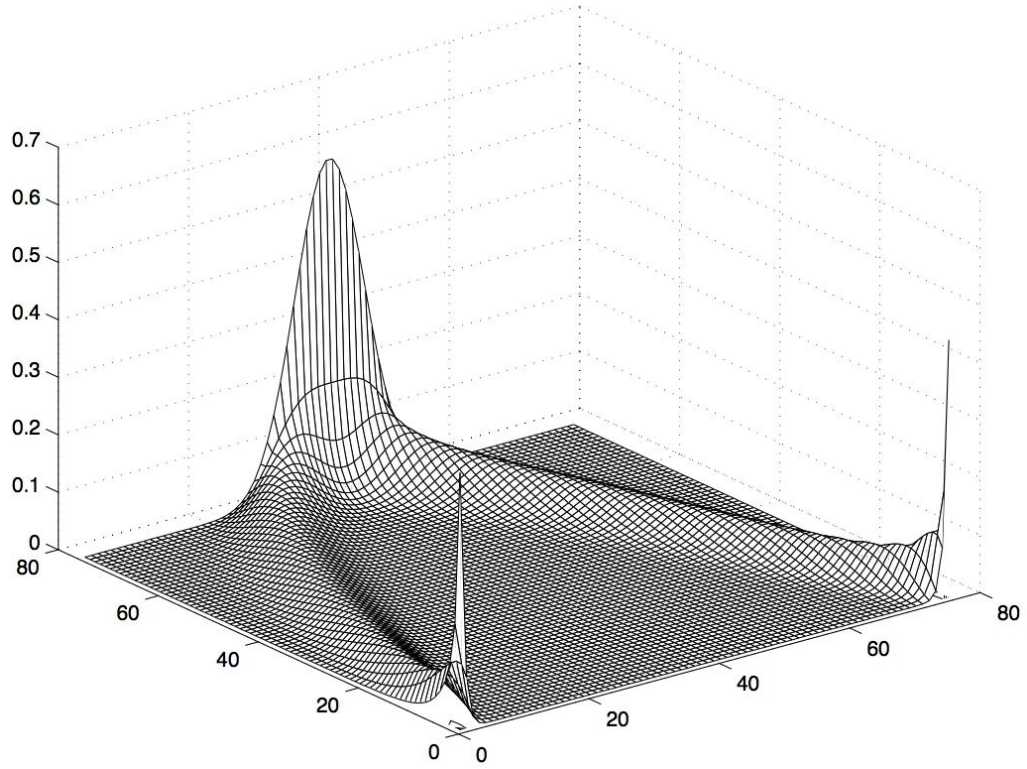


Figure 3. *The stochastic matrix T_{150} shown, from (36), exactly represents the stochastic FP operator of the stochastic tent map (47) with trig noise (37) and basis set (40) using $N = 150$. Note that $T^{(150)}$ is a matrix of size $N/2 + 1 = 150/2 + 1 = 76$ square. Compare to the matrices in (48) of smaller N.*

There is an interesting correspondence between the spectra of eigenvalues of the two matrices D and T. Since T is stochastic, its largest eigenvalue is equal to unity. Moreover, it is the only eigenvalue with modulus one, which follows from the fact that the kernel $\mathcal{P}(x, y)$ vanishes only for $x - y = 1/2 \pmod 1$, and the two-step probability function is everywhere positive:

$$\int_M \mathcal{P}(x, z)\mathcal{P}(z, y)dz > 0 \text{ for } x, y \in M \tag{46}$$

(see [16, Th. 5.7.4]). A particularly useful consequence and simplification is that the eigenstate corresponding to the largest eigenvalue of the matrix represents the invariant density of the system, $\rho_* = P_f(\rho_*)$; this can be easily found numerically by diagonalizing T.

All of the other eigenvalues are included inside the unit circle and their moduli $|\lambda_i|$ characterize the decay rates. It is worth emphasizing that the spectra of both matrix representations of the FP operator, by matrices D of size $(N + 1) \times (N + 1)$ used in [20, 21, 22] and the stochastic T matrices of size $(N/2+1) \times (N/2+1)$ developed here, coincide up to the additional $N/2$ eigenvalues which are equal to zero; see the appendix for details.

For concreteness let us discuss an exemplary one-dimensional dynamical system, a tent map:

$$\tau(x) := \begin{cases} 2x & \text{if} \quad 0 \le x \le 1/2, \\ 2(1-x) & \text{if} \quad 1/2 \le x \le 1. \end{cases} \tag{47}$$

Simple integration allows us to obtain the analytic form of the transition matrix $T^{(N)}$ for the tent map (47) perturbed by additive noise characterized by small values of N,

(48)

$$T^{(2)} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad T^{(4)} = \frac{1}{24}\begin{bmatrix} 11 & 3 & 11 \\ 6 & 6 & 6 \\ 7 & 15 & 7 \end{bmatrix}, \quad T^{(6)} = \frac{1}{320}\begin{bmatrix} 145 & 25 & 25 & 145 \\ 69 & 45 & 45 & 69 \\ 51 & 75 & 75 & 51 \\ 55 & 175 & 175 & 55 \end{bmatrix}.$$

In the simplest case $N = 2$ the transition matrix is bistochastic, but it is not so for larger N. However, for this system, the matrix $T^{(N)}$ is of rank one for arbitrary value of the noise parameter N. The spectrum of T contains one eigenvalue equal to unity and all others equal to zero. This implies that every initial density is projected onto an invariant density already after the first iteration of the map. This is not the case for other dynamical systems τ, including the logistic map $\tau_r(x) = rx(1-x)$, for which the spectrum contains several resonances—eigenvalues of moduli smaller than one—which describe the decaying modes of the system [21].

In this way we have established a relation between a sequence of noisy systems τ_N and the deterministic dynamical system τ. A stochastic system (18) with the noise profile (37) for a fixed noise parameter N is described by a stochastic matrix $T^{(N)}$ of size $K = N/2 + 1$ and acts in the Hilbert space $\mathcal{H}_K$.

We have shown that the sequence of transition matrices $T^{(N)}$ *corresponds* to the dynamical system τ in the sense that the sequence μ_N of the invariant measures of $T^{(N)}$ converges weakly to the τ-invariant measure μ in the deterministic limit $N \to \infty$. Furthermore, for any initial density ρ the sequence of vectors ρ'_N transformed by $P_{\mathcal{P}_N}$ converges weakly to the density transformed by the FP operator associated with τ. Observe that the above property holds not only for one-dimensional systems but also for dynamical system τ in higher dimensional measure spaces.

6. General example. In this section we construct a fairly rich family of transition densities satisfying the assumptions (20). Let $\{g_N\}_{N\ge1}$ be a sequence of C^2 (this condition can be weakened) nonnegative functions with support in $[-1/2, 1/2]$ such that $g_N(-1/2) = g_N(1/2)$ for all $N \ge 1$ and which converges to Dirac's delta δ_0 as $N \to \infty$.

Each g_N, which can be also seen as a 1-periodic function on the whole real line, can be approximated by its partial Fourier sum arbitrarily close in the supremum norm. Let

$$h_N(\xi) = c_{S(N)} + a_{0,N} + 2\sum_{s=1}^{S(N)} \left(a_{s,N}\cos(2s\pi\xi) + b_{s,N}\sin(2s\pi\xi)\right) \tag{49}$$

be an approximation obtained from Fourier approximation by shifting it up by a small constant $c_{S(N)}$ to ensure $h_N \ge 0$ on $[-1/2, 1/2]$. We have $c_{S(N)} \to 0$ as $S(N) \to \infty$. We can make the

functions h_N also converge to Dirac's delta δ_0 as $N \to \infty$. Using the functions h_N we define a family of densities:

$$\mathcal{P}_N(\xi) = h_N(\xi) / \int_{-1/2}^{1/2} h_N(t)dt, \quad N = 1, 2, 3, \ldots, \tag{50}$$

and then a family of transition densities

$$\mathcal{P}_N(x, y) = \mathcal{P}_N(x - y), \quad N = 1, 2, 3, \ldots.$$

Since

$$\begin{aligned}\cos(2s\pi(x-y)) &= \cos(2s\pi x)\cos(2s\pi y) + \sin(2s\pi x)\sin(2s\pi y),\\ \sin(2s\pi(x-y)) &= \sin(2s\pi x)\cos(2s\pi y) - \cos(2s\pi x)\sin(2s\pi y),\end{aligned}$$

it is clear that the assumptions (20)(c) are satisfied with $u_n(x)$ equal to $\cos(2s\pi x)$ or $\sin(2s\pi x)$ and $v_m(y)$ equal to $\cos(2s\pi y)$ or $\sin(2s\pi y)$ for $0 \le s \le S(N)$. It is also clear that for each $x \in [0, 1]$, $\mathcal{P}_N(x, \cdot)$ converges to Dirac's delta δ_x as $N \to \infty$. To have the condition (3) of the next section satisfied it is enough to start with even functions g_N.

Example. Let $g(\xi) = (0.2 + x^2)\exp(-x^2)$ and $g_N(\xi) = Ng(N\xi)$, restricted to $[-1/2, 1/2]$ and extended periodically to the whole real line, $N \ge 1$. Then, the g_N's are positive and converge to Dirac's δ_0 as $N \to \infty$. In particular, let us consider g_6. Its Fourier approximation, with $S(6) = 5$, is

$$\begin{aligned}1.24032 + 1.14838\cos(2\pi\xi) - 0.470309\cos(4\pi\xi) - 0.530699\cos(6\pi\xi)\\ - 0.163161\cos(8\pi\xi) - 0.0225748\cos(10\pi\xi).\end{aligned}$$

We have used such a poor approximation to make the example simpler. We can choose constant $c_{S(6)} = 0$ and after normalization we obtain

$$\begin{aligned}\mathcal{P}_6(\xi) = 1 + 0.92587\cos(2\pi\xi) - 0.37918\cos(4\pi\xi) - 0.42787\cos(6\pi\xi)\\ - 0.131547\cos(8\pi\xi) - 0.01820\cos(10\pi\xi).\end{aligned}$$

See the transition density in Figure 4.

We have

$$\begin{aligned}\mathcal{P}_6(x-y) = 1 &+ 0.92587\cos(2\pi x)\cos(2\pi y) + 0.92587\sin(2\pi x)\sin(2\pi y)\\ &- 0.37918\cos(4\pi x)\cos(4\pi y) - 0.37918\sin(4\pi x)\sin(4\pi y)\\ &- 0.42787\cos(6\pi x)\cos(6\pi y) - 0.42787\sin(6\pi x)\sin(6\pi y)\\ &- 0.131547\cos(8\pi x)\cos(8\pi y) - 0.131547\sin(8\pi x)\sin(8\pi y)\\ &- 0.01820\cos(10\pi x)\cos(10\pi y) - 0.01820\sin(10\pi x)\sin(10\pi y).\end{aligned}$$

See the transition kernel in Figure 5.

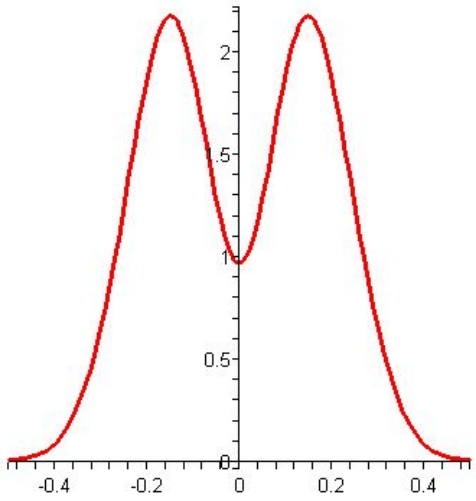


Figure 4. *The transition density* $P_6(\xi)$.

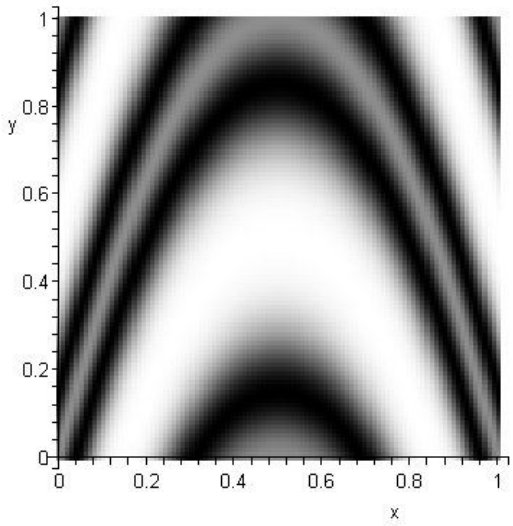


Figure 5. *The transition kernel* $\mathcal{P}_6(\tau(x), y)$ *for the logistic map* $\tau(x) = 4x(1-x)$, *with* $S(6) = 5$.

In the notation of section 4 let us define

$$\begin{aligned}
u_0(x) &= 1, \\
u_{2s+1}(x) &= \cos(2(s+1)s\pi x), \quad s = 0, 1, 2, 3, 4, \\
u_{2s}(x) &= \sin(2s\pi x), \quad s = 1, 2, 3, 4, 5, \\
v_0(y) &= 1, \\
v_{2s+1}(y) &= \cos(2(s+1)\pi y), \quad s = 0, 1, 2, 3, 4, \\
v_{2s}(y) &= \sin(2s\pi y), \quad s = 1, 2, 3, 4, 5.
\end{aligned}$$

Then, matrix $A = (A_{mn})_{0\le m,n\le 10}$ is the diagonal matrix with the diagonal

$$[1, 0.92587, 0.92587, -0.37918, -0.37918, -0.42787, -0.42787, -0.131547, -0.131547, -0.01820, -0.01820],$$

and we have

$$\tilde{v}_m = A_{mm} v_m, \quad m = 0, 1, \dots, 10.$$

Let us consider the dynamics given by the logistic map $\tau : x \mapsto 4x(1-x)$. Matrix D defined in (27) and representing FP operator $P_{\mathcal{P}_6}$ is

(51)
$$D = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2441 & -0.3096 & 0 & -0.08859 & 0 & -0.1209 & 0 & -0.01639 & 0 & -0.0007850 & 0 \\
-0.1717 & 0.2016 & 0 & -0.1313 & 0 & 0.00547 & 0 & 0.01382 & 0 & 0.001494 & 0 \\
0.1752 & -0.1940 & 0 & -0.09397 & 0 & 0.06079 & 0 & 0.02395 & 0 & -0.002347 & 0 \\
-0.1372 & 0.1658 & 0 & -0.0159 & 0 & 0.06947 & 0 & -0.02997 & 0 & -0.004869 & 0 \\
0.1436 & -0.1506 & 0 & -0.07445 & 0 & 0.08111 & 0 & -0.006781 & 0 & -0.002915 & 0 \\
-0.1178 & 0.1397 & 0 & 0.00971 & 0 & 0.02136 & 0 & -0.02983 & 0 & 0.001684 & 0 \\
0.1246 & -0.1268 & 0 & -0.06176 & 0 & 0.07547 & 0 & -0.01636 & 0 & -0.0003029 & 0 \\
-0.1051 & 0.1223 & 0 & 0.01799 & 0 & -0.00139 & 0 & -0.01922 & 0 & 0.002802 & 0 \\
0.1116 & -0.1116 & 0 & -0.05331 & 0 & 0.06756 & 0 & -0.01849 & 0 & 0.001108 & 0 \\
-0.09589 & 0.1098 & 0 & 0.02112 & 0 & -0.01226 & 0 & -0.01162 & 0 & 0.002350 & 0
\end{bmatrix}.$$

The eigenvalues of D are

$$1, -.4086428809, 0.0800412, -0.0117582, 0.00331119, -0.000739709, 0, 0, 0, 0, 0.$$

Although in this case all eigenvalues of D are real, in general they are complex. Since matrix D is real, the eigenvalues are placed symmetrically with respect to the real axis. The eigenvector for eigenvalue 1 is

$$w = [1, 0.164834, -0.154604, 0.139482, -0.107445, 0.116956, -0.0938748, 0.102202, -0.0843296, 0.0918404, -0.0772548],$$

and it provides a rough approximation $\sum_{m=0}^{10} w[m]\tilde{v}_m(\xi)$ to the τ-invariant density. A much better approximation shown in Figure 6 is obtained by taking the same noise profile for $N = 40$ and $S = 30$, which results in matrix D of size $2S + 1 = 61$.

For a comparison we performed Ulam's approximation of the invariant density of τ using an $N \times N$ matrix with $N = 61$. For $\phi_m = N \cdot \mathbf{1}_{[(m-1)/N, m/N]}$, Ulam's probabilistic matrix $U = \{U_{ij}\}$ can be obtained by putting

$$U_{ij} = (1/N) \int_0^1 \phi_i(t)\phi_j(\tau(t))dt, \quad 1 \le i, j \le N.$$

We found the 1-eigenvector w of U, and the function $f_N = \sum_{m=1}^{N} w[m]\phi_m$ is Ulam's approximation to τ-invariant density. It is shown in Figure 7.

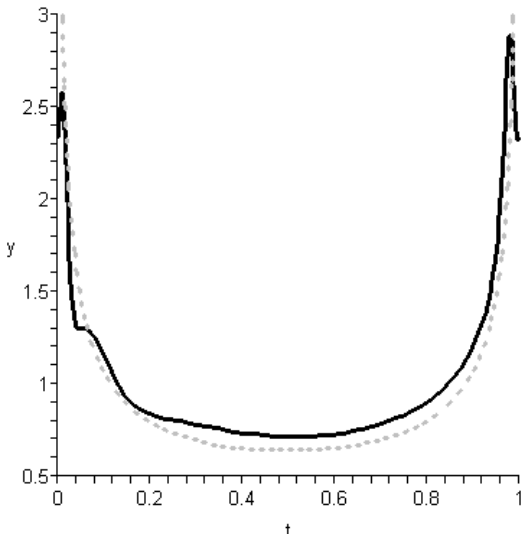


Figure 6. *An approximation to the invariant density of the logistic map (dashed line) obtained as an invariant density of transition matrix D of size 61×61 (solid line).*

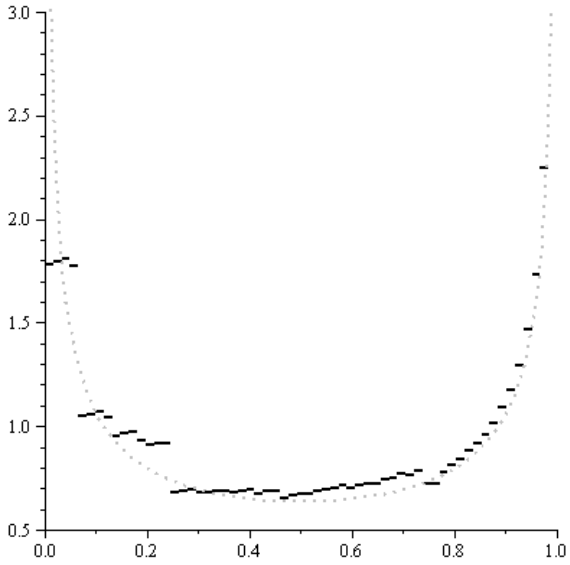


Figure 7. *An approximation to the invariant density of the logistic map (dashed line) obtained by Ulam's method with 61×61 matrix (solid line).*

The L^1 errors of approximation were comparable: 0.17 for Ulam's method and 0.20 for our method. Our method produces a smooth approximating function which is nicer for a smooth invariant density. Our method is also more general in the sense that it can be used to approximate not only the invariant density itself but also the invariant density of a random

perturbation of a map by a possibly very strange perturbing distribution. This was shown in the example above. On the other hand, Ulam's method is definitely simpler and its theoretical properties are well studied.

7. Approximation by basis Markov maps. While not all maps and noise profiles allow for the map to be basis Markov, in this section we will show that a non–basis Markov map may be weakly well approximated by basis Markov maps. In this sense, the finite approximations offered by basis Markov maps can be thought of as a good description of the general behavior, since the invariant measures of the finite approximations due to the basis Markov maps have weak-* limits to the invariant measures of the general maps.

Let us consider a family of the transition probabilities $\mathcal{P}_N(\cdot,\cdot)$ such that, for each $x \in M$, $\mathcal{P}_N(x,\cdot)$ converges to Dirac's delta δ_x as $N \to \infty$.

We require the following assumptions about the transition probabilities $\mathcal{P}_N(\cdot,\cdot)$:

1. $\mathcal{P}_N(\cdot,\cdot)$ is measurable as a function of two variables.
2. For every x we have $\int_M \mathcal{P}_N(x,y)dy = 1$.
3. For every $y \in M$ we have
$$\int_M \mathcal{P}_N(x,y)dx = 1.$$
4. Let $B(x,r) = \{y : |x-y| < r\}$ and
$$p_N(x,r) = \int_{M\setminus B(x,r)} \mathcal{P}_N(x,y)dy. \tag{52}$$
Then, for any $r > 0$,
$$p_N(r) = \sup_{x\in M} p_N(x,r) \to 0 \quad \text{as} \quad N \to +\infty.$$

Assumptions 1–3 are typical for probability measures, while assumption 4 is also rather mild, and it is easy to check that all four assumptions are satisfied by the cosine noise (37).

Under these assumptions, the following can be easily proved.

Proposition. *Let $M = [0,1]$. For any $\rho \in L^1(M)$ we have*
$$\int_M \rho(x)\mathcal{P}_N(x,y)dx \to \rho(y) \text{ as } N \to \infty \tag{53}$$
in $L^1(M)$.

In Theorem 1 below we assume that the transformation $\tau : [0,1] \to [0,1]$ is continuous. This assumption can be weakened (say, to piecewise continuous) if we impose additional restrictions on the transition probabilities $\mathcal{P}_N$ (say, such that all measures μ_N and their weak limits are continuous measures; see, for example, [8]).

Theorem 1. *Let the transformation τ be continuous. Under the assumptions 1, 2, and 4, it follows that if μ_N is an invariant measure of the stochastic perturbation of transformation f defined by the transition probability $\mathcal{P}_N$, then every weak-$*$ limit point of the set $\{\mu_N : N \geq 1\}$ is an f-invariant measure.*

This theorem can be proved following the ideas of Khasminskii [12].

A more precise result can be proved under more restrictive assumptions on the transformation τ.

Theorem 2. *Let the transformation τ be piecewise C^2 and piecewise expanding, i.e., $|\tau'| > 2$, where it exists. Then, under the assumptions 1–4, every weak-$*$ limit point of the set $\{\mu_N : N \geq 1\}$ is a τ-invariant absolutely continuous measure.*

This result was proved in Theorem I.B. of [8]. The perturbations we consider are of "convolution type" and since we treat an interval as a circle an extra factor of 2 does not occur. The example of the famous W-map [11] shows that the condition $|\tau'| > 2$ cannot be weakened.

8. Concluding remarks. In this work we have introduced the concept of basis Markov stochastic systems, for which the associated FP operator is finite. This property resembles the class of deterministic systems with a Markov partition. However, the Markov partition is characteristic to a very special class of deterministic systems, while the basis Markov property is related to the kind of stochastic perturbation. It holds for any deterministic system τ, subjected to an additive noise with a profile satisfying the separability condition (20). In this way such a random dynamical system can be described by a stochastic transition matrix of a finite size K, which diverges in the deterministic limit.

We have shown an intimate relationship between the sequence of stochastic matrices which act in the space of K-point probability distributions and the FP operator P_τ of the deterministic system, which acts in the infinite dimensional space: In the deterministic limit $K \to \infty$ the invariant densities of stochastic matrices converge in a weak sense to the invariant measure of the deterministic system τ. Thus, constructing the transition matrices T and decreasing the noise strength (and increasing the dimensionality K), one may construct arbitrary approximations of the FP operator P_τ.

Some discussion regarding generality is in order. While it is not clear at this stage how many families of examples exist that satisfy the properties in (20), we presented a concrete example in section 5, the cosine noise example, (37), with corresponding basis functions (40). We find this example instructive due to its general appearance as similar to the familiar Gaussian distribution and the fact that it provides a finite representation of the FP operator P_τ by a stochastic transition matrix T. Furthermore, in section 6 we presented a general technique of designing one-dimensional noise profiles which satisfy the separability conditions (20).

Note that the described method is not restricted to one-dimensional systems. On the contrary, the entire construction can be directly applied to a general case of multidimensional dynamical systems. In particular, the definition (20)(c) of separable noise profiles works for the case of an L-dimensional system, provided the variables x and y represent vectors with L components each.

If the dynamical system acts on the L-torus, for example, $M = [0,1]^L$, one can take the Cartesian product of the cosine noise (37) setting

$$\mathcal{P}_N(\xi_1, \ldots, \xi_L) = C_N^L \cos^N(\pi\xi_1) \cos^N(\pi\xi_2) \cdots \cos^N(\pi\xi_L), \tag{54}$$

where $\xi_k = x_k - y_k$ and $k = 1, \ldots, L$. This form of the additive noise was used in [20] to analyze a two-dimensional system (a variant of the baker map) and to compare the spectral properties of the FP operator associated with the classical stochastic system with properties

of the propagator of the corresponding quantum evolution. In such a case the deterministic limit of the classical noisy system, $K \to \infty$, is related to the classical limit, $\hbar \to 0$, of the corresponding quantum dynamics.

Note that for basis Markov stochastic systems, the transition matrices T *exactly* describe the action of the dynamical system with additive noise on densities. Thus our construction differs from an approach applied in [13, 19, 28], where a finite dimensional description of the density dynamics of a deterministic system was achieved by truncation of an infinite transition operator P_τ to the finite dimension K. The effect of such a truncation may also be regarded as a kind of noise depending on the matrix size K and the base, in which P_τ is represented. On the other hand, in our case a suitable choice of the noise profile added to the deterministic system distinguishes a relevant basis, in which the FP operator of the perturbed system is finite.

Appendix. Isospectral matrices. In this appendix we show that the matrix D defined by (27) and used in [20, 21, 22] to represent the FP operator and the stochastic transition matrix T share the same nonzero part of the spectrum. We make use of the following algebraic result.

Lemma. *Let A be a square matrix of size $N \times N$ and $\vec{s}$ a vector of length N containing only nonzero entries. Then the matrix*

$$B_{jk} \equiv A_{jk} \frac{s_j}{s_k} \tag{55}$$

has the same spectrum as A.

(There is no summation over repeating indices.)

Proof. To study equation $\det(B - \lambda \mathbb{1}) = 0$ we start analyzing an exemplary term P^B of the determinant. It consists of a product of N elements $B_{i,\sigma(j)}$, where $\sigma(i)$ stands for a certain permutation of the indices. The product of N factors of the type $s_i/s_{\sigma(i)}$ is equal to unity so that

$$P^B_\sigma = \prod_i B_{i,\sigma(i)} = \prod_i B_{i,\sigma(i)} \frac{s_1 s_2 \cdots s_N}{s_1 s_2 \cdots s_N} = \prod_i A_{i,\sigma(i)}. \tag{56}$$

Thus every term contributing to the free coefficient of the characteristic equation will be the same, $P^B_\sigma = P^A_\sigma$; hence these coefficients for both matrices A and B are equal. Since the diagonal elements of both matrices coincide, $B_{jj} = A_{jj}$, all terms forming the coefficients standing by an arbitrary power of λ are the same for both matrices. Therefore, characteristic equations for both matrices are equal and so are their spectra. ■

Treating all nonzero elements of the vector s_k, $k = 1, \ldots, K$, as vector $\vec{s}$, we may apply the lemma to (36) and obtain equivalence of the spectrum of T and the nonzero part of the spectrum of D. Since integrals (25) vanish for odd values of m, every second column of D is equal to zero, and the remaining $N/2$ eigenvalues of D are equal to zero.

Acknowledgment. We would like to thank W. Słomczyński for several fruitful discussions.

REFERENCES

[1] L. Arnold, *Random Dynamical Systems*, Springer-Verlag, New York, 1998.

[2] C. Beck and F. Schlögl, *Thermodynamics of Chaotic Systems*, Cambridge University Press, Cambridge, UK, 1993.

[3] E. Bollt and J. Skufca, *Markov partitions*, in Encyclopedia of Nonlinear Science, A. Scott, ed., Routledge, New York, 2005, pp. 557–559.

[4] R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Springer-Verlag, Berlin, 1975.

[5] A. Boyarsky and P. Góra, *Laws of Chaos: Invariant Measures and Dynamical Systems in One Dimension*, Birkhäuser Boston, Boston, 1997.

[6] R. Fischer, *Sofic systems and graphs*, Monatsh. Math., 80 (1975), pp. 179–186.

[7] P. Gaspard, *Chaos, Scattering and Statistical Mechanics*, Cambridge University Press, Cambridge, UK, 1998.

[8] P. Góra, *On small stochastic perturbations of mappings of the unit interval*, Colloq. Math., 49 (1984), pp. 73–85.

[9] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1991.

[10] A. Katok and B. Hasselblat, *Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.

[11] G. Keller, *Stochastic stability in some chaotic dynamical systems*, Monatsh. Math., 94 (1982), pp. 313–333.

[12] R. Z. Khasminskii, *Ergodic properties of recurrent diffusion processes and stabilization of the solution to the Cauchy problem for parabolic equations*, Theory Probab. Appl., 5 (1960), pp. 179–196.

[13] M. Khodas and S. Fishman, *Relaxation and diffusion for the kicked rotor*, Phys. Rev. Lett., 84 (2000), pp. 2837–2840.

[14] B. P. Kitchens, *Symbolic Dynamics, One-Sided, Two-Sided and Countable State Markov Shifts*, Springer-Verlag, New York, 1998.

[15] K. Krzyżewski, *On connection between expanding mappings and Markov chains*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 19 (1971), pp. 291–293.

[16] A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise*, 2nd ed., Springer-Verlag, New York, 1994.

[17] T.-Y. Li, *Finite approximation for the Frobenius-Perron operator. A solution to Ulam's conjecture*, J. Approximation Theory, 17 (1976), pp. 177–186.

[18] W. de Melo and S. van Strien, *One-Dimensional Dynamics*, Springer-Verlag, New York, 1993.

[19] S. Nonnenmacher, *Spectral properties of noisy classical and quantum propagators*, Nonlinearity, 16 (2003), pp. 1685–1713.

[20] A. Ostruszka, Ch. Manderfeld, K. Życzkowski, and F. Haake, *Quantization of classical maps with tunable Ruelle-Pollicott resonances*, Phys. Rev. E (3), 68 (2003), 056201.

[21] A. Ostruszka, P. Pakoński, W. Słomczyński, and K. Życzkowski, *Dynamical entropy for systems with stochastic perturbations*, Phys. Rev. E (3), 62 (2000), pp. 2018–2029.

[22] A. Ostruszka and K. Życzkowski, *Spectrum of Frobenius-Perron operator for systems with stochastic perturbation*, Phys. Lett. A, 289 (2001), pp. 306–312.

[23] C. Robinson, *Dynamical Systems, Stability, Symbolic Dynamics, and Chaos*, 2nd ed., CRC Press, Boca Raton, FL, 1999.

[24] Ye. A. Sandler, *A direct algorithm for imitation of Markov sequences*, Engrg. Cybernetics, 11 (1973), pp. 355–358.

[25] Ja. G. Sinaĭ, *Markov partitions and U-diffeomorphisms*, Funkcional. Anal. i Priložen, 2 (1968), pp. 64–89 (in Russian).

[26] R. R. Tucci, *Entanglement of Formation and Conditional Information Transmission*, preprint, http://arxiv.org/abs/quant-ph/0010041, 2000.

[27] S. Ulam, *Problems in Modern Mathematics*, Interscience Publishers, New York, 1960.

[28] J. Weber, F. Haake, P. A. Braun, C. Manderfeld, and P. Šeba, *Resonances of the Frobenius-Perron operator for a Hamiltonian map with a mixed phase space*, J. Phys. A, 34 (2001), pp. 7195–7211.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 361–420
© 2008 Society for Industrial and Applied Mathematics

Mixed-Mode Oscillations in Three Time-Scale Systems: A Prototypical Example*

Martin Krupa†, Nikola Popović‡, and Nancy Kopell§

Abstract. Mixed-mode dynamics is a complex type of dynamical behavior that is characterized by a combination of small-amplitude oscillations and large-amplitude excursions. Mixed-mode oscillations (MMOs) have been observed both experimentally and numerically in various prototypical systems in the natural sciences. In the present article, we propose a mathematical model problem which, though analytically simple, exhibits a wide variety of MMO patterns upon variation of a control parameter. One characteristic feature of our model is the presence of three distinct time-scales, provided a singular perturbation parameter is sufficiently small. Using geometric singular perturbation theory and geometric desingularization, we show that the emergence of MMOs in this context is caused by an underlying canard phenomenon. We derive asymptotic formulae for the return map induced by the corresponding flow, which allows us to obtain precise results on the bifurcation (Farey) sequences of the resulting MMO periodic orbits. We prove that the structure of these sequences is determined by the presence of secondary canards. Finally, we perform numerical simulations that show good quantitative agreement with the asymptotics in the relevant parameter regime.

Key words. mixed-mode oscillations, canard mechanism, singular perturbations, three time-scales, geometric desingularization

AMS subject classifications. 34E10, 34A26, 34D15, 34C20

DOI. 10.1137/070688912

1. Introduction.

Mixed-mode dynamics is a complex type of dynamical behavior that is characterized by a combination of small-amplitude oscillations and large-amplitude excursions of relaxation type. *Mixed-mode oscillations* (MMOs) are frequently encountered in multiscale dynamical systems, i.e., in systems of differential equations in which the relevant variables evolve over several distinct scales. Consequently, typical MMO patterns in such systems consist of oscillatory sequences in which amplitudes of different orders of magnitude alternate. Historically, MMOs were first observed in experiments on the well-known Belousov–Zhabotinsky reaction [38]. They have since been found both experimentally and numerically in numerous other contexts in the natural sciences. Examples include prototypical systems from chemical kinetics, electrocardiac dynamics, neuronal modeling, and laser dynamics, as

*Received by the editors April 20, 2007; accepted for publication (in revised form) by B. Ermentrout January 10, 2008; published electronically April 23, 2008.
http://www.siam.org/journals/siads/7-2/68891.html
†Department of Mathematical Sciences, New Mexico State University, P.O. Box 30001 Department 3MB, Las Cruces, NM 88003 (mkrupa@nmsu.edu). The research of this author was supported in part by NSF grant DMS-0406608.
‡School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, United Kingdom (nikola.popovic@ed.ac.uk). The research of this author was supported by NSF grants DMS-0109427 (to N.K.), DMS-0211505 (to N.K.), and DMS-0406608 (to M.K.).
§Center for BioDynamics and Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215 (nk@math.bu.edu). The research of this author was supported in part by NSF grant DMS-0211505.

well as from various other disciplines; see, e.g., [10, 16, 23, 27, 29, 30, 32, 31] for details and references.

Among the various mechanisms which have been proposed to explain the occurrence of MMOs are the break-up of an invariant torus [21] and the loss of stability of a Shilnikov homoclinic orbit [16]. MMOs have also been linked to slow passage through a delayed Hopf bifurcation (cf., e.g., [22]) as well as to the subcritical Hopf-homoclinic bifurcation [12, 13]. In the present article, we consider another explanation for the emergence of MMOs, namely, the so-called *canard mechanism*. To the best of our knowledge, this idea was first brought forward by Milik et al. [27]. More recently, in [2], it was extended to accommodate more general classes of systems that exhibit canard dynamics.

The classical canard phenomenon [1, 5, 8, 9] was first described in the framework of two-dimensional fast-slow systems, i.e., of systems with one fast and one slow variable; a prototypical example is the system of equations given by

$$v' = -z + f_2 v^2 + f_3 v^3, \tag{1.1a}$$

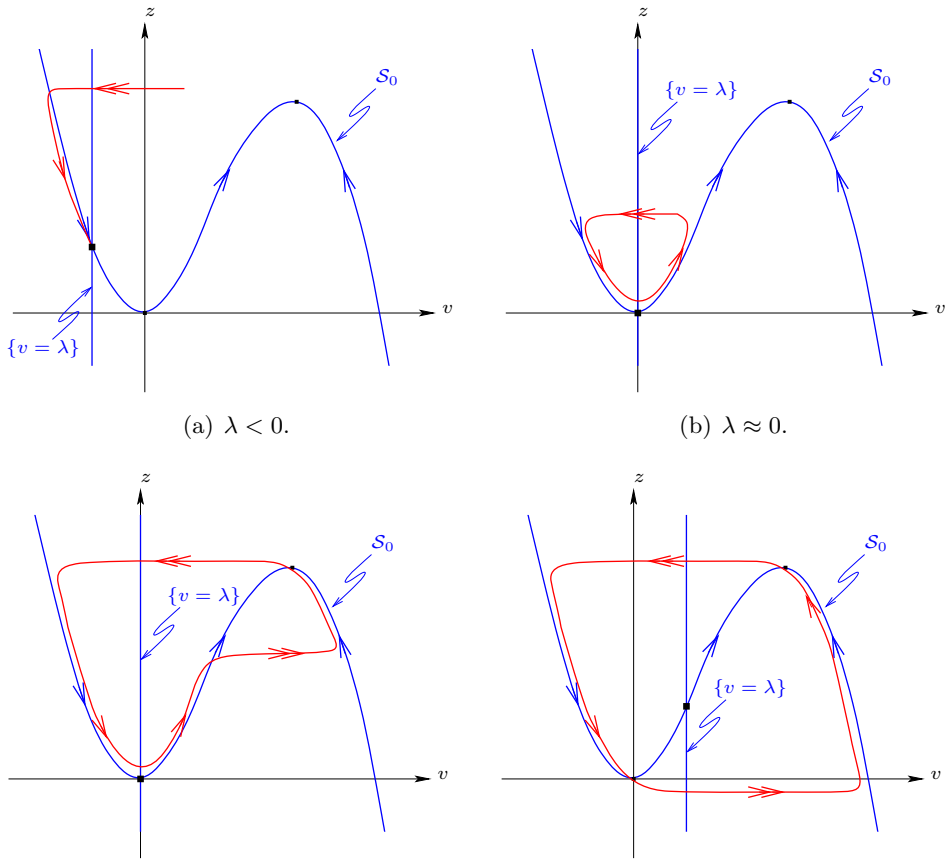
$$z' = \varepsilon(v - \lambda). \tag{1.1b}$$

(Here, $f_2 > 0$ and $f_3 < 0$ are real constants, $0 < \varepsilon \ll 1$ and λ are small parameters, and the prime denotes differentiation with respect to time t.)

The term *canard explosion* [20] is customarily used to denote a transition in (1.1) from a stable equilibrium through a family of small-amplitude cycles and subsequently to a large-amplitude relaxation oscillation. Notably, this transition occurs within an exponentially small range (in ε) of the relevant control parameter, λ. The basic mechanism of a canard explosion can be described as follows: under the above assumptions, the "fast nullcline" $\mathcal{S}_0$ for (1.1), which is given by $z = f(v) := f_2 v^2 + f_3 v^3$, is an S-shaped curve. Moreover, $\mathcal{S}_0$ is a curve of equilibria for the *layer problem* obtained for $\varepsilon = 0$ in (1.1) and is (normally) hyperbolic away from the two *fold points* where $f'(v) = 0$; in particular, the origin is one such point. Rewriting (1.1) in terms of the slow time $\tau = \varepsilon t$, one finds that the corresponding "slow nullcline" is given by $v = \lambda$. As λ passes through 0, this slow nullcline moves through the lower fold point of $\mathcal{S}_0$ at the origin, which triggers the onset of the canard explosion; see Figure 1. Finally, for $\lambda > 0$ sufficiently "large," the dynamics of (1.1) enters the relaxation regime.

One important notion that arises in the study of a canard explosion in (1.1) (as well as in other, related systems) is that of a *maximal canard*. In general, a canard is a solution of (1.1) which originates in the attracting portion of the fast nullcline $\mathcal{S}_0$ and which then crosses over to the repelling one; cf. again Figure 1. Maximal canards are canard trajectories that remain $\mathcal{O}(\varepsilon)$-close to the unstable part of $\mathcal{S}_0$ until they reach the upper fold; they mark the transition from small-amplitude (nonrelaxation) oscillations to large-amplitude oscillations of relaxation type during a canard explosion.

One of the main goals in this article is to show how systems that exhibit mixed-mode-type behavior can be constructed from systems that undergo a canard explosion by replacing the parameter moving the slow nullcline with a dynamical variable. In other words, we will argue that the emergence of MMOs in such systems is triggered by a "slow passage through a canard explosion." More specifically, consider a system of the form



(a) $\lambda < 0$. (b) $\lambda \approx 0$.

(c) $\lambda \approx 0$. (d) $\lambda > 0$.

Figure 1. *Nullcline movement leading to a canard explosion: As the slow nullcline passes through the origin, one observes a transition from* (a) *a stable equilibrium via a family of canard solutions (*(b) *"headless canard,"* (c) *"canard with head") to* (d) *a full-scale relaxation oscillation.*

$$v' = -z + f_2 v^2 + f_3 v^3, \tag{1.2a}$$

$$z' = \varepsilon(v - w), \tag{1.2b}$$

$$w' = \varepsilon\big(\mu + \phi(v, z, w)\big), \tag{1.2c}$$

where $\mu > 0$ and $\phi = \mathcal{O}(v, z, w)$ is a smooth function that will be specified in the following, and note that the new slow variable w in (1.2) assumes the role of λ in (1.1). Let $\mathcal{S}_0$ denote the (two-dimensional) *critical manifold* for (1.2), which is defined by the constraint $z = f(v)$. Finally, let $\ell^- = \{(0, 0, w)\}$ and $\ell^+ = \{(-\frac{2f_2}{3f_3}, 0, w)\}$ denote the lower and upper fold lines for (1.2), respectively, and note that $\ell^\pm$ are determined by imposing $f'(v) = 0$, in addition to $z = f(v)$. Away from these fold lines, $\mathcal{S}_0$ is normally hyperbolic; it consists of the two attracting sheets

$$\mathcal{S}_0^{a-} = \big\{(v, z, w) \,\big|\, v < 0,\, z < 0\big\} \quad \text{and} \quad \mathcal{S}_0^{a+} = \big\{(v, z, w) \,\big|\, v > -\tfrac{2f_2}{3f_3},\, z > \tfrac{4f_2^3}{27f_3^2}\big\}, \tag{1.3}$$

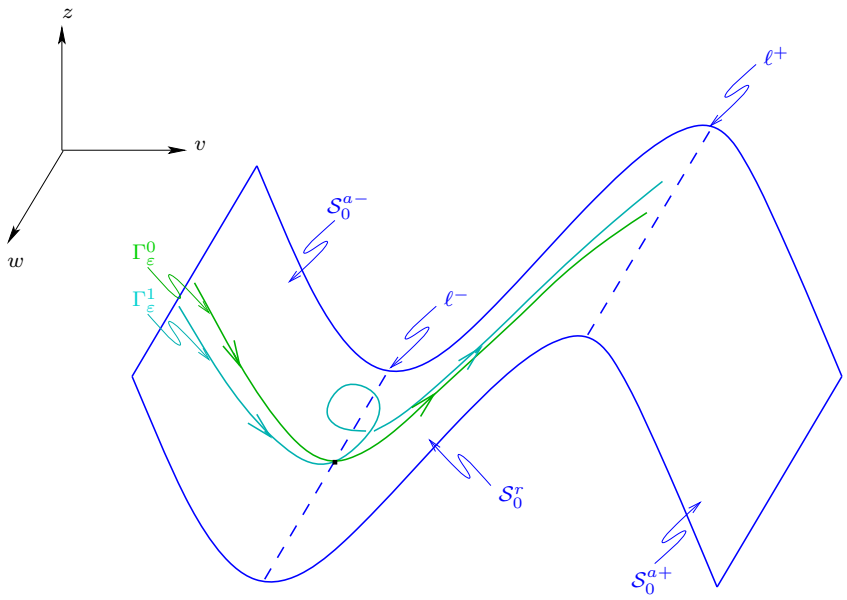


Figure 2. *The geometry of system (1.2): Critical manifold $\mathcal{S}_0$ with sheets $\mathcal{S}_0^{a\pm}$ and $\mathcal{S}_0^r$, fold lines $\ell^\pm$, and canard trajectories Γ_ε^0 and Γ_ε^1.*

as well as of a repelling sheet which is given by

$$\mathcal{S}_0^r = \left\{(v,z,w) \;\middle|\; 0 < v < -\tfrac{2f_2}{3f_3},\ 0 < z < \tfrac{4f_2^3}{27f_3^2}\right\}; \tag{1.4}$$

see Figure 2 for an illustration. (Note that, due to $f_2 > 0$ and $f_3 < 0$, there holds $-\frac{2f_2}{3f_3} > 0$ in (1.3) and (1.4).) The *singular limit* of $\varepsilon = 0$ in (1.2) is described by the dynamics of the *reduced problem* on the critical manifold $\mathcal{S}_0$.

By standard Fenichel theory [11], for $\varepsilon > 0$ sufficiently small and (v,z,w) in some bounded subset of $\mathbb{R}^3$, the critical manifold will perturb to a slow manifold $\mathcal{S}_\varepsilon$ away from $\ell^\pm$. We will denote the sheets of $\mathcal{S}_\varepsilon$ corresponding to $\mathcal{S}_0^{a-}$, $\mathcal{S}_0^{a+}$, and $\mathcal{S}_0^r$ by $\mathcal{S}_\varepsilon^{a-}$, $\mathcal{S}_\varepsilon^{a+}$, and $\mathcal{S}_\varepsilon^r$, respectively.

In analogy to the maximal canard encountered in (1.1), we define the so-called *strong canard* Γ_ε^0 for (1.2) as follows: once the two sheets $\mathcal{S}_\varepsilon^{a-}$ and $\mathcal{S}_\varepsilon^r$ are chosen, they are unique up to exponentially small terms in ε [11]. Then, Γ_ε^0 can be defined, for $\varepsilon > 0$ small, as the intersection of the continuation of these two sheets into the fold region. Moreover, as we will show in section 2, this intersection is transverse, which implies that Γ_ε^0 is well defined. It was postulated in [27] that the strong canard forms the boundary between two regions of very different dynamical behavior, in that it separates small-amplitude oscillations from large oscillations of relaxation type. We will confirm this postulate in the context of (1.2); in that sense, Γ_ε^0 can be interpreted as the "organizing center" for the emergence of MMOs in (1.2).

The detailed structure of the MMO trajectories that will be observed in (1.2) depends strongly on certain features of the specific equations under consideration. One important aspect concerns the properties of the global return mechanism, defined by the interplay of μ and ϕ in (1.2c), and in particular how far back the value of w is reset by that return.

If, during the return phase, w becomes $\mathcal{O}(1)$ and negative (i.e., if $\mu+\phi$ is not close to zero), the dynamics of (1.2) in the initial phase of the passage near the lower fold is of "node type," which means that there is strong contraction without any oscillatory behavior. That initial contractive phase is followed by oscillatory dynamics which can give rise to MMOs; however, most of the resulting oscillations are of very small amplitude. The class of these so-called *canards of folded-node type* is rather well understood and was analyzed in detail in [36].

By contrast, we will discuss a case where the global return mechanism is relatively weak in the sense that $\mu+\phi$ is $\mathcal{O}(1)$. Note that this case differs from that of the so-called *folded saddle-node* [31, 6] in that not only is μ assumed to be small, but ϕ is, too, and that the weakness of the return mechanism introduces an additional, "superslow" time-scale into the problem. In that sense, the folded saddle-node can be regarded as an intermediate case between the folded node and the situation in (1.2). (Note, however, that (1.2) could alternatively be classified as a "folded saddle-node of type II with weak global return" [35].)

The basic dynamics of (1.2) can be characterized as follows: given (v, z, w) small, the system will pass through the small-amplitude phase, where the variable w can grow slightly and become positive. Then, during the subsequent relaxation phase, w is reset to a small (negative or positive) value, and the cycle can start anew. Hence, the fact that w is always close to zero implies that there is no nonoscillatory contraction, contrary to the case of a folded node. Moreover, due to the three time-scale structure of (1.2), no slow passage through a Hopf bifurcation is observed, contrary to the case of a folded saddle-node. This distinction will be made more precise in the following; see also the discussion in section 4 below.

As we will show in this article, it is the interplay between the two main ingredients of the dynamics, the local flow close to the strong canard and the global return, that underlies the basic canard mechanism for the emergence of MMOs in (1.2). This mechanism can be generalized to other classes of systems; see, e.g., [17, 2] for details. In the following, we will refer to a combination of local, dynamical passage through a canard point and a suitably defined global return as the *generalized canard mechanism*. In other words, (1.2) represents only one specific realization of that very general mechanism. Moreover, as will follow from our analysis, (1.2) is a normal form for this class of three time-scale systems, in the sense that the addition of higher-order terms in (1.2) will not fundamentally influence the resulting dynamics.

Another aspect of the mixed-mode dynamics in (1.2), in addition to the return mechanism, is the family of so-called *secondary canards*. In the context of (1.2), we define the kth secondary canard Γ_ε^k as a trajectory that undergoes k small (nonrelaxation) rotations, or "loops," during its passage "near" the lower fold ℓ^- and that then remains $\mathcal{O}(\varepsilon)$-close to the critical manifold $\mathcal{S}_0$ until it reaches the $\mathcal{O}(\varepsilon^{\frac{1}{3}})$-vicinity of the upper fold ℓ^+ [34]. Note that the strong canard Γ_ε^0 passes through the vicinity of ℓ^- without undergoing any rotation at all, which corresponds to $k = 0$. As we will show, the existence of secondary canards in (1.2) is guaranteed by the fact that they can be defined as trajectories lying in the intersection of $\mathcal{S}_\varepsilon^r$ with subsequent iterates of $\mathcal{S}_\varepsilon^{a-}$ under the return map Π induced by the flow of (1.2); cf. section 3 below. This will allow us to give a precise asymptotic description of these canards; to the best of our knowledge, comparable results have so far been obtained only in the folded-node case [36], via a combination of asymptotics and numerics. For a qualitative illustration of the canard trajectories Γ_ε^0 and Γ_ε^1 in (1.2), cf. again Figure 2.

The notion of secondary canards leads to another important concept in this context, namely, that of the corresponding *sectors of rotation*, which are defined as (two-dimensional) portions of $\mathcal{S}_\varepsilon^{a-}$ in the fold region that are bounded by the secondary canards. These sectors, which we denote by RS^k, have the following property: trajectories starting in the kth sector undergo k small rotations near ℓ^-. Given that all MMO trajectories pass exponentially close to $\mathcal{S}_\varepsilon^{a-}$ in their relaxation phase, they must enter one of the sectors upon their return to the fold region. This fact can be exploited to reduce the corresponding (two-dimensional) return map Π for (1.2), which is a priori defined on an appropriate section for the corresponding flow, to a one-dimensional map Φ. Moreover, as we will show, the width of RS^k is $\mathcal{O}(\varepsilon^{\frac{3}{2}}\sqrt{-\ln\varepsilon})$, independent of k to leading order; see section 3. Hence, the canard phenomenon occurs rather "robustly" in the context of (1.2) in the sense that the relevant parameter intervals are not exponentially small in ε, as in the classical two-dimensional case [20].

Finally, with each MMO trajectory one can associate a sequence $L_0^{k_0}L_1^{k_1}\ldots$, called the *Farey sequence* [4], which describes the succession of large relaxation excursions and small (nonrelaxation) oscillations (loops): the segment $L_j^{k_j}$ corresponds to L_j relaxation oscillations followed by k_j small loops. (In the following, we will focus primarily on the case when $L_j = 1$.) As we will show, the Farey sequence of each trajectory is completely determined by the succession of the sectors of rotation visited by the trajectory. A natural question that arises in this context is which Farey sequences are admissible in a system of the form (1.2) and which μ-intervals they correspond to. This question is intimately related to the size of the sectors RS^k themselves, to the distance from the return point on $\mathcal{S}_\varepsilon^{a-}$ to the strong canard Γ_ε^0 after relaxation, and to the contractive (or expansive) properties of the flow induced by Π. These and similar issues will be discussed in detail in sections 3 and 4.

For the sake of definiteness, we will restrict ourselves to the more specific class of systems of the form

$$v' = -z + f_2v^2 + f_3v^3, \tag{1.5a}$$

$$z' = \varepsilon(v - w), \tag{1.5b}$$

$$w' = \varepsilon^2(\mu - g_1z) \tag{1.5c}$$

in the following, with $g_1 > 0$ constant. Note that (1.5) can be understood as a special case of (1.2), with μ rescaled by ε and $\phi(v, z, w) \equiv \phi(z) = \varepsilon g_1 z$. (Other choices of ϕ can be treated in a similar manner; see, e.g., [18].) This specific scaling of μ implies that the dynamics of (1.5) evolves on three distinct time-scales, a fast scale, a slow scale, and a "superslow" scale. Given that the flow of (1.5c) is governed by that slowest scale, w cannot vary too much, implying that trajectories cannot be reset very far back (in w) during the global return. Consequently, they will return close to the strong canard Γ_ε^0 of (1.5) after relaxation; equivalently, recalling the analogy between w and the parameter λ in (1.1), one could say that the return is close to the maximal canard of the (v, z)-subsystem in (1.5).

As we will show in section 3, it is the "lowest" sectors of rotation that will be immediately adjacent to the strong canard. Hence, only a few successive small oscillations will be observed in a typical time series of (1.5); moreover, these oscillations are relatively large in amplitude. Since the relevant parameter intervals will turn out to be relatively small, the corresponding dynamics is very sensitive to variations of μ. Also, since the stability intervals of "regular,"

L^k-type orbits (i.e., of MMO trajectories with Farey sequence $\{L^k\}$) are smaller still, the time series can be quite irregular; furthermore, there can be many relaxation cycles occurring in succession before the system returns to the small-oscillation phase. (By contrast, in the folded-node case, regular 1^k-type orbits are predicted to be stable for most μ-values; cf. [36].)

Moreover, as μ is varied in (1.5), one observes a passage through neighboring sectors of rotation: for increasing μ, the dynamics of (1.5) will be restricted to lower and lower sectors, admitting fewer and fewer small-amplitude oscillations, until eventually only relaxation cycles are seen. In other words, one observes the unfolding of a family of MMOs, including trajectories that pass through all sectors of rotation, on a fairly small parameter set. (In a folded-node system, on the other hand, such an unfolding can be expected over a μ-interval whose length is bounded below by a constant [36].) Also, numerical evidence suggests that only the Farey sequences predicted in section 3 will "generically" occur in a three time-scale system of the type of (1.5). Therefore, we conjecture that (1.5) can be interpreted as a "canonical form" for this particular class of three-dimensional systems. However, a rigorous, analytical justification of this claim is beyond the scope of this work.

In the remainder of this article, we analyze the "canonical" system (1.5) in detail, using a wide range of techniques. One of our aims is to derive asymptotic formulae for the return map induced by the flow of (1.5). To that end, we combine various methods from dynamical systems theory and, in particular, from geometric singular perturbation theory. To approximate the flow away from the fold lines $\ell^{\pm}$, we employ standard results due to Fenichel [11]. Upon entry into the neighborhood of $\ell^{\pm}$, normal hyperbolicity breaks down, and Fenichel's results are no longer applicable, which necessitates a detailed analysis of the dynamics there. We are especially interested in the lower fold ℓ^{-}, since it is there that the canard phenomenon occurs. To describe the dynamics close to ℓ^{-}, we make use of the near-integrable structure of the equations in (1.5). To access that structure, we introduce a rescaling that is akin to the blow-up transformation customarily used in this context; see, e.g., [7, 19] for details. While each of the parts of our analysis taken by itself is rather standard, the combination of the different approaches in the present context is new. In particular, by combining the leading-order global dynamics with detailed local asymptotics, we are able to obtain a closed-form description of the return map Π for (1.5) and, hence, to describe the resulting mixed-mode dynamics in detail.

This article is organized as follows. In section 2, we prove that the return map Π is well defined under an appropriate choice of sections for the flow of (1.5), and we derive precise asymptotic estimates for Π by desingularizing the dynamics of (1.5) in the fold region and by making use of the near-integrability of the resulting equations. Section 3 contains the centerpiece of our analysis in that we show how the "full," two-dimensional map Π can be reduced to a simpler, one-dimensional map Φ. This reduction is accurate with at most an exponentially small error (in ε) and is carried out in two steps: in a first step, Π is restricted from a two-dimensional section to the union of appropriately defined, one-dimensional curves, which allows us to describe the family of secondary canards, as well as the corresponding sectors of rotation, for (1.5). Then, in a second step, the map Π is further reduced and is restricted to a map Φ that is defined on a single curve. The dynamics of this map is analyzed in detail to make quantitative predictions on the relevant parameter regimes and the associated bifurcation (Farey) sequences in (1.5). In section 4, we summarize our results, and we relate

them to other mechanisms that have been proposed to explain MMOs. Moreover, we illustrate various properties of the "reduced" flow under Φ, and we compare them numerically to the "full" dynamics of (1.5). In sum, we thus obtain a fairly complete picture of the mixed-mode dynamics of (1.5), both qualitatively and quantitatively. Moreover, in doing so, we provide a framework for an even more detailed analysis of systems of the type of (1.5): once the dynamics of such a system is accurately reduced to that of a one-dimensional map, the well-developed theory of unimodal maps [25] can be applied. Our results on Φ are a first step in this direction in that there is potential for a more rigorous investigation along the lines of section 3.

Finally, we note that our analysis of (1.5) was inspired by a more specific problem, a model for the dynamics of the dopaminergic neuron that was proposed by Wilson and Callaway [37]. This model, which consists of a system of N strongly electrically coupled oscillators, was analyzed in [24] as well as in [23] (in a slightly different form) via a combination of asymptotic analysis and numerical techniques. One salient feature of the Wilson–Callaway model is precisely the unfolding of a family of MMO periodic orbits upon variation of one control parameter. In an upcoming companion paper [18], we will show how the Wilson–Callaway model can be fitted into the framework of (1.5) and how the results obtained here can be applied to study its dynamics.

2. The canonical system (1.5).

In this section, we discuss the system of equations (1.5) or, equivalently, the system obtained by rewriting (1.5) in terms of the slow time $\tau = \varepsilon t$,

$$\varepsilon \dot{v} = -z + f_2 v^2 + f_3 v^3, \tag{2.1a}$$

$$\dot{z} = v - w, \tag{2.1b}$$

$$\dot{w} = \varepsilon(\mu - g_1 z). \tag{2.1c}$$

Here, the overdot denotes differentiation with respect to τ, $f_2 > 0$, $f_3 < 0$, and $g_1 > 0$ are $\mathcal{O}(1)$ coefficients, $0 < \varepsilon \ll 1$ is small, and μ is the "free" (bifurcation) parameter; note the presence of three time-scales in (2.1).

Let $\mathcal{S}_0$ denote the critical manifold for (2.1), as before, and recall that $\mathcal{S}_0$ is given by $z = f(v) = f_2 v^2 + f_3 v^3$; cf. section 1. Moreover, recall the definition of $\mathcal{S}_0^{a\pm}$ and $\mathcal{S}_0^r$ in (1.3) and (1.4), respectively, and let $\mathcal{S}_\varepsilon^{a\pm}$ and $\mathcal{S}_\varepsilon^r$ denote the corresponding sheets of the slow manifold for $\varepsilon > 0$ sufficiently small. Finally, the upper and lower fold lines in (2.1) are again denoted by $\ell^\pm$.

2.1. Sections for the flow of (1.5).

To derive asymptotic formulae for the return of trajectories under the flow of (1.5), we will define the corresponding return map on suitable sections for the flow, which we introduce below. In the course of our analysis, we will show that the small-amplitude oscillations observed in (1.5) are due to the fact that, in the parameter regime under consideration, the system passes slowly through a canard explosion about the origin in (v, z, w)-space. The large-amplitude components of the mixed-mode time series are generated by the global return mechanism, which takes trajectories back to the fold line ℓ^- after the passage past the origin has been completed. Combining these two aspects of the dynamics will allow us to describe in detail how MMOs can arise in (1.5).

The dynamics of (1.5) can be broken down into the following four components:

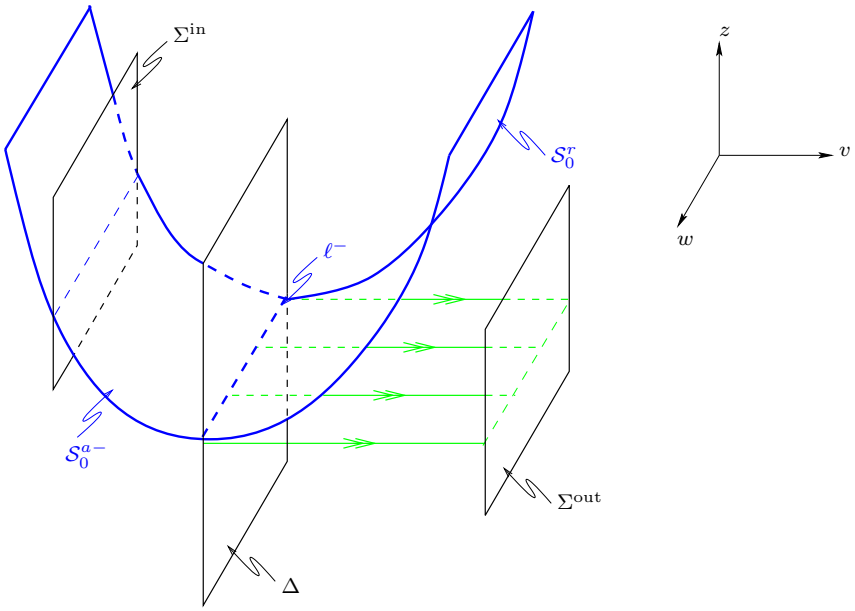


Figure 3. *The sections* Σ^{in}, Δ, *and* Σ^{out} *for the flow of* (1.5).

(i) the flow in a neighborhood of the fold line ℓ^- (section 2.2);
(ii) the entry into the fold region (section 2.3);
(iii) the exit from the fold region (section 2.4); and, finally,
(iv) the global return mechanism (section 2.5).

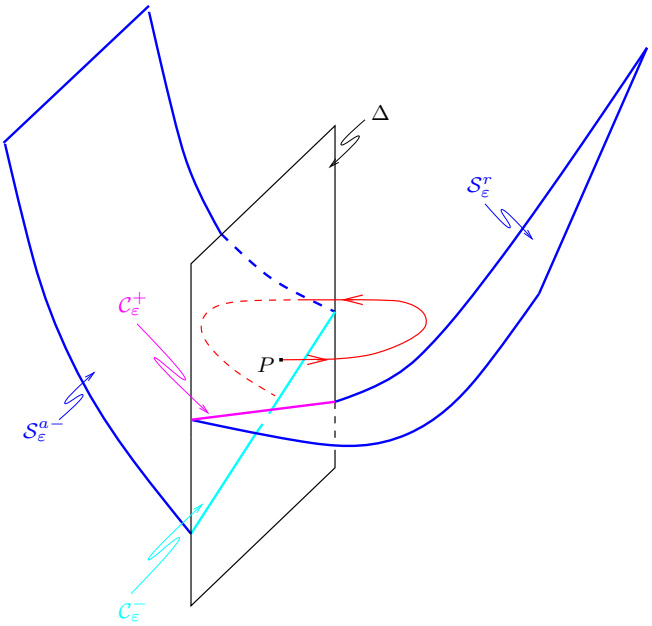
We will construct transition maps for each of the above components of the flow. The desired global return map, which we denote by Π, will then be obtained via the composition of these individual maps.

We begin by introducing *sections* for the flow of (1.5): we will require

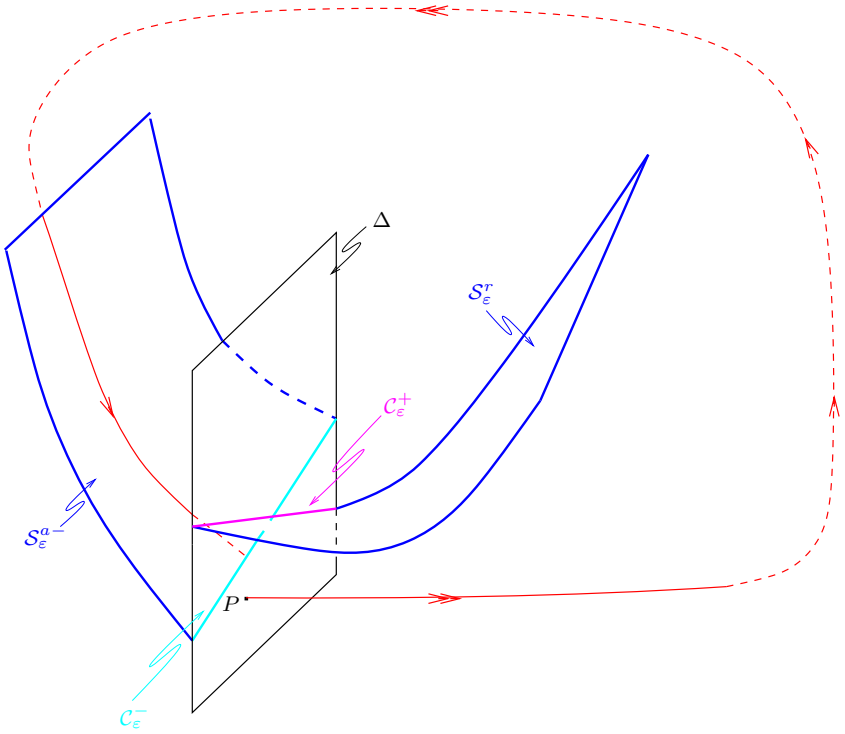
(i) a section Σ^{in} across the attracting branch $\mathcal{S}_0^{a-}$ of the critical manifold $\mathcal{S}_0$, which is given by $v = -\rho$, with $|z|$ and $|w|$ bounded;
(ii) a section Δ, which is defined by $v = 0$, with $|z|$ and $|w|$ bounded, implying that Δ lies in the (z, w)-plane and that it bisects the critical manifold $\mathcal{S}_0$ along ℓ^- (the w-axis); and
(iii) a section Σ^{out} across the fast foliation of $\mathcal{S}_0$, with $v = \delta$ and $|z|$ and $|w|$ bounded.

Here, $\rho, \delta > 0$ are small but fixed (ε-independent) constants; see Figure 3 for an illustration. The section Δ will turn out to be especially important in the following, since the global return map Π will be defined on Δ. (Note that this particular choice of Poincaré section has previously been made by Dumortier and Roussarie in their analysis of canard cycles; see, e.g., [8].)

Next, we introduce two subsets of Δ that will play a crucial role in the description of Π. We first define $\mathcal{C}_\varepsilon^-$ as follows: a point $P \in \Delta$ is an element of $\mathcal{C}_\varepsilon^-$ if P is the endpoint of a segment of trajectory that originates in $\mathcal{S}_\varepsilon^{a-}$. The set $\mathcal{C}_\varepsilon^+$ is defined analogously, with $\mathcal{S}_\varepsilon^{a-}$ replaced by $\mathcal{S}_\varepsilon^r$ and the time reversed; see Figure 4. The sets $\mathcal{C}_\varepsilon^-$ and $\mathcal{C}_\varepsilon^+$ have the following properties:



(a) Trajectory of $P \in \Delta$ above $\mathcal{C}_\varepsilon^+$.



(b) Trajectory of $P \in \Delta$ below $\mathcal{C}_\varepsilon^+$.

Figure 4. *The sets $\mathcal{C}_\varepsilon^-$ and $\mathcal{C}_\varepsilon^+$.*

(i) If $P \in \Delta$ is above $\mathcal{C}_\varepsilon^+$, the trajectory of P is blocked by $\mathcal{S}_\varepsilon^r$ from entering relaxation. Depending on the position of P, the initial motion may be toward $\mathcal{S}_\varepsilon^r$, but the trajectory must eventually turn toward $\mathcal{S}_\varepsilon^{a-}$ (under the fast flow) and return to Δ, having undergone a small-amplitude oscillation (or loop); see Figure 4(a).

(ii) If $P \in \Delta$ is below $\mathcal{C}_\varepsilon^+$, the trajectory of P must leave the vicinity of the fold in the direction of the fast flow and may re-enter only through a global return mechanism, since no trajectories can pass through $\mathcal{S}_\varepsilon^r$; see Figure 4(b).

(iii) Any trajectory that is attracted to $\mathcal{S}_\varepsilon^{a-}$ will be exponentially close to $\mathcal{C}_\varepsilon^-$ when it hits Δ.

Remark 1. Since $\mathcal{S}_\varepsilon^{a-}$ and $\mathcal{S}_\varepsilon^r$ are unique only up to exponentially small terms in ε, the sets $\mathcal{C}_\varepsilon^\pm$ are, strictly speaking, "strips" rather than curves. However, since our construction of Π will rely on leading-order ε-asymptotics throughout, this nonuniqueness will not influence our results.

A proof of these claims will be given in section 3 below. We now proceed with the derivation of the four components of the return map, as outlined above. The description of the dynamics in the fold region is the centerpiece of our analysis and will be discussed first.

2.2. Dynamics in the fold region. Our goal in this subsection is to analyze the flow in the region of the phase space of (1.5) where small-amplitude oscillations (loops) can occur. To describe these loops, we have to study the equations in (1.5) in an $\mathcal{O}(\sqrt{\varepsilon})$-vicinity of the fold line ℓ^- and, specifically, of the origin in (u, v, w)-space. Recall that under our assumptions on (1.5), ℓ^- is given by the w-axis.

To investigate the dynamics of (1.5) close to ℓ^-, we define the rescaling

$$v = \sqrt{\varepsilon}\bar{v}, \quad z = \varepsilon\bar{z}, \quad w = \sqrt{\varepsilon}\bar{w}, \quad \text{and} \quad t = \frac{\bar{t}}{\sqrt{\varepsilon}}. \tag{2.2}$$

In terms of the new "barred" variables in (2.2), (1.5) becomes

$$\bar{v}' = -\bar{z} + f_2\bar{v}^2 + \sqrt{\varepsilon} f_3 \bar{v}^3, \tag{2.3a}$$

$$\bar{z}' = \bar{v} - \bar{w}, \tag{2.3b}$$

$$\bar{w}' = \varepsilon(\mu - g_1\varepsilon\bar{z}), \tag{2.3c}$$

where the prime now denotes differentiation with regard to the new rescaled time $\bar{t}$. Note that (2.3) is a fast-slow system, with two fast variables $\bar{v}$ and $\bar{z}$ and one slow variable $\bar{w}$. In other words, the scale separation between v and z has vanished after the rescaling, whereas $\bar{w}$ is still slow and constant to leading order. Hence, we can interpret $\bar{w}$ as a slowly varying parameter.

For $\varepsilon = 0$, the equations in (2.3) reduce to

$$\bar{v}' = -\bar{z} + f_2\bar{v}^2, \tag{2.4a}$$

$$\bar{z}' = \bar{v} - \bar{w}, \tag{2.4b}$$

$$\bar{w}' = 0. \tag{2.4c}$$

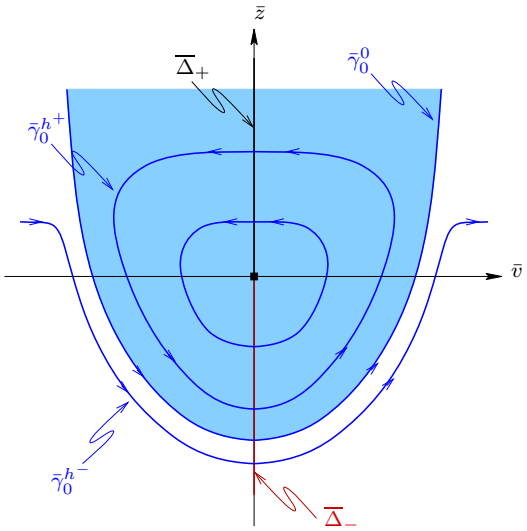


Figure 5. *Typical integral curves of H, with $h^- < 0 < h^+$. (The region where $h > 0$ is shaded.)*

Note that, up to various rescalings, (2.4) is of the form

$$x' = -y + x^2,$$
$$y' = x - \lambda,$$

which is a prototypical system for the occurrence of a canard explosion (at $\lambda = 0$) [20]; see also (1.1). In the following, we will describe how the equations in (2.3) fit into the framework of [20], where the classical two-dimensional scenario is analyzed using geometric singular perturbation theory. The role of the bifurcation parameter λ is taken by $\bar{w}$ in our case. For $\bar{w} = 0$, (2.4) is an integrable system, with constant of motion given by

$$H(\bar{v}, \bar{z}) = \frac{1}{2} \mathrm{e}^{-2f_2\bar{z}} \Big(-\bar{v}^2 + \frac{\bar{z}}{f_2} + \frac{1}{2f_2^2} \Big). \tag{2.5}$$

The equations in (2.4) have a continuous family of periodic orbits which are most conveniently described via the level curves of H; these are defined by $H(\bar{v}, \bar{z}) = h$ for h constant. The corresponding (time-parametrized) solution curves will be denoted by $\bar{\gamma}_0^h(t) = (\bar{v}_0^h, \bar{z}_0^h)(t)$ in the following.

We first note that $(\bar{v}, \bar{z}) = (0, 0)$ lies on the curve defined by $H(\bar{v}, \bar{z}) = h_0 := (4f_2^2)^{-1}$. For $h > h_0$, there exist no real solutions to $H(\bar{v}, \bar{z}) = h$. Hence, without loss of generality, we consider $h \le h_0$ now, and we note that $h_0 > 0$. For $h = 0$ in (2.5), we obtain the special solution $\bar{\gamma}_0^0$ of (2.4), with

$$\bar{\gamma}_0^0(t) = (\bar{v}_0^0, \bar{z}_0^0)(t) = \Big(\frac{1}{2f_2} t, \frac{1}{4f_2} t^2 - \frac{1}{2f_2} \Big). \tag{2.6}$$

Note that (2.6) defines an invariant parabola that separates the closed level curves of H, which are obtained for $h > 0$, from the open ones, with $h < 0$; see Figure 5 for an illustration. Since

the two branches of this parabola correspond to $\mathcal{S}_0^{a-}$ and $\mathcal{S}_0^r$ for $\bar{w}=0$, after the rescaling in (2.2), $\bar{\gamma}_0^0$ is a "singular canard solution," i.e., a solution of (2.3) that connects $\mathcal{S}_\varepsilon^{a-}$ and $\mathcal{S}_\varepsilon^r$ in the singular limit as $\varepsilon \to 0$. (In fact, as we will see in (2.16) below, the orbit determined by $\bar{\gamma}_0^0$ is precisely the strong canard Γ_ε^0 in this singular limit.)

Let $\overline{\Delta}$ denote the section that corresponds to Δ in the "barred" variables; i.e., let $\overline{\Delta} = \{\bar{v}=0\}$, with $|\bar{z}|$ and $|\bar{w}|$ bounded. For h fixed, let $\bar{z}^h$ be the corresponding value of $\bar{z}$ in $\overline{\Delta}$, with $H(0,\bar{z}^h)=h$. (In particular, by (2.6), there holds $\bar{z}^0=-(2f_2)^{-1}$.) Our first result is a direct consequence of the above discussion; see [20] for details.

Proposition 2.1. *To any $h \le h_0$, with $h_0=(4f_2^2)^{-1}>0$, there corresponds precisely one value $\bar{z}^h \le 0$ of $\bar{z}$ in $\overline{\Delta}$. Moreover, $\bar{z}^h$ is an increasing function of h.*

Since the limiting equations obtained for $\bar{w}=0=\varepsilon$ in (2.3) are integrable, we will refer to the original, "perturbed" dynamics as "near-integrable." (A related treatment of a more general family of near-integrable systems can be found in [15].) The near-integrability of (2.3) will allow us to analyze the dynamics of the equations using a perturbation analysis, and to approximate the return map from $\overline{\Delta}$ to itself, which we refer to as $\overline{\Pi}$, to leading order. Naturally, the closed level curves of H will turn out to be the singular "templates" for the small-amplitude component of the mixed-mode dynamics observed in (2.1). Moreover, as we will show, it is the bifurcation structure of $\overline{\Pi}$ that is responsible for the emergence of secondary canards in (2.3); these canards, in turn, determine the qualitative structure of the resulting MMO patterns. In that sense, the rescaling in (2.2) will enable us to access the near-integrable structure of (1.5) close to ℓ^-.

We will define the return map $\overline{\Pi}$ on $\overline{\Delta}_- \subset \overline{\Delta}$, which is the portion of $\overline{\Delta}$ where $\bar{z}<0$. Although $\overline{\Pi}$ is a priori a function of $(\bar{z},\bar{w})$, it is more convenient to parametrize $\bar{z}$ by h and to describe the asymptotics of $\overline{\Pi}$ in terms of h and $\bar{w}$ in the following. For $h \le h_0$, with h_0 as above, let $\bar{z}^h$ again denote the corresponding unique value of $\bar{z} \in \overline{\Delta}_-$, and note that we will sometimes identify $\bar{z}^h$ with its associated h-value. Moreover, let $\bar{\gamma}_\varepsilon^h(t)$ be the solution to (2.3) emanating from $(0,\bar{z}^h,\bar{w})$, where the time parametrization is chosen so that $\bar{\gamma}_\varepsilon^h(0)$ is contained in $\overline{\Delta}_+ := \overline{\Delta}\backslash\overline{\Delta}_-$. Then, we define $T_-^h(\bar{w})<0$ and $T_+^h(\bar{w})>0$ by requiring that $\bar{\gamma}_\varepsilon^h(T_\pm^h(\bar{w})) \in \overline{\Delta}_-$. Moreover, we assume that $T_\pm^h(\bar{w})$ are the times of the first such intersection. Let $T^h : \overline{\Delta}_- \to \overline{\Delta}_-$ denote the return time of solutions under the flow of (2.3), and note that, by definition, $T^h(\bar{w}) = T_+^h(\bar{w}) - T_-^h(\bar{w})$. Let $\hat{h}$ be defined by the requirement that $\bar{z}^{\hat{h}}$ is the $\bar{z}$-coordinate of $\bar{\gamma}_\varepsilon^h(T^h(\bar{w})) \in \overline{\Delta}_-$; an illustration of these definitions is given in Figure 6. Finally, for $\bar{w}=0$, we write $T^h := T_+^h(0)$, which, together with $T_-^h(0)=-T_+^h(0)$, implies

$$T^h(0) = T_+^h(0) - T_-^h(0) = 2T^h. \tag{2.7}$$

We now make the following assumption on $\bar{w}$, which will be verified a posteriori for the parameter regime we are interested in.

Assumption 1. For fixed, real $f_2>0$, $f_3<0$, $\mu>0$, and $g_1>0$ and $0<\varepsilon\ll 1$ sufficiently small in (2.3), $\bar{w}=\mathcal{O}(\sqrt{\varepsilon})$ uniformly in $\bar{t}$.

It will follow from our analysis that Assumption 1 defines an invariant region for the return map Π which roughly corresponds to the regime where $\bar{w}=\mathcal{O}(\sqrt{\varepsilon})$. More precisely, if an initial condition for (2.3) satisfies the assumption, it will be satisfied along the entire corresponding trajectory of (2.3). Finally, since $w=\sqrt{\varepsilon}\bar{w}$, Assumption 1 implies that $w=\mathcal{O}(\varepsilon)$ must hold in (2.1), uniformly in τ.

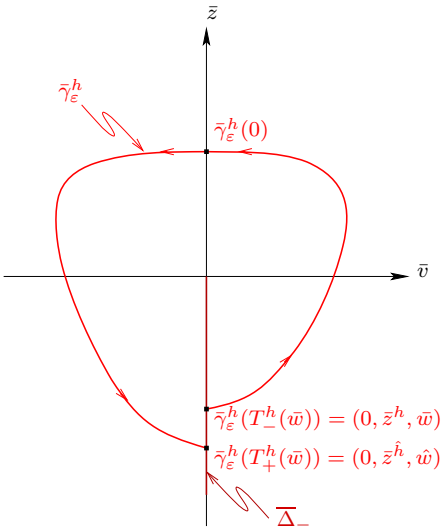


Figure 6. *The geometry of system* (2.3).

We state our next result in a slightly more general context than that of (2.3). The reason for this generalization is that we will modify (2.3) later to simplify our estimates of the return time $T^h(\bar{w})$. Thus, instead of (2.3), we now consider the following generalized system of equations:

$$\bar{v}' = -\bar{z} + f_2\bar{v}^2 + \sqrt{\varepsilon} f_3 \bar{v}^3 + \sqrt{\varepsilon} F(\bar{w}, \sqrt{\varepsilon}) + \bar{w} G(\bar{w}, \sqrt{\varepsilon}), \tag{2.8a}$$

$$\bar{z}' = \bar{v} - \bar{w} + \mathcal{O}(\varepsilon), \tag{2.8b}$$

$$\bar{w}' = \varepsilon\big(\mu - g_1 \varepsilon \bar{z} + \mathcal{O}(\varepsilon)\big), \tag{2.8c}$$

where F and G are assumed to be $\mathcal{C}^n$-smooth for $n \geq 1$ sufficiently large in both $\bar{w}$ and $\sqrt{\varepsilon}$. Note that all the definitions and notation introduced in the context of (2.3) extend without modification to (2.8).

Proposition 2.2. *Let $\overline{\Pi} : \overline{\Delta}_- \to \overline{\Delta}_-$ and $\bar{\gamma}_\varepsilon^h$ be defined as above, and let $(h, \bar{w}) \in \overline{\Delta}_-$. Suppose that $h > 0$, with $h = \mathcal{O}(\varepsilon^M)$ for some $M > 0$ and $\varepsilon > 0$ sufficiently small, and that the trajectory starting at $(h, \bar{w})$ undergoes a small oscillation ("loop") before returning to $\overline{\Delta}_-$. Then,*

$$(\hat{h}, \hat{w}) := \overline{\Pi}(h, \bar{w}) = \big(h + \sqrt{\varepsilon} d^h_{\sqrt{\varepsilon}} + \bar{w} d^h_{\bar{w}} + \mathcal{O}\big((\sqrt{\varepsilon} + \bar{w})^2\big),\ \bar{w} + \varepsilon\mu T^h(\bar{w}) + \mathcal{O}(\varepsilon^2)\big), \tag{2.9}$$

where the coefficients $d^h_{\sqrt{\varepsilon}}$ and $d^h_{\bar{w}}$ are defined as

$$d^h_{\sqrt{\varepsilon}} = \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot \big(f_3 \bar{v}_0^h(t)^3, 0\big)^T \, dt \tag{2.10}$$

and

$$d^h_{\bar{w}} = \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (0, -1)^T \, dt, \tag{2.11}$$

respectively, and $\bar{\gamma}_0^h(t) = (\bar{v}_0^h, \bar{z}_0^h)(t)$ *denotes the solution to* (2.4) *with* $H(\bar{v}_0^h, \bar{z}_0^h) = h$.

Remark 2. Given Assumption 1, as well as the fact that $T^h(\bar{w}) \sim \sqrt{-2\ln h}$ by Lemma A.2, $h = \mathcal{O}(\varepsilon^M)$ implies $\varepsilon T^h(\bar{w}) = \mathcal{O}(\varepsilon\sqrt{-\ln\varepsilon})$ in (2.9) for any $M > 0$. Moreover, the expansion for $\hat{h}$ remains valid even if $M > \frac{1}{2}$, i.e., when the leading-order term in ε is given by $\sqrt{\varepsilon}d^h_{\sqrt{\varepsilon}}$. Hence, it follows that (2.9) describes the map $\overline{\Pi}$ up to an $\mathcal{O}(\varepsilon)$-error.

Proof. We only sketch the proof here and refer the reader to [19] for details.

To derive the expression for $\hat{w}$, one makes use of the near-integrability of (2.8) as well as of regular perturbation theory.

To prove the assertion for $\hat{h}$, we first note that

$$\hat{h} - h := H(0, \hat{z}^{\hat{h}}) - H(0, \bar{z}^h) = \int_{T^h_-(w)}^{T^h_+(w)} \frac{d}{dt} H(\bar{\gamma}^h_\varepsilon(t))\, dt. \tag{2.12}$$

Since, to lowest order,

$$\frac{d}{dt} H(\bar{\gamma}^h_\varepsilon(t)) = \nabla H(\bar{\gamma}^h_\varepsilon(t)) \cdot (\bar{v}', \bar{z}')^T\big|_{\bar{\gamma}^h_\varepsilon} = \nabla H(\bar{\gamma}^h_0(t)) \cdot (\bar{v}', \bar{z}')^T\big|_{\bar{\gamma}^h_0},$$

and since H is a constant of motion, it follows with (2.8a) and (2.8b) that

$$\begin{aligned}\hat{h} - h = &\int_{-T^h}^{T^h} \nabla H(\bar{\gamma}^h_0(t)) \cdot \big(f_3 \bar{v}^h_0(t)^3 + F(0,0), 0\big)^T\, dt\, \sqrt{\varepsilon} \\ &+ \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}^h_0(t)) \cdot (G(0,0), -1)^T\, dt\, \bar{w} + \mathcal{O}(2);\end{aligned} \tag{2.13}$$

see also [20]. (Here, $\mathcal{O}(2)$ denotes terms of at least second order in $\sqrt{\varepsilon}$ and $\bar{w}$.) Since, however, $(\bar{v}^h_0, \bar{z}^h_0)(-t) = (-\bar{v}^h_0, \bar{z}^h_0)(t)$ on $\bar{\gamma}^h_0$ by symmetry, a change of variables via $t \mapsto -t$ in combination with (2.5) shows

$$\begin{aligned}\int_{-T^h}^{T^h} \frac{\partial H}{\partial \bar{v}}(\bar{\gamma}^h_0(t))\, dt &= -\int_{-T^h}^{T^h} \bar{v}^h_0(t) \mathrm{e}^{-2f_2 \bar{z}^h_0(t)}\, dt = -\int_{T^h}^{-T^h} \bar{v}^h_0(-t) \mathrm{e}^{-2f_2 \bar{z}^h_0(-t)}\, d(-t) \\ &= -\int_{-T^h}^{T^h} \frac{\partial H}{\partial \bar{v}}(\bar{\gamma}^h_0(t))\, dt.\end{aligned}$$

Therefore, the latter integral must be zero, which implies

$$\int_{-T^h}^{T^h} \frac{\partial H}{\partial \bar{v}}(\bar{\gamma}^h_0(t))\, F(0,0)\, dt = 0 \quad \text{and} \quad \int_{-T^h}^{T^h} \frac{\partial H}{\partial \bar{v}}(\bar{\gamma}^h_0(t))\, G(0,0)\, dt = 0.$$

It follows that (2.13) reduces to

$$\hat{h} - h = d^h_{\sqrt{\varepsilon}} \sqrt{\varepsilon} + d^h_{\bar{w}} \bar{w} + \mathcal{O}(2),$$

with the coefficients $d^h_{\sqrt{\varepsilon}}$ and $d^h_{\bar{w}}$ as defined in (2.10) and (2.11). This completes the proof. ∎

Remark 3. Note that the functions $T^h_\pm(\bar{w})$ and $T^h(\bar{w})$ depend very sensitively on h, $\bar{w}$, and $\sqrt{\varepsilon}$; in fact, since $\lim_{(h,\bar{w},\varepsilon)\to(0,0,0)} T^h(\bar{w}) = \infty$, $T^h(\bar{w})$ has a singularity at the origin. For

this reason, it is not immediately obvious that $T^h_+(\bar{w})$ and $T^h_-(\bar{w})$ can be replaced by T^h and $-T^h$, respectively, in (2.12). However, the arguments in [19] can easily be extended to justify this point.

Given Proposition 2.2, we make the following observations:

(i) Observe that, for $\mu = 0$, the equations in (2.3) have an equilibrium point at the origin. The linearization of (2.3) about this equilibrium has a pair of purely imaginary eigenvalues, as well as a simple eigenvalue 0. The corresponding steady-state Hopf-type interactions mark the onset of small-amplitude oscillations in (2.3); see also [20, Theorem 3.1]. (Note that the presence of the zero eigenvalue, which is due to the absence of a linear $\bar{w}$-term in (2.3c), introduces a degeneracy at the origin in (2.3).)

(ii) In order to obtain periodic orbits in (2.3), we have to require $h = \hat{h}$ and $\bar{w} = \hat{w}$; see the definition of $\overline{\Pi}$ in (2.9). Hence, to leading order, we must impose the condition

$$d^h_{\sqrt{\varepsilon}}\sqrt{\varepsilon} + d^h_{\bar{w}}\bar{w} = 0 \tag{2.14}$$

on $(h, \bar{w})$. To show that (2.14) can be solved for h and $\bar{w}$, we have to find the next-order correction to $\hat{w}$ in (2.9): integrating (2.3c), we obtain

$$\hat{w} = \bar{w} + 2\varepsilon\mu T^h - g_1\varepsilon^2 \int_{-T^h}^{T^h} \bar{z}(t)\,dt,$$

to leading order. Using (2.3a) to express $\bar{z}$ in terms of $\bar{v}$, we find

$$\hat{w} \sim \bar{w} + \varepsilon\left(2\mu T^h - g_1\varepsilon \int_{-T^h}^{T^h} \left(-\bar{v}'(t) + f_2\bar{v}(t)^2\right) dt\right).$$

(Here and in the following, the tilde indicates a leading-order asymptotic approximation.) Since, moreover, $\bar{v}(-T^h) = 0 = \bar{v}(T^h)$ by definition, and since (2.3b) implies $\frac{d\bar{z}}{dt} \sim \bar{v}$ by Assumption 1, it follows that

$$\hat{w} \sim \bar{w} + 2\varepsilon\left(\mu T^h - f_2 g_1 \varepsilon \int_{\xi^h}^{\zeta^h} \bar{v}(\bar{z})\,d\bar{z}\right).$$

Here, $\xi^h = \bar{z}(-T^h)$ and $\zeta^h = \bar{z}(0)$ denote the $\bar{z}$-values in $\overline{\Delta}$ corresponding to $\bar{\gamma}^h_\varepsilon(-T^h)$ and $\bar{\gamma}^h_\varepsilon(0)$, respectively. In sum, the requirement that $\bar{w} = \hat{w}$ gives

$$\mu T^h - f_2 g_1 \varepsilon \int_{\xi^h}^{\zeta^h} \bar{v}(\bar{z})\,d\bar{z} = 0 \tag{2.15}$$

to lowest order. Since $\frac{1}{T^h}\int_{\xi^h}^{\zeta^h} \bar{v}(\bar{z})\,d\bar{z}$ increases monotonically in h as $h \to 0$ (see [19]), it follows that, for ε and μ small and fixed, one can find h such that (2.15) holds. Given that h-value, one can use (2.14) to determine the associated value of $\bar{w}$.

(iii) For μ and ε sufficiently small in (2.1), there exists a canard trajectory lying in the intersection of the manifolds $\mathcal{S}^{a-}_\varepsilon$ and $\mathcal{S}^r_\varepsilon$; this trajectory is the strong canard Γ^0_ε. Since $\mathcal{S}^{a-}_\varepsilon$ and $\mathcal{S}^r_\varepsilon$ intersect transversely, as we will show in section 2.3 below, Γ^0_ε is

well defined; moreover, it is unique once specific sheets of $\mathcal{S}_\varepsilon$ have been chosen. The associated *canard critical value* $\bar{w}^{\mathrm{c}}$, i.e., the value of $\bar{w}$ in the rescaled system (2.3) that corresponds to Γ_ε^0, is given by

$$\bar{w}^{\mathrm{c}} = -\frac{d^0_{\sqrt{\varepsilon}}}{d^0_{\bar{w}}}\sqrt{\varepsilon} + \mathcal{O}(\varepsilon), \tag{2.16}$$

where $d^0_{\sqrt{\varepsilon}}$ and $d^0_{\bar{w}}$ are obtained from (2.10) and (2.11) in the limit as $h \to 0$ [19]. In particular, since $\bar{w}^{\mathrm{c}} \to 0$ for $\varepsilon \to 0$, (2.16) yields precisely the singular canard solution $\bar{\gamma}_0^0$ in this limit; cf. (2.6). Hence, as $h \to 0$, (2.3) undergoes a classical (two-dimensional) canard explosion at $\bar{w} = 0 = \varepsilon$ [20].
To evaluate (2.16), note that (2.5) implies

$$\frac{\partial H}{\partial \bar{v}} = -\bar{v}\mathrm{e}^{-2f_2\bar{z}} \quad \text{and} \quad \frac{\partial H}{\partial \bar{z}} = (f_2\bar{v}^2 - \bar{z})\mathrm{e}^{-2f_2\bar{z}}. \tag{2.17}$$

Using the parametrization of $\bar{\gamma}_0^0$ in (2.6) and taking into account that $T^0 = \infty$, one finds as in [19] that

$$d^0_{\sqrt{\varepsilon}} = -\frac{3f_3}{16f_2^4}\sqrt{2\pi}\mathrm{e} \quad \text{and} \quad d^0_{\bar{w}} = -\frac{1}{2f_2}\sqrt{2\pi}\mathrm{e}; \tag{2.18}$$

see Appendix A for details. Therefore, for given μ, the corresponding value of $\bar{w}^{\mathrm{c}}$ can be obtained from

$$\bar{w}^{\mathrm{c}} = -\frac{3f_3}{8f_2^3}\sqrt{\varepsilon} + \mathcal{O}(\varepsilon); \tag{2.19}$$

note that $\bar{w}^{\mathrm{c}} > 0$ due to $f_2 > 0$ and $f_3 < 0$.
These observations combined suggest the following: for $\varepsilon > 0$ fixed, system (2.3) undergoes a Hopf bifurcation at the origin for $\mu = 0$ by (i); this bifurcation gives rise to small-amplitude limit cycles in (2.3). These cycles will persist as long as both (2.14) and (2.15) can be satisfied, as shown in (ii). In that case, $\mu = \mathcal{O}(\varepsilon)$ must hold, since $T^h = \mathcal{O}(\sqrt{-\ln\varepsilon})$, $\zeta^h = \mathcal{O}(\sqrt{-\ln\varepsilon})$, and $\xi^h = \mathcal{O}(1)$ by Appendix A, while $\bar{v}$, f_2, and g_1 are $\mathcal{O}(1)$ by assumption. Hence, for μ sufficiently small, the dynamics of (2.3) will be dominated by 0^k-type orbits, i.e., by MMO trajectories with Farey sequence $\{0^k\}$. As μ is increased, the evolution of $\bar{w}$ in (2.3c) is governed by the positive, μ-dependent drift, with $\bar{w}' \sim \varepsilon\mu$. Since $\bar{z}$ decreases with increasing $\bar{w}$ (see (2.3b)), it follows that h must also decrease by Proposition 2.1. In other words, $h \to 0$ with increasing μ, and the system moves closer and closer toward a canard explosion, as discussed in (iii). Finally, for $\mu = \mu^{\mathrm{c}}$ large enough, the $\bar{w}$-drift is sufficiently strong for the dynamics of (2.3) to bypass the fold region and enter the relaxation regime. (The corresponding "critical" μ-value μ^{c} will be discussed in detail in section 2.5 below.)

In our analysis, we will focus primarily on the regime where μ is sufficiently large for 0^k-type orbits not to dominate the dynamics of (2.3) anymore. Since these orbits can occur only when (2.3) is close to Hopf bifurcation (i.e., as long as $\mu = \mathcal{O}(\varepsilon)$ and, hence, $\bar{w}' \sim 0$), the degeneracy of the equations at the Hopf point will not be of relevance to us. On the other

hand, we will assume that $\mu < \mu^c$, i.e., that μ is not large enough for (2.3) to have entered the relaxation regime, which is characterized by L^0-type orbits (trajectories with Farey sequence $\{L^0\}$).

As we will show, this "intermediate" regime corresponds precisely to the nontrivial mixed-mode dynamics of (1.5), with orbits of the type $\{L_j^{k_j}\}$ for $L_j, k_j \geq 1$. Correspondingly, h will have to be small in the sense that $|h| = \mathcal{O}(\varepsilon^M)$ for some $M > 0$ "large"; however, h cannot be exponentially small in ε, since trajectories must stay away from the strong canard Γ_ε^0. The statement of Proposition 2.2 pertains exactly to that intermediate case.

Finally, we remark that we will restrict ourselves to a leading-order description of the return map Π in the following, as we did in the proof of Proposition 2.2. The resulting approximation will remain consistent as long as $h = \mathcal{O}(\varepsilon^M)$ is not "too large," i.e., if $\varepsilon > 0$ is sufficiently small or if $M > 0$ is large enough: due to $T^h(\bar{w}) \sim \sqrt{-2\ln h} \sim \sqrt{-2M\ln\varepsilon}$, the $\varepsilon T^h(\bar{w})$-term in (2.9) will dominate the neglected terms of order $\mathcal{O}(\varepsilon)$ in that case. These considerations will be made more explicit in Proposition 3.4 below.

2.3. The transition from Σ^{in} to $\overline{\Delta}_-$. Let Π^{in} denote the transition map from Σ^{in} to $\overline{\Delta}_-$; see sections 2.1 and 2.2 for the definitions of Σ^{in} and $\overline{\Delta}_-$. Moreover, let us introduce an intermediate section $\overline{\Delta}^{\text{in}}$ for the rescaled equations in (2.3), with $\overline{\Delta}^{\text{in}} = \{(\bar{v}, \bar{z}, \bar{w}) \mid \bar{v} = -\alpha\}$, and let Δ^{in} denote the corresponding section in (v, z, w)-space. (Here, $0 < \alpha < \rho$ is some arbitrary constant.) Then, we have the following result on the transition from Σ^{in} to $\overline{\Delta}_-$.

Proposition 2.3. *Let $(z^{\text{in}}, w^{\text{in}}) \in \Sigma^{\text{in}}$. Then, for $\varepsilon > 0$ sufficiently small,*

$$(h^-, \bar{w}^-) := \Pi^{\text{in}}(z^{\text{in}}, w^{\text{in}}) = \Big(\sqrt{\varepsilon} d^-_{\sqrt{\varepsilon}} + \frac{w^{\text{in}}}{\sqrt{\varepsilon}} d^-_{\bar{w}} + \mathcal{O}\big((\sqrt{\varepsilon} + w^{\text{in}})^2\big),\ \frac{w^{\text{in}}}{\sqrt{\varepsilon}} + w^{\text{in}} f_2 \mu \sqrt{\varepsilon}\ln\varepsilon + \mathcal{O}(\sqrt{\varepsilon})\Big), \tag{2.20}$$

where $d^-_{\sqrt{\varepsilon}}$ and $d^-_{\bar{w}}$ are defined by

$$d^-_{\sqrt{\varepsilon}} = \int_{-\infty}^0 \nabla H(\bar{\gamma}_0^0(t)) \cdot \big(f_3 \bar{v}_0^0(t)^3, 0\big)^T\, dt \tag{2.21}$$

and

$$d^-_{\bar{w}} = \int_{-\infty}^0 \nabla H(\bar{\gamma}_0^0(t)) \cdot (0, -1)^T\, dt, \tag{2.22}$$

respectively (see (2.10) and (2.11)), and $\bar{\gamma}_0^0(t) = (\bar{v}_0^0, \bar{z}_0^0)(t)$, as in (2.6).

Remark 4. Since $w^{\text{in}} = \mathcal{O}(\varepsilon)$ by Assumption 1, it follows that $\frac{w^{\text{in}}}{\sqrt{\varepsilon}}$ in (2.20) remains bounded as $\varepsilon \to 0$.

Proof. We first analyze the transition from Σ^{in} to Δ^{in}. To that end, we desingularize the reduced problem associated with (1.5) following the ideas in [2]; see also the derivation of (2.44) in section 2.4. First, we approximate z by $f(v)$; i.e., we restrict ourselves to the critical manifold $\mathcal{S}_0^{a-}$ to leading order. The resulting "reduced" problem for (2.1) has the form

$$f'(v)\dot{v} = v - w, \tag{2.23a}$$

$$\dot{w} = \varepsilon\big(\mu - g_1 f(v)\big). \tag{2.23b}$$

(Note that this approximation is reasonable due to the form of (2.23b): since ε multiplies the entire right-hand side in (2.23b), the $\mathcal{O}(\varepsilon)$-correction to $z = f(v)$ will be $\mathcal{O}(\varepsilon^2)$ for the dynamics.) The desingularized version of (2.23) is obtained by multiplying the right-hand sides by $-f'(v) = -(2f_2v + 3f_3v^2)$:

$$\dot{v} = -(v - w), \tag{2.24a}$$

$$\dot{w} = -\varepsilon\big(\mu - g_1 f(v)\big) f'(v). \tag{2.24b}$$

We now introduce a new variable

$$W = \frac{w}{v}$$

in (2.24). (The introduction of W corresponds to a projectivization of the vector field in (2.24) that desingularizes the dynamics close to the origin.) After the transformation to the variables (v, W), system (2.24) becomes

$$\dot{v} = -v(1 - W), \tag{2.25a}$$

$$\dot{W} = W(1 - W) - \varepsilon\big(\mu - g_1(f_2v^2 + f_3v^3)\big)(2f_2 + 3f_3v). \tag{2.25b}$$

Since we are not interested in the (time-parametrized) solutions of (2.25) but only in the corresponding orbits, we can rescale time by dividing out a factor of $1 - W$ from both right-hand sides in (2.25). Moreover, since we consider $v \in [-\rho, -\alpha\sqrt{\varepsilon}]$ (by the definition of Σ^{in} and Δ^{in}) and $w = \mathcal{O}(\varepsilon)$ (see Assumption 1), W is small. Hence, we can expand $(1 - W)^{-1} = 1 + W + \mathcal{O}(W^2)$ and neglect terms of second order and upward in (v, W) in (2.25b), approximating the resulting equations by

$$\frac{dv}{d\tilde{t}} = -v, \tag{2.26a}$$

$$\frac{dW}{d\tilde{t}} = (1 - 2f_2\mu\varepsilon)W - 2f_2\mu\varepsilon - 3f_3\mu\varepsilon v. \tag{2.26b}$$

(Here, $\tilde{t}$ denotes the new rescaled time.)

Let $\widetilde{T}$ be the transition time from Σ^{in} to Δ^{in} under the flow of (2.26), and recall that $v = -\rho$ in Σ^{in} and $v = -\alpha\sqrt{\varepsilon}$ in Δ^{in}, respectively. Then, a simple computation using (2.26a) shows that $\widetilde{T}$ satisfies the identity

$$\mathrm{e}^{\widetilde{T}} = \frac{\rho}{\alpha}\frac{1}{\sqrt{\varepsilon}}. \tag{2.27}$$

(In particular, (2.27) implies that $\widetilde{T}$ depends only on α, ρ, and ε but not on the specific choice of trajectory in (2.26).) By a direct integration of (2.26b), it follows with $\varepsilon v(\widetilde{T}) = -\varepsilon\rho\mathrm{e}^{-\widetilde{T}} = \mathcal{O}(\varepsilon\sqrt{\varepsilon})$ that

$$W(\widetilde{T}) = (W^{\text{in}} - 2f_2\mu\varepsilon)\mathrm{e}^{(1-2f_2\mu\varepsilon)\widetilde{T}} + 2f_2\mu\varepsilon + \mathcal{O}(\varepsilon\sqrt{\varepsilon}), \tag{2.28}$$

where $W^{\text{in}} = -\frac{w^{\text{in}}}{\rho}$ is the value of W in Σ^{in}. The geometry of (2.26) is illustrated in Figure 7.

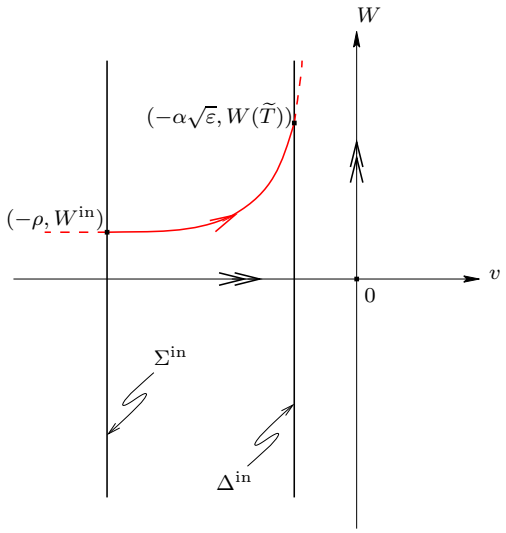


Figure 7. *The geometry of system* (2.26).

Now, note that $w = -\alpha\sqrt{\varepsilon}W(\widetilde{T})$ holds in Δ^{in} for the w-value corresponding to $W(\widetilde{T})$. Hence, expanding the exponential in (2.28), we obtain

$$w(T) = w^{\text{in}} + w^{\text{in}} f_2 \mu\varepsilon \ln\varepsilon + \mathcal{O}(\varepsilon), \tag{2.29}$$

where T denotes the transition time from Σ^{in} to Δ^{in} in the original system (2.24).

To complete the proof, we have to describe the second part of the transition, from Δ^{in} to $\overline{\Delta}_-$. To that end, we slightly modify the ideas of section 2.2. Recall the rescaled equations in (2.3), as well as the singular version obtained for $\varepsilon = 0$ (cf. (2.4)) and the parametrization of the $\bar{z}$-coordinate therein by h. (For $\bar{z}$ fixed, the corresponding (unique) value of h is determined from $H(0, \bar{z}) = h$; cf. (2.5).) Also recall that, for $h = 0$, there exists a parabolic level curve for H which corresponds to the special (singular canard) solution $\bar{\gamma}_0^0$ to (2.4) and which acts as a separatrix between the closed level curves (where $h > 0$) and the open ones (with $h < 0$).

Let $\overline{\Pi}^{\text{in}}$ denote the transition map from $\overline{\Delta}^{\text{in}}$ to $\overline{\Delta}_-$, and let $(\bar{z}, \bar{w}) \in \overline{\Delta}^{\text{in}}$. Since we are interested in describing the dynamics close to $\mathcal{S}_0^{a-}$, we may assume that $(-\alpha, \bar{z}, \bar{w})$ is the endpoint of a trajectory originating in $\mathcal{S}_\varepsilon^{a-}$. We claim that

$$(h^-, \bar{w}^-) = \overline{\Pi}^{\text{in}}(\bar{z}, \bar{w}) = \big(\sqrt{\varepsilon} d^-_{\sqrt{\varepsilon}} + \bar{w} d^-_{\bar{w}} + \mathcal{O}(2),\ \bar{w} + 2\alpha f_2 \mu\varepsilon + \mathcal{O}(\varepsilon^2)\big), \tag{2.30}$$

where $d^-_{\sqrt{\varepsilon}}$ and $d^-_{\bar{w}}$ are defined as in (2.21) and (2.22), respectively, and $\mathcal{O}(2) = \mathcal{O}((\sqrt{\varepsilon} + \bar{w})^2)$, as before.

To derive the expression for $\bar{w}^-$ in (2.30), we simply integrate the $\bar{w}$-equation in (2.3) to obtain $\bar{w}^- = \bar{w} + \varepsilon\mu\overline{T}^{\text{in}} + \mathcal{O}(\varepsilon^2)$, where $\overline{T}^{\text{in}}$ denotes the transition time from $\overline{\Delta}^{\text{in}}$ to $\overline{\Delta}$ in (2.4). Then, by integrating (2.4) directly from $\bar{v} = -\alpha$ to $\bar{v} = 0$ along $\bar{\gamma}_0^0$, we find $\overline{T}^{\text{in}} = 2\alpha f_2$.

The expression for h^- is obtained from the near-integrability of (2.3) and from the analysis in [20]; see also the proof of Proposition 2.2. More specifically, the condition for $(\bar{v}, \bar{z}, \bar{w})$ to

be on a trajectory originating in $\mathcal{S}_\varepsilon^{a-}$ is

$$h^- = \sqrt{\varepsilon} d^-_{\sqrt{\varepsilon}} + \bar{w} d^-_{\bar{w}} + \mathcal{O}(2), \tag{2.31}$$

which proves (2.30). (Here, the limits of integration in the definition of $d^-_{\sqrt{\varepsilon}}$ and $d^-_{\bar{w}}$ follow from the fact that $T^h_+(0) \to \infty$ as $h \to 0$.)

Finally, the assertion of the proposition follows by combining (2.29) and (2.30), taking into account that $\bar{w} = \frac{w}{\sqrt{\varepsilon}}$. ■

Remark 5. Note that to the order considered here, the definition of the intermediate section $\overline{\Delta}^{\text{in}}$ does not influence the asymptotics of $\overline{\Pi}^{\text{in}}$, as expected.

Proposition 2.3 has the following important implication: recall the set $\mathcal{C}_\varepsilon^- \subset \overline{\Delta}_-$ consisting of the endpoints of trajectories starting in $\mathcal{S}_\varepsilon^{a-}$. Then, it follows from (2.31) that $\mathcal{C}_\varepsilon^-$ can be represented as the graph of a function $h^-(\bar{w}, \sqrt{\varepsilon})$ satisfying

$$h^-(\bar{w}, \sqrt{\varepsilon}) = \sqrt{\varepsilon} d^-_{\sqrt{\varepsilon}} + \bar{w} d^-_{\bar{w}} + \mathcal{O}(2). \tag{2.32}$$

In analogy to (2.21) and (2.22), one can define the coefficients

$$d^+_{\sqrt{\varepsilon}} = -\int_0^\infty \nabla H(\bar{\gamma}_0^0(t)) \cdot \left(f_3 \bar{v}_0^0(t)^3, 0\right)^T dt \tag{2.33}$$

and

$$d^+_{\bar{w}} = -\int_0^\infty \nabla H(\bar{\gamma}_0^0(t)) \cdot (0, -1)^T dt \tag{2.34}$$

to describe the leading-order dynamics on $\mathcal{S}_\varepsilon^r$. Hence, it follows that the set $\mathcal{C}_\varepsilon^+$ can also be represented as the graph of a function $h^+(\bar{w}, \sqrt{\varepsilon})$ satisfying

$$h^+(\bar{w}, \sqrt{\varepsilon}) = \sqrt{\varepsilon} d^+_{\sqrt{\varepsilon}} + \bar{w} d^+_{\bar{w}} + \mathcal{O}(2). \tag{2.35}$$

Note that $d^\pm_{\sqrt{\varepsilon}} = \mp\frac{1}{2} d^0_{\sqrt{\varepsilon}}$ and, similarly, $d^\pm_{\bar{w}} = \mp\frac{1}{2} d^0_{\bar{w}}$ by symmetry, where $d^0_{\sqrt{\varepsilon}}$ and $d^0_{\bar{w}}$ are defined in (2.18).

Given the above representation of $\mathcal{C}_\varepsilon^\mp$, we make the following observations:

(i) Due to $d^-_{\bar{w}} < 0$ and $d^+_{\bar{w}} > 0$, (2.32) and (2.35) imply that $\mathcal{C}_\varepsilon^-$ and $\mathcal{C}_\varepsilon^+$ intersect transversely for $\bar{w} = \bar{w}^{\text{c}}$, with $\bar{w}^{\text{c}}$ as in (2.19). Hence, the strong canard Γ_ε^0 is indeed well defined; recall the discussion in section 1. In particular, the resulting geometry justifies the heuristic picture sketched in Figure 4; cf. Figure 8.

(ii) Similarly, the representations in (2.32) and (2.35) will be used in the definition of secondary canards Γ_ε^j for $j \geq 1$ as the transverse intersection of subsequent iterates of $\mathcal{C}_\varepsilon^-$ under $\overline{\Pi}$ with $\mathcal{C}_\varepsilon^+$; see section 3.3 for details.

2.4. The transition from $\overline{\Delta}_-$ to Σ^{out}. We now discuss the behavior of trajectories that exit the fold region in the direction of positive v and that then undergo relaxation. We begin by making a change of coordinates which transforms $\mathcal{C}_\varepsilon^+$ to the plane $\bar{z} = \bar{z}^0$, where $\bar{z}^0$ denotes the $\bar{z}$-value corresponding to $h = 0$ in (2.3). To that end, we define

$$\Delta\bar{z}(\bar{w}, \sqrt{\varepsilon}) = \bar{z}^0 - \bar{z}^{h^+(\bar{w}, \sqrt{\varepsilon})}, \tag{2.36}$$

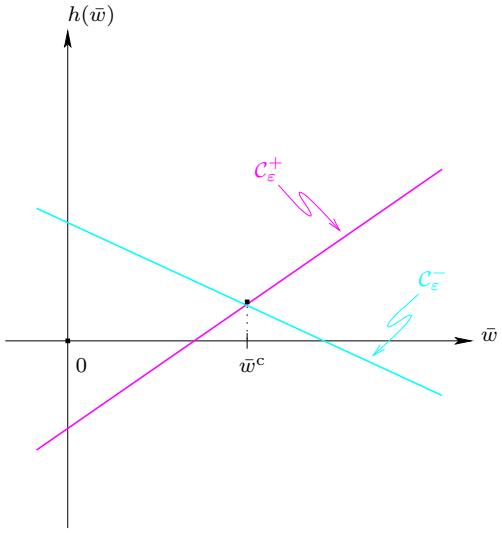


Figure 8. *The curves $\mathcal{C}_\varepsilon^-$ and $\mathcal{C}_\varepsilon^+$.*

where h^+ is as in (2.35), and we let

$$\tilde{z} = \bar{z} + \Delta\bar{z}(\bar{w}, \sqrt{\varepsilon}). \tag{2.37}$$

The transformation in (2.37) is introduced to "flatten" the repelling sheet $\mathcal{S}_\varepsilon^r$ of $\mathcal{S}_\varepsilon$ in Δ for $\varepsilon > 0$ sufficiently small: by (2.36), the $\bar{z}$-value corresponding to h^+, $\bar{z}^{h^+}$, is transformed into $\bar{z}^{h^+} + \bar{z}^0 - \bar{z}^{h^+} = \bar{z}^0$; hence, $\mathcal{C}_\varepsilon^+$ is represented as the graph of the zero function after the transformation:

$$\mathcal{C}_\varepsilon^+ = \big\{(0, \bar{w}) \,\big|\, \bar{w} = \mathcal{O}(\sqrt{\varepsilon})\big\}. \tag{2.38}$$

Recall that in the singular limit of $\varepsilon = 0 = \bar{w}$, $h = 0$ separates the small-oscillation regime in (2.3), where $h > 0$, from the relaxation regime (with $h < 0$); see Proposition 2.1. By introducing $\tilde{z}$, as defined in (2.37), we extend this characterization to the case where ε (and, hence, also $\bar{w}$) is positive but small: given (2.38), trajectories with $h < 0$ will end up "below" $\mathcal{S}_\varepsilon^r$ in Δ^-, implying that they will leave the fold region and undergo relaxation; trajectories with $h > 0$, on the other hand, will remain trapped "above" $\mathcal{S}_\varepsilon^r$ and will therefore stay in the small-oscillation regime close to ℓ^-. (This fact will simplify the following analysis and, in particular, the study of secondary canards in section 3.3, since it will facilitate the evaluation of the conditions that define these canard trajectories.)

In analogy to h^+, the function h^- in (2.32) is mapped to

$$\begin{aligned} h^0(\bar{w}) \equiv h^0(\bar{w}, \sqrt{\varepsilon}) &= h^-(\bar{w}, \sqrt{\varepsilon}) - h^+(\bar{w}, \sqrt{\varepsilon}) \\ &= \sqrt{\varepsilon}\big(d^-_{\sqrt{\varepsilon}} - d^+_{\sqrt{\varepsilon}}\big) + \bar{w}(d^-_{\bar{w}} - d^+_{\bar{w}}) + \mathcal{O}(2) \\ &= \sqrt{\varepsilon} d^0_{\sqrt{\varepsilon}} + \bar{w} d^0_{\bar{w}} + \mathcal{O}(2) \end{aligned} \tag{2.39}$$

by (2.37), where we suppress the $\sqrt{\varepsilon}$-dependence of h^0 for brevity. Hence, after performing

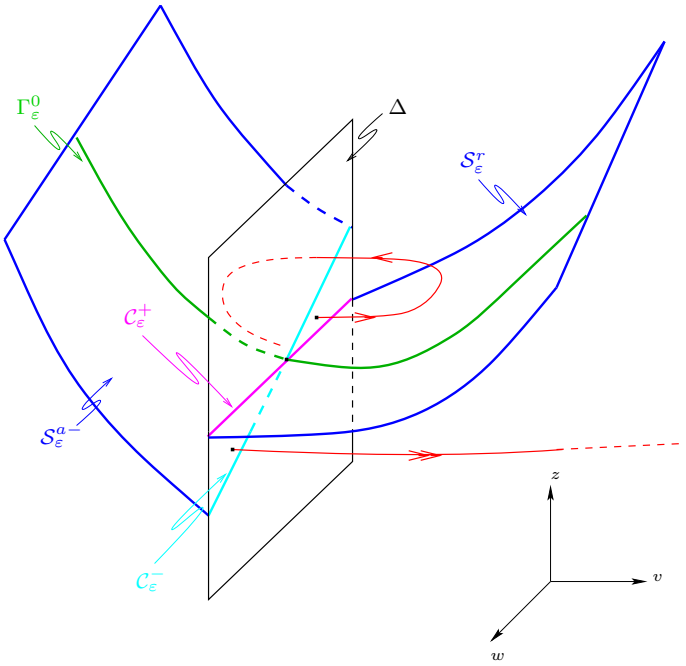


Figure 9. *The curves $\mathcal{C}_\varepsilon^-$ and $\mathcal{C}_\varepsilon^+$ after the transformation in (2.35).*

the coordinate transformation in (2.37), we find that $\mathcal{C}_\varepsilon^-$ is given by

$$\mathcal{C}_\varepsilon^- = \left\{(h^0(\bar{w}), \bar{w}) \;\middle|\; \bar{w} = \mathcal{O}(\sqrt{\varepsilon})\right\}. \tag{2.40}$$

The situation is illustrated in Figure 9; note the change from Figure 4 in that $\mathcal{C}_\varepsilon^+$ is now parallel to the w-axis, with $\mathcal{C}_\varepsilon^-$ "tilted" accordingly.

Next, we note that the higher-order terms that are introduced into (2.3) by the transformation in (2.37) are precisely of the form $\mathcal{O}(\bar{w}, \sqrt{\varepsilon})$. Hence, the resulting, transformed system is of the form (2.8), and the results of Proposition 2.2 can be applied directly to it.

Finally, in analogy to the transition times $T^h(\bar{w})$ defined for $h > 0$ above, we now define

$$T^{h,\mathrm{out}}(\bar{w}) = -T_-^{-h}(\bar{w}) \tag{2.41}$$

for $h < 0$. We have the following result on the transition from $\overline{\Delta}_-$ to Σ^{out}.

Proposition 2.4. *Let $(h, \bar{w}) \in \overline{\Delta}_-$ with $h < 0$ and $h = \mathcal{O}(\varepsilon^M)$ for some $M > 0$ and $\varepsilon > 0$ sufficiently small. Then,*

$$(z^{\mathrm{out}}, w^{\mathrm{out}}) := \Pi^{\mathrm{out}}(h, \bar{w}) = \left(\varepsilon \tilde{z}^{\mathrm{out}} + \mathcal{O}(\varepsilon \ln \varepsilon),\ \sqrt{\varepsilon}\bar{w} + \varepsilon\sqrt{\varepsilon} T^{h,\mathrm{out}}(\bar{w})\mu + \mathcal{O}(\varepsilon\sqrt{\varepsilon})\right), \tag{2.42}$$

where $\tilde{z}^{\mathrm{out}}$ is the $\tilde{z}$-value corresponding to $h^{\mathrm{out}} = h + \sqrt{\varepsilon} d_{\sqrt{\varepsilon}}^{\mathrm{out}} + \bar{w} d_{\bar{w}}^{\mathrm{out}}$, with $d_{\sqrt{\varepsilon}}^{\mathrm{out}}$ and $d_{\bar{w}}^{\mathrm{out}}$ defined by

$$d_{\sqrt{\varepsilon}}^{\mathrm{out}} = -\int_0^{T^{h,\mathrm{out}}(\bar{w})} \nabla H(\bar{\gamma}_0^h(t)) \cdot \left(f_3 \bar{v}_0^h(t)^3, 0\right)^T dt$$

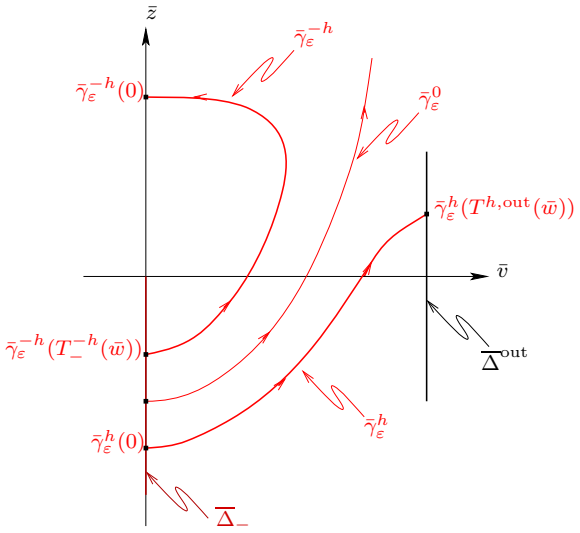


Figure 10. *The definition of $T^{h,\mathrm{out}}(\bar{w})$ for $h < 0$.*

and

$$d_{\bar{w}}^{\mathrm{out}} = -\int_0^{T^{h,\mathrm{out}}(\bar{w})} \nabla H(\bar{\gamma}_0^h(t)) \cdot (0,-1)^T \, dt,$$

respectively (see (2.10) and (2.11)).

Proof. For $(h,\bar{w}) \in \overline{\Delta}_-$ with $h < 0$ and $h = \mathcal{O}(\varepsilon^M)$, let $\overline{\Pi}^{\mathrm{out}}$ denote the time-$T^{h,\mathrm{out}}(\bar{w})$ transition map for (2.8), i.e., for the system obtained from (2.3) after the transformation to $\tilde{z}$. Moreover, let $\overline{\Delta}^{\mathrm{out}} := \overline{\Pi}^{\mathrm{out}}(\overline{\Delta}_-)$, which implies that the definition of the intermediate section $\overline{\Delta}^{\mathrm{out}}$ is now "implicit" ($\bar{w}$-dependent); cf. Figure 10. Then, it follows as in the proof of Proposition 2.2 that

$$(h^{\mathrm{out}}, \bar{w}^{\mathrm{out}}) := \overline{\Pi}^{\mathrm{out}}(h,\bar{w}) = \big(h + \sqrt{\varepsilon} d_{\sqrt{\varepsilon}}^{\mathrm{out}} + \bar{w} d_{\bar{w}}^{\mathrm{out}} + \mathcal{O}(2),\ \bar{w} + \varepsilon T^{h,\mathrm{out}}(\bar{w})\mu + \mathcal{O}(\varepsilon\sqrt{\varepsilon})\big), \tag{2.43}$$

where again $\mathcal{O}(2) = \mathcal{O}((\sqrt{\varepsilon}+\bar{w})^2)$, $T^{h,\mathrm{out}}(\bar{w})$ is, by the definition of $\overline{\Delta}^{\mathrm{out}}$, the transition time from $\overline{\Delta}$ to $\overline{\Delta}^{\mathrm{out}}$ in (2.8), and $d_{\sqrt{\varepsilon}}^{\mathrm{out}}$ and $d_{\bar{w}}^{\mathrm{out}}$ are defined as above.

To study the second part of the transition, from $\overline{\Delta}^{\mathrm{out}}$ to Σ^{out}, we introduce a new variable Z in the original (unmodified) system (1.5), where Z is defined by $z = v^2 Z$. This transformation serves to desingularize (1.5) close to the origin for v positive and small: in terms of (v,Z,w), (1.5) becomes

$$v' = v^2(-Z + f_2 + f_3 v), \tag{2.44a}$$

$$Z' = -2Zv(-Z + f_2 + f_3 v) + \frac{\varepsilon}{v}\Big(1 - \frac{w}{v}\Big), \tag{2.44b}$$

$$w' = \varepsilon^2(\mu - g_1 v^2 Z). \tag{2.44c}$$

Now, let $\Psi(v,Z) = v^2(-Z + f_2 + f_3 v)$; then, dividing the right-hand sides of (2.44) by $\Psi(v,Z)$,

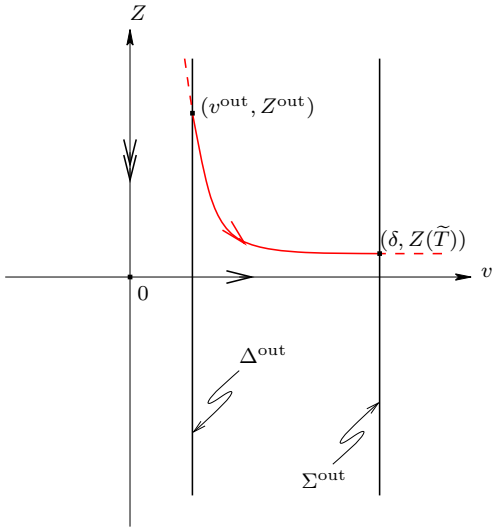


Figure 11. *The geometry of system* (2.46).

we find

$$\frac{dv}{d\hat{t}} = 1, \tag{2.45a}$$

$$\frac{dZ}{d\hat{t}} = -\frac{2}{v}Z + \frac{\varepsilon}{v\Psi(v,Z)}\Big(1 - \frac{w}{v}\Big), \tag{2.45b}$$

$$\frac{dw}{d\hat{t}} = \frac{\varepsilon^2}{\Psi(v,Z)}(\mu - g_1 v^2 Z); \tag{2.45c}$$

here, $\hat{t}$ denotes the new, rescaled time.

We first investigate the dynamics of Z in the transition. Let Δ^{out} denote the section in (v, Z, w)-space corresponding to $\overline{\Delta}^{\text{out}}$. Given an initial v-value v^{out} for (1.5) in Δ^{out}, it then follows that $v^{\text{out}} = \mathcal{O}(\sqrt{\varepsilon}T^{h,\text{out}}(\bar{w})) = \mathcal{O}(\sqrt{-\varepsilon \ln \varepsilon})$ must hold, which, together with (2.45a) and $w = \mathcal{O}(\varepsilon)$ (see Assumption 1), implies that $\frac{w}{v}$ is small throughout. Since, moreover, $\frac{dw}{d\hat{t}} = \mathcal{O}(\varepsilon(\ln \varepsilon)^{-1})$ by (2.45c), w remains almost constant, and we can neglect its evolution.

Hence, expanding Ψ in (2.45b) and truncating the resulting equation, we find that to leading order,

$$\begin{aligned} \frac{dv}{d\hat{t}} &= 1, \\ \frac{dZ}{d\hat{t}} &= -\frac{2}{v}Z + \frac{\varepsilon}{f_2 v^3}\big(1 + \mathcal{O}(v,Z)\big). \end{aligned} \tag{2.46}$$

The transition from Δ^{out} to Σ^{out} under the flow of (2.46) is illustrated in Figure 11. Now, for $(v^{\text{out}}, Z^{\text{out}}) \in \Delta^{\text{out}}$, we can solve (2.46) explicitly to leading order by variation of constants, which gives

$$Z(v) = \frac{(v^{\text{out}})^2 Z^{\text{out}}}{v^2} + \frac{\varepsilon}{f_2 v^2} \ln \frac{v}{v^{\text{out}}} + \mathcal{O}(\varepsilon). \tag{2.47}$$

Here, we have neglected the effect of the inhomogeneous $\mathcal{O}(v, Z)$-terms in (2.46), since it can be shown that these contribute only terms of order $\mathcal{O}(\varepsilon)$ in (2.47). Now, the corresponding expression in the original variable z is given by

$$z(v) = \varepsilon \tilde{z}^{\text{out}} + \frac{\varepsilon}{f_2} \ln \frac{v}{v^{\text{out}}} + \mathcal{O}(\varepsilon)$$

for $\tilde{z}^{\text{out}}$ in $\overline{\Delta}^{\text{out}}$, where we have used $v^2 Z = z = \varepsilon \tilde{z}$. Recalling that $v^{\text{out}} = \mathcal{O}(\sqrt{-\varepsilon \ln \varepsilon})$ as well as that $v = \delta$ in Σ^{out}, we find

$$z(T) = \varepsilon \tilde{z}^{\text{out}} + \mathcal{O}(\varepsilon \ln \varepsilon); \tag{2.48}$$

here, T denotes the transition time from Δ^{out} to Σ^{out}.

We now use the estimate for $z(T)$ in (2.48) to derive an estimate for $w(T)$. Since $dv = \Psi(v, Z)\, dt$ (see (2.44a)) and since, moreover, $Z = \mathcal{O}(1)$, there certainly holds $\frac{1}{2}\Psi(v, 0) \leq \Psi(v, Z)$. Hence, it follows that T satisfies the inequality

$$T \leq 2 \int_{v^{\text{out}}}^{\delta} \frac{dv}{\Psi(v, 0)}, \tag{2.49}$$

to leading order. The integral on the right-hand side of (2.49) can be evaluated explicitly, giving

$$T \leq \frac{2}{f_2 v^{\text{out}}} - \frac{f_3}{f_2^2} \ln \varepsilon + \mathcal{O}(1).$$

Integrating the w-equation (2.44c) directly and taking into account (2.43) as well as $v^{\text{out}} = \mathcal{O}(\sqrt{-\varepsilon \ln \varepsilon})$ and $w = \sqrt{\varepsilon}\bar{w}$, we obtain

$$\begin{aligned} w(T) &= \sqrt{\varepsilon}\bar{w}^{\text{out}} + \varepsilon^2 \left(\mu T - g_1 \int_{v^{\text{out}}}^{\delta} \frac{z(v)}{\Psi(v, 0)}\, dv \right) + \mathcal{O}(\varepsilon^3) \\ &= \sqrt{\varepsilon}\bar{w} + \varepsilon\sqrt{\varepsilon} T^{h,\text{out}}(\bar{w})\mu + \mathcal{O}(\varepsilon\sqrt{\varepsilon}). \end{aligned} \tag{2.50}$$

To complete the proof, it remains to collect the above estimates: with $\tilde{z}^{\text{out}}$ the $\tilde{z}$-value corresponding to h^{out} (see (2.43)), we find the desired expression for z^{out} in (2.42) from (2.48). The estimate for the w^{out}-component of Π^{out} follows directly from (2.50). ■

2.5. The global return mechanism. In this subsection, we describe the global mechanism that determines the return of trajectories of (1.5) from Σ^{out} back to Σ^{in}. The corresponding return map will be denoted by Π^{ret}. Since the necessary analysis is largely based on standard geometric singular perturbation (Fenichel) theory [11], we do not discuss it in full detail here; moreover, for the sake of exposition, we will make a number of additional, simplifying assumptions throughout this subsection. As it turns out, the resulting leading-order asymptotics of Π^{ret} will still give an approximation for the composite return map Π that is consistent to the order considered here; cf. section 4 below.

In a first approximation, we may assume that $z = f(v)$ is satisfied; i.e., for $\varepsilon > 0$ sufficiently small, we may restrict ourselves to the singular dynamics of (1.5) on $\mathcal{S}_0$. We recall the

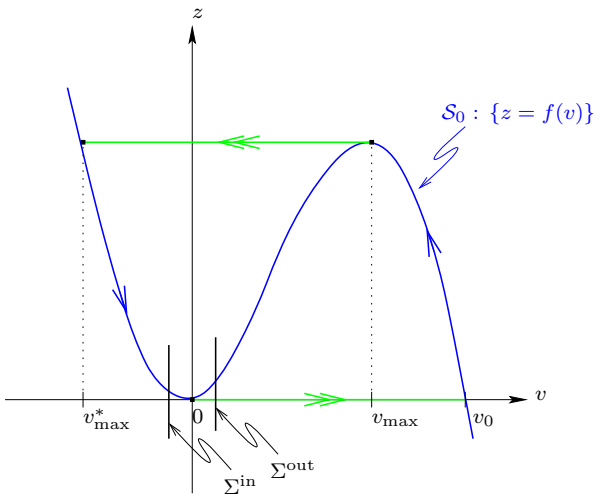


Figure 12. *The geometry of the global return mechanism.*

definition of the corresponding reduced system from (2.24):

$$\dot{v} = -(v - w), \tag{2.51a}$$

$$\dot{w} = -\varepsilon\big(\mu - g_1 f(v)\big) f'(v). \tag{2.51b}$$

Moreover, we can safely neglect the w-term on the right-hand side of (2.51a), since this term is assumed to be small throughout; see Assumption 1. Then, we rewrite (2.51) with v as the independent variable; i.e., we divide (2.51b) by (2.51a), which gives

$$\frac{dw}{dv} = \varepsilon\big(\mu - g_1 f(v)\big)\frac{f'(v)}{v}. \tag{2.52}$$

Given an initial v-value v^* on $\mathcal{S}_0$, (2.52) can be integrated explicitly as follows:

$$w(v) - w(v^*) = \varepsilon \mathcal{G}(v^*, v, \mu) := \varepsilon \int_{v^*}^{v} \big(\mu - g_1 f(\sigma)\big)\frac{f'(\sigma)}{\sigma}\, d\sigma. \tag{2.53}$$

To describe the return of trajectories from Σ^{out} to Σ^{in} under the flow of (2.52) on $\mathcal{S}_0$, we need to consider two separate parts of the transition, namely, the parts where v evolves along $\mathcal{S}_0^{a+}$ and $\mathcal{S}_0^{a-}$, respectively. (Note that by restricting ourselves to the slow flow on $\mathcal{S}_0$, we are implicitly neglecting the transition from ℓ^- to $\mathcal{S}_0^{a+}$ and from ℓ^+ to $\mathcal{S}_0^{a-}$, respectively, under the fast flow of (1.5), since, by standard Fenichel theory [11], the corresponding contributions to Π^{ret} are of higher order; cf. Figure 12.) The relevant integrals in (2.53) are given by

$$\mathcal{G}(v_0, v_{\max}, \mu) = \int_{v_0}^{v_{\max}} \big(\mu - g_1 f(\sigma)\big)\frac{f'(\sigma)}{\sigma}\, d\sigma$$

and

$$\mathcal{G}(v^*_{\max}, -\rho, \mu) = \int_{v^*_{\max}}^{-\rho} \big(\mu - g_1 f(\sigma)\big)\frac{f'(\sigma)}{\sigma}\, d\sigma,$$

respectively. Here, $v_{\max}$ is the value of v for which f attains its local maximum, $v^*_{\max} < 0$ is defined by the requirement that $f(v^*_{\max}) = f(v_{\max})$, and $v_0 > 0$ is the second (nontrivial) zero of f, with $f(v_0) = 0$; see again Figure 12. To facilitate further the evaluation of these integrals, we will approximate $\mathcal{G}(v^*_{\max}, -\rho, \mu)$ by $\mathcal{G}(v^*_{\max}, 0, \mu)$; i.e., we will evaluate the integral over $\mathcal{S}_0^{a-}$ down to and including ℓ^-. (In fact, a straightforward though lengthy computation shows that this approximation will offset precisely the part of the $\mathcal{O}(\sqrt{\varepsilon})$-error term in Π^{in} that is independent of w^{in}; cf. Proposition 2.3.)

Hence, in sum, it follows that the w-component $\hat{w}$ of $\Pi^{\text{ret}} : \Sigma^{\text{out}} \to \Sigma^{\text{in}}$ is given by

$$\hat{w} = w + \varepsilon\big(\mathcal{G}(v_0, v_{\max}, \mu) + \mathcal{G}(v^*_{\max}, 0, \mu)\big), \tag{2.54}$$

to lowest order. In particular, note that (2.54) determines the global "amount of return" of w after one relaxation cycle, expressed as a function of the parameter μ. (This fact will prove especially useful in section 3 below.) Let

$$D_\mu = \frac{d}{d\mu}\big(\mathcal{G}(v_0, v_{\max}, \mu) + \mathcal{G}(v^*_{\max}, 0, \mu)\big),$$

and observe that the rate of change of the return point with respect to μ is given by $D_\mu\varepsilon$. From the above, it follows that D_μ can easily be approximated to lowest order in terms of the function $\mathcal{G}$: by the definition of $\mathcal{G}$ and making use of the fact that

$$v_{\max} = -\frac{2f_2}{3f_3}, \quad v^*_{\max} = \frac{f_2}{3f_3}, \quad \text{and} \quad v_0 = -\frac{f_2}{f_3},$$

we obtain

$$\mathcal{G}(v_0, v_{\max}, \mu) + \mathcal{G}(v^*_{\max}, 0, \mu) = \frac{g_1}{18}\frac{f_2^5}{f_3^3} - \mu\frac{f_2^2}{f_3}. \tag{2.55}$$

Differentiating (2.55) with respect to μ, we find $D_\mu = -\frac{f_2^2}{f_3}$.

Similarly, the critical value μ^{c} of μ for which MMOs cease to exist in (1.5) is to leading order determined by requiring $\hat{w} = w$ in (2.54) or, alternatively, by finding μ such that (2.55) equals zero; again, a simple computation shows

$$\mu^{\text{c}} = \frac{g_1}{18}\frac{f_2^3}{f_3^2}. \tag{2.56}$$

For $\mu > \mu^{\text{c}}$, the dynamics of (1.5) is in the pure relaxation regime in the sense that the only admissible periodic trajectories are those with Farey sequence $\{L^0\}$.

Remark 6. Note that the fold line ℓ^+ will in general contribute logarithmic terms (in ε) to (2.56); see, e.g., [34]. In our case, however, these terms can be shown to be of higher order and are hence negligible.

2.6. Summary: The return map $\Pi : \overline{\Delta}_- \to \overline{\Delta}_-$. Given the analysis of the previous subsections, we can now define the composite return map $\Pi : \overline{\Delta}_- \to \overline{\Delta}_-$. We note that the definition of Π will depend on the sign of h: if $h > 0$, the corresponding trajectory of (1.5) will

remain in the fold region, i.e., in the small-oscillation regime, and undergo another "loop." Hence, the return to $\overline{\Delta}_-$ is described by $\overline{\Pi}$ in that case; cf. Proposition 2.2. If, on the other hand, $h < 0$, the trajectory will exit the fold region and undergo relaxation; i.e., it will leave $\overline{\Delta}_-$ in the direction of the fast flow of (1.5), move "up" the slow manifold $\mathcal{S}_\varepsilon^{a+}$ under the slow flow until it reaches ℓ^+, "jump" to $\mathcal{S}_\varepsilon^{a-}$, and move "down" that manifold until it re-enters a neighborhood of ℓ^-; cf., e.g., Figure 12. Therefore, the return to $\overline{\Delta}_-$ is described by the composition of Π^{out}, Π^{ret}, and Π^{in} in that case; see Propositions 2.3 and 2.4 as well as the discussion in section 2.5. Hence, in sum, the desired expression for Π is given as follows:

$$\Pi(h,\bar{w}) = \begin{cases} \overline{\Pi}(h,\bar{w}) & \text{if } h > 0, \\ \Pi^{\text{in}} \circ \Pi^{\text{ret}} \circ \Pi^{\text{out}}(h,\bar{w}) & \text{if } h < 0. \end{cases} \tag{2.57}$$

3. Partial dimension reduction for the map Π. In this section, we show how the two-dimensional return map Π formulated in section 2.6 can be accurately approximated by an appropriately defined one-dimensional map, which we denote by Φ. More precisely, we will prove that the resulting approximation error will be exponentially small in ε. The reduction itself is carried out in two steps: first, the map Π is restricted from the two-dimensional section $\overline{\Delta}_-$ to a union of one-dimensional curves $\cup\mathcal{C}_\varepsilon^j$, to be specified in section 3.1. In the second step, this restricted map is reduced further, in section 3.4, to a map Φ that is defined on the single curve $\mathcal{C}_\varepsilon^-$. For a detailed study of the dynamics of Φ, we require some preparatory analysis: in section 3.2, we approximate the derivative $\frac{d\Pi}{d\bar{w}}$, which, in turn, allows us to derive estimates for $\frac{d\Phi}{d\bar{w}}$ in section 3.5. The latter are needed for analyzing the contractive (or expansive) properties of the reduced flow under Φ. In section 3.3, we characterize the secondary canards introduced in section 1 above: we derive the defining conditions for these trajectories, and we use those conditions to describe the family of the associated sectors of rotation. Finally, in section 3.6, we study the dynamics of Φ on these sectors by combining the results of sections 3.3 and 3.5, and we derive precise asymptotic estimates for the bifurcation structure of the resulting mixed-mode dynamics in (1.5).

3.1. The curves $\mathcal{C}_\varepsilon^j$. In this subsection, we perform the first step in our exponentially accurate reduction of Π to a one-dimensional map Φ. More precisely, we show how Π can be restricted from $\overline{\Delta}_-$ to a union of one-dimensional curves $\cup\mathcal{C}_\varepsilon^j$ that will be defined below.

Recall the definition of the curves $\mathcal{C}_\varepsilon^-$ and $\mathcal{C}_\varepsilon^+$ from section 2.1, as well as the fact that $\mathcal{C}_\varepsilon^-$ can be represented as the graph of the function $h^0(\bar{w})$ defined in (2.39); see (2.40). For $j \geq 1$, we now make the inductive definition

$$\mathcal{C}_\varepsilon^j = \overline{\Pi}\big(\{(h,\bar{w}) \in \mathcal{C}_\varepsilon^{j-1} \mid h > 0\}\big),$$

where we define $\mathcal{C}_\varepsilon^0 \equiv \mathcal{C}_\varepsilon^-$ for the zeroth iterate of $\mathcal{C}_\varepsilon^-$ under $\overline{\Pi}$. Next, we show that for $j \geq 1$, each set $\mathcal{C}_\varepsilon^j$ can be written as the graph of a function $h^j(\bar{w})$, in analogy to the representation of $\mathcal{C}_\varepsilon^-$ given in (2.40). We first consider the case when $j = 1$. Note that by Proposition 2.2,

$$(h^1,\bar{w}^1) = \overline{\Pi}(h^0(\bar{w}),\bar{w}) = \big(h^0(\bar{w}) + \sqrt{\varepsilon}d_{\sqrt{\varepsilon}}^{h^0(\bar{w})} + \bar{w}d_{\bar{w}}^{h^0(\bar{w})} + \mathcal{O}(2),\ \bar{w} + \varepsilon\mu T^{h^0(\bar{w})} + \mathcal{O}(\varepsilon^2)\big),$$

where $\mathcal{O}(2) = \mathcal{O}((\sqrt{\varepsilon}+\bar{w})^2)$, as before. Since $h^0 = \sqrt{\varepsilon}d_{\sqrt{\varepsilon}}^0 + \bar{w}d_{\bar{w}}^0 + \mathcal{O}(2)$ by (2.39) and since

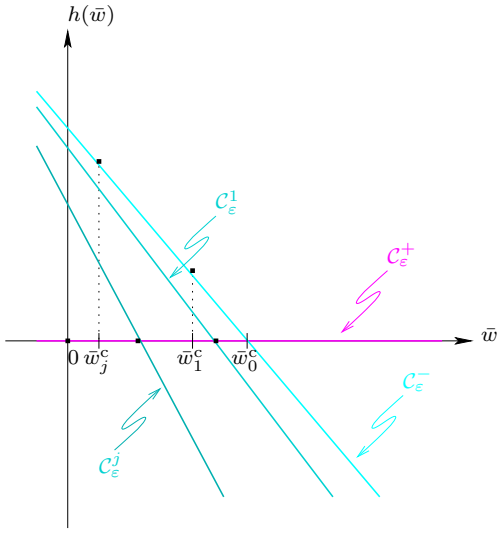


Figure 13. *The curves $\mathcal{C}_\varepsilon^j$ for $j \geq 0$, where $\mathcal{C}_\varepsilon^- \equiv \mathcal{C}_\varepsilon^0$.*

$d^{h^0}_{\sqrt{\varepsilon}} \sim d^0_{\sqrt{\varepsilon}}$ and $d^{h^0}_{\bar{w}} \sim d^0_{\bar{w}}$, respectively, it follows that

$$h^1(\bar{w}) = 2\sqrt{\varepsilon} d^0_{\sqrt{\varepsilon}} + 2\bar{w} d^0_{\bar{w}} + \mathcal{O}(2). \tag{3.1}$$

Similarly, for higher iterates of $\overline{\Pi}$, there holds

$$\begin{aligned}(h^j, \bar{w}^j) &= \overline{\Pi}^j(h^0(\bar{w}), \bar{w}) \\ &= \left(h^0(\bar{w}) + \sqrt{\varepsilon} \sum_{i=0}^{j} d^{h^i(\bar{w})}_{\sqrt{\varepsilon}} + \bar{w} \sum_{i=0}^{j} d^{h^i(\bar{w})}_{\bar{w}} + \mathcal{O}(\varepsilon),\ \bar{w} + 2\varepsilon\mu \sum_{i=0}^{j} T^{h^i(\bar{w})} + \mathcal{O}(\varepsilon^2) \right)\end{aligned} \tag{3.2}$$

and, therefore,

$$h^j(\bar{w}) = (j+1)\sqrt{\varepsilon} d^0_{\sqrt{\varepsilon}} + (j+1)\bar{w} d^0_{\bar{w}} + \mathcal{O}(2). \tag{3.3}$$

This gives the desired representation of $\mathcal{C}_\varepsilon^j$ as the graph of the function $h^j(\bar{w})$ in (3.3), with

$$\mathcal{C}_\varepsilon^j = \left\{ (h^j(\bar{w}), \bar{w}) \;\middle|\; \bar{w} = \mathcal{O}(\sqrt{\varepsilon}) \right\} \tag{3.4}$$

for $j \geq 1$; cf. Figure 13.

Finally, we prove that the map Π can be restricted from $\overline{\Delta}_-$ to the union of the set of curves $\mathcal{C}_\varepsilon^j$ with an only exponentially small error; here, Π^j denotes the jth iterate of the map Π defined in (2.57).

Proposition 3.1. *Let $(h, \bar{w}) \in \overline{\Delta}_-$, and fix $\varepsilon > 0$ sufficiently small. Then, there exists $k > 0$ such that, for $1 \leq j \leq k$, $\Pi^j(h, \bar{w})$ is exponentially close (in ε) to $\cup_{j=1}^k \mathcal{C}_\varepsilon^j$.*

Proof. First, observe that all trajectories must become exponentially close to $\mathcal{S}_\varepsilon^{a-}$ after relaxation; consequently, they must return to $\overline{\Delta}_-$ exponentially close to $\mathcal{C}_\varepsilon^-$. This is equivalent to saying that, for any $(h, \bar{w})$ with $h < 0$, $\Pi(h, \bar{w})$ is exponentially close to $\mathcal{C}_\varepsilon^-$.

We now prove that $\Pi^2(h,\bar{w})$ must be exponentially close to $\mathcal{C}_\varepsilon^- \cup \mathcal{C}_\varepsilon^1$. Let $(h^1,\bar{w}^1) = \Pi(h,\bar{w})$, and note that if $h^1 < 0$, the forward trajectory of $(h^1,\bar{w}^1)$ must undergo relaxation. Hence, by the above argument, $\Pi(h^1,\bar{w}^1) = \Pi^2(h,\bar{w})$ is exponentially close to $\mathcal{C}_\varepsilon^-$ in that case. Let us suppose that $h^1 \geq 0$ now and consider $h^1 = \mathcal{O}(\varepsilon)$ first, which is in the domain of $\overline{\Pi}$. Since the map $\overline{\Pi}$ is induced by the flow of (2.8) and since $T^{h^1}(\bar{w}^1) = \mathcal{O}(\sqrt{-\ln\varepsilon})$ for the return time to $\overline{\Delta}$ (cf. Appendix A), the expansion that can be incurred during that return is of at most algebraic order in ε. Consequently, $\Pi(h^1,\bar{w}^1)$ must be exponentially close to $\mathcal{C}_\varepsilon^1$. If, on the other hand, h^1 is exponentially small, i.e., if $\mathcal{O}(\mathrm{e}^{-\frac{\kappa}{\varepsilon}})$ for some $\kappa > 0$, the argument from the first part of the proof can be applied to show that $\Pi(h^1,\bar{w}^1)$ is again exponentially close to $\mathcal{C}_\varepsilon^-$.

The transitional regime between $h^1 = \mathcal{O}(\varepsilon)$ and exponentially small h^1 is more difficult to describe. This issue is addressed in detail in [20], where it is shown, roughly speaking, that the contraction and expansion in the z-direction cancel each other out to leading order near the fold. An analogous property can be proven to hold in our case, which allows us to conclude that $\Pi(h^1,\bar{w}^1)$ is exponentially close to $\mathcal{C}_\varepsilon^1$ even in that transitional regime. Finally, by an iteration of the above argument, it follows that $\Pi^3(h,\bar{w})$ must be exponentially close to $\mathcal{C}_\varepsilon^- \cup \mathcal{C}_\varepsilon^1 \cup \mathcal{C}_\varepsilon^2$, and so on.

To conclude the proof, we note that there exists a finite number k such that for any point $(h,\bar{w})$ with $h > 0$, there is $1 \leq j \leq k$ such that the h-coordinate of $\Pi^j(h,\bar{w})$ is negative, so that $\Pi^{j+1}(h,\bar{w})$ must again be close to $\mathcal{C}_\varepsilon^-$. (Note that k gives the maximum possible number of small oscillations a trajectory can undergo.) It follows that for any $(h,\bar{w}) \in \overline{\Delta}_-$, the trajectory of $(h,\bar{w})$ under Π must be exponentially close to the union of the sets $\mathcal{C}_\varepsilon^j$, $j = 1,\dots,k$. ∎

In the following, we will assume that the points on a trajectory of Π are on $\mathcal{C}_\varepsilon^-$ or on one of the curves $\mathcal{C}_\varepsilon^j$. By Proposition 3.1, this assumption incurs at most an exponentially small error. To find the restriction of Π to $\mathcal{C}_\varepsilon^j$, we recall that $\mathcal{C}_\varepsilon^j$ can be represented as the graph of a function $h^j(\bar{w})$; see (3.4). In analogy to the definition of Π in (2.42), we again have to distinguish between $h^j > 0$ and $h^j < 0$ here. In the former case, Π reduces to $\overline{\Pi}$, whereas in the latter case, we have to take the composition of Π^{out}, Π^{ret}, and Π^{in} to describe the return to $\cup\mathcal{C}_\varepsilon^j$; see the discussion in section 2.6 for details. Moreover, since $\mathcal{C}_\varepsilon^j$ is parametrized by $\bar{w}$ (cf. (3.4)), it is natural to consider Π as a function of $\bar{w}$. Hence, combining the definition of Π in (2.42) with (2.9) for $h^j > 0$ and with the estimates in (2.20), (2.42), and (2.54) for $h^j < 0$, respectively, we finally obtain

(3.5)
$$\Pi(\bar{w}) \equiv \Pi(h^j(\bar{w}),\bar{w}) = \begin{cases} \bar{w} + \varepsilon\mu T^{h^j(\bar{w})}(\bar{w}) + \mathcal{O}(\varepsilon^2) & \text{if } h^j(\bar{w}) > 0, \\ \bar{w} + \varepsilon\mu T^{h^j(\bar{w}),\mathrm{out}}(\bar{w}) + \bar{w} f_2 \mu\varepsilon \ln\varepsilon & \\ \quad + \sqrt{\varepsilon}\big(\mathcal{G}(0,v_{\max},\mu) + \mathcal{G}(v^*_{\max},v_0,\mu)\big) + \mathcal{O}(\varepsilon) & \text{if } h^j(\bar{w}) < 0. \end{cases}$$

3.2. The derivative of Π. To estimate the contractive (or expansive) properties of the flow induced by Π on $\cup\mathcal{C}_\varepsilon^j$, we need to estimate the derivative $\frac{d\Pi}{d\bar{w}}$ of Π. Given (3.5), it follows

that the following approximation holds to leading order, i.e., up to an $\mathcal{O}(\varepsilon)$-error:

$$\frac{d\Pi}{d\bar{w}} \sim \begin{cases} 1+\varepsilon\mu\dfrac{dT^{h^j(\bar{w})}(\bar{w})}{d\bar{w}} & \text{if } h^j(\bar{w})>0, \\ 1+\varepsilon\mu\dfrac{dT^{h^j(\bar{w}),\mathrm{out}}(\bar{w})}{d\bar{w}}+f_2\mu\varepsilon\ln\varepsilon & \text{if } h^j(\bar{w})<0. \end{cases} \tag{3.6}$$

(Here, we have used the fact that the function $\mathcal{G}$ is independent of $\bar{w}$; see (2.53).) Now, recall that $T^h(0)=2T^h$ by (2.7), and note that, for $(h,\bar{w})$ small, $T^h(\bar{w})$ depends much more sensitively on h than on $\bar{w}$. Therefore, to evaluate (3.6), we can in a first approximation neglect the $\bar{w}$-dependence of $T^h(\bar{w})$ and write

$$\frac{dT^{h^j(\bar{w})}(\bar{w})}{d\bar{w}} \sim 2\frac{dT^{h^j(\bar{w})}}{d\bar{w}}.$$

Due to $T^h\sim(-2\ln h)^{\frac{1}{2}}$ (see Appendix A), it follows that

$$\frac{dT^{h^j(\bar{w})}(\bar{w})}{d\bar{w}} \sim -\frac{2}{h^j(\bar{w})}\frac{1}{\sqrt{-2\ln h^j(\bar{w})}}(h^j)'(\bar{w}); \tag{3.7}$$

similarly, we can use the definition of $T^{h,\mathrm{out}}(\bar{w})$ in (2.41) to conclude

$$\frac{dT^{h^j(\bar{w}),\mathrm{out}}(\bar{w})}{d\bar{w}} \sim -\frac{1}{h^j(\bar{w})}\frac{1}{\sqrt{-2\ln h^j(\bar{w})}}(h^j)'(\bar{w}). \tag{3.8}$$

To complete the computation of the derivative of Π, we require approximate formulae for the derivatives of h^j with respect to $\bar{w}$: by (3.3), it follows that

$$(h^j)'(\bar{w}) = (j+1)d_{\bar{w}}^0+\mathcal{O}(\sqrt{\varepsilon},\bar{w}). \tag{3.9}$$

Combining (3.7) and (3.9), we finally obtain

$$\frac{dT^{h^j(\bar{w})}(\bar{w})}{d\bar{w}} \sim -\frac{2}{h^j(\bar{w})}\frac{1}{\sqrt{-2\ln h^j(\bar{w})}}(j+1)d_{\bar{w}}^0 \tag{3.10}$$

as well as

$$\frac{dT^{h^j(\bar{w}),\mathrm{out}}(\bar{w})}{d\bar{w}} \sim -\frac{1}{h^j(\bar{w})}\frac{1}{\sqrt{-2\ln h^j(\bar{w})}}(j+1)d_{\bar{w}}^0, \tag{3.11}$$

which can be substituted into (3.6) to obtain a more explicit expression for $\frac{d\Pi}{d\bar{w}}$.

3.3. Secondary canards and sectors of rotation. Recall the definition of the jth secondary canard Γ_ε^j as a trajectory of (1.2) that undergoes j small oscillations (loops) during its passage through the fold region. In this subsection, we derive the conditions on the rescaled equations (2.8) by which these trajectories are defined. The corresponding analysis will require us to refine the results of Proposition 2.2; see Proposition 3.2 below. Given the family

of secondary canards $\{\Gamma_\varepsilon^j\}$ for $j = 0, \dots, k$, we will define the corresponding family of sectors of rotation, $\{RS^j\}$. We will then analyze the geometry of these sectors; in particular, we will estimate the sector width (Proposition 3.3), and we will show that it is independent of j to lowest order. Here, we note that the family $\{RS^j\}$ will be crucial for the reduction of the (two-dimensional) map Π to the (one-dimensional) map Φ in section 3.4 below. Finally, in Proposition 3.4, we discuss the uniform validity of our asymptotic estimates.

Let $(h^0(\bar{w}), \bar{w}) \in \mathcal{C}_\varepsilon^-$, as before, and recall the definition of the transition map $\overline{\Pi}$ for (2.8); see Proposition 2.2. Moreover, let P_h and $P_{\bar{w}}$ denote the projections onto the h-coordinate and the $\bar{w}$-coordinate, respectively. Then, the defining condition for the jth secondary canard is given by

$$P_h \overline{\Pi}^j (h^0(\bar{w}), \bar{w}) = 0; \tag{3.12}$$

i.e., the h-coordinate of the jth iterate of $(h^0(\bar{w}), \bar{w})$ under $\overline{\Pi}$ has to be zero. In other words, we are interested in finding the points of intersection of subsequent iterates of $\mathcal{C}_\varepsilon^-$ under $\overline{\Pi}$ (i.e., of $\mathcal{C}_\varepsilon^j$) with $\mathcal{C}_\varepsilon^+$. For $j \geq 1$ fixed, let $\bar{w}_j^{\mathrm{c}}$ denote the corresponding solution of (3.12). Then, $\bar{w}_j^{\mathrm{c}}$ fixes a point in $\mathcal{C}_\varepsilon^-$ that will determine the location of the jth secondary canard Γ_ε^j; see Figure 13 for an illustration. In particular, for the first secondary canard, we have the requirement that

$$P_h \overline{\Pi}(h^0(\bar{w}_1^{\mathrm{c}}), \bar{w}_1^{\mathrm{c}}) = 0.$$

Remark 7. Recall that $\mathcal{C}_\varepsilon^-$ corresponds to the intersection of the locally invariant slow manifold $\mathcal{S}_\varepsilon^{a-}$ in (1.5) with Δ, before the rescaling. Since the critical manifold $\mathcal{S}_0$ for (1.5) is normally hyperbolic away from $\ell^\pm$, it follows that the slow manifold $\mathcal{S}_\varepsilon$ is unique up to exponentially small terms [11, 14]. Once the corresponding sheets of $\mathcal{S}_\varepsilon^{a-}$ and $\mathcal{S}_\varepsilon^r$ are chosen, the strong canard Γ_ε^0 is uniquely determined. Similarly, since the jth secondary canard Γ_ε^j, with $j \geq 1$, is defined as the trajectory lying in the intersection of the jth iterate of $\mathcal{S}_\varepsilon^{a-}$ under Π with $\mathcal{S}_\varepsilon^r$, all secondary canards will originate in the same sheet of $\mathcal{S}_\varepsilon^{a-}$. Thus, we can restrict ourselves to $\mathcal{C}_\varepsilon^-$ when studying secondary canards.

Given the asymptotics of the return map $\overline{\Pi} : \overline{\Delta}_- \to \overline{\Delta}_-$, as derived in Proposition 2.2 (cf. (2.9)), we can write

$$\overline{\Pi}(h, \bar{w}) = \overline{\Pi}_0(h, \bar{w}) + \mathcal{O}(\varepsilon), \tag{3.13}$$

where $\overline{\Pi}_0(h, \bar{w})$ denotes the return map for the system

$$\begin{aligned} \bar{v}' &= -\bar{z} + f_2 \bar{v}^2 + \sqrt{\varepsilon} f_3 \bar{v}^3 + \sqrt{\varepsilon} F(0,0) + \bar{w} G(0,0), \\ \bar{z}' &= \bar{v} - \bar{w}, \\ \bar{w}' &= 0. \end{aligned} \tag{3.14}$$

We begin by showing that the leading-order approximation $\overline{\Pi}_0$, which is obtained by omitting the $\mathcal{O}(\varepsilon)$-terms in (3.13), is not sufficiently accurate to give nontrivial solutions of (3.12), i.e., solutions that are not exponentially close (in ε) to the canard critical value $\bar{w}^{\mathrm{c}}$ for (3.14). (Recall that $\bar{w}^{\mathrm{c}}$ is the $\bar{w}$-value corresponding to the strong canard Γ_ε^0, after the rescaling in (2.2), with

$$\bar{w}^{\mathrm{c}} = \frac{d^0_{\sqrt{\varepsilon}}}{d^0_{\bar{w}}} \sqrt{\varepsilon} + \mathcal{O}(\varepsilon)$$

by (2.16).) Since Γ_ε^0 itself is only unique up to exponentially small terms, we conclude that the map $\overline{\Pi}_0$ will admit no secondary canards.

The argument goes as follows: to determine the $\bar{w}$-value corresponding to the first secondary canard Γ_ε^1 from $\overline{\Pi}_0$, one would have to solve $P_h\overline{\Pi}_0(h^0(\bar{w}),\bar{w}) = 0$. Solutions of this equation are obtained by applying the implicit function theorem about $(0,\bar{w}^{\mathrm{c}})$. (Here, we have taken into account that $h^0(\bar{w}^{\mathrm{c}}) = 0$ by the definition of $\bar{w}^{\mathrm{c}}$; cf. again (2.16).) However, since $\bar{w}^{\mathrm{c}}$ corresponds precisely to the critical value of the canard parameter $\bar{w}$ in the classical (two-dimensional) scenario, it can be shown [20] that $P_h\overline{\Pi}_0(0,\bar{w}_0^{\mathrm{c}})$ is exponentially small. By the implicit function theorem, it follows that any solution $\bar{w}^*$ of the equation $P_h\overline{\Pi}_0(h^0(\bar{w}),\bar{w}) = 0$ close to $\bar{w}^{\mathrm{c}}$ must be such that $|\bar{w}^* - \bar{w}^{\mathrm{c}}|$ is exponentially small. (Note that this is exactly the situation encountered in a two-dimensional canard explosion; see again [20].)

Hence, in order to find secondary canards, we must refine our analysis and include additional terms in the description of the "local" return map $\overline{\Pi}$. In the following, we will use the partially decoupled truncated system

$$\bar{v}' = -\bar{z} + f_2\bar{v}^2 + \sqrt{\varepsilon}f_3\bar{v}^3 + \sqrt{\varepsilon}F(0,0) + \bar{w}G(0,0), \tag{3.15a}$$

$$\bar{z}' = \bar{v} - \bar{w}, \tag{3.15b}$$

$$\bar{w}' = \varepsilon\mu \tag{3.15c}$$

as the basis for our computation. As it turns out, this refinement will suffice to solve (3.12) for $\bar{w}$, in a nontrivial fashion, to leading order. Note that the only difference between (3.14) and (3.15) lies in the $\bar{w}$-equation: instead of keeping $\bar{w}$ constant to lowest order, we let it evolve in (3.15c), according to the leading-order approximation obtained for $\bar{w}'$ from (2.8c), $\bar{w}' = \varepsilon(\mu - g_1\varepsilon\bar{z} + \mathcal{O}(\varepsilon)) \sim \varepsilon\mu$.

The relevant result on the refined asymptotics of $\overline{\Pi}$ is obtained as follows.

Proposition 3.2. *Let $\overline{\Pi} : \overline{\Delta}^- \to \overline{\Delta}_-$ denote the return map for (3.15), and fix $\varepsilon > 0$ sufficiently small. Then,*

$$\overline{\Pi}(h,\bar{w}) = \begin{pmatrix} P_h\overline{\Pi}_0(h,\bar{w}) + \varepsilon\mu\mathcal{K}(h) + \mathcal{O}(\varepsilon^2) \\ \bar{w} + 2\varepsilon\mu T^h + \mathcal{O}(\varepsilon^2) \end{pmatrix}, \tag{3.16}$$

where $\overline{\Pi}_0$ denotes the return map for (3.14) and $\mathcal{K}$ is defined via

$$\mathcal{K}(h) = \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (G(0,0),-1)^T(t+T^h)\,dt.$$

Proof. Let $\bar{w}_0^{\mathrm{c}}$ denote the critical $\bar{w}$-value for the "refined" system (3.15). We begin by showing that, to leading order, $\bar{w}_0^{\mathrm{c}}$ equals $\bar{w}^{\mathrm{c}}$, which is again the corresponding $\bar{w}$-value determined from $\overline{\Pi}_0$; cf. (2.16). Suppose that $\bar{w}$ is given and that we wish to find h^- such that $(h^-,\bar{w}) \in \mathcal{C}_\varepsilon^-$ holds. Solving (3.15c), we obtain $\bar{w}(t) = \bar{w} + \varepsilon\mu t$, which we then substitute into (3.15a) and (3.15b):

$$\begin{aligned} \bar{v}' &= -\bar{z} + f_2\bar{v}^2 + \sqrt{\varepsilon}f_3\bar{v}^3 + \sqrt{\varepsilon}F(0,0) + (\bar{w} + \varepsilon\mu t)G(0,0), \\ \bar{z}' &= \bar{v} - \bar{w} - \varepsilon\mu t. \end{aligned} \tag{3.17}$$

Fix $\bar{w}$, and suppose that h_0^- is the h-value obtained from (3.14) such that $(h_0^-, \bar{w}) \in \mathcal{C}_\varepsilon^-$; see the proof of Proposition 2.2. Then, it follows that

$$h^- = h_0^- - \varepsilon\mu \int_{-\infty}^{0} \frac{\partial H}{\partial \bar{z}}(\bar{\gamma}_0^0(t))G(0,0)t\,dt, \tag{3.18}$$

again by the proof of Proposition 2.2. A similar computation shows that

$$h^+ = h_0^+ + \varepsilon\mu \int_{0}^{\infty} \frac{\partial H}{\partial \bar{z}}(\bar{\gamma}_0^0(t))G(0,0)t\,dt, \tag{3.19}$$

where h^+ and h_0^+ are defined by the requirement that $(h^+, \bar{w}) \in \mathcal{C}_\varepsilon^+$ and $(h_0^+, \bar{w}) \in \mathcal{C}_\varepsilon^+$ in (3.15) and (3.14), respectively. By symmetry, we find that

$$\int_{-\infty}^{0} -\frac{\partial H}{\partial z}(\bar{\gamma}_0^0(t))t\,dt = \int_{0}^{\infty} \frac{\partial H}{\partial z}(\bar{\gamma}_0^0(t))t\,dt. \tag{3.20}$$

It follows that the defining condition for the strong canard in (3.15), which, for (2.8), is given by $h^- = h^+$, reduces to $h_0^- = h_0^+ + \mathcal{O}(\varepsilon^2)$ and, hence, that the corresponding critical values of $\bar{w}$ are indeed the same to leading order.

Finally, the approximation for $\overline{\Pi}$ in (3.16) is derived as in the proof of Proposition 2.2, where we note that the additional $\mathcal{K}$-term is due to the fact that $h \mapsto h^+ - h^- = h^0 + \varepsilon\mu\mathcal{K}(h)$, by (3.18), (3.19), and (3.20). ■

Remark 8. It can be shown that the inclusion of additional (higher-order) terms in (3.15) will not alter the result of Proposition 3.2, since these terms will either drop out by symmetry, as in the proof of Proposition 2.2, or contribute only terms of higher order in (3.16).

The asymptotics of $\mathcal{K}$ are studied in Appendix A, where we show that $\mathcal{K}(h) = 2d_{\bar{w}}^0 T^h + \mathcal{O}(1)$; see Lemma A.5. Therefore, the defining condition for the first secondary canard, $P_h\overline{\Pi}(h^0(\bar{w}), \bar{w}) = 0$, can be written as

$$P_h\overline{\Pi}_0(h^0(\bar{w}), \bar{w}) = -\varepsilon\mu\mathcal{K}(h^0(\bar{w})) + \mathcal{O}(\varepsilon^2), \tag{3.21}$$

to leading order. Moreover, recalling that $\bar{w}_1^c$ denotes the value of $\bar{w}$ that solves (3.21), we write $\bar{w}_1^c = \bar{w}_0^c + \Delta\bar{w}$. Then, we have the following estimate for the width $\Delta\bar{w}$ of the first sector of rotation.

Proposition 3.3. *With $\Delta\bar{w}$ defined as above, there holds*

$$\Delta\bar{w} = -2\varepsilon\mu\sqrt{-2\ln\varepsilon} + \mathcal{O}(\varepsilon) \tag{3.22}$$

for $\varepsilon > 0$ sufficiently small.

Proof. Making use of the definition of Π_0 (see Proposition 2.2), we first rewrite $P_h\overline{\Pi}_0(h^0(\bar{w}), \bar{w})$ as

$$\begin{aligned}
P_h\overline{\Pi}_0(h^0(\bar{w}), \bar{w}) &= P_h\overline{\Pi}_0(0, \bar{w}_0^c) + P_h\overline{\Pi}_0(h^0(\bar{w}), \bar{w}) - P_h\overline{\Pi}_0(0, \bar{w}_0^c) \\
&= P_h\overline{\Pi}_0(0, \bar{w}_0^c) + d_{\bar{w}}^{h^0(\bar{w})}\bar{w} - d_{\bar{w}}^0\bar{w}_0^c + \sqrt{\varepsilon}\big(d_{\sqrt{\varepsilon}}^{h^0(\bar{w})} - d_{\sqrt{\varepsilon}}^0\big) + \mathcal{O}(\varepsilon, \sqrt{\varepsilon}\Delta\bar{w}, \Delta\bar{w}^2) \\
&= P_h\overline{\Pi}_0(0, \bar{w}_0^c) + \big(d_{\bar{w}}^{h^0(\bar{w})} - d_{\bar{w}}^0\big)\bar{w} + d_{\bar{w}}^0\Delta\bar{w} + \sqrt{\varepsilon}\big(d_{\sqrt{\varepsilon}}^{h^0(\bar{w})} - d_{\sqrt{\varepsilon}}^0\big) \\
&\qquad + \mathcal{O}(\varepsilon, \sqrt{\varepsilon}\Delta\bar{w}, \Delta\bar{w}^2);
\end{aligned}$$

see the discussion in section 3.1 as well as (2.39). Now, recall that $\bar{w} = \mathcal{O}(\sqrt{\varepsilon})$ by Assumption 1, and note that one can estimate $d_{\bar{w}}^{h^0(\bar{w})} - d_{\bar{w}}^0 = \mathcal{O}(h^0(\bar{w})\ln(-h^0(\bar{w}))^{\frac{3}{2}})$; cf. (A.9). Also, since $h^0(\bar{w}_0^{\mathrm{c}}) = 0$ by (2.16) and (2.39), a Taylor expansion shows

$$h^0(\bar{w}) = d_{\bar{w}}^0 \Delta\bar{w} + \mathcal{O}(\Delta\bar{w}^2), \tag{3.23}$$

which implies in sum

$$P_h\overline{\Pi}_0(h^0(\bar{w}),\bar{w}) = P_h\overline{\Pi}_0(0,\bar{w}_0^{\mathrm{c}}) + d_{\bar{w}}^0\Delta\bar{w} + \sqrt{\varepsilon}\big(d_{\sqrt{\varepsilon}}^{h^0(\bar{w})} - d_{\sqrt{\varepsilon}}^0\big) + \mathcal{O}\big(\varepsilon, \sqrt{\varepsilon}(-\ln\varepsilon)^{\frac{3}{2}}\Delta\bar{w}, \Delta\bar{w}^2\big).$$

Using the fact that $P_h\overline{\Pi}_0(0,\bar{w}_0^{\mathrm{c}}) = \mathcal{O}(\mathrm{e}^{-\frac{\kappa}{\varepsilon}})$ for some $\kappa > 0$ as well as the estimates from (A.9) and Lemma A.5, we conclude that the $\bar{w}$-value corresponding to the first secondary canard, $\bar{w}_1^{\mathrm{c}}$, is determined from $d_{\bar{w}}^0\Delta\bar{w} = -2\varepsilon\mu d_{\bar{w}}^0 T^{h^0(\bar{w})} + \mathcal{O}((\sqrt{\varepsilon} + \Delta\bar{w})^2)$. Hence, we obtain

$$\bar{w}_1^{\mathrm{c}} = \bar{w}_0^{\mathrm{c}} - 2\varepsilon\mu T^{h^0(\bar{w}_1^{\mathrm{c}})} + \mathcal{O}(\varepsilon), \tag{3.24}$$

which implies in particular $|\bar{w}_1^{\mathrm{c}} - \bar{w}_0^{\mathrm{c}}| \gtrsim \varepsilon$, i.e., $|\bar{w}_1^{\mathrm{c}} - \bar{w}_0^{\mathrm{c}}| > \varepsilon$ as well as $|\bar{w}_1^{\mathrm{c}} - \bar{w}_0^{\mathrm{c}}| \sim \varepsilon$. Due to $h^0(\bar{w}_0^{\mathrm{c}}) = 0$ and $\frac{dh^0}{d\bar{w}} \sim d_{\bar{w}}^0$, it follows from the intermediate value theorem that $h^0(\bar{w}_1^{\mathrm{c}}) \gtrsim \varepsilon$, which, together with Lemma A.2, shows that the desired estimate for the size of the first sector of rotation is given by

$$\bar{w}_1^{\mathrm{c}} - \bar{w}_0^{\mathrm{c}} = \Delta\bar{w} = -2\varepsilon\mu\sqrt{-2\ln\varepsilon} + \mathcal{O}(\varepsilon). \tag{3.25}$$

This completes the proof. ∎

Let $k > 1$, and consider $j = 0,\dots,k$. We now set out to find an analogue of condition (3.21) for the kth secondary canard Γ_ε^k. Let $\bar{w}_k^{\mathrm{c}}$ again denote the corresponding $\bar{w}$-value, consider an initial condition $(h^0(\bar{w}),\bar{w}) \in \mathcal{C}_\varepsilon^-$, and let

$$\bar{w}^j = P_{\bar{w}}\overline{\Pi}^j(h^0(\bar{w}),\bar{w}),$$

as before. Note that $\bar{w}_k^{\mathrm{c}}$ must be a solution of the equation

$$P_h\overline{\Pi}(h^{k-1}(\bar{w}^{k-1}),\bar{w}^{k-1}) = 0$$

or, equivalently, of

$$P_h\overline{\Pi}_0(h^{k-1}(\bar{w}^{k-1}),\bar{w}^{k-1}) = -\varepsilon\mu\mathcal{K}(h^{k-1}(\bar{w}^{k-1})) + \mathcal{O}(\varepsilon^2). \tag{3.26}$$

Observe that the condition in (3.26) is analogous to (3.21), with h^0 replaced by h^{k-1}; hence, the structure of (3.21) is replicated at higher orders. Note also that it follows from (3.24) that $\bar{w}_1^{\mathrm{c},1} = \bar{w}_0^{\mathrm{c}} + \mathcal{O}(\varepsilon)$, where $\bar{w}_1^{\mathrm{c},1}$ is the first iterate of $\bar{w}_1^{\mathrm{c}}$ under $\overline{\Pi}$. This estimate, in turn, implies that $h^1(\bar{w}_0^{\mathrm{c}}) = \mathcal{O}(\varepsilon)$; see (3.1). An argument analogous to the derivation of (3.24) now leads to the estimate

$$\bar{w}_2^{\mathrm{c},1} = \bar{w}_0^{\mathrm{c}} - 2\varepsilon\mu T^{h(\bar{w}_2^{\mathrm{c},1})} + \mathcal{O}(\varepsilon)$$

or, equivalently, to

$$\bar{w}_2^{\mathrm{c}} = \bar{w}_0^{\mathrm{c}} - 2\varepsilon\mu\big(T^{h(\bar{w}_2^{\mathrm{c}})} + T^{h(\bar{w}_2^{\mathrm{c},1})}\big) + \mathcal{O}(\varepsilon). \tag{3.27}$$

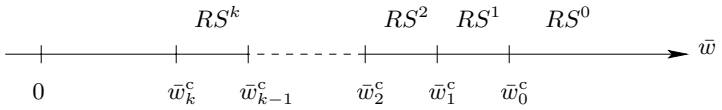


Figure 14. *The sectors of rotation, RS^j.*

Proceeding inductively, we obtain $h^{k-1}(\bar{w}_0^c) = \mathcal{O}(\varepsilon)$ and

$$\bar{w}_k^c = \bar{w}_0^c - 2\varepsilon\mu\left(\sum_{j=0}^{k-1} T^{h(\bar{w}_k^{c,j})}\right) + \mathcal{O}(\varepsilon), \tag{3.28}$$

where we define $\bar{w}_k^{c,0} \equiv \bar{w}_k^c$. Finally, using Lemma A.2 to again approximate $T^{h(\bar{w}_k^{c,j})}$ by $\sqrt{-2\ln\varepsilon} + \mathcal{O}(1)$ (nonuniformly in k), we obtain in analogy to (3.25) that

$$\bar{w}_k^c = \bar{w}_{k-1}^c - 2\varepsilon\mu\sqrt{-2\ln\varepsilon} + \mathcal{O}(\varepsilon), \tag{3.29}$$

which, in conjunction with (2.19), verifies Assumption 1 above.

One question that naturally arises in this context is whether $k > 1$ can be chosen arbitrarily large. Here, we show that our analysis, and, in particular, the estimate in (3.28), does not hold uniformly in k with respect to ε; rather, (3.28) is valid for k fixed and ε sufficiently small. This is due to the fact that the contributions coming from $T^{h(\bar{w}_k^{c,j})}$ become increasingly smaller with k: since $\bar{w}_k^c$ decreases with k and since $h(\bar{w}) \sim \sqrt{\varepsilon}d^0_{\sqrt{\varepsilon}} + \bar{w}d^0_{\bar{w}}$ with $d^0_{\bar{w}} < 0$ (see (2.18)), it follows that h increases with k. Therefore, $T^h \sim \sqrt{-2\ln h}$ decreases, and the $\mathcal{O}(\varepsilon)$-terms can come to dominate the $2\varepsilon\mu(\cdots)$-terms in (3.28) if k is sufficiently large. However, in our analysis, we had to assume that these terms are uniformly of lower order than ε, starting with the leading-order approximation for $\overline{\Pi}$ in Proposition 2.2. In summary, for k "large," ε thus has to be chosen small enough to ensure that the estimate in (3.28) remains consistent.

Proposition 3.4. *Fix any integer $K > 0$. Then, there exists an $\varepsilon > 0$ sufficiently small such that the estimate in (3.28) holds for $k \leq K$.*

For $j = 1, \ldots, k$, we now define the jth sector of rotation RS^j as follows:

$$RS^j = \left\{(h^0(\bar{w}), \bar{w}) \in \mathcal{C}_\varepsilon^- \;\middle|\; \bar{w}_j^c \leq \bar{w} < \bar{w}_{j-1}^c\right\}.$$

This definition provides a connection between the family of secondary canards $\{\Gamma_\varepsilon^j\}$ and the corresponding sectors of rotation: the jth sector, RS^j, is bounded by the secondary canards Γ_ε^{j-1} and Γ_ε^j in that the corresponding points $\bar{w}_{j-1}^c$ and $\bar{w}_j^c$ on $\mathcal{C}_\varepsilon^-$ define the boundaries of RS^j.

For notational purposes, we also introduce the zeroth sector RS^0 via

$$RS^0 = \left\{(h, \bar{w}) \in \mathcal{C}_\varepsilon^- \;\middle|\; \bar{w}_0^c \leq \bar{w}\right\},$$

and we note that this definition is equivalent to requiring that $h < 0$; see (2.37). An illustration of these sectors of rotation is given in Figure 14. In particular, since $\bar{w}_j^c < \bar{w}_{j-1}^c$ for any $j \geq 1$, the sector RS^j lies further "to the left" of RS^0 with increasing j.

It follows from the preceding analysis that all of the sectors RS^j are of equal size to leading order; cf. (3.29). (However, we conjecture that due to higher-order corrections, the sector size actually decreases as j increases.) Moreover, for ε and μ fixed, the number of sectors of rotation RS^j, and, hence, also the number of corresponding secondary canards Γ_ε^j, has to be finite: note that the frequency of the small-oscillation component in any mixed-mode time series in (1.5) is globally bounded, with the bound given approximately by the frequency determined by the Hopf bifurcation around the origin in (1.5). Additionally, the speed of the drift in $\bar{w}$ in (2.3c) is always positive for nonzero ε and μ, which implies that $\bar{w} > \bar{w}_0^c$ in finite time. Hence, trajectories of (1.5) can undergo only a finite number of small-amplitude oscillations before entering the relaxation regime, which implies that there can be only a finite number of sectors of rotation lying in $(0, \bar{w}_0^c)$; see again Figure 14.

Remark 9. It follows from Proposition 3.4 that both the number of secondary canards and that of the corresponding sectors of rotation must go to infinity as $\varepsilon \to 0$. However, it is important to note that Proposition 3.4 gives no bound on the total number of secondary canards for $\varepsilon > 0$. Rather, the integer K can be chosen arbitrarily large provided ε is small enough, implying that our analysis is then valid for all $k \leq K$.

Finally, we observe that the definition of RS^j can be extended to a small neighborhood of $\mathcal{C}_\varepsilon^-$ by the flow of (2.8) and, hence, that the sectors of rotation can be interpreted as two-dimensional subsets of $\mathcal{S}_\varepsilon^{a-}$.

3.4. The return map Φ to $\mathcal{C}_\varepsilon^-$. To show how the "full" map Π (which a priori has to be interpreted as a map that is defined on $\bigcup \mathcal{C}_\varepsilon^j$) can be approximated accurately by a "simplified" map, we introduce $\Phi : \mathcal{C}_\varepsilon^- \to \mathcal{C}_\varepsilon^-$ as follows. Let $k \geq 0$, and recall the definition of the kth sector of rotation, RS^k, from the previous subsection. Then, we define Φ via

$$\Phi(\bar{w}) = P_{\bar{w}}\big(\Pi^{\text{in}} \circ \Pi^{\text{ret}} \circ \Pi^{\text{out}} \circ \overline{\Pi}^k(h^0(\bar{w}), \bar{w})\big) \qquad \text{if } (h^0(\bar{w}), \bar{w}) \in RS^k. \tag{3.30}$$

Note that Φ is a reinterpretation of Π in that it is a composition of the same components that were used in the definition of Π in (3.5). However, it is defined on a different domain: the definition in (3.30) reduces the analysis of the flow induced by (1.5) to that of a one-dimensional map that is defined on the single curve $\mathcal{C}_\varepsilon^-$, which will allow us to study the recurrent dynamics on RS^k in considerable detail. Moreover, we note that Φ is still an exponentially accurate approximation for the full, two-dimensional return map Π, which is again due to the fact that all trajectories must return exponentially close to $\mathcal{C}_\varepsilon^-$ after relaxation, i.e., after application of Π^{ret}; cf. the proof of Proposition 3.1. One drawback of this simplification, however, lies in the fact that the defining formula (3.30) for Φ is k-dependent; in other words, the definition of Φ changes with the sector of rotation under consideration. This k-dependence will have to be taken into account throughout the subsequent analysis.

Finally, we remark that the map Φ is smooth on each of the sectors RS^k but that it has discontinuities at the points $\bar{w}_k^c$ and $\bar{w}_{k-1}^c$. We will not study the nature of these discontinuities in detail, since we are not attempting to analyze the dynamics of Φ "very close" to the secondary canards. Rather, we will restrict ourselves to describing Φ on the interior of the individual sectors RS^k.

3.5. The derivative of Φ. In this subsection, we derive estimates for the derivative $\Phi'(\bar{w}) := \frac{d\Phi}{d\bar{w}}$ of Φ on the kth sector of rotation, RS^k. We then investigate some of the

properties of Φ'. The resulting estimates are needed for the analysis of the dynamics of Φ in section 3.6 below and will allow us to characterize the admissible Farey sequences in (1.5) as well as to describe the corresponding parameter intervals.

Let $\bar{w}$ be such that $(h^0(\bar{w}), \bar{w}) \in RS^k$, and let $\bar{w}^j = P_{\bar{w}}\overline{\Pi}^j(h^0(\bar{w}^0), \bar{w}^0)$, where we set $\bar{w}^0 \equiv \bar{w}$. Given the definition of Φ in (3.30), we have the following result.

Lemma 3.5. *To leading order, there holds*

(3.31)
$$\frac{d\Phi(\bar{w})}{d\bar{w}} = 1 - \varepsilon\mu d^0_{\bar{w}}\left(\sum_{j=0}^{k-1} 2(j+1)\frac{1}{h^j(\bar{w}^j)}\frac{1}{\sqrt{-2\ln h^j(\bar{w}^j)}} + (k+1)\frac{1}{h^k(\bar{w}^k)}\frac{1}{\sqrt{-2\ln h^k(\bar{w}^k)}}\right) + \mathcal{O}(\varepsilon\ln\varepsilon)$$

for the derivative of Φ on RS^k.

Proof. By the chain rule and taking into account the definitions of $\overline{\Pi}$, Π^{in}, and Π^{out}, as well as of Π^{ret} in Propositions 2.2, 2.3, and 2.4 as well as in (2.54), respectively, we have

$$\begin{aligned}\frac{d\Phi(\bar{w})}{d\bar{w}} &= \prod_{j=0}^{k-1}\left(1 + 2\varepsilon\mu\frac{dT^{h^j(\bar{w})}}{d\bar{w}}\right)\left(1 + \varepsilon\mu\frac{dT^{h^k(\bar{w}),\text{out}}}{d\bar{w}}\right) + \mathcal{O}(\varepsilon\ln\varepsilon)\\ &= 1 + \varepsilon\mu\sum_{j=0}^{k-1} 2\frac{dT^{h^j(\bar{w})}}{d\bar{w}} + \varepsilon\mu\frac{dT^{h^k(\bar{w}),\text{out}}}{d\bar{w}} + \mathcal{O}(\varepsilon\ln\varepsilon)\\ &= 1 - \varepsilon\mu d^0_{\bar{w}}\left(\sum_{j=0}^{k-1} 2(j+1)\frac{1}{h^j(\bar{w}^j)}\frac{1}{\sqrt{-2\ln h^j(\bar{w}^j)}} + (k+1)\frac{1}{h^k(\bar{w}^k)}\frac{1}{\sqrt{-2\ln h^k(\bar{w}^k)}}\right)\\ &\quad + \mathcal{O}(\varepsilon\ln\varepsilon),\end{aligned}$$

where the last step follows from (3.10) and (3.11). ∎

Since we assume that $h^j(\bar{w}^j) = \mathcal{O}(\varepsilon\sqrt{-\ln\varepsilon})$ (see the proof of Proposition 3.3 above), we can write

$$\frac{1}{\sqrt{-2\ln h^j(\bar{w}^j)}} = \frac{1}{\sqrt{-2\ln\varepsilon}}\big(1 + o(1)\big).$$

This gives a somewhat less accurate but more concise estimate for the derivative of Φ:

$$\frac{d\Phi(\bar{w})}{d\bar{w}} \sim 1 - \varepsilon\mu d^0_{\bar{w}}\frac{1}{\sqrt{-2\ln\varepsilon}}\left(\sum_{j=0}^{k-1}\frac{2(j+1)}{h^j(\bar{w}^j)} + \frac{k+1}{h^k(\bar{w}^k)}\right). \tag{3.32}$$

(Note that again due to $h^j(\bar{w}^j) = \mathcal{O}(\varepsilon\sqrt{-\ln\varepsilon})$, the $\mathcal{O}(\varepsilon)$-correction in (3.32) will actually be of the order $(\ln\varepsilon)^{-1}$ and that we can therefore neglect the $\mathcal{O}(\varepsilon\ln\varepsilon)$-terms in (3.31).)

To simplify this estimate further, we have to distinguish between different k-values in (3.32). We first focus on the case where $k > 0$; the case when $k = 0$ will be discussed separately.

Given $k > 0$, fix an initial condition $\bar{w}^0 \in RS^k$, and let $\bar{w}^1, \bar{w}^2, \ldots, \bar{w}^k$ be defined as in section 3.4 above; i.e., let $\bar{w}^j$ be the jth iterate of $\bar{w}^0$ under $\overline{\Pi}$. Then, it follows directly from

(3.32) that $\Phi' < 1$ if $\bar{w}^0 \approx \bar{w}^{\mathrm{c}}_k$, respectively, that $\Phi' > 1$ if $\bar{w}^0 \approx \bar{w}^{\mathrm{c}}_{k-1}$. We are interested in approximating more precisely the size of the $\bar{w}$-intervals where Φ' is less than 1 and greater than 1, respectively.

To that end, let $\Delta\bar{w}^j = \bar{w}^{\mathrm{c}}_{j-1} - \bar{w}^{\mathrm{c}}_j$ be the width of the jth sector of rotation RS^j, and recall that we have the estimate

$$\Delta\bar{w}^j \sim 2\varepsilon\mu\sqrt{-2\ln\varepsilon},$$

independent of j to leading order. Given any $\bar{w}^0 \in RS^k$, we can write $\bar{w}^0 = \bar{w}^{\mathrm{c}}_k + \nu\Delta\bar{w}^k$ for some $\nu \in [0,1]$; i.e., the sector RS^k will be parametrized by the variable ν in the following. Moreover, for any $j \geq 0$, we have the following estimates:

$$\begin{aligned}
\bar{w}^{\mathrm{c}}_j &\sim \bar{w}^{\mathrm{c}}_0 - 2j\varepsilon\mu\sqrt{-2\ln\varepsilon},\\
\bar{w}^j &\sim \bar{w}^{\mathrm{c}}_0 - 2\big((k-j)-\nu\big)\varepsilon\mu\sqrt{-2\ln\varepsilon},\\
h^j(\bar{w}^j) &\sim -2d^0_{\bar{w}}(j+1)\big((k-j)-\nu\big)\varepsilon\mu\sqrt{-2\ln\varepsilon},
\end{aligned}$$

where the last expression is a consequence of (3.3). Using (3.32), we obtain

$$\frac{d\Phi(\bar{w})}{d\bar{w}} \sim 1 - \frac{\omega_k(\nu)}{4\ln\varepsilon}, \tag{3.33}$$

where the function ω_k is defined via

$$\omega_k(\nu) = \sum_{j=0}^{k-1} \frac{1}{(k-j)-\nu} - \frac{1}{2\nu}. \tag{3.34}$$

Finally, we consider the case where $k = 0$. For any initial condition $\bar{w}^0 \in RS^0$, we can write $\bar{w}^0 = \bar{w}^0_c + 2\nu\varepsilon\mu\sqrt{-2\ln\varepsilon}$, where ν is now some positive number. Then,

$$\frac{d\Phi(\bar{w})}{d\bar{w}} \sim 1 - \frac{\omega_0(\nu)}{4\ln\varepsilon}, \tag{3.35}$$

with $\omega_0(\nu) = -\frac{1}{2\nu}$. Observe that, clearly, $\Phi'(\bar{w}) < 1$ for any $\bar{w} \in RS^0$.

Remark 10. Note that for $k \geq 0$, the function $\omega_k(\nu)$ defined in (3.34) is increasing on $[0,1]$ and that ω_k changes sign exactly once if $k > 0$; see Figure 15.

The zeros of $\omega_k(\nu)$, $k > 0$, give the approximate sizes of the subintervals of RS^k where Φ' is greater than 1 and less than 1, respectively. More precisely, we have proven the following result.

Proposition 3.6. *For $k > 0$ and $\varepsilon > 0$ sufficiently small, the subinterval of RS^k on which $\Phi'(\bar{w}) < 1$ is approximately given by $(\bar{w}^{\mathrm{c}}_k, \bar{w}^{\mathrm{c}}_k + 2\nu^k_0\mu\varepsilon\sqrt{-2\ln\varepsilon})$, where ν^k_0 denotes the unique zero of ω_k on RS^k.*

3.6. The dynamics of Φ. In this subsection, we analyze the dynamics of the reduced map Φ in more detail, combining the results obtained so far in section 3. The aim of our analysis is to relate the properties of Φ to the resulting mixed-mode dynamics in (1.5) and to estimate the relevant parameter (μ-)range corresponding to this dynamics. Our first result

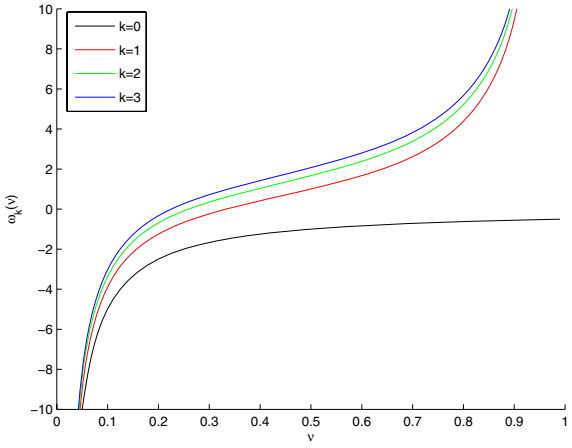


Figure 15. *The function* ω_k *on* $[0,1]$ *for* $k = 0, \dots, 3$.

(Theorem 3.7) concerns the existence and stability of 1^k-type orbits, i.e., of periodic orbits with symbolic (Farey) sequence $\{1^k\}$; these orbits correspond to the recurrent dynamics of (1.5) on the kth sector of rotation, RS^k. Then, in Theorem 3.9, we derive conditions for when a given orbit will pass through RS^k. In Theorem 3.10, Proposition 3.11, and Corollary 3.12, we apply these conditions to classify the periodic orbits of the more general type $\{L_j^{k_j}\}$, with $L_j, k_j \geq 1$, that can "typically" occur in (1.5).

We start by summarizing some of the features of Φ which follow directly from the results of sections 3.3 and 3.5; see Figure 16 for a qualitative illustration.

(i) Φ must be decreasing close to the left boundary of RS^k and increasing on most of RS^k, with $\bar{w}_k^{\mathrm{c}} + \nu_0^k \Delta\bar{w}^k$ giving an estimate of the point where Φ' becomes greater than 1; cf. Proposition 3.6.

(ii) The derivative Φ' must change sign near $\bar{w}_{\min}^k := \bar{w}_k^{\mathrm{c}} + \nu_{\min}^k \Delta\bar{w}^k$, with $\nu_{\min}^k$ determined by the condition that $\omega_k(\nu_{\min}^k) = 4\ln\varepsilon$. This implies in particular that $\nu_{\min}^k = \mathcal{O}((\ln\varepsilon)^{-1})$ and, hence, that $\bar{w}_{\min}^k \approx \bar{w}_k^{\mathrm{c}}$. (Note that our analysis does not prove the uniqueness of this minimum, though.)

(iii) A simple computation along the lines of section 3.5 shows that $\Phi(\bar{w}_{\min}^k) = \Phi_{\min} + \mathcal{O}(\varepsilon)$ is independent of k to lowest order, where

$$\Phi_{\min} := \bar{w}_0^{\mathrm{c}} + \sqrt{\varepsilon}\big(\mathcal{G}(v_0, v_{\max}, \mu) + \mathcal{G}(v_{\max}^*, 0, \mu)\big) + \varepsilon\mu\sqrt{-2\ln\varepsilon}; \tag{3.36}$$

cf. (2.54). Indeed, given the formula for Π in (3.5), as well as $P_{\bar{w}}\Pi^k(\bar{w}_{\min}^k) \sim \bar{w}_{\min}^0$, it follows with (3.30) that

$$\begin{aligned}\Phi(\bar{w}_{\min}^k) &\sim \Phi \circ P_{\bar{w}}\Pi^k(\bar{w}_{\min}^k)\\ &\sim \bar{w}_{\min}^0 + \varepsilon\mu T^{h(\bar{w}_{\min}^0),\mathrm{out}} + \bar{w}_{\min}^0 f_2 \mu\varepsilon\ln\varepsilon + \sqrt{\varepsilon}\big(\mathcal{G}(v_0, v_{\max}, \mu) + \mathcal{G}(v_{\max}^*, 0, \mu)\big).\end{aligned}$$

Since $\nu_{\min}^0 \gtrsim (\ln\varepsilon)^{-1}$ implies $\bar{w}_{\min}^k \sim \bar{w}_0^{\mathrm{c}}$ and since $T^{h(\bar{w}_{\min}^0),\mathrm{out}} \sim \sqrt{-2\ln\varepsilon}$, one obtains (3.36).

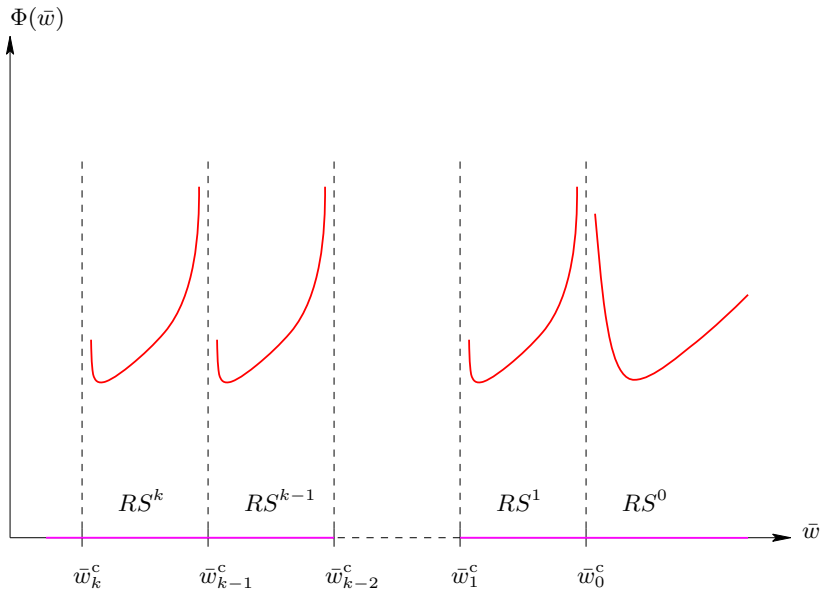


Figure 16. *A qualitative illustration of the map* Φ.

By definition, fixed points of Φ on RS^k correspond to periodic 1^k-type orbits in (2.1). We are interested in estimating the parameter range (i.e., the μ-interval) in which such orbits can be observed.

Theorem 3.7. *For* $\varepsilon > 0$ *sufficiently small, the periodic orbit of type* 1^k, $k \geq 1$, *exists and is stable on a* μ*-interval of the form* $(\underline{\mu}^k, \overline{\mu}^k)$, *with*

$$\Delta\mu^k := \overline{\mu}^k - \underline{\mu}^k = -\frac{\underline{\mu}^k}{\sqrt{2}D_\mu}\frac{\sqrt{\varepsilon}}{\sqrt{-\ln\varepsilon}}\int_{\nu_{-2}^k}^{\nu_0^k}\omega_k(\nu)\,d\nu + \mathcal{O}\big(\sqrt{\varepsilon}(-\ln\varepsilon)^{-1}\big); \tag{3.37}$$

here, ν_{-2}^k *denotes the* ν*-value that solves* $\omega_k(\nu) = 8\ln\varepsilon$.

Proof. Note that MMO orbits with Farey sequence $\{1^k\}$ correspond to solutions of the equation

$$\Phi(\bar{w}, \bar{\mu}) = \bar{w} \tag{3.38}$$

with $\bar{w} \in RS^k$, where we have now included explicitly the μ-dependence of Φ. We are interested in determining μ in (3.38) so that the corresponding fixed point of Φ will be stable. To that end, let ν_{-2}^k be defined as in the statement of the theorem, and note that for $\bar{w} \in RS^k$, the leading term of $\Phi'(\bar{w}, \mu)$ satisfies $|\Phi'(\bar{w}, \mu)| < 1$ if and only if $\bar{w} = \bar{w}_k^c + \nu\Delta\bar{w}^k$ with $\nu \in (\nu_{-2}^k, \nu_0^k)$; cf. (3.33).

Now, if (3.38) is interpreted as defining implicitly a function $\mu = \mu(\bar{w})$, we can set $\underline{\mu}^k = \mu(\bar{w}_k^c + \nu_{-2}^k\Delta\bar{w}^k)$ and $\overline{\mu}^k = \mu(\bar{w}_k^c + \nu_0^k\Delta\bar{w}^k)$. We will use the fundamental theorem of calculus to estimate $\Delta\mu^k = \overline{\mu}^k - \underline{\mu}^k$. Applying implicit differentiation to (3.38), we obtain

$$\frac{d\mu}{d\bar{w}} = -\frac{\frac{\partial}{\partial\bar{w}}\Phi(\bar{w}, \mu) - 1}{\frac{\partial}{\partial\mu}\Phi(\bar{w}, \mu)}.$$

Since

$$\frac{\partial}{\partial \mu}\Phi(\bar{w},\mu) \sim D_\mu \sqrt{\varepsilon}$$

(recall the discussion in section 2.5), it follows that

$$\frac{d\mu}{d\bar{w}} \sim \frac{1}{4D_\mu \sqrt{\varepsilon}\ln\varepsilon}\omega_k(\nu)$$

for $\bar{w} = \bar{w}_k^c + \nu\Delta\bar{w}^k$ with $\nu \in (\nu_{-2}^k, \nu_0^k)$; see (3.33). Therefore, using $\frac{d\bar{w}}{d\nu} \sim \Delta\bar{w}^k$, we find

$$\Delta\mu^k = \overline{\mu}^k - \underline{\mu}^k = \int_{\bar{w}(\underline{\mu}^k)}^{\bar{w}(\overline{\mu}^k)} \frac{d\mu}{d\bar{w}}\, d\bar{w} \sim \frac{1}{4D_\mu\sqrt{\varepsilon}\ln\varepsilon}\Delta\bar{w}^k \int_{\nu_{-2}^k}^{\nu_0^k} \omega_k(\nu)\, d\nu. \tag{3.39}$$

Given that $\Delta\bar{w}^k = 2\varepsilon\mu\sqrt{-2\ln\varepsilon} + \mathcal{O}(\varepsilon)$, the result follows. ∎

Observe that by the definition of ν_0^k, $\overline{\mu}^k$ marks the value of μ for which the orbit of type 1^k disappears in a saddle-node bifurcation of Φ, since $\Phi' = 1$ there. In the following, we summarize a few additional observations which follow from Theorem 3.7:

(i) Note that

$$\int \omega_k(\nu)\, d\nu = \ln|\Gamma(-k+\nu)| - \ln|\Gamma(1+\nu)| - \frac{1}{2}\ln\nu = -\frac{1}{2}\ln\nu + \mathcal{O}(1),$$

where Γ denotes the standard Gamma function. Since the leading-order contribution to the corresponding definite integral in (3.39) comes from $\nu_{-2}^k = \mathcal{O}((\ln\varepsilon)^{-1})$, one can show that, for ε sufficiently small,

$$\Delta\mu^k = \frac{\underline{\mu}^k}{\sqrt{2}D_\mu}\frac{\sqrt{\varepsilon}\ln(\sqrt{-\ln\varepsilon})}{\sqrt{-\ln\varepsilon}} + \mathcal{O}\big(\sqrt{\varepsilon}(-\ln\varepsilon)^{-\frac{1}{2}}\big).$$

Given that the double logarithmic term is "almost constant" (at least if ε does not vary over too many orders of magnitude), it follows that $\Delta\mu^k$ is roughly of the order $\sqrt{\varepsilon}(-\ln\varepsilon)^{-\frac{1}{2}}$ as $\varepsilon \to 0$.

(ii) The estimate in (3.37) implies that, for ε fixed, the ratio of the widths of the stability intervals of "adjacent" periodic orbits (i.e., of orbits of the types 1^{k+1} and 1^k) is approximately given by the ratio of the corresponding integrals of ω_{k+1} and ω_k. Since $\{\nu_0^k\}$ decays faster with k than $\{\nu_{-2}^k\}$ (see Figure 15), it follows that $\int_{\nu_{-2}^k}^{\nu_0^k}\omega_k(\nu)\,d\nu$ decreases. Hence, the sequence $\{\Delta\mu^k\}$ is decreasing with k.

(iii) The well-developed theory of unimodal maps [25] implies that the μ-interval for which there is an attractor for Φ in RS^k is also of size $(\underline{\mu}^k, \overline{\mu}^k)$, to lowest order. Hence, for μ in any interval given approximately by $(\overline{\mu}^k, \underline{\mu}^{k+1})$, the dynamics of Φ must involve at least two different sectors.

Next, we derive a set of conditions under which a given periodic orbit will have to pass through the kth sector of rotation RS^k. For the remainder of this subsection, we will consider only points $\bar{w} \in RS^k$, $k \geq 1$, for which $\bar{w} = \bar{w}_k^c + \nu\Delta\bar{w}^k$, with

$$\nu \in \left(\frac{1}{(-\ln\varepsilon)^p}, 1 - \frac{1}{(-\ln\varepsilon)^p}\right) \tag{3.40}$$

for some fixed integer $p > 1$. Note that the condition in (3.40) is "generic" in that it covers "most of" RS^k; to put it differently, only $\bar{w}$-values that are "very close" to the boundary points $\bar{w}_k^c$ and $\bar{w}_{k-1}^c$ are excluded by (3.40).

We begin by proving a simple preparatory result.

Lemma 3.8. *Consider $\bar{w} = \bar{w}_j^c + \nu\Delta\bar{w}^j \in RS^j$ for some $j \geq 1$, and assume that (3.40) holds. Then, if $\bar{w} \leq \bar{w}_{\min}^j$,*

$$|\Phi(\bar{w}) - \Phi_{\min}| = \mathcal{O}\Big(\frac{\ln(-\ln\varepsilon)}{-\ln\varepsilon}\Big)\Delta\bar{w}^j, \tag{3.41}$$

whereas if $\bar{w} > \bar{w}_{\min}^j$,

$$|\Phi(\bar{w}) - \Phi_{\min}| \lesssim \left(1 + \mathcal{O}\Big(\frac{\ln(-\ln\varepsilon)}{-\ln\varepsilon}\Big)\right)\Delta\bar{w}^j. \tag{3.42}$$

Proof. Let $\nu_{\min}^j$ be the ν-value corresponding to $\bar{w}_{\min}^j$. By the fundamental theorem of calculus, we have

$$\Phi(\bar{w}) - \Phi_{\min} = \Delta\bar{w}^j \int_{\nu_{\min}^j}^{\nu} \Big(1 - \frac{\omega_j(\eta)}{4\ln\varepsilon}\Big)\, d\eta = \Delta\bar{w}^j \left(\nu - \nu_{\min}^j - \frac{1}{4\ln\varepsilon}\int_{\nu_{\min}^j}^{\nu} \omega_j(\eta)\, d\eta\right).$$

Since ν is constrained by condition (3.40), we find

$$\int_{\nu_{\min}^j}^{\nu} \omega_k(\eta)\, d\eta = \mathcal{O}(\ln\nu) + \mathcal{O}(\ln\nu_{\min}^j) = \mathcal{O}\big(\ln(-\ln\varepsilon)\big);$$

see also the proof of Theorem 3.7. Now, if $\nu \leq \nu_{\min}^j$, then $\nu_{\min}^j - \nu = \mathcal{O}((-\ln\varepsilon)^{-1})$, and the estimate in (3.41) follows. If, on the other hand, $\nu > \nu_{\min}^j$, then $\nu - \nu_{\min}^j < 1$, which implies (3.42). ∎

Next, we show that orbits satisfying the generic condition in (3.40) will typically pass through the kth sector of rotation if, additionally, $\Phi_{\min} \in RS^k$ holds.

Theorem 3.9. *Assume that $\Phi_{\min} \in RS^k$ and that, for some q satisfying $0 < q < \frac{1}{2}$,*

$$\bar{w}_{k-1}^c - \Phi_{\min} \lesssim \frac{1}{(-\ln\varepsilon)^q}\Delta\bar{w}^k \quad \text{and} \quad \Phi_{\min} - \bar{w}_k^c \lesssim \frac{1}{(-\ln\varepsilon)^q}\Delta\bar{w}^k. \tag{3.43}$$

Consider a periodic orbit $\{\bar{w}^0, \dots, \bar{w}^j\}$, with $\Phi(\bar{w}^\ell) = \bar{w}^{\ell+1}$ for $\ell = 0, \dots, j-1$, and let $\{\nu^0, \dots, \nu^j\}$ be the corresponding values of ν. Assume that (3.40) holds. Then, the orbit in question must pass through RS^k provided $\varepsilon > 0$ is sufficiently small.

Proof. We will assume that $k \geq 2$ in the following and will omit the remaining cases for the sake of brevity.

First, note that Lemma 3.8 and the assumption in (3.43) imply that, for any $\bar{w} \in \mathcal{C}_\varepsilon^-$, $\Phi(\bar{w}) \in RS^k \cup RS^{k-1} \cup RS^{k-2}$. This follows from the estimates below, which are a straightforward consequence of (3.41), (3.42), and (3.43): we begin by assuming that $\bar{w} \in RS^k$;

then,

$$\begin{aligned}\Phi(\bar{w}) - \bar{w}^{\mathrm{c}}_{k-1} &= \Phi(\bar{w}) - \Phi_{\min} + \Phi_{\min} - \bar{w}^{\mathrm{c}}_{k-1} \\ &\lesssim \Phi_{\min} - \bar{w}^{\mathrm{c}}_{k-1} + \Delta\bar{w}^{k}\Big(1 + \mathcal{O}\Big(\frac{\ln(-\ln\varepsilon)}{-\ln\varepsilon}\Big)\Big) \\ &\lesssim \Delta\bar{w}^{k}\big(1 + \mathcal{O}\big((-\ln\varepsilon)^{-q}\big)\big),\end{aligned}$$

which implies that $\Phi(\bar{w})$ can be no higher than RS^{k-2} in that case.

Similarly, for $\bar{w} \in RS^{k-1} \cup RS^{k-2}$, we have the estimate

$$\begin{aligned}\Phi(\bar{w}) - \bar{w}^{\mathrm{c}}_{k-2} &= \Phi(\bar{w}) - \Phi_{\min} + \Phi_{\min} - \bar{w}^{\mathrm{c}}_{k-2} \\ &\lesssim \Phi_{\min} - \bar{w}^{\mathrm{c}}_{k-2} + \Delta\bar{w}^{k-1}\Big(1 + \mathcal{O}\Big(\frac{\ln(-\ln\varepsilon)}{-\ln\varepsilon}\Big)\Big) \\ &\lesssim \mathcal{O}\big((-\ln\varepsilon)^{-q}\big)\Delta\bar{w}^{k-1};\end{aligned}$$

see (3.41) as well as (3.43). It follows that for $\bar{w} \in RS^{k-1} \cup RS^{k-2}$, $\Phi(\bar{w})$ can be no higher than RS^{k-2}.

Finally, for any point $\bar{w} \in RS^{k-2}$ which is contained in the image of Φ, there holds

$$\bar{w} - \bar{w}^{\mathrm{c}}_{k-2} \lesssim \mathcal{O}\big((-\ln\varepsilon)^{-q}\big)\Delta\bar{w}^{k-1}$$

and, consequently,

$$\Phi(\bar{w}) \lesssim \bar{w}^{\mathrm{c}}_{k-1} + \mathcal{O}\big((-\ln\varepsilon)^{-q}\big)\Delta\bar{w}^{k-2} \tag{3.44}$$

by (3.41). It follows that any recurrent set, including the periodic orbit $\{\bar{w}^0, \dots, \bar{w}^j\}$, is contained in $RS^k \cup RS^{k-1} \cup RS^{k-2}$.

Now, suppose that such a periodic orbit is given, and note that there is an unstable fixed point $\bar{w}^*$ of Φ in RS^{k-1} close to $\bar{w}^{\mathrm{c}}_{k-2}$. Assume that $\bar{w}^0 > \bar{w}^*$. Then, the trajectory of $\bar{w}^0$ under Φ must eventually enter RS^{k-2}; moreover, by (3.44), it must terminate at a point $\bar{w}^j$ with $\bar{w}^j < \bar{w}^*$.

Looking at the forward trajectory of $\bar{w}^j$, we see that it is decreasing until it falls below $\bar{w}^{k-1}_{\min}$. In other words, there exists $\ell \geq 0$ such that $\bar{w}^j, \bar{w}^{j+1}, \dots, \bar{w}^{j+\ell-1}$ are greater than or equal to $\bar{w}^{k-1}_{\min}$ and $\bar{w}^{j+\ell}$ is less than or equal to $\bar{w}^{k-1}_{\min}$. Hence, we conclude that either $\bar{w}^{j+\ell} \in RS^k$ or, by combining (3.41) and (3.43), $\bar{w}^{j+\ell+1} \in RS^k$. ∎

It remains to comment briefly on the assumption put forward in (3.43): given that $\Phi(\bar{w}^k_{\min}) \sim \Phi_{\min}$ (cf. (3.36)) as well as that necessarily $\bar{w}^{\mathrm{c}}_k \lesssim \Phi_{\min} \lesssim \bar{w}^{\mathrm{c}}_{k-1}$ by (3.43), one can show that, to lowest order,

$$(2k-1)\frac{\mu^{\mathrm{c}}}{D_\mu}\sqrt{\varepsilon}\sqrt{-2\ln\varepsilon} \leq \mu^{\mathrm{c}} - \mu \leq (2k+1)\frac{\mu^{\mathrm{c}}}{D_\mu}\sqrt{\varepsilon}\sqrt{-2\ln\varepsilon}$$

must hold for (3.43) to be true, with μ^{c} defined as in (2.56). This condition is consistent with the estimate for $\Delta\mu^k$, e.g., given after the proof of Theorem 3.7, and will typically be satisfied if q is not "too large."

Remark 11. The restriction to $q < \frac{1}{2}$ in (3.43) is made to ensure that $\Phi_{\min} \in RS^k$ will imply $\Phi(\bar{w}^j_{\min}) \in RS^k$ for $0 \le j \le k-1$, since we can a priori conclude only $\Phi_{\min} - \Phi(\bar{w}^j_{\min}) = \mathcal{O}(\varepsilon)$ from (3.36).

One important consequence of Theorem 3.9 is that it allows us to give a precise qualitative description of the segments that the symbolic sequence of a given periodic orbit can contain. For any such orbit, let $k \ge 1$ be the largest integer such that the segment 1^k is contained in the corresponding Farey sequence. With this convention, $k = 1$ implies that the sequence can contain only the segments 1^1 and 1^0; restrictions on the sequences that can occur when $k \ge 2$ are given in the following theorem.

Theorem 3.10. *Assume that $k \ge 2$. Then, a periodic orbit can occur if its sequence consists of segments of the form 1^k (some number of times in succession), 1^{k-1} (some number of times in succession), and 1^{k-2} (preceded by 1^k and followed by 1^{k-1} or 1^k).*

Proof. First, let us assume that (3.43) is satisfied. Then, the result already follows from the proof of Theorem 3.9.

Now, suppose that (3.43) does not hold as well as that $\Phi_{\min} \sim \bar{w}^c_{k-1}$. Then, the steps given in the proof of Theorem 3.9 can be retraced until almost the very end, namely, up to the statement that $\bar{w}^{j+\ell}$ will be less than or equal to $\bar{w}^{k-1}_{\min}$ for some $\ell \ge 0$. Instead, if $\bar{w}^{j+\ell} \in RS^{k-1}$, we can now conclude only that $\bar{w}^{j+\ell+1} \in RS^k \cup RS^{k-1}$. The orbit can then either remain in RS^{k-1} or enter RS^k and subsequently jump back to either RS^{k-1} or RS^{k-2}. If, on the other hand, $\Phi_{\min} \sim \bar{w}^c_k$, the same kind of sequences can occur, with k shifted upward by 1. This completes the proof. ∎

Given the result of Theorem 3.10, a natural question that arises is how many times in succession a given segment can occur.

Proposition 3.11. *Let $k \ge 2$. If a periodic orbit involves all of the segments 1^{k-2}, 1^{k-1}, and 1^k, then both 1^{k-2} and 1^k can occur at most once in succession.*

Proof. First, note that an orbit can contain all of the segments 1^{k-2}, 1^{k-1}, and 1^k only if $\Phi_{\min} \sim \bar{w}^c_{k-1}$; see the proofs of Theorems 3.9 and 3.10. It follows that any point on the orbit that lies in RS^{k-2} must lie close to $\bar{w}^{k-1}_{\min}$ and, hence, that it must be mapped to $RS^{k-1} \cup RS^k$ under Φ. Similarly, any point on the orbit in RS^k must be close to $\bar{w}^c_{k-1}$ and therefore must be mapped to $RS^{k-2} \cup RS^{k-1}$. ∎

Finally, Theorem 3.10 allows us to make a precise statement on the periodic orbits of the type $\{L_j^{k_j}\}$ that can be observed for $L_j \ge 2$.

Corollary 3.12. *For $k \ge 2$, $L \ge 2$, and $L+k \ge 5$, there are no periodic orbits which contain the segment L^k and which pass through the part of RS^k defined by (3.40).*

Proof. Since the segment L^k corresponds to k small loops followed by L large relaxation excursions, this segment can also be written in the form $1^k(1^0)^{L-1}$. If $k = 0$ or $k = 1$, Theorem 3.10 places no restrictions on the existence of such segments. Furthermore, Theorem 3.10 implies that the only remaining admissible k-value is 2 and that $L - 1 = 1$ must hold in that case, implying $L = 2$. ∎

To put it differently, one will not "generically" observe Farey sequences of the form $\{L_j^{k_j}\}$ if $L_j \ge 3$; if $L_j = 2$, only segments of the form 2^1 or 2^2 will occur. The segment L_j^1, however, is admissible for any $L_j \ge 1$; this is due to "leakage" from RS^0 in the sense that $\Phi(\bar{w}) \lesssim \bar{w}$ for $\bar{w} \approx \bar{w}^c_0$, implying that trajectories can "drift" back into RS^1.

Finally, we note that we make no assumptions about the stability of the periodic orbits under consideration, neither in Theorem 3.10 nor in Corollary 3.12; indeed, our results apply to any orbit for which the condition in (3.40) is satisfied.

4. Conclusions and discussion. In the present article, we have studied mixed-mode oscillations (MMOs) in a three-dimensional model system of ordinary differential equations with three distinct time-scales; see (1.5). Here, the "superslow" variable w has been playing the role of a "dynamical parameter" which makes the (v, z)-subsystem of (1.5) move slowly through a canard explosion. One major advantage of our modeling ansatz is the fact that the resulting system dynamics is "almost" two-dimensional in the sense that the integrable structure close to a canard explosion can be exploited to derive the return map Π for the induced flow.

We are aware of two specific examples of three time-scale systems which exhibit mixed-mode dynamics akin to that studied here. One is a compartmental model for the dopaminergic neuron, first derived by Wilson and Callaway [37] and subsequently analyzed in [23] and [24], which in fact served as our motivation for formulating the simplified model system considered in this article. The other example is a model for a chemical reaction, discussed by Moehlis [28]. Although these two systems are not exactly analogous to the one studied here, they do share many of the underlying features and can be analyzed in a similar manner; see also the upcoming article [18].

The three time-scale model studied in this article is one realization of a more general canard mechanism that has been put forward to explain the mixed-mode dynamics often observed in multiscale dynamical systems [36, 2]. This generalized canard mechanism is defined as a combination of dynamical (local) passage through a canard point and a (global) return that resets the system dynamics after the passage has been completed; cf. also section 1. Other mechanisms that do not explicitly involve canards have been proposed to explain MMOs; examples include break-up of an invariant torus [21], loss of stability of a Shilnikov orbit [16], slow passage through Hopf bifurcation [22], and subcritical Hopf-homoclinic bifurcation [12, 13]. While these other mechanisms are consistent with some of the characteristic features of MMOs, they cannot typically explain all of them; see [2]. On the other hand, the generalized canard mechanism is consistent with most examples known to us of systems exhibiting mixed-mode-type behavior [2, 17]. In particular, we note that both the Shilnikov and the delayed Hopf mechanisms can be realized as an aspect of it. These and similar questions are the topic of ongoing research; see, e.g., the forthcoming article [3].

An explanation of mixed-mode dynamics based on the Shilnikov mechanism has been suggested by a number of authors (cf. [16] and the references therein) and is based on the similarities between the respective bifurcation sequences as well as on the presence of Shilnikov-type equilibria in systems that exhibit mixed-mode-type behavior. Roughly speaking, the Shilnikov phenomenon is the unfolding of a homoclinic orbit to an equilibrium of saddle type with a one-dimensional stable manifold and a two-dimensional unstable manifold of spiral focus type. Since Shilnikov-type equilibria are present in canard-based systems that involve a so-called *folded saddle-node (of type* II*)*, we propose that the latter systems do realize a "suitably modified" Shilnikov mechanism; cf. [3]. Similarly, a case of slow passage through Hopf bifurcation is seen in the dynamics near a folded saddle-node (of type II) and plays an important role there. This observation was made already in [26] and will also be fully

elucidated in [3].

Finally, it is important to note that the equations in (1.2) are neither of Shilnikov type nor of slow-passage-through-Hopf-bifurcation type, though they clearly realize the generalized canard mechanism. Moreover, due to our assumption that $\mu + \phi = \mathcal{O}(1)$ in (1.2), the mixed-mode dynamics analyzed in this article is neither of folded-node type nor of folded saddle-node type; cf. section 1. More precisely, in a folded-node system, i.e., for $\mu+\phi = \mathcal{O}(1)$ and negative in (1.2), the dynamics in the fold region would be strongly contractive and not oscillatory. Furthermore, this dynamics would be transient, since μ would cause w to increase until the relaxation regime in (1.1) is reached. The MMO patterns observed in this case would be regular and robust; irregular time series with two or more successive relaxation cycles would rarely occur upon variation of μ only. In a folded saddle-node system with μ small but ϕ large, on the other hand, one would typically observe slow passage through a Hopf bifurcation; moreover, the resulting mixed-mode dynamics would again be fairly regular in the sense that trajectories would generically consist of one relaxation excursion followed by a large number of "loops"; the amplitudes of these loops would be relatively small. This distinction is clearly reflected in the dynamics of (1.5), as predicted analytically in section 3 and verified numerically below.

Some of our findings on the mixed-mode dynamics of (1.5) are summarized and discussed in detail in the subsequent paragraphs.

A principal result of our analysis is the accurate reduction of the global return map Π (which is defined as a two-dimensional map on the Poincaré section $\overline{\Delta}_-$) to a one-dimensional map Φ which can be studied in a standard, straightforward way.

The first step of this reduction entails the restriction of Π from $\overline{\Delta}_-$ to the union of a set of (one-dimensional) intersecting curves. (These curves, which we have denoted by $\mathcal{C}_\varepsilon^j$, are defined recursively, with $\mathcal{C}_\varepsilon^0 \equiv \mathcal{C}_\varepsilon^-$ the flow image of the attracting slow manifold $\mathcal{S}_\varepsilon^{a-}$ in $\overline{\Delta}$ and $\mathcal{C}_\varepsilon^j = \Pi(\mathcal{C}_\varepsilon^{j-1})$, $j \geq 1$.) Most importantly, by Proposition 3.1, this reduction incurs an only exponentially small error; i.e., the sequence $\{\mathcal{C}_\varepsilon^j\}$ very accurately approximates the attractor of Π.

Then, in a second step, another reduction is performed, which yields a one-dimensional map Φ that is defined on the curve $\mathcal{C}_\varepsilon^-$. This map again gives an exponentially accurate approximation, this time for the $(k+1)$th iterate of Π on the kth sector of rotation, RS^k. (In other words, Φ restricted to RS^k describes the recurrent dynamics on RS^k with an exponentially small error.) Even though the map Φ is multimodal and possibly discontinuous at the boundaries of RS^k, it is one-dimensional and thus can be analyzed using techniques from one-dimensional discrete dynamics. It is interesting to note that, conceptually, the reduction to Φ is valid for any finite k, since the return of trajectories under Π will always eventually be to $\mathcal{C}_\varepsilon^-$. However, given the nonuniformity of our results in k (Proposition 3.4), one might have to consider higher-order terms (in ε) or, alternatively, take ε "very small" to describe the asymptotics accurately for "very large" k.

Some authors [23, 27] postulate a reduction to the dynamics of an *interval* map that would capture the properties of MMOs in systems of the type of (1.2). The fact that all MMO trajectories must pass extremely close to $\mathcal{S}_\varepsilon^{a-}$ is a strong indication that the system dynamics of (1.2) is almost two-dimensional in nature. Similarly, one might expect that the corresponding return map Π is almost one-dimensional. However, our results imply that a straightforward reduction of Π to a one-dimensional map defined on a single interval is not possible, whereas

the one-dimensional map Φ, which is defined on a *set of intervals* corresponding to the sectors of rotation, approximates Π with an only exponentially small error. By contrast, in [23], the return is approximated by a piecewise linear map, with a jump discontinuity corresponding to the strong canard, that admits a large variety of potential Farey sequences. Our analysis, on the other hand, resolves precisely the rich bifurcation structure of Φ close to the strong canard of (1.2), allowing us to characterize exactly which Farey sequences will actually be observed in (1.2), as well as to give accurate estimates of the relevant parameter intervals. (It is important to note, though, that the analysis in [23] does not focus primarily on resolving the canard structure in detail; rather, it is concerned with the system dynamics close to Hopf bifurcation which we do not analyze in detail here.)

The properties of Φ on RS^k directly determine those of the corresponding MMO trajectories of type 1^k, i.e., of periodic orbits for (1.5) which pass through the kth sector of rotation. Hence, a large part of our analysis is devoted to establishing the qualitative and quantitative asymptotics of the reduced return map Φ. More specifically, our results on the bifurcation structure of Φ as well as on the Farey sequences $L_0^{k_0} L_1^{k_1} \dots$ of the corresponding MMO trajectories include a proof of the existence and stability of 1^k-type orbits (Theorem 3.7), a precise description of the ordering of the Farey sequences that will "generically" occur for $L_j \equiv 1$ (Theorem 3.10 and Proposition 3.11), as well as a statement on the "improbability" of observing orbits with symbolic sequence $\{L_j^{k_j}\}$ when $L_j \geq 3$ (Corollary 3.12). It is important to note that these restrictions on the dynamics of Φ are by no means exhaustive; rather, they provide a sample of the types of results that can be proved using the techniques of section 3. A more comprehensive analysis, however, is beyond the scope of this work.

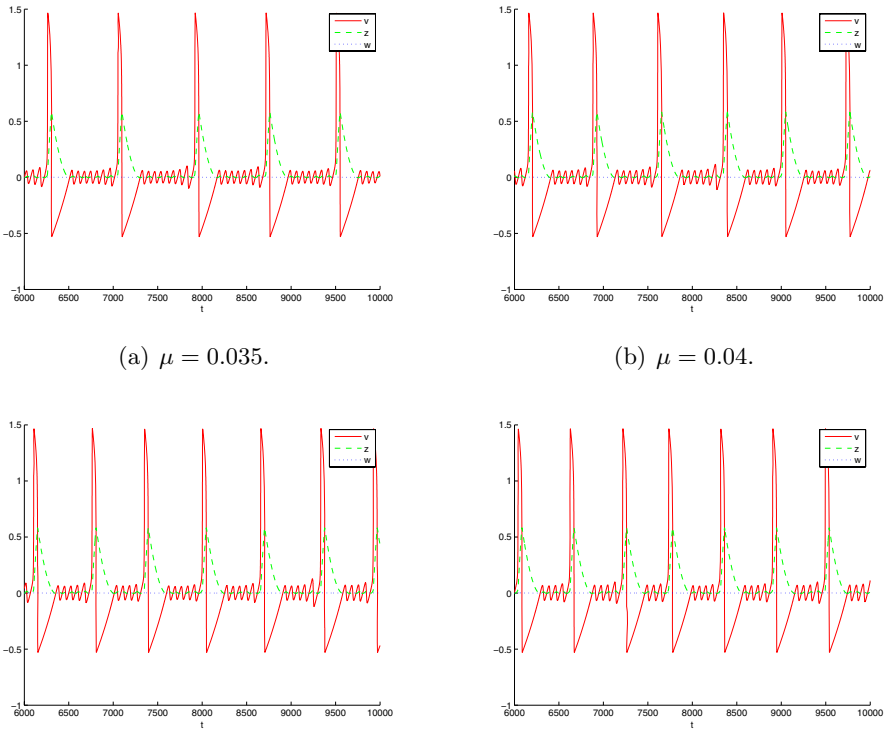
Another important aspect of the generalized canard mechanism is the asymptotic structure of secondary canards, as well as of the corresponding sectors of rotation. To date, rigorous results in this direction have only been obtained by Wechselberger [36] for systems of general folded-node type, via a bifurcation analysis of resonances. To the best of our knowledge, no comparable analysis has been available so far for other realizations of the generalized canard mechanism. The three time-scale structure of our problem in combination with the resulting near-integrability, however, allows us to obtain rather specific results; in particular, it enables us to derive a more or less explicit asymptotic estimate for the sector size: given the definition of the critical canard value $\bar{w}_0^{\mathrm{c}}$, as well as of the $\bar{w}$-value $\bar{w}_k^{\mathrm{c}}$ corresponding to the kth secondary canard Γ_ε^k, it follows with $w = \sqrt{\varepsilon}\bar{w}$ that $w^{\mathrm{c}} = \mathcal{O}(\varepsilon)$ after "blow-down," as well as that

$$\Delta w^k := \sqrt{\varepsilon}\Delta\bar{w}^k \sim 2\mu\varepsilon^{\frac{3}{2}}\sqrt{-2\ln\varepsilon}$$

is the width of $RS^k \subset \mathcal{C}_\varepsilon^-$, independent of k to leading order. This estimate confirms the well-known fact [33, 36] that the canard phenomenon is fairly "robust" in three dimensions in the sense that the relevant parameter intervals are relatively large, whereas in two dimensions, they are only exponentially small [20]: in our case, the width of the relevant w-interval will roughly be $\mathcal{O}(\varepsilon)$.

Finally, given the above discussion, our partly rigorous and partly heuristic conclusions on the bifurcation (Farey) structure of the mixed-mode dynamics which will typically be observed in (1.5) can be summed up as follows:

(i) Symbolic sequences of the form $\{1^k\}$ and $\{1^k 1^{k-1}\}$ dominate the stable dynamics; such



(a) $\mu = 0.035$. (b) $\mu = 0.04$.

(c) $\mu = 0.045$. (d) $\mu = 0.05$.

Figure 17. *The time series of v, z, and w in (1.5) for $f_2 = 1.5$, $f_3 = -1$, $g_1 = 0.5$, and $\varepsilon = 0.01$. As μ increases from* (a) 0.035 *via* (b) 0.04 *and* (c) 0.045 *to* (d) 0.05*, one observes a transition from $1^7 1^6$ via $1^6 1^5$ and $1^5 1^4$ to $1^4 1^3$ in the resulting Farey sequences.*

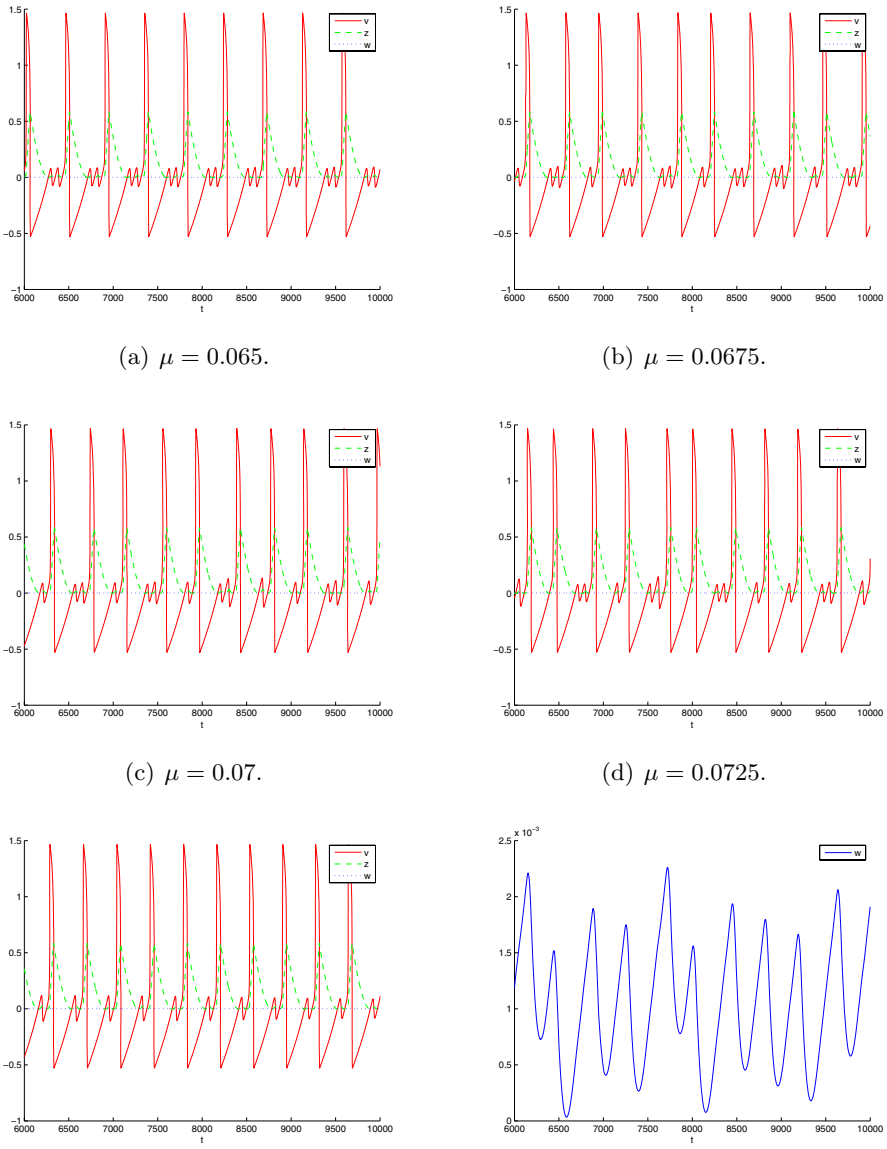
sequences correspond to MMO trajectories that visit only one sector of rotation and two adjacent sectors, respectively; see Figure 17.

(ii) Stable 1^k-type orbits are observed in a relatively small parameter range. Consequently, non-1^k orbits (i.e., orbits that are not periodic with Farey sequence $\{1^k\}$) dominate a significant portion of the parameter space. Moreover, they occur more frequently with increasing k, since the 1^k-stability intervals decrease in size as k increases; cf. Figures 17 and 18.

(iii) For $L_j \geq 2$, segments of the form $L_j^{k_j}$ are not generically observed when $k_j \geq 2$, except for the segment 2^2. The segment L_j^1, on the other hand, is possible for any $L_j \geq 1$; see Figure 19.

(iv) As μ increases, the Farey sequences observed in the transition are roughly of the form $\ldots \to 1^k \to 1^k 1^{k-1} \to 1^{k-1} \to \ldots$; in particular, all sectors of rotation are "swept through" until $\mu > \mu^c$, when the dynamics finally enters the relaxation regime (cf. Figures 18 and 20).

(v) The local dynamics depends quite sensitively on the curvature of $f(v)$, i.e., on the coefficient f_2; in particular, 1^k-type orbits become increasingly harder to observe with



(a) $\mu = 0.065$.

(b) $\mu = 0.0675$.

(c) $\mu = 0.07$.

(d) $\mu = 0.0725$.

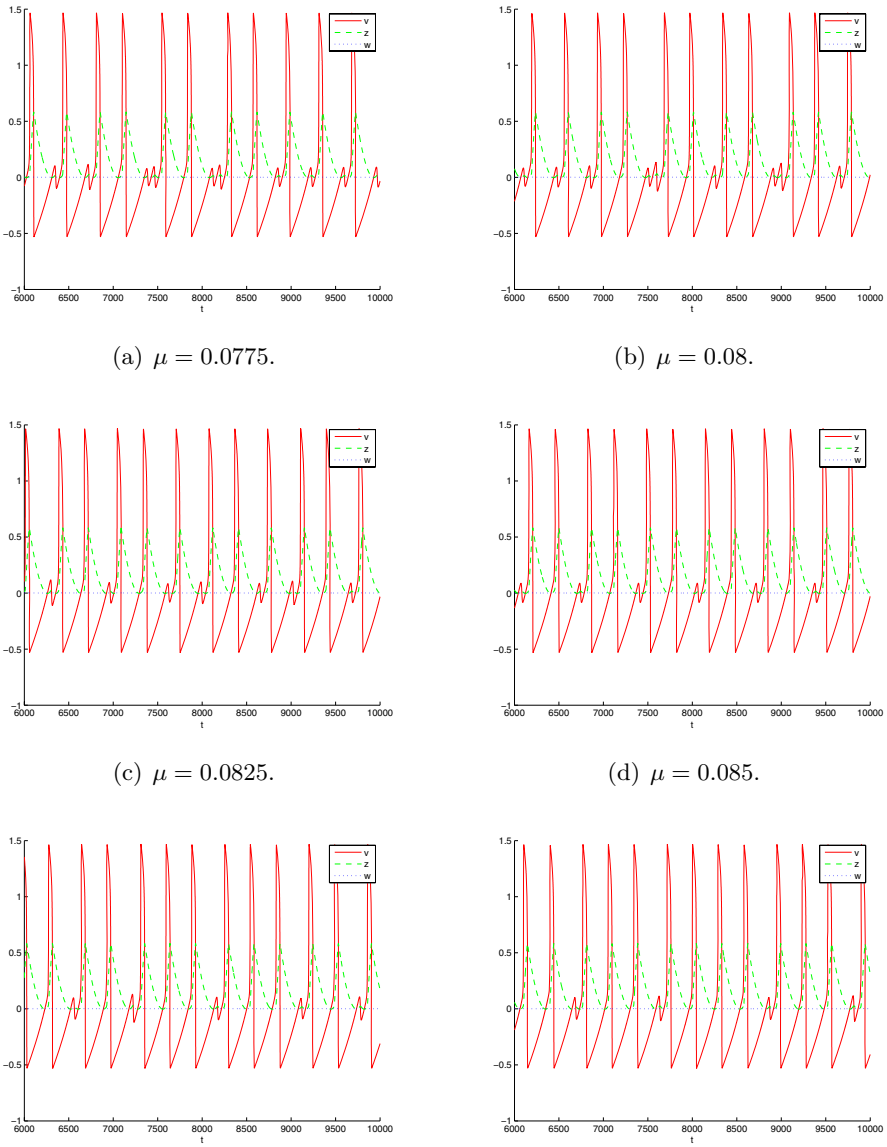
(e) $\mu = 0.075$.

(f) $\mu = 0.0725$.

Figure 18. *The time series of v, z, and w in (1.5) for $f_2 = 1.5$, $f_3 = -1$, $g_1 = 0.5$, and $\varepsilon = 0.01$. As μ increases from* (a) 0.065 *via* (b) 0.0675, (c) 0.07, *and* (d) 0.0725 *to* (e) 0.075, *one observes a transition from 1^2 to 1^1 in the resulting Farey sequences, with transitory sequences which contain mixed segments of the form $1^2 1^1$ as well as $2^2 1^2 1^1$. Panel* (f) *shows a zoom on the time series of w for $\mu = 0.0725$; clearly, $w = \mathcal{O}(\varepsilon)$, in accordance with Assumption 1 (cf. also section 3.3).*

growing f_2; see Figure 21(a).

(vi) The number of sectors visited is also influenced by the strength of the global dynamics, i.e., by how far "back" w is reset after relaxation: the smaller the parameter g_1 is, the



(a) $\mu = 0.0775$. (b) $\mu = 0.08$.

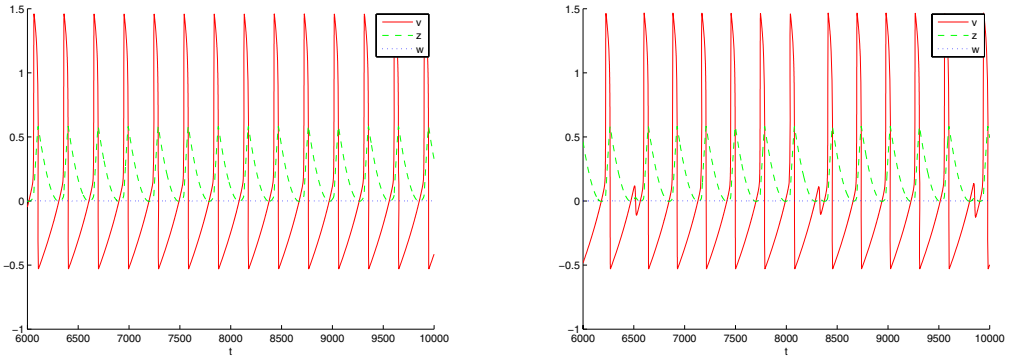
(c) $\mu = 0.0825$. (d) $\mu = 0.085$.

(e) $\mu = 0.0875$. (f) $\mu = 0.09$.

Figure 19. *The time series of v, z, and w in (1.5) for $f_2 = 1.5$, $f_3 = -1$, $g_1 = 0.5$, and $\varepsilon = 0.01$. As μ increases from* (a) 0.0775 *via* (b) 0.08 *to* (c) 0.0825, *one observes a variety of complex Farey sequences, with segments containing 1^1, 2^2, and 2^1 as well as repetitions thereof. As μ is increased further to* 0.09, *one observes a transition from* (c) $1^1 2^1$ *via* (d) 2^1 *and* (e) $2^1 3^1$ *to* (f) $3^1 4^1$, *as predicted analytically in section 3.5.*

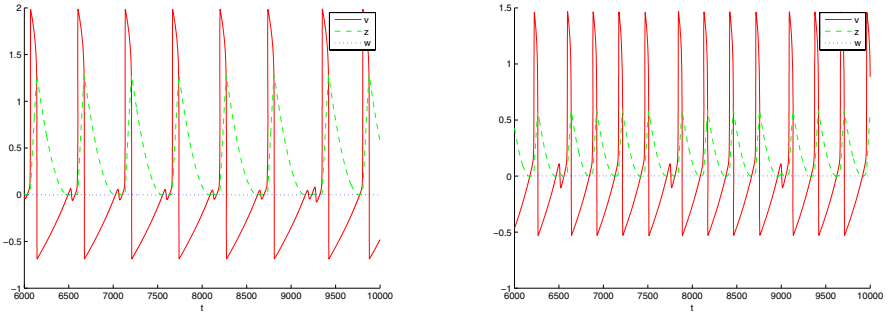
closer to the strong canard trajectories will return after relaxation, and the smaller the relevant μ-interval will be; cf. Figure 21(b).

(vii) Since $w = \mathcal{O}(\varepsilon)$ throughout (see Figure 18(f)), the global return point will be $\mathcal{O}(\varepsilon)$-close (in w) to the strong canard. This implies that only the "lower" sectors will



(a) $\mu = 0.095$. (b) $\mu = 0.0925$.

Figure 20. *The time series of v, z, and w for $\mu = 0.095$ and $\mu = 0.0925$. Clearly, the system is in the pure relaxation regime in* (a)*, whereas in* (b)*, one observes already mixed-mode dynamics, in agreement with the theoretical prediction that the critical μ-value should be $\mu^c \approx 0.0938$, up to an $\mathcal{O}(\varepsilon)$-error.*



(a) $\mu = 0.019$, $f_2 = 2$, $g_1 = 0.5$. (b) $\mu = 0.045$, $f_2 = 1.5$, $g_1 = 0.25$.

Figure 21. *The effects of a change of* (a) f_2 *and* (b) g_1 *on the dynamics of* (1.5)*. As f_2 is increased from* 1.5 *to* 2*, the stability interval of 1^1-type orbits decreases, since $D_\mu = 2.25$ is replaced by $D_\mu = 4$ and since $\Delta\mu^1 \propto D_\mu^{-1}$; see Theorem* 3.7*. As g_1 is decreased from* 0.5 *to* 0.25*, the dynamics recurs to lower sectors of rotation; cf. Figure* 17(c)*.*

typically be involved in the dynamics, resulting in MMO trajectories with a submaximum number of small oscillations.

(viii) As k increases or, alternatively, as μ decreases, the sectors of rotation decrease in size. Overall, however, the dynamics seems to become less expanding with higher k, making it less likely for sequences containing segments of the form $1^k1^{k-\ell}$, $\ell > 1$, to occur.

With the exception of the conjecture in (viii), these observations are reflected by our numerical findings; see Figures 17 to 21 as referred to in the individual items. Figure 17 shows a sample of regular 1^k1^{k-1}-type orbits for $k = 4, \ldots, 7$, while Figure 18 illustrates the transition from 1^2 to 1^1 via mixed transitory segments of the form $2^21^21^1$; Figure 19 indicates how Farey sequences with mixed segments containing 1^1, 2^2, and 2^1, as well as L_j^1-type sequences with

$L_j \geq 1$, can arise; Figure 20 illustrates the transition from mixed-mode dynamics to the pure relaxation regime at $\mu = \mu^c$ in (1.5); finally, in Figure 21, (a) and (b) exemplify the effects of a change in f_2 and g_1, respectively, on the dynamics of (1.5). In each case, the relevant parameter regimes are specified in detail in the corresponding captions. All numerical simulations were performed in MATLAB using the predefined routine `ode23tb` with absolute and relative accuracies 10^{-10} and 10^{-8}, respectively. For clarity, the results are illustrated starting at $t = 6000$, after initial transients have subsided.

Appendix A. Some asymptotic results. In this appendix, we summarize a few results on the asymptotics of the rescaled system (2.3), as well as of its generalization in (2.8). Recall that the equations in (2.3) are given by

$$\bar{v}' = -\bar{z} + f_2\bar{v}^2 + \sqrt{\varepsilon} f_3 \bar{v}^3, \tag{A.1a}$$

$$\bar{z}' = \bar{v} - \bar{w}, \tag{A.1b}$$

$$\bar{w}' = \varepsilon(\mu - g_1\varepsilon\bar{z}) \tag{A.1c}$$

as well as that they reduce, for $\varepsilon = 0$, to

$$\bar{v}' = -\bar{z} + f_2\bar{v}^2, \tag{A.2a}$$

$$\bar{z}' = \bar{v} - \bar{w}, \tag{A.2b}$$

$$\bar{w}' = 0; \tag{A.2c}$$

cf. (2.4). For $\bar{w} = 0$, the system in (A.2) is integrable. Moreover, given the constant of motion

$$H(\bar{v}, \bar{z}) = \frac{1}{2}\mathrm{e}^{-2f_2\bar{z}}\Big(-\bar{v}^2 + \frac{\bar{z}}{f_2} + \frac{1}{2f_2^2}\Big) \tag{A.3}$$

as defined in (2.5), the orbits of (A.2) correspond in a unique fashion to the level curves of H with $H = h$ constant; cf. section 2. More precisely, to any $h < h_0 = (4f_2^2)^{-1}$, we can assign a unique $\bar{z}$-value $\bar{z}^h$ in $\overline{\Delta}_-$. For any such point $(0, \bar{z}^h, \bar{w}) \in \overline{\Delta}_-$, we denote the corresponding solution to (A.1) by $\bar{\gamma}_\varepsilon^h$. Here, we assume the parametrization to be such that $\bar{\gamma}_\varepsilon^h(-T^h(\bar{w})) = (0, \bar{z}^h, \bar{w})$ holds and that $\bar{\gamma}_\varepsilon^h(T^h(\bar{w}))$ is the point of first return to $\overline{\Delta}_-$; recall Figure 6.

In the particular case when $\bar{w} = 0$, we write $T^h = T^h(0)$. Let $h > 0$ be fixed, and let $\bar{\gamma}_0^h$ denote the corresponding (periodic) solution of (A.2). For convenience, we denote the $\bar{z}$-coordinates of the two points of intersection of $\bar{\gamma}_0^h$ with $\overline{\Delta}$ by ξ^h and ζ^h, respectively; see Figure 22.

Lemma A.1. *There holds* $\zeta^h = \frac{1}{2f_2}(-\ln h) + \mathcal{O}(1)$ *and* $\xi^h = -\frac{1}{2f_2} + \mathcal{O}(h)$.

Proof. The assertion follows from (A.3): note that $\bar{v} = 0$ in $\overline{\Delta}_-$, and expand $\frac{\bar{z}}{f_2} + \frac{1}{2f_2^2} = h\frac{1}{2}\mathrm{e}^{2f_2\bar{z}}$ for $\bar{z}$ large, respectively, for $\bar{z}$ (asymptotically) constant, to obtain the expansions for ζ^h and ξ^h, respectively. ∎

Given Lemma A.1, we have the following result on the asymptotics of T^h.

Lemma A.2. *There holds* $T^h = \sqrt{2}(-\ln h)^{\frac{1}{2}} + \mathcal{O}(1)$.

Proof. Given (A.3), we first express $\bar{v}$ via

$$\bar{v} = \sqrt{\frac{\bar{z}}{f_2} + \frac{1}{2f_2^2} - 2h\mathrm{e}^{2f_2\bar{z}}} = \sqrt{\frac{\bar{z}}{f_2}}\sqrt{1 + \frac{1}{2f_2\bar{z}} - 2f_2 h\frac{\mathrm{e}^{2f_2\bar{z}}}{\bar{z}}} \tag{A.4}$$

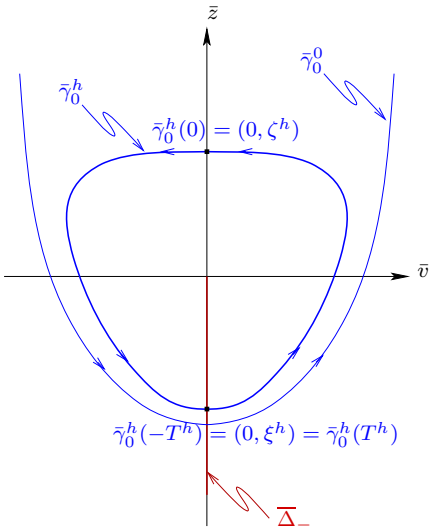


Figure 22. *A typical solution of* (A.2).

and then make use of $\bar{v} = \bar{z}' = \frac{d\bar{z}}{dt}$ for $\bar{w} = 0$ (see (A.4)) to obtain

$$\int_{\xi^h}^{\zeta^h} \frac{d\bar{z}}{\sqrt{\frac{\bar{z}}{f_2}}\sqrt{1 + \frac{1}{2f_2\bar{z}} - 2f_2 h \frac{\mathrm{e}^{2f_2\bar{z}}}{\bar{z}}}} = \int_{-T^h}^{0} dt.$$

Integrating the left-hand side by parts, we find that

$$\text{(A.5)}\quad T^h = 2\sqrt{f_2\bar{z}}\Big(1 + \frac{1}{2f_2\bar{z}} - 2f_2 h\frac{\mathrm{e}^{2f_2\bar{z}}}{\bar{z}}\Big)^{-\frac{1}{2}}\Big|_{\xi^h}^{\zeta^h} + \int_{\xi^h}^{\zeta^h} \sqrt{f_2\bar{z}}\Big(1 + \frac{1}{2f_2\bar{z}} - 2f_2 h\frac{\mathrm{e}^{2f_2\bar{z}}}{\bar{z}}\Big)^{-\frac{3}{2}}\Big(-\frac{1}{2f_2\bar{z}^2} - 2f_2 h\frac{\mathrm{e}^{2f_2\bar{z}}}{\bar{z}^2}(2f_2\bar{z} - 1)\Big)\, d\bar{z}.$$

From Lemma A.1, it follows that the leading-order contribution in the first term on the right-hand side of (A.5) comes from the evaluation at the upper limit ζ^h. Moreover, by expanding the integrand in the second term, one can check that the corresponding integral will contribute only terms of $\mathcal{O}(1)$. Hence, again by Lemma A.1, $T^h \sim 2\sqrt{f_2\zeta^h} \sim \sqrt{2}(-\ln h)^{\frac{1}{2}}$. This concludes the proof. ■

Recall the definitions of $d^h_{\sqrt{\varepsilon}}$ and $d^h_{\bar{w}}$ in (2.10) and (2.11), respectively:

$$\text{(A.6a)}\qquad d^h_{\sqrt{\varepsilon}} = \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot \big(f_3\bar{v}_0^h(t)^3, 0\big)^T\, dt,$$

$$\text{(A.6b)}\qquad d^h_{\bar{w}} = \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (0, -1)^T\, dt.$$

For a numerical evaluation of the transition map $\overline{\Pi} : \overline{\Delta}_- \to \overline{\Delta}_-$ (as defined in section 2.2), it is convenient to express $d^h_{\sqrt{\varepsilon}}$ and $d^h_{\bar{w}}$ as follows.

Lemma A.3. *Let the integrals $\mathcal{I}_1$ and $\mathcal{I}_2$ be defined by*

$$\mathcal{I}_1(h) := 2\int_{\xi^h}^{\zeta^h} \mathrm{e}^{-2f_2\bar{z}}\bar{v}_0^h(\bar{z})\,d\bar{z} \quad \textit{and} \quad \mathcal{I}_2(h) := 2\int_{\xi^h}^{\zeta^h} \mathrm{e}^{-2f_2\bar{z}}\bar{v}_0^h(\bar{z})^3\,d\bar{z},$$

respectively, with ξ^h and ζ^h as above. Then, there holds

$$d_{\bar{w}}^h = -2f_2\mathcal{I}_1(h) \quad \textit{and} \quad d_{\sqrt{\varepsilon}}^h = -f_3\mathcal{I}_2(h). \tag{A.7}$$

Proof. We will verify the assertion for $d_{\sqrt{\varepsilon}}^h$ first: since

$$\frac{\partial H}{\partial \bar{v}} = -\bar{v}\mathrm{e}^{-2f_2\bar{z}} \quad \text{and} \quad \frac{\partial H}{\partial \bar{z}} = (f_2\bar{v}^2 - \bar{z})\mathrm{e}^{-2f_2\bar{z}},$$

it follows that $\nabla H \cdot (f_3\bar{v}^3, 0)^T = -f_3\bar{v}^4\mathrm{e}^{-2f_2\bar{z}}$. To replace the t-integration in (A.6a) by an integration with respect to $\bar{z}$, we make use of the fact that $\frac{d\bar{z}}{dt} = \bar{z}' = \bar{v}$ for $\bar{w} = 0$. Then,

$$d_{\sqrt{\varepsilon}}^h = -2f_3\int_{\xi^h}^{\zeta^h} \mathrm{e}^{-2f_2\bar{z}}\bar{v}_0^h(\bar{z})^3\,d\bar{z},$$

since $(\bar{v}_0^h, \bar{z}_0^h)(-t) = (-\bar{v}_0^h, \bar{z}_0^h)(t)$ on $\bar{\gamma}_0^h$. To evaluate $d_{\bar{w}}^h$, note that the corresponding integrand in (A.6b) is given by $-f_2\bar{v}^2 + \bar{z}$. Also, it follows from (A.4) that $\bar{v}$ and $\bar{z}$ are related via $\bar{z} = -\bar{v}' + f_2\bar{v}^2$. The result then follows from an integration by parts, since $\bar{v}_0^h(\xi^h) = 0 = \bar{v}_0^h(\zeta^h)$ by definition. ∎

In general, for $h \neq 0$, the integrals $\mathcal{I}_1$ and $\mathcal{I}_2$ cannot be computed analytically but have to be approximated numerically. However, for $h = 0$, one can evaluate $\mathcal{I}_1$ and $\mathcal{I}_2$ exactly by integrating by parts repeatedly. Recalling the definition of $\bar{\gamma}_0^0$ in (2.6), one finds, for instance,

$$\mathcal{I}_1(0) = \frac{\mathrm{e}}{4f_2^2}\int_{-\infty}^{\infty} t^2\mathrm{e}^{-\frac{t^2}{2}}\,dt = \frac{\mathrm{e}}{4f_2^2}\left(-t\mathrm{e}^{-\frac{t^2}{2}}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} t^2\mathrm{e}^{-\frac{t^2}{2}}\,dt\right) = \frac{\mathrm{e}\sqrt{2\pi}}{4f_2^2}.$$

Similarly, one can show $\mathcal{I}_2(0) = \frac{3\mathrm{e}\sqrt{2\pi}}{16f_2^4}$; see also [19, 20]. In particular, this implies

$$d_{\bar{w}}^0 = -\frac{1}{2f_2}\sqrt{2\pi}\mathrm{e} < 0 \quad \text{and} \quad d_{\sqrt{\varepsilon}}^0 = -\frac{3f_3}{16f_2^4}\sqrt{2\pi}\mathrm{e} > 0.$$

We require the following result on the asymptotics of $\mathcal{I}_2(h)$ for h small.

Lemma A.4. *There holds $\mathcal{I}_2(h) = \mathcal{I}_2(0) - \frac{\sqrt{2}}{f_2^2}h(-\ln h)^{\frac{3}{2}} + \mathcal{O}(h(-\ln h)^{\frac{1}{2}})$.*

Proof. We will prove the assertion by first determining the leading-order behavior of $\frac{d\mathcal{I}_2}{dh}$: given $\bar{v}^2 = \frac{\bar{z}}{f_2} + \frac{1}{2f_2^2} - h\mathrm{e}^{2f_2\bar{z}}$, we obtain by implicit differentiation that $\frac{\partial\bar{v}}{\partial h} = -\bar{v}^{-1}\mathrm{e}^{2f_2\bar{z}}$ and, hence, that

$$\frac{d\mathcal{I}_2(h)}{dh} \sim -6\int_{\xi^h}^{\zeta^h} \bar{v}_0^h(\bar{z})\,d\bar{z}.$$

As in the proof of Lemma A.1, we now make use of (A.4) and then perform an integration by parts to find

$$\frac{d\mathcal{I}_2(h)}{dh} = -\frac{4}{\sqrt{f_2}}(\zeta^h)^{\frac{3}{2}} + \mathcal{O}\big((\zeta^h)^{\frac{1}{2}}\big) = -\frac{\sqrt{2}}{f_2^2}(-\ln h)^{\frac{3}{2}} + \mathcal{O}\big((-\ln h)^{\frac{1}{2}}\big). \tag{A.8}$$

The assertion follows by integrating (A.8) with respect to h, to leading order. ∎

Given Lemma A.4, one can write

$$d^h_{\sqrt{\varepsilon}} = d^0_{\sqrt{\varepsilon}} + \mathcal{R}(h) = d^0_{\sqrt{\varepsilon}} + h\widetilde{\mathcal{R}}(h), \tag{A.9}$$

where $\mathcal{R}(h)$ denotes the corresponding remainder term and $\widetilde{\mathcal{R}}(h) = -f_3\frac{d\mathcal{I}_2}{dh}$ is the first-order coefficient in the Taylor expansion of $d^h_{\sqrt{\varepsilon}}$ about $h = 0$. Note that the leading-order asymptotics of $d^h_{\bar{w}}$ can be obtained in a similar manner.

Finally, recall the generalized system of equations from (2.8), as well as the definition of the corresponding return map in (3.16),

$$\overline{\Pi}(h,\bar{w}) = \begin{pmatrix} P_h\overline{\Pi}_0(h,\bar{w}) + \varepsilon\mu\mathcal{K}(h) + \mathcal{O}(\varepsilon^2) \\ \bar{w} + 2\varepsilon\mu T^h + \mathcal{O}(\varepsilon^2) \end{pmatrix},$$

where $\overline{\Pi}_0$ denotes the return map for (3.14) and $\mathcal{K}$ is defined via

$$\mathcal{K}(h) = \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (G(0,0),-1)^T(t+T^h)\,dt.$$

An estimate for $\mathcal{K}$ is derived as follows.

Lemma A.5. *There holds*

$$\mathcal{K}(h) = 2d^0_{\bar{w}}T^h + \mathcal{O}(1). \tag{A.10}$$

Proof. Recall that, by definition, we have

$$\int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (G(0,0),-1)^T\,dt = \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (0,-1)^T\,dt = d^h_{\bar{w}};$$

see (A.6b) and the proof of Proposition 2.2. It follows that

$$\mathcal{K}(h) = d^h_{\bar{w}}T^h + \int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (G(0,0),-1)^T t\,dt.$$

To estimate the above integral, note that

$$\begin{aligned}\int_{-T^h}^{T^h} \nabla H(\bar{\gamma}_0^h(t)) \cdot (G(0,0),-1)^T t\,dt &= -G(0,0)\int_{-T^h}^{T^h} \bar{v}_0^h(t)\mathrm{e}^{-2f_2\bar{z}_0^h(t)}t\,dt \\ &\quad - \int_{-T^h}^{T^h} \big(f_2\bar{v}_0^h(t)^2 - \bar{z}_0^h(t)\big)\mathrm{e}^{-2f_2\bar{z}_0^h(t)}t\,dt.\end{aligned} \tag{A.11}$$

Since the integrand in the second integral on the right-hand side of (A.11) is odd in t, that integral vanishes. Hence, it remains to estimate the first integral: using integration by parts, we obtain

$$\int_{-T^h}^{T^h} \bar{v}_0^h(t)\mathrm{e}^{-2f_2\bar{z}_0^h(t)}t\,dt = T^h\int_{-T^h}^{T^h} \bar{v}_0^h(t)\mathrm{e}^{-2f_2\bar{z}_0^h(t)}\,dt - \int_{-T^h}^{T^h}\int_{-T^h}^{t} \bar{v}_0^h(s)\mathrm{e}^{-2f_2\bar{z}_0^h(s)}\,ds\,dt. \tag{A.12}$$

The first integral on the right-hand side of (A.12) is again zero, since the corresponding integrand is odd in t. Next, we recall that $(\bar{v}_0^h, \bar{z}_0^h)(t)$ is a solution of (A.2) for $\bar{w} = 0$ and, hence, that

$$\bar{v}_0^h(s)\mathrm{e}^{-2f_2\bar{z}_0^h(s)} = -\frac{1}{2f_2}\frac{d}{ds}\big(\mathrm{e}^{-2f_2\bar{z}_0^h(s)}\big).$$

Consequently,

$$\int_{-T^h}^{t} \bar{v}_0^h(s)\mathrm{e}^{-2f_2\bar{z}_0^h(s)}\,ds = -\frac{1}{2f_2}\big(\mathrm{e}^{-2f_2\bar{z}_0^h(t)} - \mathrm{e}^{-2f_2\xi^h}\big),$$

where $\xi^h = \bar{z}_0^h(\pm T^h)$, as before. Now, since

$$\int_{-T^h}^{T^h} \mathrm{e}^{-2f_2\bar{z}_0^h(t)}\,dt$$

is bounded, i.e., $\mathcal{O}(1)$, we conclude that

$$\int_{-T^h}^{T^h} \bar{v}_0^h(t)\mathrm{e}^{-2f_2\bar{z}_0^h(t)}t\,dt \sim \frac{1}{f_2}T^h\mathrm{e}^{-2f_2\xi^h},$$

and it remains only to estimate $G(0,0)$: indeed, by (2.36) and (2.37), there holds

$$G(0,0) = -\frac{d\bar{z}^{h^+(\bar{w},\varepsilon)}}{d\bar{w}}(0,0).$$

Recalling that the relationship between $\bar{z}^h$ and h is given implicitly by

$$\frac{1}{2f_2}\mathrm{e}^{-2f_2\bar{z}^h}\left(\bar{z}^h + \frac{1}{2}\right) = h$$

(cf. (A.3)), we find from an implicit differentiation that

$$\frac{d\bar{z}^h}{dh} = 2f_2\mathrm{e}^{2f_2\bar{z}^h} + \mathcal{O}(h).$$

Since $\xi^h = \bar{z}^h$, (2.35) shows that

$$\frac{dh^+(\bar{w},\sqrt{\varepsilon})}{d\bar{w}}(0,0) = d_{\bar{w}}^+ = -\frac{1}{2}d_{\bar{w}}^0$$

and, hence, that $G(0,0) = d_{\bar{w}}^0 f_2\mathrm{e}^{2f_2\xi^h}$. The result follows. ∎

Acknowledgments. The authors are grateful to Horacio Rotstein for his involvement during the early stages of this work, as well as to Alexey Kuznetsov, Georgi Medvedev, and Martin Wechselberger for valuable discussions and comments, and to Heidi Lyons and Paola Malerba for their careful reading of (parts of) the original manuscript.

REFERENCES

[1] E. Benoit, J.-L. Callot, F. Diener, and M. Diener, *Chasse au canard*, Collect. Math., 32 (1981), pp. 37–119.

[2] M. Brøns, M. Krupa, and M. Wechselberger, *Mixed mode oscillations due to the generalized canard phenomenon*, in Bifurcation Theory and Spatio-Temporal Pattern Formation, Fields Inst. Commun. 49, AMS, Providence, RI, 2006, pp. 39–63.

[3] M. Brøns, M. Krupa, and M. Wechselberger, *Dynamics Near a Folded Saddle-Node of Type* II, in preparation.

[4] K. M. Brucks and C. Tresser, *A Farey tree organization of locking regions for simple circle maps*, Proc. Amer. Math. Soc., 124 (1996), pp. 637–647.

[5] M. Diener, *The canard unchained or how fast/slow dynamical systems bifurcate*, Math. Intelligencer, 6 (1984), pp. 38–49.

[6] J. Drover, J. Rubin, J. Su, and B. Ermentrout, *Analysis of a canard mechanism by which excitatory synaptic coupling can synchronize neurons at low firing frequencies*, SIAM J. Appl. Math., 65 (2004), pp. 69–92.

[7] F. Dumortier, *Techniques in the theory of local bifurcations: Blow-up, normal forms, nilpotent bifurcations, singular perturbations*, in Bifurcations and Periodic Orbits of Vector Fields, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 408, D. Schlomiuk, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 19–73.

[8] F. Dumortier and R. Roussarie, *Canard cycles and center manifolds*, Mem. Amer. Math. Soc., 121 (577) (1996).

[9] W. Eckhaus, *Relaxation oscillations including a standard chase on French ducks*, in Asymptotic Analysis, II, Lecture Notes in Math. 985, Springer-Verlag, Berlin, 1983, pp. 449–494.

[10] I. R. Epstein and K. Showalter, *Nonlinear chemical dynamics: Oscillations, patterns, and chaos*, J. Phys. Chem., 100 (1996), pp. 13132–13147.

[11] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[12] J. Guckenheimer, R. Harris-Warwick, J. Peck, and A. Wilms, *Bifurcation, bursting and frequency spike adaptation*, J. Comput. Neurosci., 4 (1997), pp. 255–277.

[13] J. Guckenheimer and A. Wilms, *Asymptotic analysis of subcritical Hopf-homoclinic bifurcation*, Phys. D, 139 (2000), pp. 196–216.

[14] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme, 1994), Lecture Notes in Math. 1609, Springer-Verlag, Berlin, 1995, pp. 44–118.

[15] N. Kopell and L. N. Howard, *Bifurcations and trajectories joining critical points*, Advances in Math., 18 (1975), pp. 306–358.

[16] M. T. M. Koper, *Bifurcations of mixed-mode oscillations in a three-variable autonomous Van der Pol-Duffing model with a cross-shaped phase diagram*, Phys. D, 80 (1995), pp. 72–94.

[17] M. Krupa, *Mixed-Mode Oscillations in Systems with More than Two Slow Dimensions*, in preparation.

[18] M. Krupa, N. Popović, N. Kopell, and H. G. Rotstein, *Mixed-mode oscillations in a three time-scale model for the dopaminergic neuron*, Chaos, to appear.

[19] M. Krupa and P. Szmolyan, *Extending geometric singular perturbation theory to nonhyperbolic points—fold and canard points in two dimensions*, SIAM J. Math. Anal., 33 (2001), pp. 286–314.

[20] M. Krupa and P. Szmolyan, *Relaxation oscillation and canard explosion*, J. Differential Equations, 174 (2001), pp. 312–368.

[21] R. Larter and C. G. Steinmetz, *Chaos via mixed-mode oscillations*, Philos. Trans. Roy. Soc. London Ser. A, 337 (1991), pp. 291–298.

[22] R. Larter, C. G. Steinmetz, and B. D. Aguda, *Fast-slow variable analysis of the transition to mixed-mode oscillations and chaos in the peroxidase reaction*, J. Phys. Chem., 89 (1988), pp. 6506–6514.

[23] G. S. Medvedev and J. E. Cisternas, *Multimodal regimes in a compartmental model of the dopamine neuron*, Phys. D, 194 (2004), pp. 333–356.

[24] G. S. Medvedev, C. J. Wilson, J. C. Callaway, and N. Kopell, *Dendritic synchrony and transient dynamics in a coupled oscillator model of the dopaminergic neuron*, J. Comput. Neurosci., 15 (2003), pp. 53–69.

[25] W. De Melo and S. Van Strien, *One-Dimensional Dynamics*, Ergeb. Math. Grenzgeb. (3) 25, Springer-Verlag, Berlin, 1993.
[26] A. Milik and P. Szmolyan, *Multiple time scales and canards in a chemical oscillator*, in Multiple-Time-Scale Dynamical Systems, IMA Vol. Math. Appl. 122, C. K. R. T. Jones and A. I. Khibnik, eds., Springer-Verlag, New York, 2001, pp. 117–140.
[27] A. Milik, P. Szmolyan, H. Löffelmann, and E. Gröller, *Geometry of mixed-mode oscillations in the 3-d autocatalator*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 8 (1998), pp. 505–519.
[28] J. Moehlis, *Canards in a surface oxidation reaction*, J. Nonlinear Sci., 12 (2002), pp. 319–345.
[29] C. A. Del Negro, C. G. Wilson, R. J. Butera, H. Rigatto, and J. C. Smith, *Periodicity, mixed-mode oscillations, and quasiperiodicity in a rhythm-generating neural network*, Biophys. J., 82 (2002), pp. 206–214.
[30] H. G. Rotstein and R. Kuske, *Localized and asynchronous patterns via canards in coupled calcium oscillators*, Phys. D, 215 (2006), pp. 46–61.
[31] H. G. Rotstein, T. Oppermann, J. A. White, and N. Kopell, *The dynamic structure underlying subthreshold oscillatory activity and the onset of spikes in a model of medial entorhinal cortex stellate cells*, J. Comput. Neurosci., 21 (2006), pp. 271–292.
[32] J. Rubin and M. Wechselberger, *Giant squid—hidden canard: The 3D geometry of the Hodgkin-Huxley model*, Biol. Cybernet., 97 (2007), pp. 5–32.
[33] P. Szmolyan and M. Wechselberger, *Singularly perturbed folds and canards in* $\mathbb{R}^3$, J. Differential Equations, 177 (2001), pp. 419–453.
[34] P. Szmolyan and M. Wechselberger, *Relaxation oscillations in* $\mathbb{R}^3$, J. Differential Equations, 200 (2004), pp. 69–104.
[35] M. Wechselberger, *personal communication*.
[36] M. Wechselberger, *Existence and bifurcation of canards in* $\mathbb{R}^3$ *in the case of a folded node*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 101–139.
[37] C. J. Wilson and J. C. Callaway, *Coupled oscillator model of the dopaminergic neuron of the substantia nigra*, J. Neurophysiol., 83 (2000), pp. 3084–3100.
[38] A. M. Zhabotinsky, *Periodic kinetics of oxidation of malonic acid in solution*, Biofizika, 9 (1964), pp. 306–311.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 7, No. 2, pp. 421–436
© 2008 Society for Industrial and Applied Mathematics

Chaotic n-Dimensional Euclidean and Hyperbolic Open Billiards and Chaotic Spinning Planar Billiards*

Ali Deniz†, Judy Kennedy‡, Şahin Koçak†, Andrei V. Ratiu§, Cevat Üstün¶, and James A. Yorke¶

Abstract. We propose a new method to handle the n-dimensional billiard problem in the exterior of a finite mutually disjoint union of convex (but not necessarily strictly convex) smooth obstacles without eclipse in the Euclidean or hyperbolic n-space, and we prove that there exist trajectories visiting the obstacles in any given doubly infinite prescribed order (with the obvious restriction of no consecutive repetition). As an interesting variant of planar billiards, we consider spinning obstacles and particles and prove that any forward sequence of obstacles has a trajectory that follows it.

Key words. chaos, billiards, dissipative, friction

AMS subject classifications. 37C, 37N15

DOI. 10.1137/060654189

1. Introduction.

The billiard problems of various types have not only physical significance but also mathematical beauty. The long history of billiards is full of mathematical gems [10]. A less investigated class of dispersing billiards is the so-called open-billiards problem, where the reflections on the boundaries of some sorts of obstacles in an infinite affine space are producing the billiard map. An interesting theorem was proven by Morita [8] in the following set-up.

Let $O_1, O_2, \dots O_K$ $(K \geq 3)$ be a finite number of mutually disjoint, closed, bounded, convex subsets of $\mathbb{R}^2$ (the so-called *obstacles*) with boundaries ∂O_j, which are simple smooth closed curves. Consider a particle in $\mathbb{R}^2$ outside $\bigcup_{j=1}^{K} O_j$ which moves along a straight line with unit speed and reflects at the boundary $\bigcup_{j=1}^{K} \partial O_j$ obeying the law of reflection; i.e., the angle of reflection coincides with the angle of incidence. Assume additionally the following two conditions.

Condition 1 (strict convexity). *The boundary curves all have nonvanishing curvature.*

Condition 2 (no eclipse). *For any triple (j_1, j_2, j_3) of distinct indices,*

$$\operatorname{conv}[O_{j_1} \cup O_{j_2}] \cap O_{j_3} = \emptyset,$$

*Received by the editors March 13, 2006; accepted for publication (in revised form) by L. Young June 12, 2007; published electronically April 30, 2008. This research was supported under NSF grants DMS 0104087.
http://www.siam.org/journals/siads/7-2/65418.html

†Department of Mathematics, Anadolu University, Yunusemre Kampusü, 26470 Eskisehir, Turkey (adeniz@anadolu.edu.tr, skocak@anadolu.edu.tr).

‡Department of Mathematical Sciences, University of Delaware, 501 Ewing Hall, Newark, DE (jkennedy@math.udel.edu).

§Department of Mathematics, İstanbul Bilgi University, Kurtuluş Deresi Cad 47, Dolapdere, 34435 Beyoğlu İstanbul, Turkey (ratiu@bilgi.edu.tr).

¶Department of Mathematics, University of Maryland, College Park, MD 20742 (ustun@vis.caltech.edu, yorke@ipst.umd.edu).

where $\mathrm{conv}[A]$ *denotes the convex hull of the set* A.

Then the following theorem holds (in $\mathbb{R}^2$).

Theorem 1 (Morita [8]). *Given an itinerary* $(O_n)_{n\in\mathbb{Z}}$ *of obstacles without consecutive repetition, there exists a unique trajectory following this itinerary.*

The setting and two assumptions of Morita go back to Ikawa [6], who proved the existence and uniqueness of arbitrary periodic trajectories in the 3-dimensional case with strictly convex obstacles without eclipse. The n-dimensional version of this problem is considered in Stoyanov [9], where important and intricate estimates for the separation of nearby trajectories are given, and very recently the problem was solved by Blokh, Misiurewicz, and Simanyi for strictly convex obstacles without eclipse in $\mathbb{R}^n$ (Theorem 2.2 in [1]), the trajectories being unique by Chernov [2].

We propose another method to solve the problem, which can be applied to other similar billiard problems, and we exemplify this for the hyperbolic case. At the same time we weaken the strict convexity assumption for obstacles and allow their boundaries to have flat parts, as suggested by one of the referees. Throughout the literature it is assumed that the obstacles are strictly convex, so we hope this generalization will be of some value. Polyhedral obstacles smoothed along a narrow region of the edges might give interesting examples. This generalization works, however, at the price of uniqueness of trajectories with a fixed itinerary. It remains to be understood to what extent uniqueness is lost.

After proving the theorem for $\mathbb{R}^n$ (see Theorem 2), we replace $\mathbb{R}^n$ by $\mathbb{H}^n$ (the n-dimensional hyperbolic space) using the conformal unit disk model $\mathbb{B}^n \subset \mathbb{R}^n$ for $\mathbb{H}^n$. The rays along which the particle moves are no longer straight lines but the geodesics of $\mathbb{H}^n$, i.e., circle arcs orthogonal to the unit sphere S^{n-1} or Euclidean straight lines going through the origin of $\mathbb{B}^n$. Theorem 3 shows that the basic result (Theorem 2) also holds in this case, taking as obstacles hyperbolically convex smooth subsets which are diffeomorphic with balls inside $\mathbb{B}^n$.

Finally, we consider another generalization of Theorem 1 on the plane, where we assume the obstacles to be (geometric) disks but allow them to spin around their fixed centers. We also allow the moving particles to spin and collide with the spinning obstacles according to the laws of physics. Under some plausible assumptions, we show that there exists a trajectory following any given (forward) sequence of obstacles.

2. n-dimensional (Euclidean) open-billiards. We define an obstacle in $\mathbb{R}^n$ to be a convex (not necessarily strictly convex) subset O (with boundary ∂O), which is diffeomorphic to the standard disk $\mathbb{D}^n$. Now, let us consider a finite number of mutually disjoint obstacles $O_1, O_2, \dots O_K$ $(K \geq 3)$ for which Condition 2 holds.

Condition 2 (no eclipse). *For any triple* (j_1, j_2, j_3) *of distinct indices,*

$$\mathrm{conv}[O_{j_1} \cup O_{j_2}] \cap O_{j_3} = \emptyset.$$

To define the billiard map, we consider the space $\mathcal{L}$ of oriented lines in $\mathbb{R}^n$ (which can be identified with the total space of the tangent bundle of the unit sphere $S^{n-1} \subset \mathbb{R}^n$; see [10]). We denote the unit orientation vector of $L \in \mathcal{L}$ by $v(L)$ and the line going through a point p and having orientation vector v by $L_p v$. We denote the ray $\{x \in L_p v \mid x = p + tv \text{ with } t \geq 0\}$ by $L_p^+ v$.

The subspace $Q \subset \mathcal{L}$ of the oriented lines intersecting at least one obstacle is compact.

We define for $i = 1, 2, \ldots, K$ the set $S_i \subset Q$ as the set of all oriented lines, which, in the direction of orientation, first hit the obstacle O_i and then, for some $j \neq i$, hit another obstacle O_j. (Thus by Condition 2, a line in S_i hits only these two obstacles O_i and O_j.) Note that the sets S_i are mutually disjoint.

Let $Q_0 = \bigcup_{i=1}^{K} S_i \subset Q$.

An oriented line L intersecting an obstacle O_i "enters" the obstacle at a point $p' \in \partial O_i$ with $\langle n_{p'}, v(L)\rangle \leq 0$ ($n_{p'}$ being the outward unit normal vector at p') and "leaves" the obstacle at a point $p \in \partial O_i$ with $\langle n_p, v(L)\rangle \geq 0$. Generically $L \cap \partial O_i$ consists of two points, but it can also be a singleton or the interval $[p', p]$.

Now we define the billiard map $f : Q_0 \to Q$.

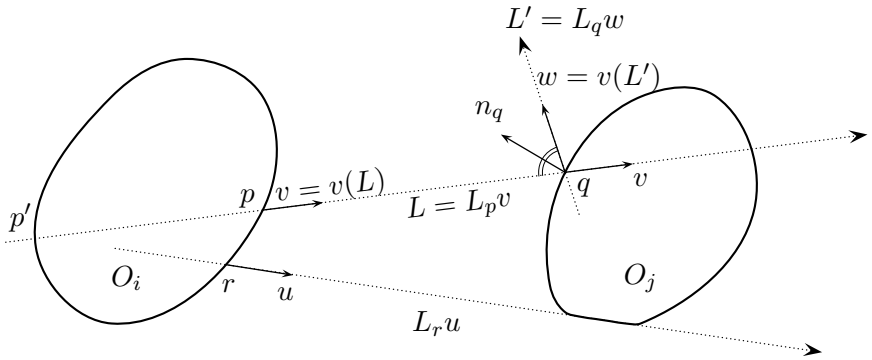


Figure 1. *The billiard map sends $L_p v$ to $L_q w$ with $w = v - 2\langle n_q, v\rangle n_q$. As a special case, $L_r u$ is sent to itself.*

Let $L \in S_i$, $L \cap O_j \neq \emptyset$, with $j \neq i$ and q the first-hit point of L on ∂O_j. We set $f(L) = L'$, where L' is the oriented line through q with the orientation vector $v(L') = v(L) - 2\langle n_q, v(L)\rangle n_q$ (see Figure 1).

We will prove the following.

Theorem 2. *Given an itinerary $(S_{i_n})_{n\in\mathbb{Z}}$ with $i_n \in \{1, 2, \ldots, K\}$ and $i_n \neq i_{n+1}$, there exists a trajectory $(L_n)_{n\in\mathbb{Z}}$ following this itinerary; i.e., $f(L_n) = L_{n+1}$ and $L_n \in S_{i_n}$.*

This means obviously that given any sequence of obstacles in $\mathbb{R}^n$, there is a billiard trajectory in the usual sense hitting every obstacle in the given order.

To prove this theorem, we will use the following special case of the result in [7].

Theorem (the chaos lemma). *Let Q be a compact metric space, $Q_0 \subset Q$ a compact subset, $f : Q_0 \to Q$ a continuous map, and $S_i \subset Q_0$ $(i = 1, 2, \ldots, K)$ pairwise disjoint compact subsets of Q_0 with $\bigcup_{i=1}^{K} S_i = Q_0$. Assume there exists a collection $\{\mathcal{E}_i\}_{i=1}^{K}$, where each $\mathcal{E}_i$ is a nonempty family of nonempty compact subsets of Q, with the following property.*

Property 1. *If $E \in \mathcal{E}_i$ and $j \neq i$, there exists a set $E_j \subset E \cap S_i$ such that $f(E_j) \in \mathcal{E}_j$.*

Then, given any bi-infinite sequence $(S_{i_n})_{n\in\mathbb{Z}}$ of the sets $\{S_1, S_2, \ldots, S_K\}$ with $i_n \in \{1, 2, \ldots, K\}$ and $i_n \neq i_{n+1}$, there exists a sequence $(x_n)_{n\in\mathbb{Z}}$ such that $x_n \in S_{i_n}$ and $f(x_n) = x_{n+1}$.

The sets S_i are called *symbol sets*, the sets $E \in \mathcal{E}_i$ are called *expanders*, and the sets $E_j \subset E \cap S_i$ with $f(E_j) \in \mathcal{E}_j$ are called *pre-expanders*.

3. Proof of Theorem 2. We shall apply the chaos lemma to the billiards setting with the notation fixed above. To make the chaos lemma work, we have to define the expander sets $\mathcal{E}_i$. For this purpose we define the notion of a dispersive vector field on a nonempty, compact subset $D \subset \partial O_i$.

Let $\sigma : D \to S^{n-1}$ be a continuous outward unit vector field; i.e., $\langle \sigma(p), n_p \rangle \geq 0$ for all $p \in D$. We call σ *dispersive* if the condition

$$\langle \sigma(p_1) - \sigma(p_2), p_1 - p_2 \rangle \geq 0 \text{ for all } p_1, p_2 \in D$$

is satisfied. See Figure 4.

Lemma 1. *For a dispersive vector field σ, either the rays $L^+_{p_1}\sigma(p_1)$ and $L^+_{p_2}\sigma(p_2)$ are disjoint or one is contained in the other.*

We omit the proof.

We call a dispersive vector field $\sigma : D \to S^{n-1}$ *exhaustive* (or call σ an exhaustively dispersive vector field) if there exists a continuous extension $\sigma^* : \partial O_i \to S^{n-1}$ of σ such that

$$\langle n_p, \sigma^*(p) \rangle \geq 0 \text{ for all } p \in \partial O_i, \text{ and}$$
$$L^+_p \sigma^*(p) \cap \partial O_j = \emptyset \text{ for all } p \in \partial O_i \backslash D,\ j \neq i.$$

(That is, σ can be extended to an outward unit vector field on ∂O_i in such a way that the rays along the new vectors outside D do not hit any of the obstacles.)

We can associate to every vector field $\sigma : D_\sigma \subset \partial O_i \to S^{n-1}$ the set $E(\sigma)$ of oriented lines defined as $E(\sigma) = \{L_p\sigma(p) \mid p \in D_\sigma\} \subset Q$.

Now we define our expanders (recall that expanders are sets of sets):

$$\mathcal{E}_i = \{E \subset Q \mid \text{there exists an exhaustively dispersive vector field}$$
$$\sigma : D_\sigma \subset \partial O_i \to S^{n-1} \text{ such that } E = E(\sigma)\}.$$

$\mathcal{E}_i$ is nonempty because the outward unit normal vector field $\mathcal{N}$ is exhaustively dispersive, and thus $E(\mathcal{N}) \in \mathcal{E}_i$.

To obtain Theorem 2 from the chaos lemma, we have to verify Property 1: For any $E \in \mathcal{E}_i$ ($E = E(\sigma)$ for some $\sigma : D_\sigma \to S^{n-1}$) and for all $j \neq i$, there exists a subset $E_j \subset E \cap S_i$ such that $f(E_j) \in \mathcal{E}_j$.

We define as a pre-expander

$$E_j = \{L_p\sigma(p) \mid p \in D_\sigma \text{ and } L^+_p\sigma(p) \cap \partial O_j \neq \emptyset\}$$

and are going to show that $f(E_j) \in \mathcal{E}_j$. To this end we first note some well-known facts.

$\sigma^* : \partial O_i \to S^{n-1}$ has degree 1 (because it is homotopic to the normal vector field) and thus it is onto. In other words, given any unit vector in $\mathbb{R}^n$, there is a point in ∂O_i at which this vector is attached. Moreover, we can state the following lemma.

Lemma 2. *Given any point $q \in \mathbb{R}^n \backslash O_i$, there exists a point $p \in \partial O_i$ such that $q \in L^+_p\sigma^*(p)$.*

Proof. Let $q \in \mathbb{R}^n \backslash O_i$. Define the map

$$\omega_q : \partial O_i \to S^{n-1},$$
$$\omega_q(p) = \frac{q-p}{\|q-p\|},$$

which assigns a unit vector in direction q for each $p \in \partial O_i$. ω_q has degree 0 because it is not onto.

Now, suppose that $\sigma^*(p) \neq \omega_q(p)$ for all $p \in \partial O_i$. Then the map

$$\begin{aligned} H : \partial O_i \times [0,1] &\to S^{n-1}, \\ H(p,t) &= \frac{t\sigma^*(p) - (1-t)\omega_q(p)}{\|t\sigma^*(p) - (1-t)\omega_q(p)\|} \end{aligned}$$

is a homotopy between $-\omega_q(p) = H(p,0)$ and $\sigma^*(p) = H(p,1)$. To see this, it is enough to show that $\|t\sigma^*(p) - (1-t)\omega_q(p)\| \neq 0$ for all $p \in \partial O_i$ and $t \in [0,1]$. If we had $\|t\sigma^*(p) - (1-t)\omega_q(p)\| = 0$ for some t, this would give $t\sigma^*(p) = (1-t)\omega_q(p)$ and thus $t = \frac{1}{2}$ (by $\|\sigma^*(p)\| = \|\omega_q(p)\| = 1$), contradicting $\sigma^*(p) \neq \omega_q(p)$.

This homotopy implies that $\mathit{degree}(\sigma^*(p)) = \mathit{degree}(-\omega_q(p))$, which is impossible, because $\mathit{degree}(-\omega_q(p)) = 0$ but $\mathit{degree}(\sigma^*(p)) = 1$. ■

As a consequence, we have $\partial O_j \subset \bigcup_{L \in E_j} L_p^+ \sigma(p)$ for any $j \neq i$.

Let $\varphi : \partial O_j \to S^{n-1}$ be defined as follows: Given $q \in \partial O_j$ there exists $p \in D_\sigma$ with $q \in L_p^+ \sigma(p)$. A point with this property might not be uniquely defined, but $\sigma(p)$ is well defined by dispersivity. So we set $\varphi(q) = \sigma(p)$. φ can be seen to be continuous.

We can now express $f(E_j)$ as $E(\tau)$ for the function

$$\begin{aligned} \tau : D_\tau &\to S^{n-1}, \\ \tau(q) &= \varphi(q) - 2\langle n_q, \varphi(q)\rangle n_q, \end{aligned}$$

where $D_\tau = \{q \in \partial O_j \mid \langle n_q, \varphi(q)\rangle \leq 0\} \subset \partial O_j$ (see Figure 2).

Lemma 3. *τ, as defined above, is exhaustively dispersive.*

Proof (τ is dispersive). We must show $\langle \tau(q_1) - \tau(q_2),\ q_1 - q_2\rangle \geq 0$ for all $q_1, q_2 \in D_\tau$. Let $q_1 = p_1 + t_1\sigma(p_1)$, $q_2 = p_2 + t_2\sigma(p_2)$ and assume $t_1 \geq t_2 > 0$. We thus have, inserting $\tau(q_\alpha) = \sigma(p_\alpha) - 2\langle \sigma(p_\alpha), n_{q_\alpha}\rangle n_{q_\alpha}$ for $\alpha = 1, 2$,

$$\begin{aligned} \langle \tau(q_1) - \tau(q_2), q_1 - q_2\rangle &= \langle \sigma(p_1) - \sigma(p_2), q_1 - q_2\rangle \\ &\quad + 2\langle [\langle \sigma(p_2), n_{q_2}\rangle n_{q_2} - \langle \sigma(p_1), n_{q_1}\rangle n_{q_1}], q_1 - q_2\rangle. \end{aligned} \tag{1}$$

We will show that both terms on the right-hand side of (1) are nonnegative. The first term satisfies

$$\begin{aligned} \langle \sigma(p_1) - \sigma(p_2), q_1 - q_2\rangle &= \langle \sigma(p_1) - \sigma(p_2), p_1 - p_2\rangle + t_2\|\sigma(p_1) - \sigma(p_2)\|^2 \\ &\quad + (t_1 - t_2)\langle \sigma(p_1) - \sigma(p_2), \sigma(p_1)\rangle. \end{aligned} \tag{2}$$

The first term on the right-hand side of (2) is nonnegative, because σ is dispersive. The other two terms are nonnegative for obvious reasons.

The second term of the right-hand side of (1),

$$\langle \sigma(p_2), n_{q_2}\rangle \langle n_{q_2}, q_1 - q_2\rangle - \langle \sigma(p_1), n_{q_1}\rangle \langle n_{q_1}, q_1 - q_2\rangle,$$

is also nonnegative because

$$\langle n_{q_2}, q_1 - q_2\rangle \leq 0 \text{ and } \langle n_{q_1}, q_1 - q_2\rangle \geq 0$$

by convexity of O_j. (The other two factors are ≤ 0 by construction.) ■

Proof (τ is exhaustive). To define the extension

$$\tau^* : \partial O_j \to S^{n-1},$$

we first note that $D_\tau = \{q \in \partial O_j \mid \langle n_q, \varphi(q)\rangle \leq 0\}$ and $D'_\tau = \{q \in \partial O_j \mid \langle n_q, \varphi(q)\rangle \geq 0\}$ are both closed.

Let

$$\tau^*(q) = \begin{cases} \varphi(q) - 2\langle \varphi(q), n_q\rangle n_q & \text{for} \quad q \in D_\tau \\ \varphi(q) & \text{for} \quad q \in D'_\tau \end{cases}$$

(see Figure 2).

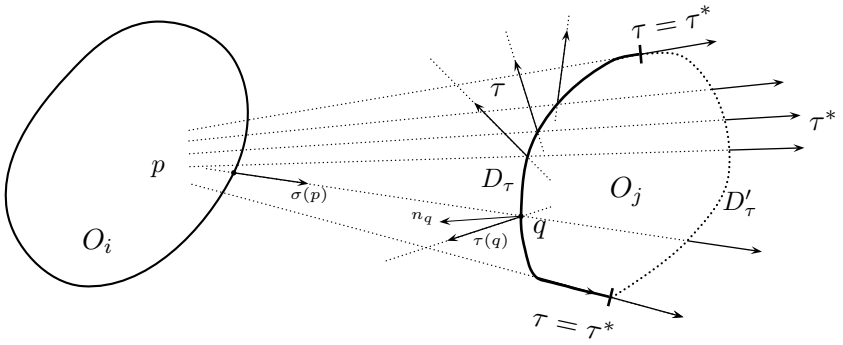


Figure 2. *The extension of* τ.

τ^* is well defined on $D_\tau \cup D'_\tau = \partial O_j$ since $\langle \varphi(q), n_q\rangle = 0$ for $q \in D_\tau \cap D'_\tau$, and consequently τ^* is continuous. τ^* is outward on D'_τ ($\langle n_q, \tau^*(q)\rangle \geq 0$ for $q \in D'_\tau$), and $L_q^+\tau^*(q) \cap \partial O_k = \emptyset$ for $k \neq j$ by Condition 2. Thus τ is exhaustive. ■

The chaos lemma now verifies Theorem 2.

4. Open billiards in hyperbolic n-space. We consider the open-billiards problem in the n-dimensional hyperbolic space $\mathbb{H}^n$ using the conformal unit-disk model $\mathbb{B}^n = \{x \in \mathbb{R}^n \mid \|x\|_{\mathbb{E}} < 1\}$ with the Riemannian metric

$$g_p(u_p, v_p) = \langle u_p, v_p\rangle_{\mathbb{H}} = \frac{4}{(1 - \|p\|^2)^2}\langle u_p, v_p\rangle_{\mathbb{E}},$$

where u_p, v_p are vectors at the point $p \in \mathbb{B}^n$ and the subscripts $\mathbb{E}$ and $\mathbb{H}$ denote the Euclidean and hyperbolic metrics, respectively. In this section we will assume the hyperbolic metric applies when the subscript is dropped. The angles between (hyperbolic) vectors are the same as in the Euclidean case; only the lengths are affected.

Given any point $p \in \mathbb{B}^n$ and any unit vector v_p at p, the geodesic going through p in the direction of v_p is a Euclidean circle-arc perpendicular to the unit sphere $S^{n-1} \subset \mathbb{R}^n$. (In the case when it goes through the origin of $\mathbb{R}^n$, it is a straight line segment.) We denote the oriented geodesic through p in the direction of v_p by $L_p v_p$, and we call the geodesic part starting at p in the direction of v_p a hyperbolic ray and denote it by $L_p^+ v_p$. As obstacles we

consider hyperbolically convex (the geodesic segment connecting any two points is contained in the obstacle) smooth subsets of $\mathbb{B}^n$ which are diffeomorphic to the standard closed disk $\mathbb{D}^n = \{x \in \mathbb{R}^n \mid \|x\|_{\mathbb{E}} \leq 1\}$.

Let $\mathcal{L}$ denote the space of oriented geodesics, and let $Q \subset \mathcal{L}$ be the compact subspace of oriented geodesics intersecting at least one obstacle. As before, let S_i be the set of all oriented geodesics which in the direction of orientation first hit the obstacle O_i and then, for some $j \neq i$, another obstacle O_j. We set $Q_0 = \bigcup_{i=1}^{K} S_i \subset Q$.

For an oriented geodesic L intersecting an obstacle O_i, there is a first-hit point $p' \in \partial O_i$ with $\langle n_{p'}, v_{p'} \rangle \leq 0$ (note that the hyperbolic normal vector is in the direction of the Euclidean normal vector, but possibly of different length) and a last-hit point $p \in \partial O_i$ with $\langle n_p, v_p \rangle \geq 0$.

The billiard map $f : Q_0 \to Q$ can be defined as before: For $L \in S_i$, $L \cap O_j \neq \emptyset$ with $j \neq i$, p the last-hit point of L on ∂O_i, q the first-hit point of L on ∂O_j, we set $f(L) = L'$, where L' is the oriented geodesic through q with the vector $w_q = P_p^q v_p - 2\langle n_q, P_p^q v_p \rangle n_q$, P_p^q denoting the parallel transportation of a vector at p, along the geodesic, to a vector at q (see Figure 3).

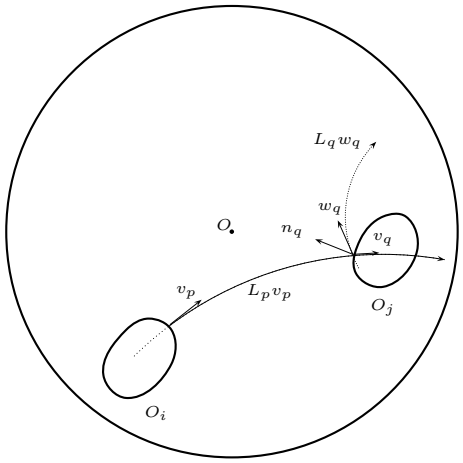


Figure 3. *The billiard map sends $L_p v_p$ to the geodesic $L_q w_q$ with $w_q = v_q - 2\langle n_q, v_q \rangle n_q$, where $v_q = P_p^q v_p$.*

Now let us consider a finite number of mutually disjoint obstacles $O_1, O_2, \ldots, O_K$ ($K \geq 3$). We again assume Condition 2.

Condition 2 (no eclipse). *For any triple (j_1, j_2, j_3) of distinct indices,*

$$\operatorname{conv}[O_{j_1} \cup O_{j_2}] \cap O_{j_3} = \emptyset.$$

Theorem 3. *Given an itinerary $(S_{i_n})_{n \in \mathbb{Z}}$ with $i_n \in \{1, 2, \ldots, K\}$ and $i_n \neq i_{n+1}$, there exists a trajectory $(L_n)_{n \in \mathbb{Z}}$ following this itinerary; i.e., $f(L_n) = L_{n+1}$ and $L_n \in S_{i_n}$.*

For proof, we again apply the chaos lemma. We will outline those points where there are slight modifications in comparison to the Euclidean case.

We denote the restriction of the hyperbolic unit tangent bundle $T_1(\mathbb{B}^n)$ to ∂O_i by Σ_i. Let $D \subset \partial O_i$ be a nonempty, compact subset of ∂O_i and $\sigma : D \to \Sigma_i$ a continuous outward vector field, i.e., $\sigma(p) = (p, \sigma_2(p))$ with $\langle n_p, \sigma_2(p) \rangle \geq 0$ for $p \in D$. Let $p_1, p_2 \in D$ and α_i denote the angle between $\sigma_2(p_i)$ and the hyperbolic segment $[p_1 p_2]$ for $i = 1, 2$. We call σ a dispersive

vector field if $\alpha_1 + \alpha_2 \geq \pi$ for all $p_1, p_2 \in D$. (This condition is equivalent to the one we used in the Euclidean case, but this formulation is more convenient for the hyperbolic setting, especially for checking the dispersiveness of reflected rays.)

Lemma 4. *For a dispersive vector field σ, either the hyperbolic rays $L^+_{p_1}\sigma_2(p_1)$ and $L^+_{p_2}\sigma_2(p_2)$ are disjoint or one is contained in the other.*

We omit the proof.

The definition of exhaustiveness remains the same. As for expanders, we take again sets of oriented geodesics determined by exhaustively dispersive vector fields. If $\sigma : D_\sigma \to \Sigma_i$, $D_\sigma \subset \partial O_i$, we set $E(\sigma) = \{L_p\sigma_2(p) \mid p \in D_\sigma\} \subset Q$ and define

$$\mathcal{E}_i = \{E(\sigma) \mid \sigma \text{ is an exhaustively dispersive vector field on } D_\sigma \subset \partial O_i\}.$$

$\mathcal{E}_i$ is nonempty because by hyperbolic convexity the segment $[p_1p_2]$ lies inside O_i for $p_1, p_2 \in \partial O_i$, making the angle with the normals n_{p_1} and n_{p_2} greater than (or equal to) $\frac{\pi}{2}$. This shows that the normal outward vector field on ∂O_i is exhaustively dispersive.

We define the pre-expanders as before: given $E \in \mathcal{E}_i$, we set

$$E_j = \{L_p\sigma_2(p) \mid p \in D_\sigma \text{ and } L^+_p\sigma_2(p) \cap \partial O_j \neq \emptyset\}.$$

We have to show that $f(E_j) \in \mathcal{E}_j$. In the Euclidean case, we used degree theory to see this. In the hyperbolic setting, degree theory can still be used with suitable modifications.

Given any section $\sigma : \partial O_i \to \Sigma_i$, for any $p \in \partial O_i$ we can parallel-transport the vector $\sigma_2(p)$ to the origin along the geodesic between p and the origin (which is a Euclidean straight line segment). We thus obtain a map $\tilde{\sigma} : \partial O_i \to S^{n-1}_{\mathbb{H}}$ (the hyperbolic unit sphere is also a Euclidean sphere) and define *degree*(σ) to be *degree*$(\tilde{\sigma})$.

If $\sigma : D_\sigma \to \Sigma_i$ is any exhaustively dispersive vector field, then we get as before *degree*$(\sigma^*) = 1$. As a consequence, given any point $q \in \mathbb{B}^n \backslash O_i$, there exists a point $p \in \partial O_i$ with $q \in L^+_p\sigma^*_2(p)$. Then we get $\partial O_j \subset \bigcup_{L \in E_j} L^+_p\sigma_2(p)$. We define as before $\varphi : \partial O_j \to \Sigma_j$ $\varphi(q) = P^q_p\sigma_2(p)$ for $q \in L^+_p\sigma_2(p)$ and $\tau : D_\tau \to \Sigma_j$ with $D_\tau = \{q \in \partial O_j \mid \langle n_q, \varphi(q)\rangle \leq 0\} \subset \partial O_j$ and

$$\tau(q) = (q, \varphi(q) - 2\langle n_q, \varphi(q)\rangle n_q).$$

With this τ it holds that $f(E_j) = E(\tau)$, and to show $f(E_j) \in \mathcal{E}_j$ we must prove that τ is continuous, dispersive, and exhaustive. Continuity and exhaustiveness go almost verbatim as in the Euclidean case, but dispersiveness requires a separate argument. We have to show that for $q_1, q_2 \in D_\tau$, the angle β_1 between $\tau_2(q_1)$ and the hyperbolic segment $[q_1q_2]$ and the angle β_2 between $\tau_2(q_2)$ and $[q_1q_2]$ satisfy the inequality $\beta_1 + \beta_2 \geq \pi$.

We first show that the angles θ_1, θ_2 between $P^{q_1}_{p_1}\sigma_2(p_1), P^{q_2}_{p_2}\sigma_2(p_2)$ and the hyperbolic segment $[q_1q_2]$ satisfy $\theta_1 + \theta_2 \geq \pi$. Then we will show $\beta_1 \geq \theta_1$ and $\beta_2 \geq \theta_2$. Now let $\measuredangle p_1p_2q_1 = \alpha'_2$, $\measuredangle q_1p_2q_2 = \alpha''_2$, $\measuredangle q_2q_1p_2 = \gamma'$, and γ be the angle between $P^{q_1}_{p_1}\sigma_2(p_1)$ and $[p_2q_1]$. By the triangle inequality for angles, we have

$$\alpha_2 \leq \alpha'_2 + \alpha''_2 \text{ and } \gamma \leq \gamma' + \theta_1.$$

On the other hand, for the hyperbolic triangle $p_1p_2q_1$, $\gamma > \alpha_1 + \alpha'_2$ (because the sum of the inner angles of a hyperbolic triangle is less than π). We thus get

$$\gamma + \alpha'_2 + \alpha''_2 \geq \alpha_1 + \alpha'_2 + \alpha_2,$$
$$\gamma + \alpha''_2 \geq \alpha_1 + \alpha_2 \geq \pi.$$

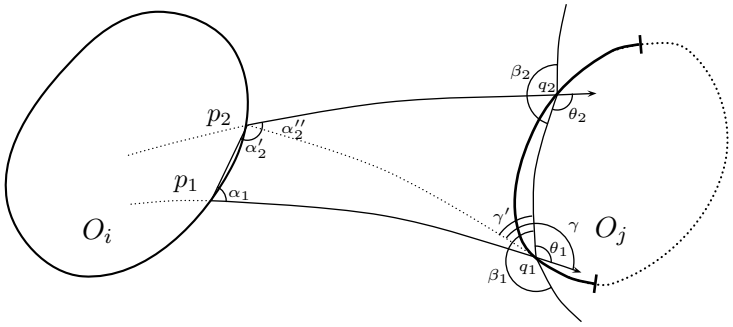


Figure 4. *τ is a dispersive vector field.*

From the hyperbolic triangle $p_2q_1q_2$ we have $\theta_2 > \gamma' + \alpha_2''$. Hence, $\theta_1 + \theta_2 \geq \theta_1 + \gamma' + \alpha_2'' \geq \gamma + \alpha_2'' \geq \alpha_1 + \alpha_2 \geq \pi$.

We now show $\beta_1 \geq \theta_1$:

$$\tau_2(q_1) = P_{p_1}^{q_1}\sigma_2(p_1) - 2\langle n_{q_1}, P_{p_1}^{q_1}\sigma_2(p_1)\rangle n_{q_1} \text{ by definition of } \tau.$$

As the scalar product on the right-hand side is negative, we have

$$\tau_2(q_1) = P_{p_1}^{q_1}\sigma_2(p_1) + \lambda n_{q_1} \text{ with } \lambda \geq 0.$$

Denote the unit vector at q_1 in the direction of q_1q_2 by u, and multiply the above equation by u:

$$\begin{aligned}\langle \tau_2(q_1), u\rangle &= \langle P_{p_1}^{q_1}\sigma_2(p_1), u\rangle + \lambda\langle n_{q_1}, u\rangle,\\ \cos\beta_1 &= \cos\theta_1 + \lambda\langle n_{q_1}, u\rangle.\end{aligned}$$

As $\langle n_{q_1}, u\rangle \leq 0$ we get $\cos\beta_1 \leq \cos\theta_1$, and thus $\beta_1 \geq \theta_1$. Similarly, $\beta_2 \geq \theta_2$, giving the dispersiveness and completing the proof.

5. Spinning planar billiards.

5.1. Introduction. Billiard dynamics has been a standard dynamics model for many years, but the model is far from a complete description of reality. Real billiard balls have spin, which is utilized by billiards players. Here we extend the model to permit the billiard to spin. See [3, 4, 5] for a discussion of the physics. We call the particle a "puck"; it is analogous to a hockey puck that can spin on the plane. In contrast, a billiard ball on a table spins in three dimensions, a case we do not consider here.

We find the new mathematics that result intriguing. The state space describing a puck is larger since it must include speed and spin rate as well as position and direction—we can no longer assume the puck's speed is constant. Our puck is assumed to have no friction between collisions, but there is friction in the collision. Friction in the collision introduces an interaction among the spin of the obstacle, the spin of the puck, and the incoming velocity of the puck. Most people have observed that a thrown rubber ball has a bounce that is affected

by the spin of the ball. A rapidly spinning ball can gain speed in the collision, translating spin energy into velocity. So-called *super balls* are resilient (i.e., they retain most of their energy in a bounce) and have a high friction coefficient. Their bounces are especially influenced by spin.

In the appendix, we discuss the equations for a collision between the puck and an obstacle, using the most commonly used rules of friction, namely, stick-slip friction. A block sliding on a surface slows down, and its deceleration is independent of its velocity, provided that the velocity is positive. The block therefore reaches 0 speed in finite time. We must compress this phenomenon into an instantaneous bounce. A key factor is the relative speeds of the points on the puck and obstacle that are in contact during the collision. For the incoming trajectory, they are unlikely to be equal, but for the outgoing trajectory, equality is quite likely if the incoming difference is small.

The stick-slip dynamics of a collision with a fixed obstacle (described in the appendix) have some easily derived properties that can be useful in modeling, though they are not necessary for the results in this section. Let Δs denote the absolute value of the difference between the incoming spin rate of a puck (just before a collision) and its spin immediately after. Similarly, let Δv denote the norm of the difference in the incoming and outgoing velocities. Let w be the norm of the normal component of the incoming velocity. Then there are constants $C_1 > 0$ and $C_2 > 0$ (that depend only on the coefficient of friction and moment of inertia of the puck) such that

$$\Delta s \le C_1 w \quad \text{and} \quad C_2 w \le \Delta v \le C_1 w.$$

In particular, in a tangential collision, $w = 0$, so the change in velocity and spin are 0.

We also note that the outgoing velocity and spin are continuous functions of initial position, velocity, and spin. Additionally, the change in the normal component of velocity depends only on w, and the outgoing velocity must be nonzero.

5.2. Spinning billiards. In our result we assume the following properties, which are much less specific than would be required by the stick-slip friction model.

Condition 2 (no eclipse). *For any triple (j_1, j_2, j_3) of distinct indices,*

$$\operatorname{conv}[O_{j_1} \cup O_{j_2}] \cap O_{j_3} = \emptyset.$$

Condition 3 (physical properties). *The obstacles are fixed disks with fixed spin rates, all having the same fixed coefficients of friction. The puck is a disk and has a fixed coefficient of restitution $e > 0$, so its outgoing speed is always positive (the puck cannot stop at a collision). The velocity and spin rate of the puck are constant between collisions.*

Condition 4 (continutity). *For each obstacle, following a collision, the puck's outward velocity and rate of spin are continuous functions of the inward velocity, rate of spin, direction of motion, and point of contact with the obstacle. Furthermore, if the puck hits an obstacle tangentially, then its velocity and rate of spin do not change.*

The ambient phase space of the dynamics (at instants of collisions) will be

$$Q = \bigcup_{i=1}^{K} (S_i^1 \times S^1 \times (0, \infty) \times \mathbb{R}),$$

where the first factor S^1_i is the boundary of the disk O_i, the second factor S^1 codes the direction of the puck, the third factor $(0,\infty)$ codes its linear speed, and the last factor $\mathbb{R}$ codes its angular velocity at the moment of leaving O_i.

The symbol sets will be

$$S_i = \{(p, v, s, r) \mid p \in S^1_i,\ \langle n_p, v\rangle \geq 0, \text{ and } \exists j \text{ such that } L^+_p v \cap S^1_j \neq \emptyset\} \subset Q,$$

where n_p and $L^+_p v$ denote as usual the outward unit normal at p and the (Euclidean) ray starting at p in direction v. Let

$$Q_0 = \bigcup_{i=1}^{K} S_i.$$

The billiard map $f : Q_0 \to Q$ is continuous by Condition 4. See Figure 5.

Theorem 4. *Assume Conditions 2, 3, and 4. Given a forward itinerary $(S_{i_n})_{n\in\mathbb{N}}$ with $i_n \in \{1, 2, \ldots, K\}$ and $i_n \neq i_{n+1}$, there exists a trajectory $(L_n)_{n\in\mathbb{N}}$ following this itinerary; i.e., $f(L_n) = L_{n+1}$ and $L_n \in S_{i_n}$.*

Since this theorem discusses only forward itineraries, the corresponding trajectories are not unique. Even if the forward and backward trajectories were specified, there is no guarantee that they would correspond to a *unique* trajectory. The approach we use is general enough that it can be applied to situations where uniqueness of trajectories does not hold.

To define the expanders in the present case, we first consider continuous maps

$$g : J \to S^1_i \times S^1 \times (0,\infty) \times \mathbb{R},$$
$$t \mapsto (p(t), v(t), s(t), r(t)),$$

where J is a compact interval in $\mathbb{R}$ or $J \subset S^1_i$ for some i and the second coordinate $v(t)$ is an outward direction at the point $p(t)$. We call such a map g *exhaustive* if it can be extended to a map g^* on a circle such that, on the complementary closed arc J', the following conditions hold:

1. For $t \in J'$, $v(t)$ is an outward direction at $p(t)$, and the ray $L^+_{p(t)}v(t)$ does not hit any disk O_j for $j \neq i$.
2. The degree of the map given by the first coordinate of g^* is ± 1. (That is, $p(t)$ winds around S^1_i once.)

The expanders are

$$\mathcal{E}_i = \{\text{Image}(g) \mid g : J \to S^1_i \times S^1 \times (0,\infty) \times \mathbb{R} \text{ is an exhaustive map}\}.$$

Each $E \in \mathcal{E}_i$ is a subset of $S^1_i \times S^1 \times (0,\infty) \times \mathbb{R} \subset Q$.

We must prove the existence of pre-expanders: Given $E = \text{Image}(g) \in \mathcal{E}_i$ and $j \neq i$, there must exist some $E_j \subset E \cap S_i$ such that $f(E_j) \in \mathcal{E}_j$.

As t traverses the interval $J = [a, b]$ from a to b, we start with the ray $L^+_{p(a)}v(a)$ which does not hit O_j. There will be t with $L^+_{p(t)}v(t)$ hitting O_j by the degree condition, and so there will be tangencies. Finally, the last ray $L^+_{p(b)}v(b)$ does not hit O_j again. This implies the existence of a subinterval $[a_j, b_j] \subset [a, b]$ sweeping O_j: $L^+_{p(a_j)}v(a_j)$ and $L^+_{p(b_j)}v(b_j)$ are tangent to O_j, $L^+_{p(t)}v(t)$ hits O_j for $t \in [a_j, b_j]$, and $O_j \subset \bigcup_{t\in[a_j,b_j]} L^+_{p(t)}v(t)$.

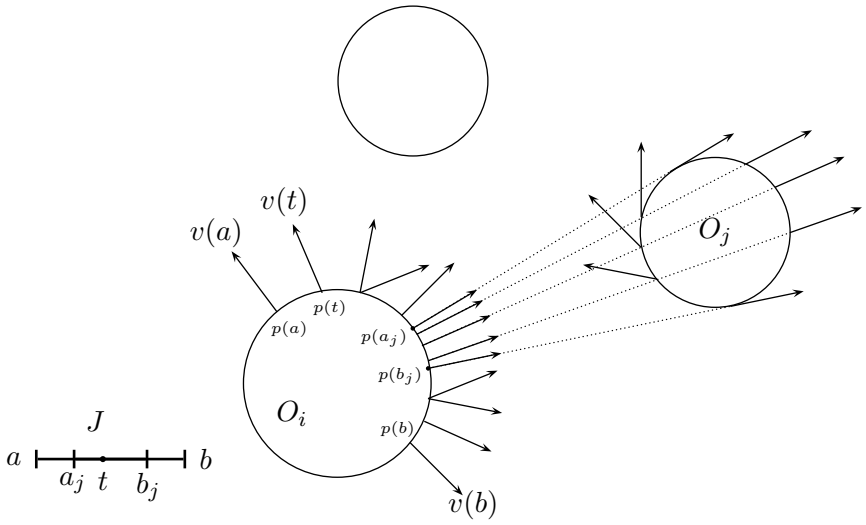


Figure 5. *Spinning billiards. Three obstacles are shown, with trajectories leaving obstacle O_i. If O_j has a strong clockwise spin, the puck's outward trajectories might be as shown. Independent of the spin of the obstacles, trajectories like $p(a_j)$ and $p(b_j)$ hit O_j tangentially and continue without deviating.*

If we set $E_j = \text{Image}\left(g|_{[a_j,b_j]}\right)$, then $E_j \subset E \cap S_i$. It can be seen by the same reasoning as at the end of the proof of Theorem 2 that $f(E_j)$ is an expander again, completing the proof that in spinning billiards in the Euclidean plane, any nonrepeating but otherwise arbitrary infinite sequence of obstacles has a trajectory that follows it.

Appendix. In this section, we develop a model for the scattering of a circular puck from an immovable circular obstacle, where the interaction between the two bodies involves friction. The obstacles' centers do not move, and each spins at a constant rate unaffected by collision with the puck.

Consider the situation shown in Figure 6. At time $t = 0$, a puck starts out at initial position $\mathbf{x_i}$ and with initial velocity $\mathbf{v}$ directed toward an obstacle situated at $\mathbf{x_o}$. The initial angular velocity of the puck is $\boldsymbol{\omega_p}$, and the obstacle is given to be always rotating with an angular velocity of $\boldsymbol{\omega_o}$. The goal is to find the final state given by $\mathbf{v}'$ and $\boldsymbol{\omega_p'}$ immediately after the collision at $\mathbf{x_c}$, the point of impact. With the exception of the angular velocities $\boldsymbol{\omega_p}$ and $\boldsymbol{\omega_o}$ which are strictly in the z direction, all vectorial quantities in this discussion are (and hence all motion is) confined to the $x - y$ plane. The dynamics are independent of the puck's mass m, which we can set to 1.

The first step is to determine if and when a collision occurs. For times t before the collision, the position vector of the puck is given by $\mathbf{x} = \mathbf{x_i} + \mathbf{v}\,t$. The collision takes place at some time t_c, when the puck is at $\mathbf{x_c}$:

$$\mathbf{x_c} = \mathbf{x_i} + \mathbf{v}\,t_c. \tag{3}$$

We assume that the collision deforms neither the puck nor the obstacle. This means that at

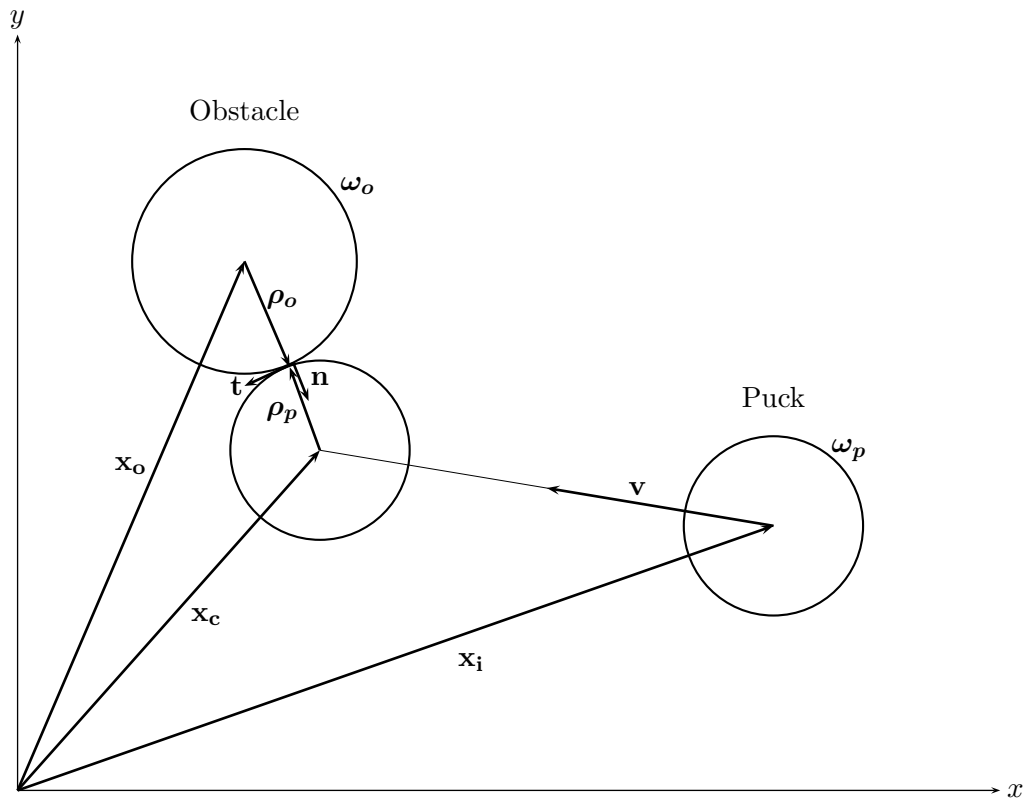


Figure 6. *A puck (smaller disk) starts at* $\mathbf{x_i}$ *and travels to* $\mathbf{x_c}$*, where it collides with an obstacle situated at* $\mathbf{x_o}$*.*

impact time,

$$|\mathbf{x_c} - \mathbf{x_o}| = R_p + R_o \tag{4}$$

will hold, where R_p and R_o are the radii of the puck and obstacle, respectively. Substituting (3) into (4) and defining $\mathbf{z} = \mathbf{x_i} - \mathbf{x_o}$ leads to

$$|\mathbf{z} + \mathbf{v}\, t_c| = R_p + R_o.$$

After squaring both sides, we get a quadratic equation for the collision time t_c of the form $a\, t_c^2 + b\, t_c + c\, t_c = 0$, where $a = v^2$, $b = 2\, \mathbf{z} \cdot \mathbf{v}$, and $c = z^2 - (R_p + R_o)^2$.

The discriminant $\Delta = b^2 - 4ac$ leads to several regimes of qualitatively different physical behavior: $\Delta < 0$ implies that the puck will miss the obstacle entirely, whereas $\Delta = 0$ means that the puck will pass tangentially to the obstacle. No interaction is assumed between the two bodies in either case. A value of $\Delta > 0$, on the other hand, means that there will be two values of t, t_+ and t_-, for which (4) is satisfied (one for each side of the obstacle). If these are positive, we pick the smaller of the two values. Of course $t_+, t_- < 0$ correspond to meaningless collisions for negative times. In short,

$$t_c = \min\{t_+, t_-\} \quad (t_\pm > 0), \tag{5}$$

and substituting this back into (3) yields the collision coordinate $\mathbf{x_c}$. For the remainder of this discussion assume that a given set of initial conditions will lead to a collision satisfying (5) above.

The second and final step is to determine what happens in a collision. Define unit vectors in the normal and tangential directions at the point of contact as

$$\mathbf{n} = (\mathbf{x_c} - \mathbf{x_o})/|\mathbf{x_c} - \mathbf{x_o}|,$$
$$\mathbf{t} = \mathbf{n} \times \mathbf{k},$$

where $\mathbf{k}$ is the unit vector in the z direction (we assume a right-handed coordinate system) and where a "$\times$" denotes the usual vector cross product. We have defined $\mathbf{n}$ to be pointing away from the obstacle. Position vectors from the centers of the puck and obstacle that extend to the point of contact are then given by $\boldsymbol{\rho_p} = -R_p\,\mathbf{n}$ and $\boldsymbol{\rho_o} = R_o\,\mathbf{n}$. These in turn allow us to write the tangential velocities of the rims of the puck and obstacle at the point of contact:

$$\mathbf{V_p} = (\mathbf{v}\cdot\mathbf{t})\,\mathbf{t} + \boldsymbol{\omega_p}\times\boldsymbol{\rho_p}, \tag{6}$$

$$\mathbf{V_o} = \boldsymbol{\omega_o}\times\boldsymbol{\rho_o}. \tag{7}$$

We now turn our attention from kinematics to dynamics. The collision is instantaneous but can be considered as the limit of a more physically realistic, very brief collision. All these approximations have the same impulse J. Throughout the collision, the obstacle will exert a normal force $N(t)$, where t lies in the small interval during which the collision takes place. The time integral of this over that small interval is the impulse $J \equiv \int N(t)dt$, which acts in the direction $\mathbf{n}$. We do not calculate $N(t)$ itself; instead, we work with J not only for motion in the normal direction but also for the tangential and rotational degrees of freedom. To begin, recall from mechanics that if $\mathbf{p} = \mathbf{v}$ is the initial momentum of the puck having mass $m = 1$ and $\mathbf{p_n} = (\mathbf{p}\cdot\mathbf{n})\mathbf{n}$ is the component of this momentum in the normal direction, then the new value of $\mathbf{p_n}$ is given by $\mathbf{p'_n} = \mathbf{p_n} + J\,\mathbf{n}$. On the other hand, the coefficient of restitution [3, 4, 5], $e \in (0, 1]$, satisfies $\mathbf{p'_n} = -e\,\mathbf{p_n}$. Therefore,

$$J = -(1+e)\mathbf{v}\cdot\mathbf{n}, \tag{8}$$

which is always positive since $\mathbf{v}\cdot\mathbf{n}$ is negative.

The impulse imparted to the puck in the tangential direction involves the force of friction, $F(t)$. If the puck and obstacle are sliding with respect to each other during the entire duration of the impact (the "sliding" regime), this force will be assumed to have the form $F(t) = \mu_s N(t)$, where μ_s is the coefficient of sliding friction. The impulse generated by $F(t)$ will then be $\mu_s J$. If, on the other hand, at some point during the collision the tangential velocities $\mathbf{V_p}$ and $\mathbf{V_o}$ are equalized due to friction, $F(t)$ has to be set to zero for the remainder of the impact (we neglect rolling friction here), and the impulse in this direction will then be correspondingly less than $\mu_s J$. Physically, this corresponds to the puck and obstacle rolling about each other (the "rolling" regime, even though the puck may have been initially sliding).

Before we discuss the details of the two regimes, we note that (6) and (7) allow us to define a unit vector in the direction of friction as $\mathbf{f} = -(\mathbf{V_p} - \mathbf{V_o})/|\mathbf{V_p} - \mathbf{V_o}|$. This simply

states that friction acts to oppose the motion of the puck relative to the obstacle at the point of contact. The force of friction is then written as $\mathbf{F}(t) = F(t)\mathbf{f}$. Since $\mathbf{f}$ and $\mathbf{t}$ have the same sense, $\mathbf{f} \cdot \mathbf{t} = \pm 1$.

The sliding regime. The velocity of the puck's center of mass after the collision is

$$\mathbf{v}' = \mathbf{v} + J\,\mathbf{n} + \mu_s J\,\mathbf{f}. \tag{9}$$

In addition to this, the presence of friction will create a torque $\boldsymbol{\rho_p} \times \mathbf{F}(t)$ on the puck, which will cause a change in its angular momentum $\mathbf{L} = I\boldsymbol{\omega_p}$, where I is the moment of inertia of the puck. The final angular velocity $\boldsymbol{\omega_p'}$ is then found to be

$$\boldsymbol{\omega_p'} = \boldsymbol{\omega_p} + (1/I)(\boldsymbol{\rho_p} \times \mathbf{f})\,\mu_s J. \tag{10}$$

Together with (8), equations (9) and (10) express the final state of the puck in the sliding regime.

The rolling regime. This case differs from the sliding friction case in that the changes in the angular and tangent velocities are smaller than those predicted by (9) and (10). As mentioned above, if rolling takes hold during a collision, the force due to friction will drop discontinuously to zero. As a result, the values of J for the puck in the tangential and rotational directions in (9) and (10) will take a new value J^L, where the superscript denotes the "locking" of the bodies. To calculate J^L we recall that the onset of rolling happens when $\mathbf{V_p'} = \mathbf{V_o}$. More explicitly, using (6) and (7),

$$(\mathbf{v}' \cdot \mathbf{t})\,\mathbf{t} + \boldsymbol{\omega_p'} \times \boldsymbol{\rho_p} = \boldsymbol{\omega_o} \times \boldsymbol{\rho_o}.$$

Substituting the previous solutions (9) and (10) into this and using the identity

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}\,(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}\,(\mathbf{a} \cdot \mathbf{b}),$$

we can solve for J^L,

$$J^L = \frac{m}{\mu_s}\frac{\alpha}{1+\alpha}|\mathbf{V_p} - \mathbf{V_o}|,$$

which is also positive and where $\alpha = I/(mR_p^2)$ ($\alpha = \frac{1}{2}$ for a solid disk). The final state for the case of rolling is then

$$\mathbf{v}' = \mathbf{v} + (1/m)J\,\mathbf{n} + (1/m)\mu_s J^L\,\mathbf{f}, \tag{11}$$

$$\boldsymbol{\omega_p'} = \boldsymbol{\omega_p} + (1/I)(\boldsymbol{\rho_p} \times \mathbf{f})\,\mu_s J^L, \tag{12}$$

where J is still given by (8).

Deciding between the regimes. The question of whether a set of initial conditions leads to sliding or rolling is easily addressed by considering the relative velocities $\mathbf{V_p} - \mathbf{V_o}$ (at collision) and $\mathbf{V_p'} - \mathbf{V_o}$ (after the collision, assuming the sliding case). If the relative motion of the puck at the rim has changed direction, i.e., if $(\mathbf{V_p} - \mathbf{V_o}) \cdot (\mathbf{V_p'} - \mathbf{V_o}) < 0$, then rolling must take place, and we use (11) and (12) instead of (9) and (10) for the final state of the puck. It is also possible that $\mathbf{V_p} - \mathbf{V_o} = 0$; in this case, the bodies start rolling immediately after impact, and we take the rolling case equations and set $J^L = 0$.

Acknowledgment. As we find is often the case, the referees' comments have allowed us to significantly improve the paper. In particular, one referee suggested that we not require strict convexity of the billiards.

REFERENCES

[1] A. Blokh, M. Misiurewicz, and N. Simanyi, *Rotation sets of billiards with one obstacle*, Comm. Math. Phys., 266 (2006), pp. 239–265.
[2] N. Chernov, *Construction of transverse fiberings in multidimensional semidispersed billiards*, Funct. Anal. Appl., 16 (1982), pp. 270–280.
[3] R. Cross, *Measurements of the horizontal coefficient of restitution for a superball and a tennis ball*, Amer. J. Phys., 70 (2002), pp. 482–489.
[4] R. Cross, *Grip-slip behavior of a bouncing ball*, Amer. J. Phys., 70 (2002), pp. 1093–1102.
[5] R. Garwin, *Kinematics of an ultraelastic rough ball*, Amer. J. Phys., 37 (1969), pp. 88–92.
[6] M. Ikawa, *Decay of solutions of the wave equation in the exterior of several convex bodies*, Ann. Inst. Fourier (Grenoble), 38 (1988), pp. 113–146.
[7] J. Kennedy, S. Koçak, and J. A. Yorke, *A chaos lemma*, Amer. Math. Monthly, 108 (2001), pp. 411–423.
[8] T. Morita, *The symbolic representation of billiards without boundary condition*, Trans. Amer. Math. Soc., 325 (1991), pp. 819–828.
[9] L. Stoyanov, *Exponential instability for a class of dispersing billiards*, Ergodic Theory Dynam. Systems, 19 (1999), pp. 201–226.
[10] S. Tabachnikov, *Geometry and Billiards*, AMS, Providence, RI, 2005.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 437–460
© 2008 Society for Industrial and Applied Mathematics

Asymptotics of Null Lie Quadratics in E^3*

Lyle Noakes†

Abstract. *Lie quadratics* are curves in Lie algebras, arising from studies in the mid-1980s of motion planning for rigid bodies. Attention has focused on Lie quadratics in Euclidean 3-space E^3 (with cross-product as Lie bracket), especially the codimension-3 subclass of *null* Lie quadratics in E^3. The present paper substantially improves known asymptotic results for this subclass, to an extent that the new results apply to asymptotic dynamics of spherically symmetric rigid balls in classical mechanics.

Key words. Lie group, Lie quadratic, asymptotic estimate, rigid body

AMS subject classifications. Primary, 70E99, 34E05; Secondary, 70E17, 70E18, 49S05, 49N99

DOI. 10.1137/070686755

1. Introduction.

Let $so(3)$ be the real Lie algebra of skew-symmetric 3×3 real matrices with respect to the Lie bracket $[A, B] := AB - BA$. For nonzero $B_0, B_1 \in so(3)$, the linear differential equation

$$\dot{z}(t) = (B_0 + tB_1)z(t) \tag{1}$$

for $z : \mathbb{R} \to E^3$ is significant for the mechanics of spherically symmetric rigid bodies.

Example 1. *Relative to some reference frame fixed at some point in a moving rigid body, let $(z_1(t), z_2(t), z_3(t))$ be the representation at time t of a fixed positively oriented orthonormal inertial frame in E^3. Let $z(t) \in SO(3)$ be the matrix whose columns are $z_1(t), z_2(t), z_3(t)$, and set $B(t) := -\dot{z}(t)z(t)^{\mathbf{T}}$, where $^{\mathbf{T}}$ means transpose. It is easily verified that $B(t) \in so(3)$. Define a linear isomorphism $\hat{} : E^3 \to so(3)$ by*

$$\hat{v}(w) := v \times w, \tag{2}$$

where $v, w \in E^3$ and $\times$ is the cross-product. Then $B(t) = \hat{\Omega}(t)$, where $\Omega : \mathbb{R} \to E^3$ is the angular velocity in body coordinates *of the rigid body. Alternatively,*

$$\dot{z}(t) = -\hat{\Omega}(t)z(t) \iff \dot{z}_i(t) = -\Omega(t) \times z_i(t) \iff \dot{z}_i(t) = -\hat{\Omega}(t)z_i(t).$$

When Ω is an affine function of t the vectors z_i satisfy a form of (1). More detail is given in section 7 (especially section 7.1). In section 7.2 the differential equation (1) arises in another situation in mechanics.

Now E^3 is also a Lie algebra, with respect to $\times$, and $\hat{}$ is a Lie isomorphism onto $so(3)$. An inner product $\langle\ ,\ \rangle$ on $so(3)$ is defined by requiring $\hat{}$ to be an isometry.

*Received by the editors March 29, 2007; accepted for publication (in revised form) by J. Meiss August 17, 2007; published electronically April 30, 2008.
http://www.siam.org/journals/siads/7-2/68675.html
†School of Mathematics and Statistics, The University of Western Australia, Nedlands, WA 6009, Perth, Australia (lyle@maths.uwa.edu.au).

Since $B_1 \neq \mathbf{0}$ in (1) we can set

$$t_0 := -\frac{\langle B_0, B_1\rangle}{\langle B_1, B_1\rangle}.$$

Then (1) is equivalent to $\dot{z}(t) = ((B_0 + tB_1) + (t - t_0)B_1)z(t)$. So after translation of t by $-t_0$ we can take $\langle B_0, B_1\rangle = 0$ in (1). Similarly, after dilation of t by $\|B_1\|$ we can suppose that B_1 has unit length. Then, applying an orthogonal change of coordinates to $z(t)$, we can suppose without loss of generality that

$$B_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\sqrt{c} \\ 0 & \sqrt{c} & 0 \end{bmatrix} \quad \text{and} \quad B_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix},$$

where $c > 0$. In other words, it suffices to study the *canonical form* of (1):

$$\dot{z}(t) = B_S(t)z(t), \quad \text{where} \quad B_S(t) := \begin{bmatrix} 0 & 0 & t \\ 0 & 0 & -\sqrt{c} \\ -t & \sqrt{c} & 0 \end{bmatrix}. \tag{3}$$

From skew-symmetry of $B_S(t)$ we see that $\langle z_i(t), z_j(t)\rangle$ is conserved, for solutions z_i, z_j of (3). Let $z_1, z_2, z_3 : \mathbb{R} \to E^3$ be the solutions of (3) satisfying $z_1(0) = (1, 0, 0)$, $z_2(0) = (0, 1, 0)$, $z_3(0) = (0, 0, 1)$. Then the z_i's map into the unit 2-sphere $S^2 \subset E^3$ and are pairwise orthogonal.

Example 2. *For $c = 3$, solutions $z_i : [-10, 10] \to E^3$ are plotted in Figures 1–3, respectively. The $z_i(t)$ appear to be asymptotic to circles in S^2 as $t \to \pm\infty$. This turns out to be true, but the asymptotic circles are not those that are apparent in Figures 1–3. In section 6 these illustrations are shown to be misleading: much longer simulations are needed to display the correct asymptotics.*

Proposition 1. *The general solution of (3) is*

$$z(t) = \frac{1}{\sqrt{c}}\begin{bmatrix} y(t) - t\dot{y}(t) & \sqrt{c}\dot{y}(t) & -\ddot{y}(t) \end{bmatrix}^{\mathbf{T}},$$

where $y : \mathbb{R} \to \mathbb{R}$ is the general solution of

$$y^{(3)}(t) = ty(t) - (t^2 + c)\dot{y}(t). \tag{4}$$

Proof. Differentiate z as given, and verify (3). ∎

Although we have superposition of solutions of linear ODEs (and not for nonlinear), a very interesting feature of (1) and (4) is that these *linear* ODEs can be effectively studied (and asymptotic solutions found) using *quadratic* ODEs (5) and (6). This is done as follows.

- In section 2, elements y_i of a basis of solutions of (4) are shown to satisfy a second order quadratic ODE (5). This nonlinear ODE is then used to prove local properties of the y_i, such as $\dot{y}_i(t) \neq 0$ for $|y_i(t)|$ large (Lemma 2).

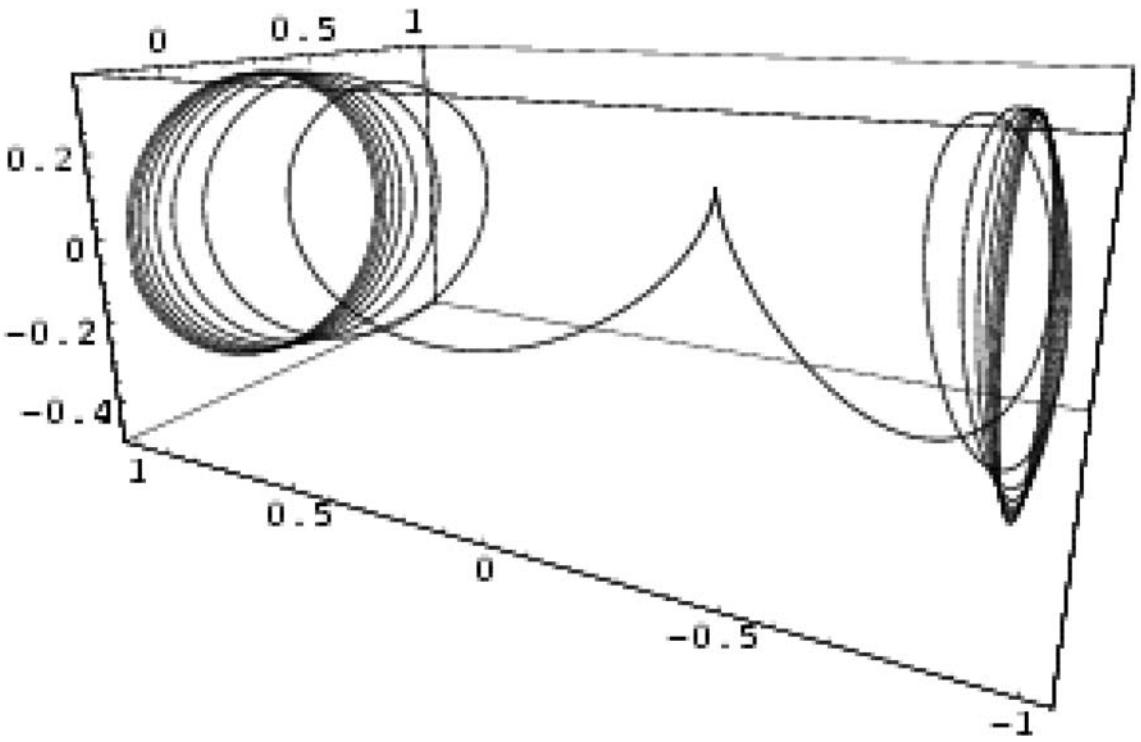


Figure 1. $z_1 : [-10, 10] \to E^3$ *with* $z_1(0) = (1, 0, 0)$ *in Example* 2.

- In section 3 background is given for the theory of Lie quadratics. Lemma 6 shows that the y_i are components of a solution (a *canonical null Lie quadratic* in E^3) of the quadratic ODE (6). All other solutions of (6) can be found from the canonical null Lie quadratic. Section 4 reviews and sharpens results of [12] on asymptotics of null Lie quadratics V in E^3. In particular, $V(t), \dot{V}(t), \ddot{V}(t)$ are estimated with errors $O(t^{-2}), O(t^{-1}), O(1)$ for $|t|$ large, corresponding (by Proposition 2) to $O(1)$ errors in estimates of solutions z_i of (3).
- We seek improved estimates of the $z_i(t)$, with only $O(t^{-1})$ errors. The improvements are significant because (3) describes dynamics of homogeneous rigid balls in simple situations (sections 7.1, 7.2), and a matrix with columns z_i is a *canonical null Riemannian cubic* in $SO(3)$ (section 7.3). Riemannian cubics are examples of higher order geodesics, used for interpolation in Riemannian manifolds. To improve the estimates of the z_i, we first find $O(t^{-3}), O(t^{-2}), O(t^{-1})$-accurate asymptotics for $V(t), \dot{V}(t), \ddot{V}(t)$ (Theorems 3, 2, 1, respectively, of section 5). New asymptotic axes $\beta_\pm$ appear in Theorem 1, after manipulations of inequalities and identities from section 4. Theorem 1 is then used to prove Theorems 2 and 3, using similar methods. This substantially improves our understanding of $V, \dot{V}, \ddot{V}$ for moderate values of t, as illustrated in Example 5.
- In section 6, the improved estimates for canonical null Lie quadratics are reexamined and applied to give $O(t^{-1})$-accurate asymptotic estimates (Theorem 4) for solutions z

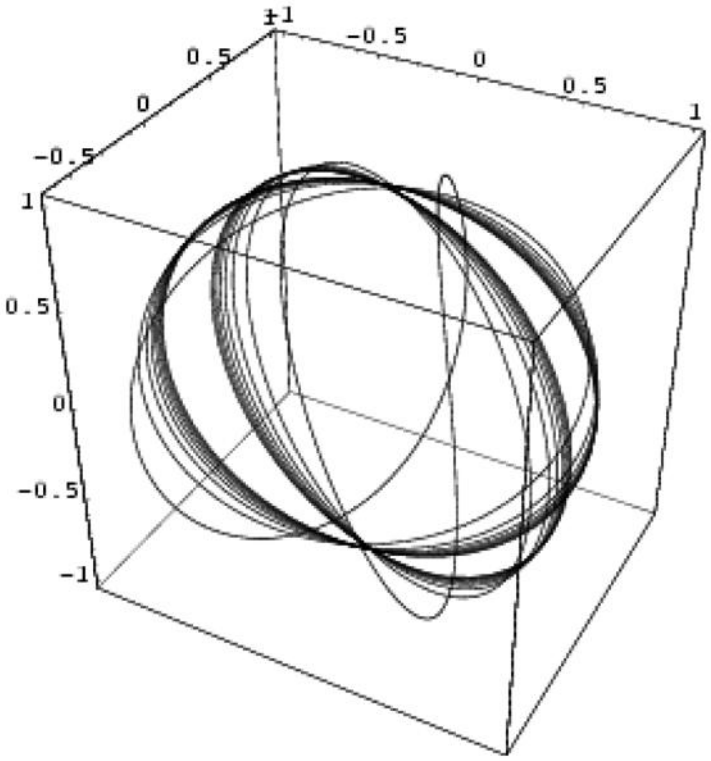


Figure 2. $z_2 : [-10, 10] \to E^3$ *with* $z_2(0) = (0, 1, 0)$ *in Example* 2.

of (3) (canonical null Riemannian cubics).

- The z_i are shown to be asymptotic to parallel pairs of circles for $i = 1, 3$, and to a single circle for $i = 2$ (Corollary 8), contradicting the apparent appearance of asymptotes in Figures 1–3 of Example 2. Simulations of $z_i(t)$ for much larger values of t confirm that the medium-term behavior is characterized by gradual shifting of apparent circular asymptotes into configurations conforming to Corollary 8.
- Section 7 discusses applications to path-planning for spherically symmetric balls.

2. The ODE (4). For $i = 1, 2, 3$ let $y_i : \mathbb{R} \to \mathbb{R}$ be the solutions of (4) corresponding by Proposition 1 to the z_i. Then

$$\begin{aligned} y_1(0) &= \sqrt{c}, \quad \dot{y}_1(0) = 0, \quad \ddot{y}_1(0) = 0, \\ y_2(0) &= 0, \quad \dot{y}_2(0) = 1, \quad \ddot{y}_2(0) = 0, \\ y_3(0) &= 0, \quad \dot{y}_3(0) = 0, \quad \ddot{y}_3(0) = -\sqrt{c}. \end{aligned}$$

From (4), y_1 and y_3 are even functions, and y_2 is odd.

Example 3. *For $c = 3$, we plot y_1, y_2, y_3 in Figures 4, 5, 6, respectively. Figure 4 resembles a parabola with a flattened base, but y_1 is not so bland as might be thought: Figure 7 shows that $\dot{y}_1$ is very convoluted. In Figure 5, y_2 also appears oscillatory with dampening as $|t|$ increases (owing to numerical error, y_2 is shown not (quite) odd). Except that it is an even function, y_3 in Figure 6 has something of the appearance of y_2.*

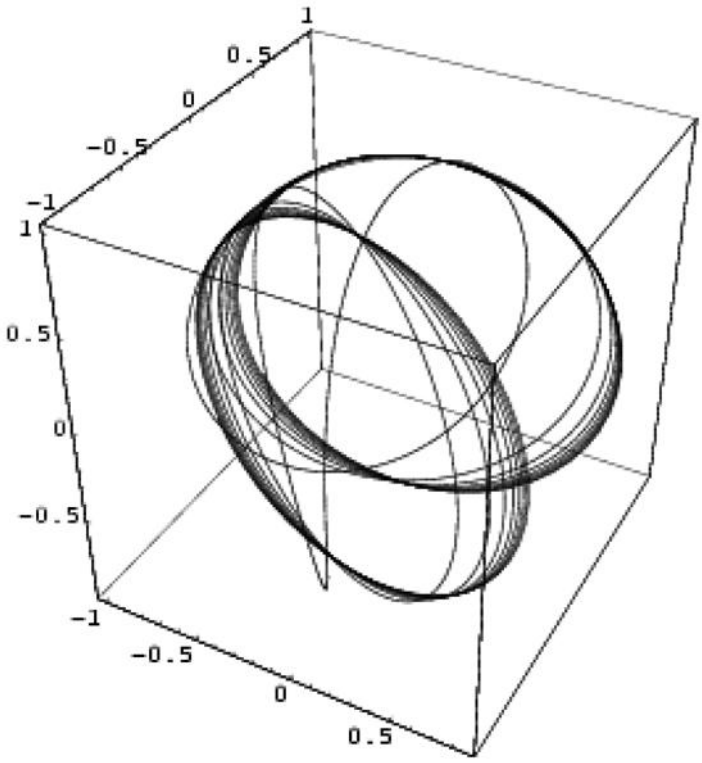


Figure 3. $z_3 : [-10, 10] \to E^3$ *with* $z_3(0) = (0, 0, 1)$ *in Example 2.*

Because $\langle z_i(t), z_j(t)\rangle$ is conserved, Proposition 1 gives the following result.

Lemma 1. *For* $i, j = 1, 2, 3$,

$$\ddot{y}_i\ddot{y}_j + c\dot{y}_i\dot{y}_j + (y_i - t\dot{y}_i)(y_j - t\dot{y}_j) = c\delta_{ij},$$

where δ_{ij} *is Kronecker's delta.*

In particular, any y_i satisfies

$$\ddot{y}^2 + c\dot{y}^2 + (y - t\dot{y})^2 = c. \tag{5}$$

For any solution y of (5) and all $t \in \mathbb{R}$

$$\begin{aligned}(t^2 + c)\dot{y}^2 - 2t\dot{y}y + y^2 &\le c,\\ c\dot{y}^2 + \ddot{y}^2 &\le c,\\ cy^2 + (t^2 + c)\ddot{y}^2 &\le c(t^2 + c),\end{aligned}$$

where the last inequality says that the discriminant of (5) as a quadratic in $\dot{y}$ is nonnegative. Also from (5), we have the following result.

Lemma 2. *For any solution* y *of* (5) *and any* $t \in \mathbb{R}$, *if* $\dot{y}(t) = 0$, *then* $y(t)^2 + \ddot{y}(t)^2 = c$.

Notice that $y(t) = \sqrt{c}$ satisfies (5), with $y(0) = \sqrt{c}$, $\dot{y}(0) = 0$. Also $y(t) = t$ satisfies (5), with $y(0) = 0$, $\dot{y}(0) = 1$. However, (5) has other solutions satisfying these initial conditions.

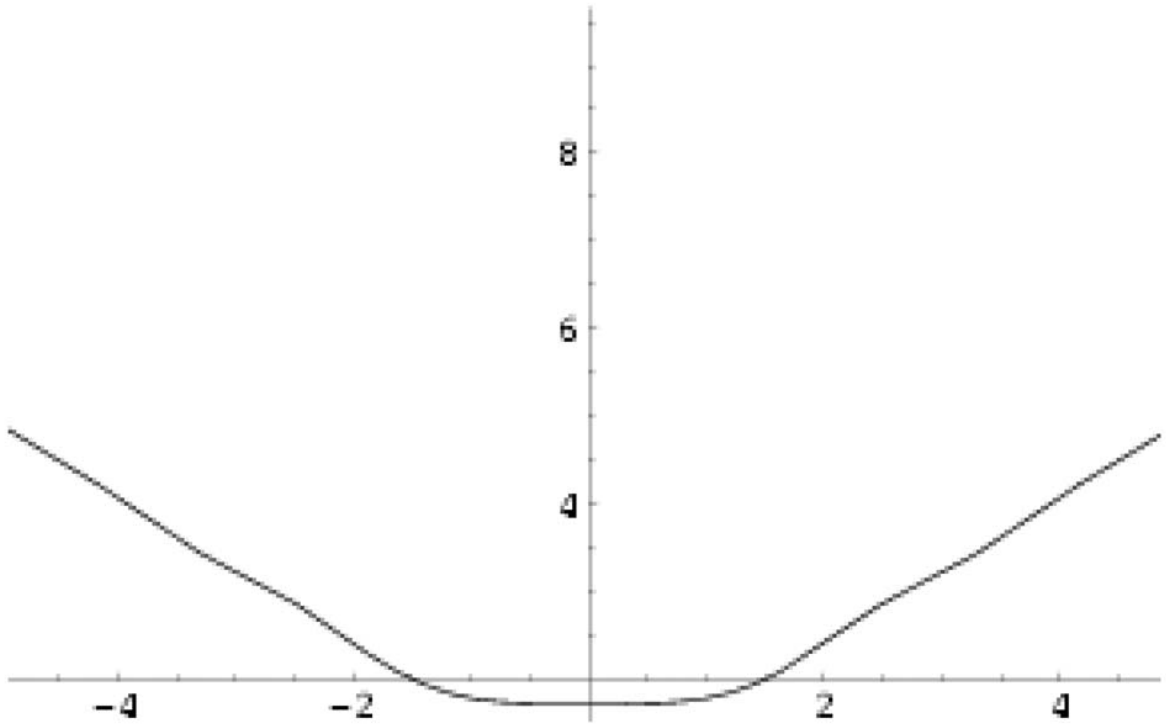


Figure 4. $y_1(0) = \sqrt{3}$, $\dot{y}_1(0) = 0$, $\ddot{y}_1(0) = 0$ *in Example* 3.

Lemma 3. $\dot{y}_1(t) > 0$ *for* $t > 0$.

Proof. We have $y_1^{(3)}(0) = 0$ and $y_1^{(4)}(0) = \sqrt{c} > 0$. So by Taylor's theorem, $\dot{y}_1(t) > 0$ for small $t > 0$. If the lemma does not hold, let t_0 be the supremum of the nonempty bounded

$$T := \{t_1 \in \mathbb{R} : \dot{y}_1(t) > 0 \ \text{ for all } t \in (0, t_1)\}.$$

Then $\dot{y}_1(t_0) = 0$ by continuity of $\dot{y}_1$. By Lemma 2, $\ddot{y}_1(t_0)^2 + y_1(t_0)^2 = c$. Now $t_0 > 0$ and therefore $y_1(t_0) > \sqrt{c}$, since y_1 is strictly increasing on $(0, t_0)$. So

$$c < y_1(t_0)^2 \leq \ddot{y}_1(t_0)^2 + y_1(t_0)^2 = c,$$

and the contradiction completes the proof. ■

From Lemma 3 the following result follows.

Proposition 2. y_1 *is an even function, strictly decreasing on* $(-\infty, 0)$, *strictly increasing on* $(0, \infty)$, *with a single point of global minimum at* $t = 0$ *and no other critical points.*

Lemma 4. *Any solution* y *of* (4) *satisfies*

$$\frac{d}{dt}\left(\ddot{y}y - \frac{1}{2}\dot{y}^2 + \frac{1}{2}(t^2 + c)y^2\right) = 2ty^2.$$

Proof. The left-hand side expands as

$$y^{(3)}y + ty^2 + (t^2 + c)\dot{y}y = y(y^{(3)} - ty + (t^2 + c)\dot{y}) + 2ty^2 = 2ty^2 \qquad \text{by (4)}. \quad ■$$

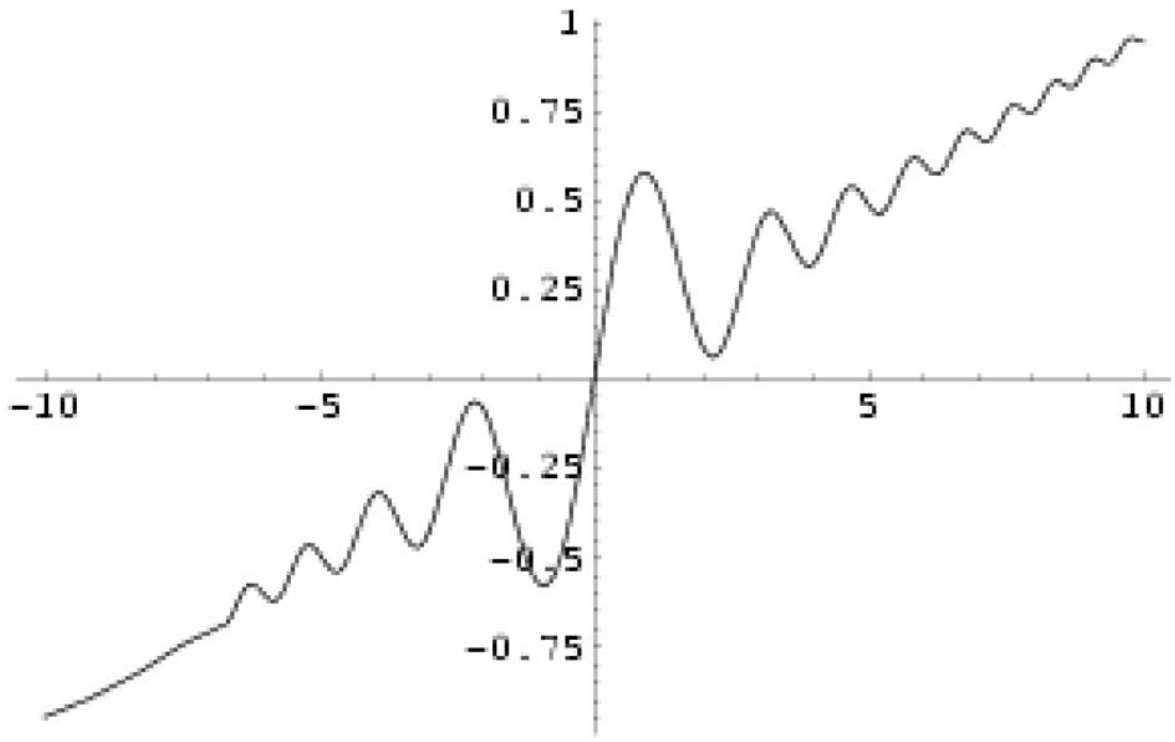


Figure 5. $y_2(0) = 0$, $\dot{y}_2(0) = 1$, $\ddot{y}_2(0) = 0$ *in Example* 3.

Proposition 3. *y_3 is an even function, strictly increasing on $(-\delta, 0)$ for some $\delta > 0$ and strictly decreasing on $(0, \delta)$. Also $y_3(0) = \dot{y}_3(0) = 0$, and $y_3(t) < 0$ for all $t \neq 0$.*

Proof. Because $c > 0$ and by Taylor's theorem, $\dot{y}_3(t) < 0$ for small $t > 0$. Assuming $y_3(t) \geq 0$ for some $t > 0$, let t_0 be the supremum of the nonempty bounded set

$$T := \{t_1 \in \mathbb{R} : y_3(t) < 0 \ \text{ for all } t \in (0, t_1)\}.$$

By continuity, $y_3(t_0) = 0$. By Lemma 4,

$$f(t) := \ddot{y}_3 y_3 - \frac{1}{2}\dot{y}_3^2 + \frac{1}{2}(t^2 + c)y_3^2$$

is strictly increasing on $(0, t_0)$, since y_3 is nowhere-zero on $(0, t_0)$. So

$$0 = f(0) < f(t_0) = -\frac{1}{2}\dot{y}_3(t_0)^2 \leq 0,$$

and the contradiction proves $y_3(t) < 0$ for $t > 0$. Because y_3 is even, this holds for all $t \neq 0$. ∎

Proposition 4. *Solutions of (4), (5) satisfy $\frac{d}{dt}(\ddot{y}\dot{y} - \frac{3t}{2}y^2 - ct) = -2(t^2 + c)\dot{y}^2 - \frac{5}{2}y^2$.*

Proof. Expanding the left-hand side gives $y^{(3)}\dot{y} + \ddot{y}^2 - 3t\dot{y}y - \frac{3}{2}y^2 - c$. Substituting for $\ddot{y}^2$ using (5), then for $y^{(3)}$ using (4), we obtain

$$\dot{y}(y^{(3)} - ty + (t^2 + c)\dot{y}) - 2(t^2 + c)\dot{y}^2 - \frac{5}{2}y^2 = -2(t^2 + c)\dot{y}^2 - \frac{5}{2}y^2. \qquad \blacksquare$$

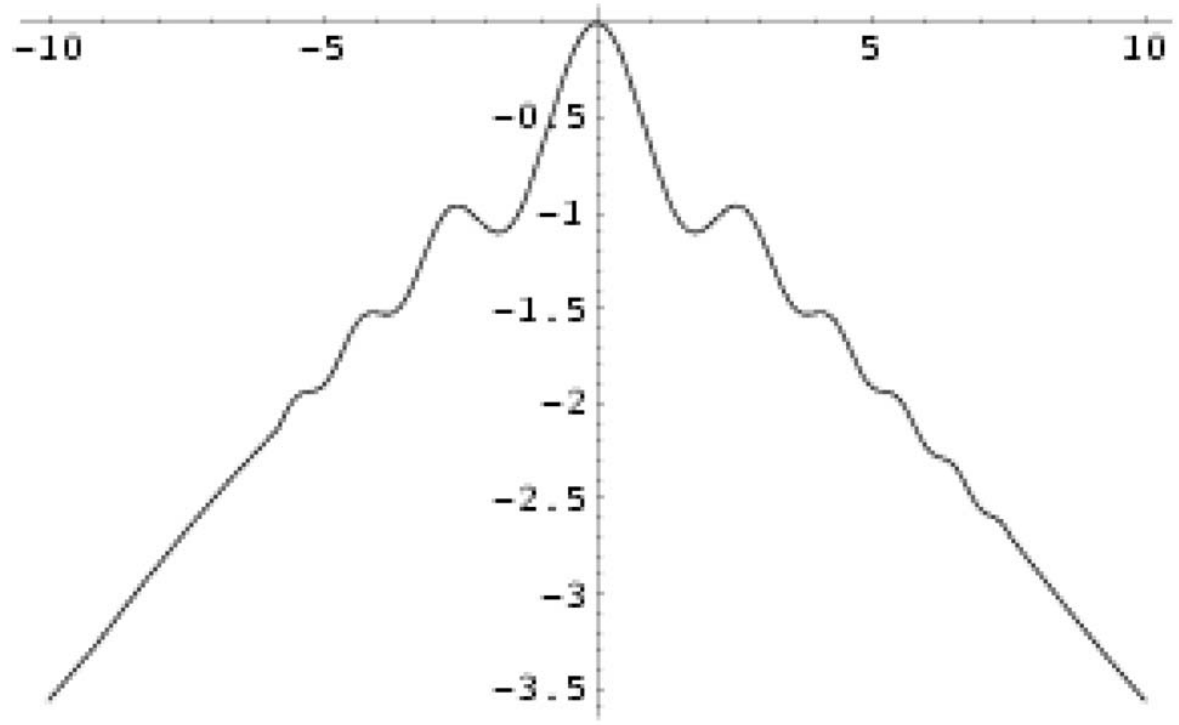


Figure 6. $y_3(0) = 0$, $\dot{y}_1(0) = 0$, $\ddot{y}_1(0) = -\sqrt{3}$ *in Example* 3.

3. Null Lie quadratics. The matrix valued function

$$x(t) := \begin{bmatrix} z_1(t) & z_2(t) & z_3(t) \end{bmatrix} \in SO(3)$$

also satisfies (3), and $x(0)$ is the identity matrix $\mathbf{1}$. Building on Definition 2, define $V : \mathbb{R} \to E^3$ by

$$\hat{V}(t) := x(t)^{-1}\dot{x}(t).$$

Lemma 5.

$$\ddot{V}(t) = \dot{V}(t) \times V(t). \tag{6}$$

Proof. By direct calculation, $\ddot{\hat{V}} = [\dot{\hat{V}}, \hat{V}]$. Also $\hat{}$ is a Lie isomorphism. ■

Curves V satisfying equations like (6) in other Lie algebras, and related curves x in Lie groups other than $SO(3)$, have been studied elsewhere.

Definition 1.

- *A smooth curve $V : \mathbb{R} \to \mathcal{G}$ in a real Lie algebra $\mathcal{G}$ is a* Lie quadratic *when*

$$\ddot{V}(t) = [\dot{V}(t), V(t)] + C$$

for some $C \in \mathcal{G}$. When $C = \mathbf{0}$ the Lie quadratic V is said to be null.

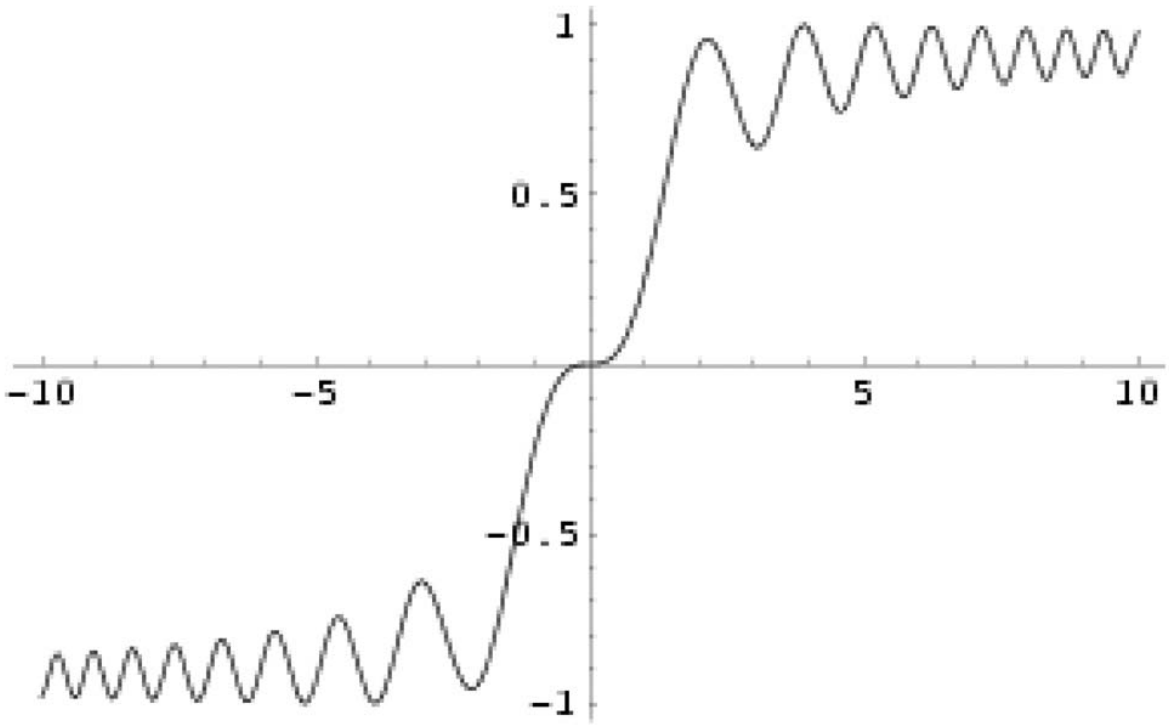


Figure 7. $\dot{y}_1$ *in Example* 3.

- *A* Riemannian cubic[1] *in a Lie group* G *is a smooth curve* $x : \mathbb{R} \to G$ *satisfying*

$$\dot{x}(t) = dL(x(t))_e V(t), \tag{7}$$

where $L(x(t))$ *is left-multiplication by* $x(t)$, $e \in G$ *is the identity, and* V *is a Lie quadratic in the Lie algebra* $\mathcal{G}$ *of* G. *The Riemannian cubic* x *is said to be* null *when* V *is a null Lie quadratic.*

By Lemma 5, our V is a null Lie quadratic in Euclidean 3-space E^3. Then, since $\hat{}$ is a Lie isomorphism, and from the definition of $V : \mathbb{R} \to E^3$, our x is a *null Riemannian cubic* in $SO(3)$. In [12] the study of null Lie quadratics is reduced to null Lie quadratics that are *canonical* in the following[2] sense.

Definition 2. *A null Lie quadratic* V *is said to be* canonical *when, for some* $c \geq 0$, $V(0) = (\sqrt{c}, 0, 0)^{\mathbf{T}}$ *and* $\dot{V}(0) = (0, 1, 0)^{\mathbf{T}}$.

Thus our V is also *canonical*. The formula for x in terms of V, given by Proposition 1, can also be obtained from [14, Theorem 5] or [15, Corollary 3.2]. An introduction to Riemannian cubics and Lie quadratics can be found in [12], where Proposition 3 says that any nonconstant

[1] In section 7.3, Riemannian cubics are defined differently, as curves in Riemannian (or even semi-Riemannian) manifolds M. The definitions are equivalent when M is a Lie group with bi-invariant Riemannian metric.

[2] Definition 2 is more restrictive than Definition 1 in [12], where it is required only that $\|V(0)\| = \sqrt{c}$ and $\|\dot{V}(0)\| = 1$.

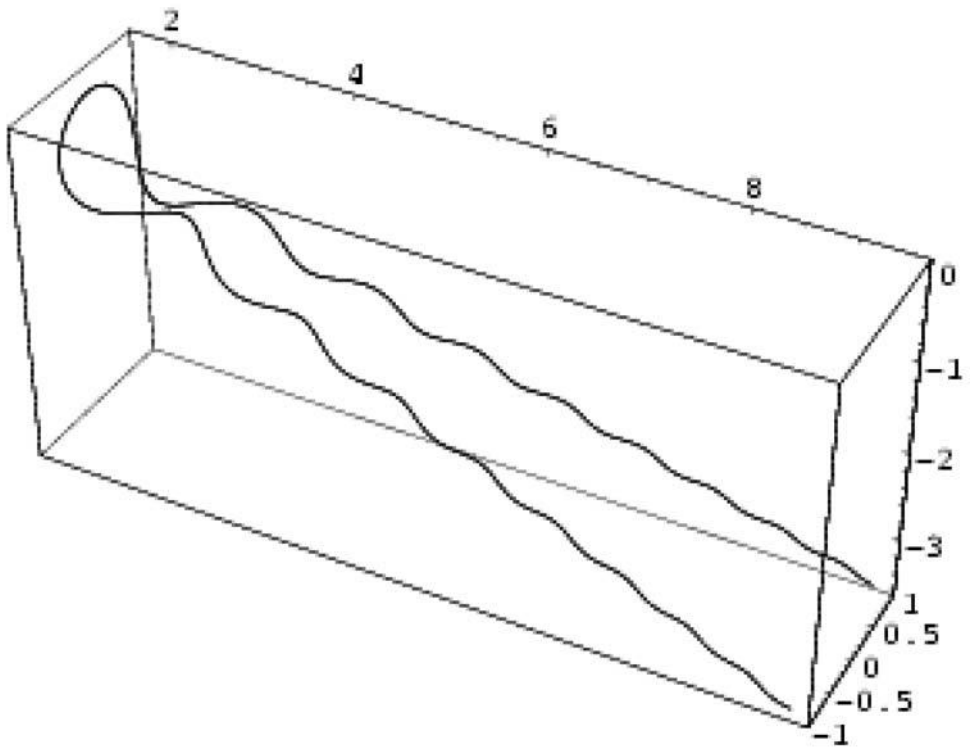


Figure 8. *The canonical null Lie quadratic V in E^3 with $c = 3$.*

null Lie quadratic can be written in terms of a canonical null Lie quadratic in the sense of [12]. The same symmetry argument proves the stronger statement where "canonical" has the meaning of Definition 2. Also in [12], canonical null Lie quadratics V in E^3 are shown to have

- constant curvature $\kappa(t) = \sqrt{c}$, and
- linear torsion $\tau(t) = -t$.

Example 4. *Figure 8 shows the canonical null Lie quadratic V with $c = 3$. The illustration certainly does not suggest constant curvature, and the illusion is accounted for by linearly varying torsion. Non-null Lie quadratics, mentioned briefly in section 7.3 (but nowhere else in the present paper), have much richer geometry [13].*

In [12] a canonical null Lie quadratic V is shown to satisfy

$$\langle V(t), V(t)\rangle = f(t)^2, \tag{8}$$

$$\langle \dot{V}(t), \dot{V}(t)\rangle = 1, \tag{9}$$

$$\langle \ddot{V}(t), \ddot{V}(t)\rangle = c, \tag{10}$$

where $f(t) := \sqrt{t^2 + c}$.

Lemma 6. $V(t) = \begin{bmatrix} y_1(t) & y_2(t) & y_3(t) \end{bmatrix}^{\mathbf{T}}$.

Proof. Differentiating the ODE in Lemma 5,

$$V^{(3)} = \ddot{V} \times V = (\dot{V} \times V) \times V = \langle \dot{V}, V\rangle V - \langle V, V\rangle \dot{V},$$

where we use the identity

$$(u \times v) \times w = \langle u, w\rangle v - \langle v, w\rangle u \quad \text{for} \quad u, v, w \in E^3. \tag{11}$$

Differentiating (8), we find $\langle \dot{V}(t), V(t)\rangle = t$. Then from (8), the components of V satisfy (4), as do the y_i. Because $V(0) = (\sqrt{c}, 0, 0)^{\mathbf{T}}$, $\dot{V}(0) = (0, 1, 0)^{\mathbf{T}}$, we have $\ddot{V}(0) = (0, 0, -\sqrt{c})^{\mathbf{T}}$. So the components of V also satisfy the initial conditions for the y_i. ■

From (8), (9), Lemma 6, and Proposition 2 we obtain the following results.

Corollary 1. *For all $t \neq 0$, $y_2(t)^2 + y_3(t)^2 < t^2$ and $\dot{y}_2(t)^2 + \dot{y}_3(t)^2 < 1$.*

Corollary 2.[3] *Let $I \subseteq \mathbb{R}$ be a nonempty open interval. Any $V : I \to E^3$ satisfying $\ddot{V} = \dot{V} \times V$ extends to a unique null Lie quadratic in E^3 defined over the whole of $\mathbb{R}$.*

Proof. As in the proof of Lemma 6, the components y_i of V are solutions of the linear ODE (4) and therefore uniquely extendible. ■

The next section reviews and sharpens some results of [12] on asymptotic properties of canonical null Lie quadratics in E^3.

4. Primary asymptotics of canonical null Lie quadratics in E^3. Notice $f\dot{f} = t$ and $\dot{f}^2 + f\ddot{f} = 1$. Thus $\ddot{f} = c/f^3$. Define $U(t) := V(t)/f(t) \in S^2$. First we have the existence of asymptotes, as follows.

Proposition 5 (see [12]). *Let V be a canonical null Lie quadratic in E^3 with $c > 0$. There exist limits*

$$\alpha_+ := \lim_{t\to\infty} U(t) \quad \textit{and} \quad \alpha_- := \lim_{t\to-\infty} U(t).$$

Unfortunately no simple formula is known giving $\alpha_\pm \in S^2$ in terms of c. However, a number of inequalities for asymptotes can be proved. Let $\rho : E^3 \to E^3$ be reflection in the hyperplane orthogonal to the second coordinate axis. Then it is known from [12] that $\alpha_- = \rho(\alpha_+)$. From Lemma 6 and Proposition 2 we find the following.

Corollary 3. $\langle \alpha_\pm, V(0)\rangle \geq 0$.

From Lemma 6 and Proposition 3 we have the next three corollaries.

Corollary 4. $\langle \alpha_\pm, \ddot{V}(0)\rangle \geq 0$.

Corollary 5. *For all $t \neq 0$, $y_1(t)^2 + y_2(t)^2 < t^2 + c$.*

Corollary 6. *For all $t \neq 0$, $-|t| < y_3(t) < 0$. If $\dot{y}_3(t) = 0$, then $y_3(t) \geq -\sqrt{c}$.*

Proof of Corollary 6. Suppose $t \neq 0$. By Corollary 1, $y_3(t)^2 < t^2$. By Proposition 3, $y_3(t) < 0$. Lemma 2 completes the proof. ■

Set $\tilde{V}(t) := V(t) + \frac{1}{f(t)^2}\ddot{V}(t)$ and $\tilde{U}(t) := \frac{f(t)^2}{\sqrt{f(t)^6+c}}\tilde{V}(t) \in S^2$. In [12] the identity

$$\frac{d}{dt}\left(U(t) + \frac{\ddot{V}(t)}{f(t)^3}\right) = -\frac{3t\ddot{V}(t)}{f(t)^5} \tag{12}$$

is used to prove the next proposition.

Proposition 6 (see [12]). *For $t \geq 0$ and $t \leq 0$, respectively,*

$$\|\tilde{V}(t) - f(t)\,\alpha_\pm\| \leq \frac{\sqrt{c}}{f(t)^2}.$$

[3] Similar results are proved in [13] for non-null Lie quadratics.

In particular, by (10), $\|V(t) - f(t)\,\alpha_\pm\| \le \frac{2\sqrt{c}}{f(t)^2}$.

Corollary 7.

$$\|\alpha_\pm - \tilde{U}(0)\| \le \frac{\sqrt{2}}{\sqrt{\sqrt{c^2+1}(\sqrt{c^2+1}+c)}} < \frac{1}{c}.$$

Proof. Squaring both sides of $\|V(0) + \frac{1}{c}\ddot{V}(0) - \sqrt{c}\alpha_\pm\| \le \frac{1}{\sqrt{c}}$ and expanding the left,

$$c + \frac{1}{c} + c - 2\sqrt{c}\langle\alpha_\pm, V(0)\rangle - \frac{2}{\sqrt{c}}\langle\alpha_\pm, \ddot{V}(0)\rangle \le \frac{1}{c},$$

namely, $\langle\alpha_\pm, \tilde{V}(0)\rangle \ge \sqrt{c}$. Since $\|\tilde{V}(0)\| = \frac{\sqrt{c^2+1}}{\sqrt{c}}$, $\langle\alpha_\pm, \tilde{U}(0)\rangle \ge \frac{c}{\sqrt{c^2+1}}$. So

$$\|\alpha_\pm - \tilde{U}(0)\|^2 = 2 - 2\langle\alpha_\pm, \tilde{U}(0)\rangle \le 2\frac{\sqrt{c^2+1}-c}{\sqrt{c^2+1}} = \frac{2}{\sqrt{c^2+1}(\sqrt{c^2+1}+c)}. \qquad \blacksquare$$

In effect, Proposition 6 calculates $V(t)$ as $f(t)\alpha_\pm + O(\frac{1}{f(t)^2})$ as $t \to \pm\infty$. This asymptotic estimate is significantly improved in section 5 below. In section 6, the improved accuracy translates into useful asymptotics for solutions z of (3).

5. Secondary asymptotics. Before estimating $V, \dot{V}, \ddot{V}$ we try to relate these as closely as possible to the asymptotes $\alpha_\pm$.

Lemma 7. *For $t \ge 0$ and $t \le 0$, respectively,* $|\langle\alpha_\pm, \ddot{V}(t)\rangle| \le \frac{2c}{f(t)^3}$.

Proof. By Proposition 6, $\|\frac{\tilde{V}(t)}{f(t)} - \alpha_\pm\| \le \frac{\sqrt{c}}{f(t)^3}$. So by the Cauchy–Schwarz inequality and (10),

$$\left|\langle\ddot{V}, \alpha_\pm\rangle - \frac{\langle\ddot{V}, \tilde{V}\rangle}{f}\right| \le \frac{c}{f^3}.$$

But by (10) and Lemma 5, $\langle\ddot{V}, \tilde{V}\rangle = \frac{c}{f^2}$. $\blacksquare$

Lemma 8. *For $t \ge 0$ and $t \le 0$, respectively,*

$$-\frac{3c}{f(t)^5} \le \langle\alpha_\pm, V(t)\rangle - f(t) \le 0.$$

Proof. Taking inner products of (12) with α_+ and integrating both sides from $s > 0$ to ∞,

$$1 - \langle U(s), \alpha_+\rangle - \frac{\langle\ddot{V}(s), \alpha_+\rangle}{f(s)^3} = -\int_s^\infty \frac{3t\langle\ddot{V}(t), \alpha_+\rangle}{(t^2+c)^{5/2}}dt.$$

So by Lemma 7, $1 - \langle U(s), \alpha_+\rangle \le \frac{2c}{f(s)^6} + \int_s^\infty \frac{6ct}{(t^2+c)^4}dt = \frac{3c}{f(s)^6}$. Replacing s by t and multiplying through by $f(t)$, the left-hand inequality holds for $t \ge 0$ and α_+. The other inequality follows from the Cauchy–Schwarz inequality. The proof for $t \le 0$ and α_- is similar. $\blacksquare$

Lemma 9. *For $t \ge 0$ and $t \le 0$, respectively,* $|\langle\alpha_\pm, \ddot{V}(t) \times V(t)\rangle| \le \frac{2c}{f(t)^2}$.

Proof. $\langle\alpha_\pm, \ddot{V}(t) \times V(t)\rangle = -\langle\alpha_\pm \times V(t), \ddot{V}(t)\rangle$. So the lemma follows from Proposition 6, the Cauchy–Schwarz inequality, and (10). $\blacksquare$

The restrictions $\hat{\alpha}'_\pm$ of $\hat{\alpha}_\pm$ to

$$H_\pm := \alpha_\pm^\perp := \{v \in E^3 : \langle v, \alpha_\pm\rangle = 0\}$$

are orthogonal endomorphisms. By (11), $\hat{\alpha}'_\pm$ is skew-adjoint, and $(\hat{\alpha}'_\pm)^2$ is multiplication by -1. Thus $\hat{\alpha}'_\pm$ defines a complex structure on $H_\pm$. As before, let $\rho : E^3 \to E^3$ be reflection in the hyperplane orthogonal to $\dot{V}(0) = (0,1,0)^{\mathbf{T}}$.

Lemma 10. *$\rho(H_+) = H_-$ and $\rho \circ \hat{\alpha}'_+ = -\hat{\alpha}'_- \circ \rho$.*

Proof. Because ρ is orthogonal and $\rho(\alpha_+) = \alpha_-$, we have for $v \in H_+$

$$0 = \langle v, \alpha_+\rangle = \langle \rho(v), \rho(\alpha_+)\rangle = \langle \rho(v), \alpha_-\rangle,$$

namely $\rho(v) \in H_-$. Because ρ is a reflection,

$$\rho \circ \hat{\alpha}'_+(v) = \rho(\alpha_+ \times v) = -(\rho(\alpha_+) \times \rho(v)) = -\alpha_- \times \rho(v). \qquad \blacksquare$$

The following function is needed to describe the asymptotics of $V, \dot{V}, \ddot{V}$.

Definition 3. $g(t) := \frac{1}{2}(|t|f(t) + c\ln(|t| + f(t)) - (c/2)\ln c)$.

Now Lemmas 7–9 are used to give an asymptotic estimate for $\ddot{V}$.

Theorem 1. *For $t \geq 0$ and $t \leq 0$, respectively, and some unit vector $\beta_\pm \in H_\pm$,*

$$\ddot{V}(t) = \sqrt{c}\exp(\mp g(t)\hat{\alpha}'_\pm)\beta_\pm + O\left(\frac{1}{f(t)}\right),$$

where $\beta_- = \rho(\beta_+)$.

Proof. By Lemma 7, $Z_\pm(t) := \ddot{V}(t) - \langle \ddot{V}(t), \alpha_\pm\rangle\alpha_\pm = \ddot{V}(t) + O(\frac{1}{f(t)^3})\alpha_\pm$ and

$$\langle Z_\pm(t), Z_\pm(t)\rangle = c - \langle \ddot{V}, \alpha_\pm\rangle^2 = c + O\left(\frac{1}{f(t)^6}\right). \tag{13}$$

Differentiating the equation in Lemma 5, $V^{(3)} = \ddot{V} \times V$, and so

$$\dot{Z}_\pm(t) = \ddot{V} \times V - \langle \ddot{V} \times V, \alpha_\pm\rangle\alpha_\pm.$$

So by Lemmas 5 and 9,

$$\begin{aligned}\langle \dot{Z}_\pm(t), \dot{Z}_\pm(t)\rangle &= \|\ddot{V} \times V\|^2 - \langle \ddot{V} \times V, \alpha_\pm\rangle^2 = cf(t)^2 - \langle \ddot{V}, \alpha_\pm \times V\rangle^2 \\ &= cf(t)^2 + O\left(\frac{1}{f(t)^4}\right). \end{aligned} \tag{14}$$

Notice that $Z_\pm(t) \in H_\pm$. By Lemmas 5 and 8,

$$\begin{aligned}\langle \dot{Z}_\pm(t), \hat{\alpha}'_\pm \circ Z_\pm(t)\rangle &= \langle \ddot{V} \times V, \alpha_\pm \times \ddot{V}\rangle = -\langle V, \alpha_\pm\rangle\langle \ddot{V}, \ddot{V}\rangle \\ &= -cf(t) + O\left(\frac{1}{f(t)^5}\right). \end{aligned} \tag{15}$$

Setting $W_\pm(t) := \dot{Z}_\pm(t) + f(t)\hat{\alpha}'_\pm \circ Z_\pm(t)$,

$$\langle W_\pm, W_\pm\rangle = \langle \dot{Z}_\pm, \dot{Z}_\pm\rangle + f^2\langle Z_\pm, Z_\pm\rangle + 2f\langle \dot{Z}_\pm, \hat{\alpha}'_\pm \circ Z_\pm\rangle = O\left(\frac{1}{f^4}\right)$$

by (13), (14), (15). So $\dot{Z}_\pm(t) + f(t)\hat{\alpha}'_\pm \circ Z_\pm(t) = W_\pm(t) = O(\frac{1}{f(t)^2})$. Since $\dot{g}(t) = f(t)$ for $t > 0$ we have, for $0 < r < s$,

$$\text{(16)}\quad \exp(g(s)\hat{\alpha}'_+)Z_+(s) - \exp(g(r)\hat{\alpha}'_+)Z_+(r) = \int_r^s \exp(g(t)\hat{\alpha}'_+)W_+(t)dt = O\left(\frac{1}{f(r)} - \frac{1}{f(s)}\right).$$

Let $\beta_+ \in H_+$ be the limit of a convergent subsequence $\{\beta_{i_j} : j \geq 1\}$ of the bounded

$$\left\{\frac{\exp(g(i)\hat{\alpha}'_+)Z_+(i)}{\sqrt{c}} : i \geq 1\right\}.$$

By (16), $\exp(g(r)\hat{\alpha}'_+)Z_+(r) = \sqrt{c}\beta_{i_j} + O(\frac{1}{f(r)})$. Taking limits as $j \to \infty$,

$$\ddot{V}(r) = Z_+(r) + O\left(\frac{1}{f(r)^3}\right)\alpha_+ = \sqrt{c}\exp(-g(r)\hat{\alpha}'_+)\beta_+ + O\left(\frac{1}{f(r)}\right),$$

proving the result for $t \geq 0$ and β_+. A symmetry argument completes the proof. ■

Now, using the asymptotic estimate for $\ddot{V}$, we are able to estimate $\dot{V}$ and V.

Theorem 2. *For $t \geq 0$ and $t \leq 0$, respectively, we have*

$$\dot{V}(t) = \frac{t}{f}\alpha_\pm + \frac{\sqrt{c}}{f}\hat{\alpha}'_\pm \exp(\mp g(t)\hat{\alpha}'_\pm)\beta_\pm + O\left(\frac{1}{f^2}\right).$$

Proof. By symmetry it suffices to consider $t \geq 0$ and β_+. By Lemma 5 and (11),

$$\ddot{V} \times V = (\dot{V} \times V) \times V = \langle \dot{V}, V\rangle V - \langle V, V\rangle \dot{V} = tV - f^2\dot{V},$$

namely $\dot{V}(t) = \frac{tV(t) - \ddot{V}(t)\times V(t)}{f(t)^2}$. So by Proposition 6, for $t \geq 0$,

$$\dot{V}(t) = \frac{tf(t)\alpha_+ - f\ddot{V}(t) \times \alpha_+}{f^2} + O\left(\frac{1}{f^3}\right)$$
$$= \frac{t}{f}\alpha_+ + \frac{1}{f}\hat{\alpha}'_+(\ddot{V}(t)) + O\left(\frac{1}{f^3}\right) = \frac{t}{f}\alpha_+ + \frac{\sqrt{c}}{f}\hat{\alpha}'_+ \exp(-g(t)\hat{\alpha}'_+)\beta_+ + O\left(\frac{1}{f^2}\right)$$

by Theorem 1. ■

Theorem 3. *For $t \geq 0$ and $t \leq 0$, respectively,*

$$V(t) = f(t)\alpha_\pm - \frac{\sqrt{c}}{f(t)^2}\exp(\mp g(t)\hat{\alpha}'_\pm)\beta_\pm + O\left(\frac{1}{f(t)^3}\right).$$

Proof. By (12) and Theorem 1, for $0 < r < s$,

$$U(s) + \frac{\ddot{V}(s)}{f(s)^3} - U(r) - \frac{\ddot{V}(r)}{f(r)^3} = -\int_r^s \frac{3\sqrt{c}t\exp(-g(t)\hat{\alpha}'_+)\beta_+}{f(t)^5}dt + O\left(\frac{1}{f(r)^4}\right)$$
$$= 3\sqrt{c}\hat{\alpha}'_+\int_r^s f(t)\hat{\alpha}'_+\exp(-g(t)\hat{\alpha}'_+)\beta_+\left(\frac{t}{f(t)^6}\right)dt + O\left(\frac{1}{f(r)^4}\right)$$
$$= 3\sqrt{c}\hat{\alpha}'_+\left(-\exp(-g(s)\hat{\alpha}'_+)\beta_+\left(\frac{s}{f(s)^6}\right) + \exp(-g(r)\hat{\alpha}'_+)\beta_+\left(\frac{r}{f(r)^6}\right)\right)$$
$$+ 3\sqrt{c}\hat{\alpha}'_+\int_r^s \exp(-g(t)\hat{\alpha}'_+)\beta_+\left(\frac{c-5t^2}{f(t)^8}\right)dt + O\left(\frac{1}{f(r)^4}\right)$$

after integration by parts using $\dot{g} = f$. Since all except the last term on the right-hand side are $O(\frac{1}{f(r)^5})$,

$$U(s) + \frac{\ddot{V}(s)}{f(s)^3} = U(r) + \frac{\ddot{V}(r)}{f(r)^3} + O\left(\frac{1}{f(r)^4}\right),$$

and, as $s \to \infty$, this gives

$$U(r) = \alpha_+ - \frac{\ddot{V}(r)}{f(r)^3} + O\left(\frac{1}{f(r)^4}\right) = \alpha_+ - \frac{\sqrt{c}}{f(r)^3}\exp(-g(r)\hat{\alpha}'_+)\beta_+ + O\left(\frac{1}{f(r)^4}\right),$$

by Theorem 1. This proves Theorem 3 for $t \geq 0$. The result follows by symmetry for $t \leq 0$. ∎

The asymptotic estimates of Theorems 1–3 can be surprisingly sharp, even for moderate values of t.

Example 5. *Figures 9 and 10 show $y_1(t)$, $\dot{y}_1(t)$, and their asymptotic estimates for $0 < t < 3$. Figure 11 does the same for $\ddot{y}_1(t)$ for $0 < t < 10$. For $|t| > 4$ the estimates are more or less indistinguishable from $y_1(t)$, $\dot{y}_1(t)$, $\ddot{y}_1(t)$ obtained by numerically solving (4). Taylor approximation about $t = 0$ suffices for $|t| \leq 4$.*

By comparison,

- *the previous best asymptotic estimates for $y_1(t)$ are linear in $|t|$ as $t \to \pm\infty$;*
- *the previous best estimate of $\dot{y}_1(t)$ is that $\lim_{t\to\pm\infty}\dot{y}_1(t)$ exist, saying nothing about the finer structure observed in Figure 10;*
- *the only asymptotic statement previously available for $\ddot{y}_1(t)$ was that it is bounded, saying nothing about the behavior (oscillations with constant amplitude and linearly increasing frequency) observed in Figure 11.*

Similar comments can be made for y_2, y_3.

6. Back to the ODE (3). For the canonical null Lie quadratic V in E^3 given by $c > 0$ Theorems 3, 2, 1 say that there are orthogonal unit vectors $\alpha, \beta \in S^2$ such that, for $t > 0$,

$$V = f\alpha - \frac{\sqrt{c}}{f^2}\exp(-gJ)\beta \quad + O\left(\frac{1}{f^3}\right), \tag{17}$$

$$\dot{V} = \frac{t}{f}\alpha + \frac{\sqrt{c}}{f}J\exp(-gJ)\beta \quad + O\left(\frac{1}{f^2}\right), \tag{18}$$

$$\ddot{V} = \sqrt{c}\exp(-gJ)\beta \quad + O\left(\frac{1}{f}\right), \tag{19}$$

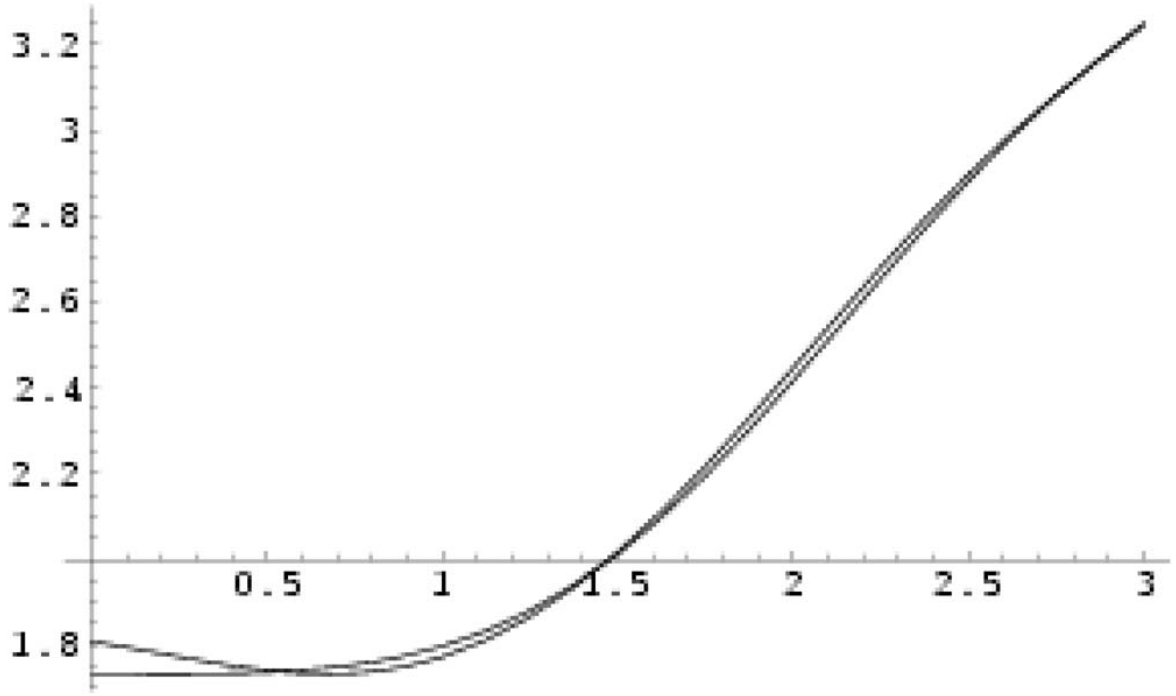


Figure 9. *$y_1(t)$ and its estimate (higher near $t = 0$) for $0 < t < 3$, with $c = 3$.*

with $\alpha = \alpha_+$, $\beta = \beta_\pm$, and J the restriction $\hat{\alpha}'$ of $\hat{\alpha}$ to $\alpha^\perp$. Defining

$$a := \langle V, \alpha\rangle, \quad b := \langle V, \exp(-gJ)\beta\rangle, \quad d := \langle V, J\exp(-gJ)\beta\rangle,$$

we have

$$\begin{bmatrix} V \\ \dot{V} \\ \ddot{V} \end{bmatrix} = \begin{bmatrix} a & b & d \\ \dot{a} & \dot{b} + df & \dot{d} - bf \\ \ddot{a} & \ddot{b} + 2\dot{d}f + d\dot{f} - bf^2 & \ddot{d} - 2\dot{b}f - b\dot{f} - df^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \exp(-gJ)\beta \\ J\exp(-gJ)\beta \end{bmatrix},$$

and consequently, by (17), (18), (19),

$$\begin{bmatrix} a \\ b \\ d \end{bmatrix} = \begin{bmatrix} f \\ -\sqrt{c}f^{-2} \\ 0 \end{bmatrix} + O\left(\frac{1}{f^3}\right), \qquad \begin{bmatrix} \dot{a} \\ \dot{b} \\ \dot{d} \end{bmatrix} = \begin{bmatrix} \dot{f} \\ 0 \\ 0 \end{bmatrix} + O\left(\frac{1}{f^2}\right), \qquad \begin{bmatrix} \ddot{a} \\ \ddot{b} \\ \ddot{d} \end{bmatrix} = O\left(\frac{1}{f}\right).$$

By Lemmas 7, 8, 9, and then (24) below for $\dot{a}$,

$$a = f + O\left(\frac{1}{f^5}\right), \quad \dot{a} = \dot{f} + O\left(\frac{1}{f^4}\right), \quad \ddot{a} = O\left(\frac{1}{f^3}\right), \quad a^{(3)} = O\left(\frac{1}{f^2}\right).$$

These lemmas also give more detailed estimates which can be useful for moderate values of t.

Example 6. *By Lemma 8, $a(t) = \sqrt{t^2 + c} - \delta(t)$, where*

$$0 \le \delta(t) \le \frac{3c}{f(t)^5}.$$

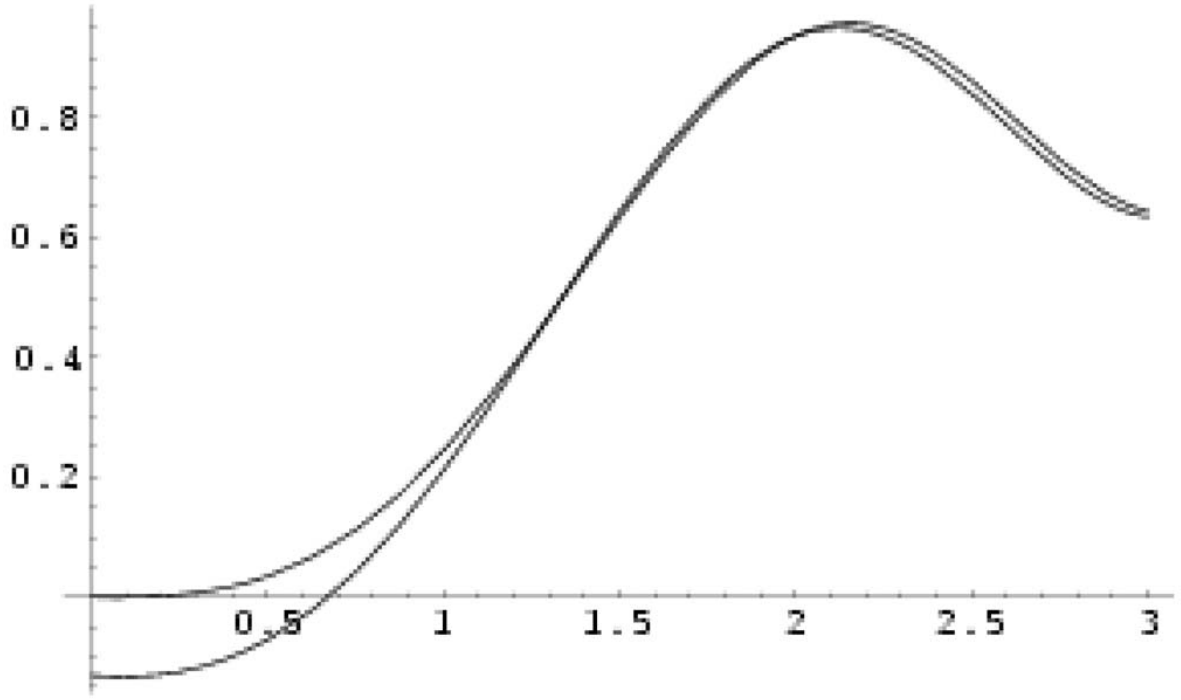


Figure 10. *$\dot{y}_1(t)$ and its estimate (lower near $t = 0$) for $0 < t < 3$, with $c = 3$.*

With $c = 3$ and $t > 2$, the bound on the right-hand side is smaller than $9/7^{5/2} \approx 0.069$. For $t > 4$ we obtain instead $9/19^{5/2} \approx 0.0057$.

Substituting for $V, \dot{V}, \ddot{V}$ in Lemma 5,

$$\ddot{V} = \ddot{a}\alpha + (\ddot{b} + 2\dot{d}f + d\dot{f} - bf^2)\exp(-gJ)\beta + (\ddot{d} - 2\dot{b}f - b\dot{f} - df^2)J\exp(-gJ)\beta$$
$$= (\dot{a}\alpha + (\dot{b} + df)\exp(-gJ)\beta + (\dot{d} - bf)J\exp(-gJ)\beta) \times (a\alpha + b\exp(-gJ)\beta + dJ\exp(-gJ)\beta)$$
$$= (-b\dot{d} + d\dot{b} + b^2 f + d^2 f)\alpha + (a\dot{d} - \dot{a}d - abf)\exp(-gJ)\beta + (\dot{a}b - a\dot{b} - adf)J\exp(-gJ)\beta,$$

and then

$$\ddot{a} = -b\dot{d} + d\dot{b} + b^2 f + d^2 f, \tag{20}$$
$$\ddot{b} = -2\dot{d}f - d\dot{f} + bf^2 + a\dot{d} - \dot{a}d - abf, \tag{21}$$
$$\ddot{d} = \quad 2\dot{b}f + b\dot{f} + df^2 + \dot{a}b - a\dot{b} - adf. \tag{22}$$

Notice also

$$a^2 + b^2 + d^2 - f^2 = 0, \tag{23}$$
$$a^{(3)} - ta + f^2\dot{a} = 0. \tag{24}$$

Similar statements hold for $t < 0$ but with different α, β, and J.

Substituting from (17), (18), (19) into Proposition 1, we obtain the following theorem.

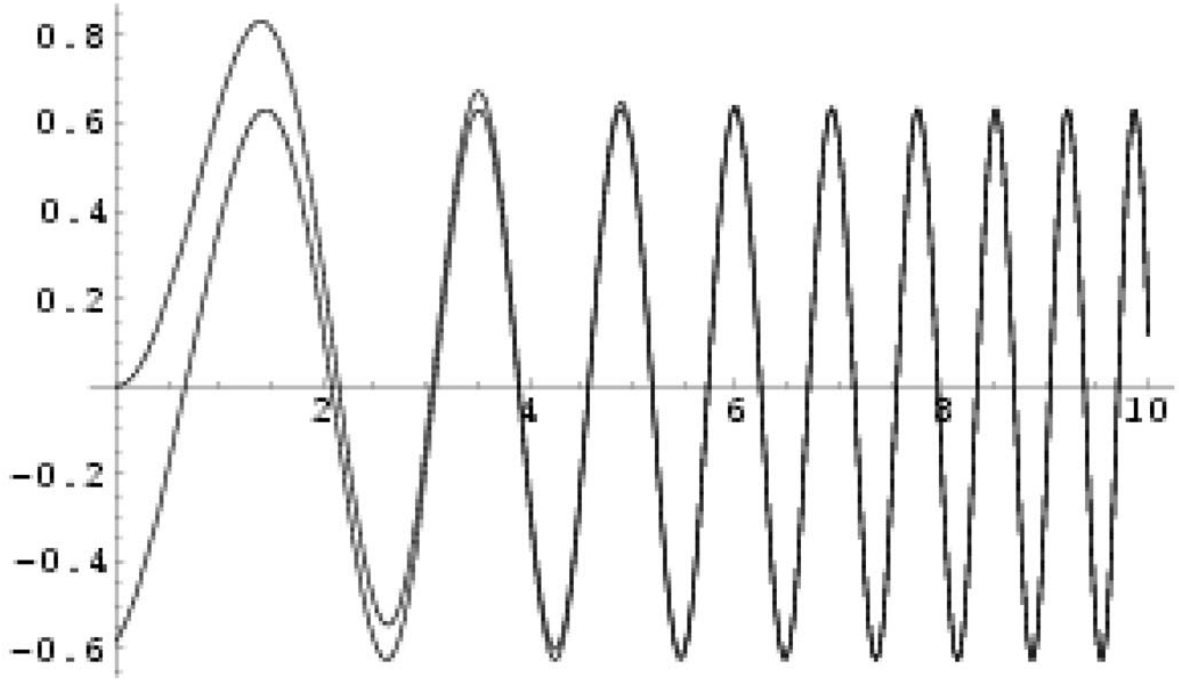


Figure 11. *$\ddot{y}_1(t)$ and its estimate (lower near $t = 0$) for $0 < t < 10$, with $c = 3$.*

Theorem 4. *As $t \to \pm\infty$, $z_1(t), z_2(t), z_3(t) \in E^3$ are the* rows *of the 3×3 matrix*

$$\left[\begin{array}{ccc} \frac{\sqrt{c}}{f}\alpha_\pm - \frac{t}{f}\hat{\alpha}'_\pm \exp(\mp g\hat{\alpha}'_\pm)\beta_\pm & \frac{t}{f}\alpha_\pm & -\exp(\mp g\hat{\alpha}'_\pm)\beta_\pm \end{array} \right] + O\left(\frac{1}{f}\right).$$

Corollary 8. *Write $\alpha_+ = \alpha = (\alpha_1, \alpha_2, \alpha_3)$. Then for $i = 1, 3$, the z_i are asymptotic to the limiting circles*

$$\{w \in E^3 : \|w\| = 1 \text{ and } w_2 = \pm\alpha_i\}$$

according as $t \to \pm\infty$, and z_2 is asymptotic to the single circle

$$\{w \in E^3 : \|w\| = 1 \text{ and } w_2 = \alpha_2\}$$

whether $t \to \pm\infty$.

In Example 2 the circles apparent in Figures 2 and 3 (especially) are not orthogonal to $(0, 1, 0)$, nor even parallel to each other, and the circles in Figure 2 appear distinct. So these cannot be the limiting circles of Corollary 8. Indeed the numerically generated $z_i : [-10, 10] \to E^3$ do not reveal the long-term asymptotics, and plots of $z_i : [-100, 100] \to E^3$ confirm Corollary 8.

As we see next, the ODE (1) is fundamental for studying

- motion of a ball under torque that is constant in body coordinates,
- rolling of a ball on an inclined plane,
- variational motion planning for a ball,

where $(z_1(t), z_2(t), z_3(t))$ is a positively oriented orthonormal frame in E^3, fixed relative to the moving ball (the configuration of the ball relative to the center of mass $c(t) \in E^3$ at time t).

7. Applications to spherically symmetric rigid bodies. Consider a rigid body $\mathbf{B}$ of mass $\mu > 0$ subject to external forces totalling $f(t)$ with total moment $n(t)$ measured in an absolute (inertial) frame. A point $Q \in \mathbf{B}$ in body coordinates with origin at the center of mass $\bar{q}(t)$ has absolute position

$$q(t) = \bar{q}(t) + x(t)Q, \quad \text{where } x(t) \in SO(3).$$

Define the *absolute angular velocity* $\omega(t) \in E^3$ by $\hat{\omega}(t) := \dot{x}(t)x(t)^{-1}$. The *body angular velocity* is $\Omega(t) := x(t)^{-1}\omega(t)$, namely $\hat{\Omega}(t) = x(t)^{-1}\dot{x}(t)$. The *body angular momentum* is $M(t) := K\Omega(t)$, where K is the inertia matrix. The *absolute angular momentum* is $m(t) := x(t)M(t)$. From Newton's second law,

$$\mu\ddot{\bar{q}}(t) = f(t), \tag{25}$$

$$\dot{m}(t) = n(t). \tag{26}$$

Substituting for $m(t)$ in (26), $\dot{x}(t)M(t)+x(t)K\dot{\Omega}(t) = n(t)$, namely $x(t)^{-1}\dot{x}(t)M(t)+K\dot{\Omega}(t) = N(t)$, where $x(t)N(t) = n(t)$. Equivalently,

$$\mu\ddot{\bar{q}}(t) = x(t)^{-1}F(t), \tag{27}$$

$$K\dot{\Omega}(t) = (K\Omega(t)) \times \Omega(t) + N(t), \tag{28}$$

where $x(t)F(t) = f(t)$.

Suppose that the mass distribution of $\mathbf{B}$ is spherically symmetric. Then (28) reads

$$\nu\dot{\Omega}(t) = N(t), \tag{29}$$

where $\nu > 0$. For a homogeneous ball $\mathbf{B}$ of radius $b > 0$ and density ρ,

$$\mu = \frac{4\pi\rho}{3}b^3 \quad \text{and} \quad \nu = \frac{8\pi\rho}{15}b^5.$$

We review some special cases.

7.1. Turning a ball in space. Let $\mathbf{B}$ be a homogeneous ball subject only to a constant torque N (turning) in body coordinates. (A ball in space might be controlled by momentum wheels with linearly increasing angular momenta.) By (29) for any given $t_0 \in \mathbb{R}$,

$$\Omega(t) = \Omega(t_0) + \frac{N}{\nu}(t - t_0)$$

so that $\Omega : \mathbb{R} \to E^3$ is affine, $x : \mathbb{R} \to SO(3)$ is given by

$$\dot{x}(t) = x(t)\hat{\Omega}(t), \tag{30}$$

and $\bar{q} : \mathbb{R} \to E^3$ is found from (27). Since $x(t)x(t)^{\mathbf{T}}$ is 3×3 the identity matrix, we obtain on differentiation

$$\dot{x}(t)x(t)^{\mathbf{T}} + x(t)\dot{x}(t)^{\mathbf{T}} = \mathbf{0} \iff x(t)^{-1}\dot{x}(t) = -\dot{z}(t)z(t)^{-1},$$

where $z(t) := x(t)^{\mathbf{T}} = x(t)^{-1}$. Thus (30) is equivalent to (1) with

$$B_0 = -\hat{\Omega}(0) = -\hat{\Omega}(t_0) + \frac{\hat{N}}{\nu}t_0 \quad \text{and} \quad B_1 = -\dot{\hat{\Omega}}(0) = -\frac{\hat{N}}{\nu}.$$

Suppose $N \neq \mathbf{0}$. As in section 1, translate and rescale t, and choose a reference frame for body coordinates, so that z satisfies the canonical form (3) of (1). Then

$$\Omega(t) = -\begin{bmatrix} \sqrt{c} & t & 0 \end{bmatrix}^{\mathbf{T}},$$

where $c \geq 0$ depends on i. Choose the inertial reference frame so that $z(0)$ is the identity matrix $\mathbf{1}$. Then $z_i(t)$ is the ith vector of the inertial frame measured in body coordinates at time t. Replacing $x(t)$ in section 3 by the present $x(t)^{-1}$, $V(t)$ becomes the negative of the absolute angular momentum $\omega(t)$ of $\mathbf{B}$. By Lemma 5, x^{-1} is a null Riemannian cubic in $SO(3)$ with associated (canonical) null Lie quadratic $-\omega$ in $E^3 \cong so(3)$. By Lemma 6, the components ω_i of ω satisfy the ODE (4) with

$$\omega(0) = -\begin{bmatrix} \sqrt{c} \\ 0 \\ 0 \end{bmatrix}, \quad \dot{\omega}(0) = -\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \ddot{\omega}(0) = \begin{bmatrix} 0 \\ 0 \\ \sqrt{c} \end{bmatrix}.$$

Remark 1. *These conclusions hold even when $c = 0$, although previously $c > 0$ throughout. For $c = 0$,*

$$z(t) = \begin{bmatrix} \cos(\frac{t^2}{2}) & 0 & \sin(\frac{t^2}{2}) \\ 0 & 1 & 0 \\ -\sin(\frac{t^2}{2}) & 0 & \cos(\frac{t^2}{2}) \end{bmatrix}.$$

Thus $\omega(t) = \begin{bmatrix} 0 & -t & 0 \end{bmatrix}^{\mathbf{T}}$ and, as viewed from $\mathbf{B}$, the inertial frame rotates with linearly increasing angular velocity in the plane orthogonal to the second axis. So, in body coordinates, the vectors z_i of the inertial frame move on great circles.

From now on suppose $c > 0$. Then we have the following:

- There are unit vectors $\alpha_\pm, \beta_\pm \in S^2$ with $\langle \alpha_\pm, \beta_\pm \rangle = 0$ such that (Theorem 4)

$$x(t) = \begin{bmatrix} \frac{\sqrt{c}}{f}\alpha_\pm - \frac{t}{f}\hat{\alpha}'_\pm \exp(\mp g\hat{\alpha}'_\pm)\beta_\pm & \frac{t}{f}\alpha_\pm & -\exp(\mp g\hat{\alpha}'_\pm)\beta_\pm \end{bmatrix} + O(t^{-1})$$

  as $t \to \pm\infty$. Here $f(t) := \sqrt{t^2 + c}$ and $g(t)$ is given in Definition 3 of section 5.
- The vectors $z_i(t)$ of the inertial frame viewed from $\mathbf{B}$ are asymptotic to (usually not great) circles as $t \to \pm\infty$ (Corollary 8).
- Because $\omega(t) = -V(t)$, the components ω_i of absolute angular momentum satisfy
  - if any $|\omega_i(t)| > \sqrt{c}$, then $\dot{\omega}_i(t) \neq 0$ (Lemma 2);
  - we have $\omega_1(t)$ an even function of t, with $\dot{\omega}_1(t) < 0$ for $t > 0$ (Lemma 3);
  - we have $\omega_2(t)$ an odd function of t;
  - we have $\omega_3(t)$ an even function of t, with $0 < \omega_3(t) < |t|$ for $t \neq 0$ (Corollary 6).

- Formulae (17), (18), (19) give

$$\begin{aligned}\omega(t) &= -f\alpha_\pm + \frac{\sqrt{c}}{f^2}\exp(-g\hat{\alpha}'_\pm)\beta_\pm \qquad + O(t^{-3}),\\ \dot{\omega}(t) &= -\frac{t}{f}\alpha_\pm - \frac{\sqrt{c}}{f}\hat{\alpha}'_\pm\exp(-g\hat{\alpha}'_\pm)\beta_\pm \quad + O(t^{-2}),\\ \ddot{\omega}(t) &= -\sqrt{c}\exp(-g\hat{\alpha}'_\pm)\beta_\pm \qquad + O(t^{-1})\end{aligned}$$

as $t \to \pm\infty$.

7.2. Ball rolling on a tilted plane. Consider next a nonholonomic circumstance in classical mechanics [2]: a ball $\mathbf{B}$ of radius $b > 0$ is subject to a constant force $\mu\gamma \in E^3$ and rolls without slipping on a fixed plane parallel to

$$w^\perp := \{v \in E^3 : \langle v, w\rangle = 0\},$$

where $w \in S^2 \subset E^3$ (measuring tilting). Rolling means $\dot{\bar{q}}(t) - b\omega(t) \times w = \mathbf{0}$. Taking moments about $\bar{q}(t)$,

$$\nu\dot{\omega}(t) = -bw \times \phi(t),$$

where the frictional force $\phi(t)$ at $p(t) := \bar{q}(t) - bw$ is

$$\mu\ddot{\bar{q}}(t) - \mu\gamma + \mu\langle\gamma, w\rangle w = \mu b\dot{\omega}(t) \times w - \mu\gamma + \mu\langle\gamma, w\rangle w.$$

Substituting for $\phi(t)$ and expanding cross-products,

$$(\nu + \mu b^2)\dot{\omega} = \mu b^2\langle\dot{\omega}, w\rangle w + \mu b w \times \gamma.$$

Taking inner products with w, $\langle\dot{\omega}, w\rangle = 0$ since $\nu > 0$. Then

$$\dot{\omega} = \frac{\mu b}{\nu + \mu b^2} w \times \gamma = -\frac{\mu b}{\nu + \mu b^2}\hat{\gamma}(w).$$

Given $t_0 \in \mathbb{R}$, we have

$$\omega(t) = \omega(t_0) - \frac{\mu b}{\nu + \mu b^2}\hat{\gamma}(w)(t - t_0)$$

so that $\omega : \mathbb{R} \to E^3$ is affine, $x : \mathbb{R} \to SO(3)$ is given by

$$\dot{x}(t) = \hat{\omega}(t)x(t), \tag{31}$$

and $\bar{q} : \mathbb{R} \to E^3$ is found from the rolling condition. Then (31) is equivalent to (1) with $z(t) := x(t)$,

$$B_0 := \hat{\omega}(t_0) + \frac{\mu b}{\nu + \mu b^2} t_0\,\widehat{(\gamma \times w)} \quad \text{and} \quad B_1 := -\frac{\mu b}{\nu + \mu b^2}\,\widehat{(\gamma \times w)}.$$

Suppose $B_1 \neq \mathbf{0}$. As in section 1, translate and rescale t, and choose a reference frame for inertial coordinates so that $z(t)$ satisfies the canonical form (3) of (1). Choose a reference frame for body coordinates so that $z(0) = \mathbf{1}$, and exclude the easy case where $B_0 = \mathbf{0}$.

Then $V(t) := \Omega(t)$ is a canonical null Lie quadratic in $E^3 \cong so(3)$, associated with the null Riemannian cubic $z = x$ in $SO(3)$. Thus the following hold:

- There are unit vectors $\alpha_\pm, \beta_\pm \in S^2$ with $\langle \alpha_\pm, \beta_\pm \rangle = 0$ such that (Theorem 4)

$$x(t) = \begin{bmatrix} \frac{\sqrt{c}}{f}\alpha_\pm - \frac{t}{f}\hat{\alpha}'_\pm \exp(\mp g\hat{\alpha}'_\pm)\beta_\pm & \frac{t}{f}\alpha_\pm & -\exp(\mp g\hat{\alpha}'_\pm)\beta_\pm \end{bmatrix}^{\mathbf{T}} + O(t^{-1})$$

  as $t \to \pm\infty$.
- The vectors $x_i(t)$ of the body reference frame are asymptotic to circles, as in Corollary 8.
- The components $\Omega_i(t)$ of body angular acceleration satisfy the results found in section 2 for the $y_i(t)$.
- Equations (17), (18), (19) give

$$\begin{aligned} \Omega(t) &= f\alpha_\pm - \frac{\sqrt{c}}{f^2}\exp(-g\hat{\alpha}'_\pm)\beta_\pm &&+ O(t^{-3}), \\ \dot{\Omega}(t) &= \frac{t}{f}\alpha_\pm + \frac{\sqrt{c}}{f}\hat{\alpha}'_\pm \exp(-g\hat{\alpha}'_\pm)\beta_\pm &&+ O(t^{-2}), \\ \ddot{\Omega}(t) &= \sqrt{c}\exp(-g\hat{\alpha}'_\pm)\beta_\pm &&+ O(t^{-1}) \end{aligned}$$

  as $t \to \pm\infty$.

7.3. Variational motion planning. For a Riemannian manifold M, given $t_0 < t_1 \in \mathbb{R}$ and $x_0, x_1 \in M$ with $v_0 \in TM_{x_0}$ and $v_1 \in TM_{x_1}$, let $X = X(t_0, t_1, x_0, x_1, v_0, v_1)$ be the space of C^1 curves $x : [t_0, t_1] \to M$ satisfying

$$x(t_0) = x_0, \quad \dot{x}(t_0) = v_0, \quad x(t_1) = x_1, \quad \dot{x}(t_1) = v_1. \tag{32}$$

For $x \in X$ set

$$J_2(x) := \int_{t_0}^{t_1} \|\nabla_{d/dt}\dot{x}(t)\|^2 dt,$$

where $\| \ \|$ denotes the Riemannian norm and ∇ is the Levi–Civita covariant derivative defined by the Riemannian metric. Then (see [8], [11]) a critical point x of J_2 is a *Riemannian cubic* in the sense that

$$\nabla^3_{d/dt}\dot{x} + R(\nabla_{d/dt}\dot{x}, \dot{x})\dot{x} = \mathbf{0}, \tag{33}$$

where R denotes Riemannian curvature. The derivation of (33) can be described much more simply than (say) in [11], as follows.

Consider smooth variations $x_{(h)} \in X$ of x with $h \in \mathbb{R}$ and $x_{(0)} = x$. Regarding $J_2(x_{(h)})$ as a function of h, the derivative at $h = 0$ is

$$2\int_{t_0}^{t_1} \langle \nabla_{d/dh}\nabla_{d/t}\dot{x}_{(h)}, \nabla_{d/dt}\dot{x}_{(h)} \rangle dt,$$

where, using the definition of Riemannian curvature, the integrand may be written as

$$\langle \nabla_{d/dt}\nabla_{d/dh}\dot{x}(h), \nabla_{d/dt}\dot{x}_{(h)} \rangle + \langle R(W, \dot{x})\dot{x}, \nabla_{d/dt}\dot{x} \rangle$$

with $W := W(x(t))$ the partial derivative with respect to h of $x_{(h)}$ evaluated at $h = 0$. Because ∇ is torsion-free we can write instead

$$\langle \nabla^2_{d/dt} W, \nabla_{d/dt}\dot{x}\rangle + \langle R(W,\dot{x})\dot{x}, \nabla_{d/dt}\dot{x}\rangle = \langle \nabla^2_{d/dt} W, \nabla_{d/dt}\dot{x}\rangle + \langle R(\nabla_{d/dt}\dot{x},\dot{x})\dot{x}, W\rangle$$

using standard symmetries of R [10, Lemma 9.3] for the Levi–Civita covariant derivative. Integrating twice by parts and taking account of boundary conditions,

$$\int_{t_0}^{t_1} \langle \nabla^3_{d/dt}\dot{x} + R(\nabla_{d/dt}\dot{x},\dot{x})\dot{x}, W\rangle dt = 0$$

for all W arising from variations $x_{(h)}$, and (33) follows.

In the special case where M is a Lie group G with bi-invariant Riemannian metric, it was first shown in [11] that the present definitions of Riemannian cubic and Lie quadratic agree with those of section 3. A more readable account is given in section 2 of [12].

Taking $G = SO(3)$, with x describing the trajectory of a spherically symmetric rigid body $\mathbf{B}$, the integrand in J_2 is proportional to the squared norm of the applied torque N in body coordinates. With few exceptions, non-null Lie quadratics in $so(3) \cong E^3$ and Riemannian cubics in $SO(3)$ are more complicated than the null objects [13], but there are significant remaining questions for the null case.

The present paper provides new asymptotic formulae (17), (18), (19) for canonical null Lie quadratics V in E^3. Moreover, for canonical null Riemannian cubics x in $SO(3)$, Theorem 4 gives

$$x(t) = \begin{bmatrix} \frac{\sqrt{c}}{f}\alpha_\pm - \frac{t}{f}\hat{\alpha}'_\pm \exp(\mp g\hat{\alpha}'_\pm)\beta_\pm & \frac{t}{f}\alpha_\pm & -\exp(\mp g\hat{\alpha}'_\pm)\beta_\pm \end{bmatrix}^{\mathbf{T}} + O(t^{-1})$$

as $t \to \pm\infty$, improving on previous results by an order of magnitude.

More background on dynamics of Riemannian cubics and integrability can be found in [12], [13], [14], [15]. For related research (including alternative methods for motion planning) see [3], [4], [5], [6], [7], [9], [16], [17], [18], [19]. Current work with Marin Kobilarov and Jerrold Marsden, building on the approach in [1], includes a practical implementation of Riemannian cubics for engineering.

8. Conclusion. The present paper builds on the theory of a class of curves $t \mapsto V(t)$ (canonical null Lie quadratics) in Euclidean 3-space E^3, to study a class of *first order linear* ODEs (1) in E^3 with coefficients *affine* in t. As shown in section 7, solutions of (1) have a simple interpretation in elementary problems in classical mechanics, and connections with approximation theory in Riemannian geometry (null Riemannian cubics). Known geometric properties of *nonlinear* ODEs are strengthened and used to prove asymptotic results for the *linear* system (1), as follows:

- We focus on a canonical form of (1), where the right-hand side depends on a single scalar parameter $c > 0$. (The case where $c = 0$ is standard.)
- (1) is shown to be solvable in terms of the general solution of a *third order linear* ODE (4) with *quadratic* coefficients. Local properties of solutions of (4) are found from a *second order quadratic* ODE (5).
- The strategy is to find global properties via a relationship (Lemma 6) between solutions of (4) and a particular solution V (a canonical null Lie quadratic) of a *second order quadratic* ODE (6) in E^3. All solutions of (6) can be given in terms of the canonical null Lie quadratic.

- Quite a lot is already known about canonical null Lie quadratics in E^3. They have constant curvature, linear torsion, and *asymptotes*. The present paper strengthens these results, giving asymptotic formulae for V and the derivatives $\dot{V}, \ddot{V}$.
- These formulae lead in turn to asymptotic expressions for *null Riemannian cubics* (a class of curves in $SO(3)$ satisfying a second order variational principle, of interest for interpolation problems in engineering).

Acknowledgments. I am grateful to Professor Jerrold Marsden for some valuable and illuminating conversations, including advice to explore links between Lie quadratics and classical mechanics. I also thank Professor Tasso Kaper, and an anonymous reviewer, for careful and constructive readings, resulting in a greatly improved and more readable paper.

REFERENCES

[1] A. H. Barr, B. Currin, S. Gabriel, and J. F. Hughes, *Smooth interpolation of orientations with angular velocity constraints using quaternions*, Comput. Graphics, 26 (1992), pp. 313–320.
[2] A. M. Bloch, J. Baillieul, P. Crouch, and J. Marsden, *Nonholonomic Mechanics and Control*, Interdiscip. Appl. Math. 24, Springer, New York, 2003.
[3] S. Buss, *Accurate and efficient simulations of rigid body rotations*, J. Comput. Phys., 164 (2000), pp. 377–406.
[4] M. Camarinha, F. Silva Leite, and P. Crouch, *On the geometry of Riemannian cubic polynomials*, Differential Geom. Appl., 15 (2001), pp. 107–135.
[5] M. Camarinha, F. Silva Leite, and P. Crouch, *Splines of class C^k on non-Euclidean spaces*, IMA J. Math. Control Inform., 12 (1995), pp. 399–410.
[6] P. Crouch and F. Silva Leite, *The dynamic interpolation problem: On Riemannian manifolds, Lie groups, and symmetric spaces*, J. Dynam. Control Systems, 1 (1995), pp. 177–202.
[7] P. Crouch, G. Kun, and F. Silva Leite, *The De Castlejau algorithm on Lie groups and spheres*, J. Dynam. Control Systems, 5 (1999), pp. 397–429.
[8] S. Gabriel and J. Kajiya, *Spline interpolation in curved space*, in State of the Art in Image Synthesis, SIGGRAPH '85 Course Notes, ACM Press, New York, 1985, pp. 1–14.
[9] F. Silva Leite, M. Camarinha, and P. Crouch, *Elastic curves as solutions of Riemannian and sub-Riemannian control problems*, Math. Control Signals Systems, 13 (2000), pp. 140–155.
[10] J. Milnor, *Morse Theory*, Ann. of Math. Stud. 51, Princeton University Press, Princeton, NJ, 1963.
[11] L. Noakes, G. Heinzinger, and B. Paden, *Cubic splines on curved spaces*, IMA J. Math. Control Inform., 6 (1989), pp. 465–473.
[12] L. Noakes, *Null cubics and Lie quadratics*, J. Math. Phys., 44 (2003), pp. 1436–1448.
[13] L. Noakes, *Non-null Lie quadratics in E^3*, J. Math. Phys., 45 (2004), pp. 4334–4351.
[14] L. Noakes, *Duality and Riemannian cubics*, Adv. Comput. Math., 25 (2006), pp. 195–209.
[15] L. Noakes, *Lax constraints in semisimple Lie groups*, Q. J. Math., 57 (2006), pp. 527–538.
[16] M. Zefran, V. Kumar, and C. Croke, *Choice of Riemannian metrics for rigid body dynamics*, in Proceedings of the ASME Design Engineering Technical Conference and Computers in Engineering Conference, Irvine, CA, 1996, pp. 1–11.
[17] M. Zefran and V. Kumar, *Planning of smooth motions on $SE(3)$*, in Proceedings of the IEEE International Conference on Robotics and Automation, Minneapolis, MN, 1996.
[18] M. Zefran and V. Kumar, *Two methods for interpolating rigid body motions*, in Proceedings of the IEEE International Conference on Robotics and Automation, Leuven, Belgium, 1996.
[19] M. Zefran and V. Kumar, *Interpolation schemes for rigid body motions*, Comput. Aided Design, 30 (1998), pp. 179–189.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 461–490
© 2008 Society for Industrial and Applied Mathematics

Normal Vectors on Critical Manifolds for Robust Design of Transient Processes in the Presence of Fast Disturbances*

Johannes Gerhard†, Wolfgang Marquardt†, and Martin Mönnigmann‡

Abstract. Information on steady-state bifurcations, most notably stability boundaries, is frequently used for the analysis and design of nonlinear systems. The bifurcation points separate regions with different dynamic behavior and thus give valuable information about nonlinear systems. They cannot, however, reflect the impact of fast disturbances on the transient behavior of nonlinear systems. The influence of fast disturbances can be addressed by bifurcation points that are defined as critical points during the transient behavior of a dynamic system in the presence of fast disturbances. Specifically, we consider two types of points—grazing points and end-points. At a grazing point the trajectory of a nonlinear system tangentially touches a hypersurface spanned by a state or output constraint. At an end-point the trajectory crosses the hypersurface at a specified final time. These critical points unfold to manifolds in the parameter space of the nonlinear system separating parts of the parameter space that admit trajectories that do not violate the constraint from those where the constraint is violated. The parametric distance between a candidate design of a nonlinear system and the critical manifold is used as a robustness measure. As the closest connection between the design and the critical manifold is along the normal direction of the critical manifold, normal vectors are used to formulate minimal-distance constraints for a nonlinear program. Thus it is possible to robustly take into account state and output constraints in the presence of fast disturbances for the design of a nonlinear system. Application of the approach to closed-loop systems allows for an integration of operating point and control design. Several case studies from chemical engineering are presented to illustrate the proposed method.

Key words. grazing bifurcation, end-point constraint, normal vector, robust optimization, disturbances

AMS subject classifications. 37N99, 90C30, 93C95

DOI. 10.1137/070698981

1. Introduction.

We consider systems of nonlinear differential algebraic equations (DAEs) of the form

$$\begin{aligned} x_t &= f(x(t), y(t), p, d(\alpha, t), t), \quad x(t_0) = x_0, \\ 0 &= g(x(t), y(t), p, d(\alpha, t), t), \end{aligned} \tag{1.1}$$

with dynamic state variables $x \in \mathbb{R}^{n_x}$, corresponding time derivatives x_t, initial conditions x_0, algebraic variables $y \in \mathbb{R}^{n_y}$, system parameters $p \in \mathbb{R}^{n_p}$, disturbances $d \in \mathbb{R}^{n_d}$ parameterized by a set of parameters $\alpha \in \mathbb{R}^{n_\alpha}$, and time $t \in \mathbb{R}$. The functions f and g are assumed to be

*Received by the editors August 1, 2007; accepted for publication (in revised form) by F. Doyle December 5, 2007; published electronically May 2, 2008. This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) under grant MA1188/22-1.

http://www.siam.org/journals/siads/7-2/69898.html

†Lehrstuhl für Prozesstechnik, RWTH Aachen University, Turmstraße 46, 52064 Aachen, Germany (johannes.gerhard@avt.rwth-aachen.de, wolfgang.marquardt@avt.rwth-aachen.de).

‡Institut für Wärme- und Brennstofftechnik, Technische Universität Braunschweig, Postfach 3329, 38023 Braunschweig, Germany (m.moennigmann@tu-bs.de).

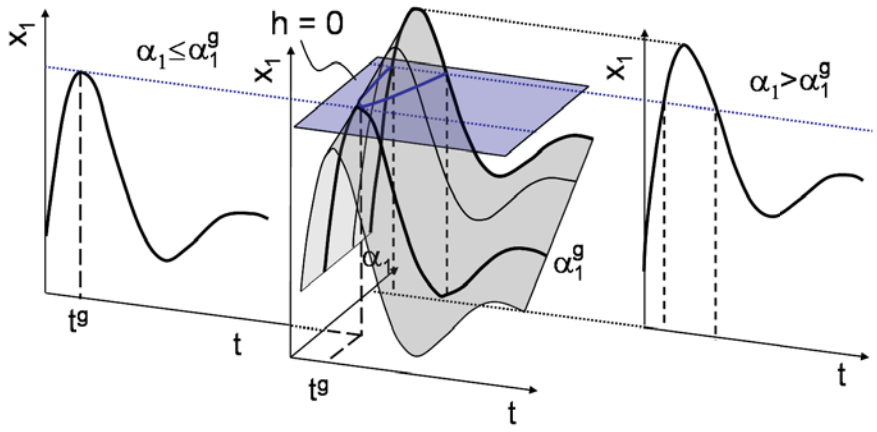


Figure 1. *Trajectories of state x_1 for different values of a step disturbance triggered at $t = 0$ and parameterized by α_1. A parabola-like curve connects the points where the trajectories cross the hypersurface spanned by constraint $h := x_{1\max} - x_1 = 0$. The grazing trajectory touches the hypersurface tangentially at $t = t^g$ for $\alpha = \alpha_1^g$.*

sufficiently smooth with respect to x, y, p, α, and t for $t > t_0$. f and g map from some open subset $U \subset \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_p} \times \mathbb{R}^{n_\alpha} \times \mathbb{R}$ into $\mathbb{R}^{n_x}$ and $\mathbb{R}^{n_y}$, respectively. We assume the Jacobian g_y of g with respect to y to have full rank n_y; i.e., the DAE system (1.1) has a differential index of at most one [50]. This implies that consistent initial conditions of the algebraic states y_0 can be computed with the algebraic equations and specified initial conditions of the dynamic states x_0. Note that (1.1) can represent both open-loop and closed-loop systems including a controller with fixed structure. For closed-loop systems, the parameter vector p refers to both system design parameters and control parameters.

In what follows we assume that there is a set of inequality constraints,

$$0 \leq h(x(t), y(t), p, d(\alpha, t), t), \tag{1.2}$$

with h sufficiently smooth mapping from $U \subset \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_p} \times \mathbb{R}^{n_\alpha} \times \mathbb{R}$ into $\mathbb{R}^{n_h}$. Constraints (1.2) may represent safety constraints for a physical system, such as an upper temperature limit in a chemical reactor, or constraints that are established due to economic reasons, such as a quality constraint on a product. For safe and economical operation of system (1.1) it is therefore important to ensure that bounds (1.2) hold despite disturbances $d(\alpha, t)$. Figure 1 displays an illustrative example of the transient behavior of state x_1 of a dynamic system after a step disturbance triggered at $t = 0$. The magnitude of the disturbance is parameterized by the parameter α_1. A simple upper bound for x_1, $h := x_{1\max} - x_1$ defines a plane in the space (α_1, x_1) which must not be crossed from below by the trajectories after the disturbance. As shown in Figure 1, the trajectories always violate the bound for some time for $\alpha_1 > \alpha_1^g$, whereas the trajectories do not cross the bounding plane for $\alpha_1 < \alpha_1^g$. Obviously, there is a critical value of $\alpha_1 = \alpha_1^g$ for which the trajectory only touches the bounding plane tangentially but does not cross it.

The point where the transient touches the bounding plane tangentially is a so-called grazing point [42]. The grazing point is of special interest as it separates those trajectories

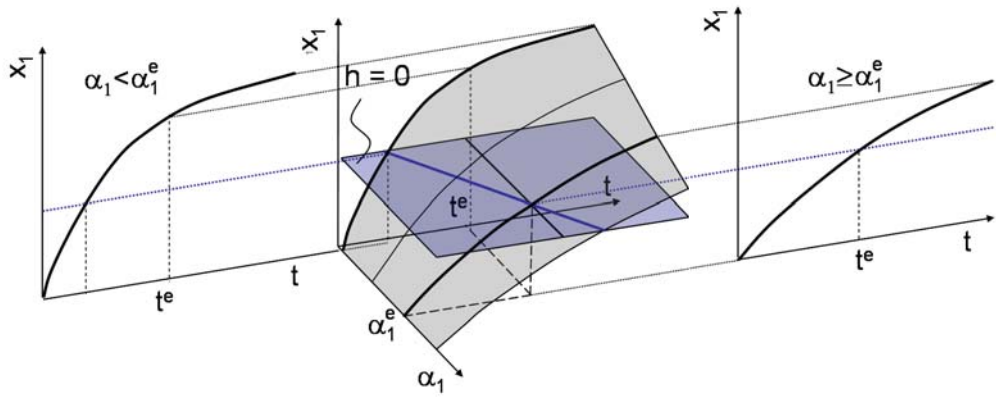


Figure 2. *End-point constraint: Trajectories of state x_1 for different values of a step disturbance triggered at $t = 0$ and parameterized by α_1. Curve on the plane $0 = h$ connects points of trajectories at which the hypersurface of the end-point constraint $h := x_{1\max} - x_1 = 0$ is crossed. The end-point constraint trajectory crosses the bound at the specified final time $t = t^e$ for $\alpha = \alpha_1^e$.*

where the disturbance does not cause a violation of the bound from those trajectories that cross the bounding plane at some point. Note that the term *grazing bifurcation* usually refers to nonsmooth dynamic systems where a bounding surface triggers a discrete event, such as a periodically forced oscillator hitting a wall [42]. Grazing bifurcations have been actively investigated in recent years for nonsmooth mechanical systems such as impact oscillators [8, 11, 42], stick slip oscillations in systems with friction [10], and rub-impact rotor systems [9], but also for switching power systems [12, 15, 16]. In this paper, we will focus on smooth systems with constraints (1.2) that must not be crossed, such as safety boundaries or product quality constraints.

Besides the grazing point we will also consider end-point constraints, referring to the trajectory which crosses a bound (1.2) at a specified final time $t = t^e$ as shown in Figure 2. This point separates those trajectories which do not cross the bound before the specified final time from those which violate the bound before the final time is reached. End-point constraints are useful, for example, to bound monotonously increasing states or outputs until a specified final time is reached. An example of a monotonously increasing output is the integrated squared error in closed-loop systems.

In this work the grazing point and the end-point constraint will be used for the robust design and optimization of dynamic systems (1.1) in the presence of parametric uncertainty according to the optimization-based approach recently presented by Mönnigmann and Marquardt [36, 37, 38, 39]. This approach uses the parametric distance between the nearest critical point and the border of a compact region of parametric uncertainty as a measure of robustness for a candidate operating point. The nearest critical point, here a grazing point or an end-point, occurs in the direction normal to the manifold of critical points, which we will refer to as a *critical manifold*. In general, the concept of critical manifolds applies to all boundaries in the parameter space at which the system behavior changes qualitatively. A parametric distance greater than zero ensures that all values the parameters may attain within the region of uncertainty are located on the side of the manifold where the system

exhibits the desired properties.

A drawback of the approach so far was the restriction to steady states and critical points of steady states. Steady states allow neither considering bounds that must hold during transient behavior of the dynamic system (1.1) nor including disturbances with dynamics that are as fast or faster than the system dynamics [37]. To overcome these limitations, we extend the existing approach to consider bounds on trajectories of nonlinear dynamic systems and the corresponding critical points. This involves the calculation of the normal direction for manifolds of grazing points and end-point constraints. The normal direction is used for the formulation of additional constraints in a nonlinear program to ensure that, at the optimal solution, no points within the region of uncertainty violate the constraints (1.2).

Independently from the work presented in this paper, the normal direction on the manifold of grazing bifurcations has also been used in a recent publication [16] to find the closest grazing bifurcation from an operating point for nonsmooth systems, specifically for switching power systems. The most important conceptual difference is that in [16] normal vectors are used for the *analysis of a fixed operating point* while the scope of this work is to use normal vectors for the *robust design of an operating point*. The point of operation is not fixed but is allowed to vary in the space of the design parameters p to minimize an objective function within an optimization problem. The normal vector direction is therefore used not only to locate the closest critical point to a fixed operating point but also to track the closest point, which will change its location with the operating point being modified by the optimizer to obtain the economically optimal point. Technical differences between the calculations presented in this paper and the derivations in [16] are briefly discussed in section 3 and in Appendix B.

As (1.1) may represent both open- and closed-loop dynamic systems, the sketched approach can be used for the simultaneous operating point and control system design of nonlinear systems to guarantee that constraints (1.2) hold despite fast disturbances. Integrated design and control is attractive as it allows incorporating operability issues directly in the system design instead of the usual consecutive approach of separated system and control design. A family of approaches addressing simultaneous design and control with output constraints is based on the robust design framework of Halemane and Grossmann [24] for steady-state design. An extension to dynamic systems to consider robustness of constraints despite disturbances was formulated by Mohideen, Perkins, and Pistikopoulos [34] using mixed-integer dynamic optimization (MIDO) programs. In this approach, first a weighted multiperiod MIDO design problem is solved for a set of fixed uncertainty realizations. A feasibility test checks for further worst-case points which are added to the design problem. Several solution strategies for the involved MIDO problems have been derived in a series of follow-up papers summarized in [46]. A similar approach is used in [2] to solve the integrated control and design problem. In all these approaches the dynamic optimization problem is solved either by complete discretization of the dynamic system by collocation, leading to a large NLP, or by bounds (1.2) being transformed to end-point constraints [46], which may cause numerical difficulties [19]. In our approach we consider special points of critical trajectories. A discretization or conversion into end-point constraints is therefore not required. Another advantage is the seamless integration of rigorous stability constraints by using the results for steady-state design with guaranteed stability derived in our previous work [36, 37]. Our approach does not rely on matrix measures [27, 35] to guarantee stability, which are known to be conservative.

Input and output constraints can also be handled by model predictive control (MPC) [33], where at each point in time a constrained optimal control problem on a finite time horizon is solved. Uncertainties and disturbances are addressed by min-max robust MPC approaches [6], where the performance is optimized for the worst possible case. To reduce the conservatism of min-max MPC approaches a family of control sequences instead of a single control input profile can be considered both for linear systems [29, 48, 53] as well as for nonlinear systems [32]. Evaluation of the control policies, though, is computationally very demanding apart from very simple problems. Linear or higher approximations of the worst-case performance can help to reduce the computational burden of nonlinear min-max robust MPC [13, 41]. The introduced approximation error might, however, lead to poor results for strongly nonlinear models and larger uncertainties [14]. Alternatively to min-max approaches robustness in MPC is addressed by methods that adapt online a nominal optimal profile in the presence of uncertainties. Examples are the use of neighboring extremals [22] or tracking of the necessary conditions of optimality [26]. There are, however, only a few approaches that use MPC or even robust MPC approaches for simultaneous system and control design. One of the early works using MPC for the integration of design and control is [5]. The complexity of the design task to find an optimal operating point is greatly increased since MPC requires the solution of an additional and computationally intensive online dynamic optimization problem. MPC for simple linear models is an exception. In this special case, the resulting optimal control laws can be written in an explicit form. In [45] control laws are evaluated offline via parametric programming. This approach, however, is not suitable for problems with many states and constraints as the number of linear control laws grows exponentially with the number of constraints. Unconstrained MPC for linear models is used in [7, 31]. This does not, however, make use of the ability of MPC to consider constraints explicitly. Additional conditions have to be introduced in [7, 31] to ensure that specified constraints hold.

Input and output constraints under uncertainty are also addressed by Lyapunov-based control design methods using invariant sets [4]. In [17, 18] an explicit Lyapunov-based control law is derived that guarantees stability and performance in the presence of input constraints and disturbances within a stability region around the operating point. A lower bound of the stability region can be estimated by computing an invariant subset of the stability region. There are, however, some drawbacks for Lyapunov-based approaches. Typically, certain assumptions, so-called matching conditions, have to hold for the uncertainties and disturbances. For the approach presented in this paper no matching conditions have to be introduced. Furthermore, we do not have to restrict nonlinear system (1.1) to be affine with respect to disturbances or control inputs. The estimation of the stability region around a nominal operating point is beyond the scope of this paper and will not be investigated.

The paper is organized as follows. In section 2 the defining equations for the grazing and end-points are presented including a short discussion on the types of disturbances which can be considered. In section 3 the corresponding normal directions of the critical manifolds are derived. In section 4 the optimization approach including the normal vector constraints is presented. Finally, illustrative case studies are shown in section 5. The notation used is summarized in Appendix A.

2. Critical manifolds due to bounds on trajectories. No matter how accurately the governing physical laws are taken into account in the mathematical modeling of a real physical system, there will always exist a mismatch between the behavior of a nominal model and the real system. Using only a nominal model for the design of an operating point will generally lead to a too optimistic design for which robustness of the design specifications cannot be guaranteed. We identify two sources for the prediction error of a nominal model:

(i) uncertainties in the model parameters that can only be determined up to a certain accuracy, and

(ii) input disturbances with dynamics that are at least as fast as the system dynamics or faster.

Uncertainty in the model parameters has been considered in previous work [36, 37, 38, 39] using critical points of steady states such as Hopf or saddle-node bifurcations. It has been shown in [37] that uncertain model parameters can also be used to model disturbances if the dynamics of the disturbances are much slower than the dynamics of the system. A slow disturbance occurs if its dominant time scale is much larger than that related to the slowest eigenvalue of the dynamic system.

Here, however, we consider fast disturbances leading to potentially hazardous transient behavior. In this case, transient behavior of the nonlinear systems cannot be neglected. In the framework of the robust design approach we are following in this work, the disturbances d have to be explicitly parameterized by parameters α. Variation of α in a compact set is introduced to bound disturbances d and to represent a rich family of disturbances. These uncertain parameters should not be confused with uncertain model parameters. In this work, parametric uncertainty always refers to the parameterization of the fast disturbances. More precisely, we assume that the parameters α are restricted to a smooth compact subset $\mathcal{A} \subset \mathbb{R}^{n_\alpha}$ and $\forall \alpha \in \mathcal{A}$, $\forall t > t_0 : d(\alpha, t) \in \mathcal{D}$, with compact subset $\mathcal{D} \subset \mathbb{R}^{n_d}$. Throughout the paper, $\mathcal{A}$ has the form

$$\mathcal{A} = \left\{ (\tilde{x}, p, \alpha) \in \tilde{U} : 0 = \hat{g}(\tilde{x}, p, \alpha),\ 0 \leq \tilde{g}(\tilde{x}, p, \alpha) \right\}, \tag{2.1}$$

with auxiliary variables $\tilde{x}$ and expressions $\hat{g}$ and $\tilde{g}$ mapping from some open subset $\tilde{U} \subset \mathbb{R}^{n_{\tilde{x}}} \times \mathbb{R}^{n_p} \times \mathbb{R}^{n_\alpha}$ into $\mathbb{R}^{n_{\tilde{x}}}$, $\mathbb{R}$, respectively. This definition includes, for example, the confidence ellipsoid for Gaussian random variables with expectation value $\overline{\alpha}$, covariance matrix Σ, and the desired confidence level $\gamma > 0$,

$$\mathcal{A} = \left\{ \alpha \in \tilde{U} : 0 \leq \gamma - (\alpha - \overline{\alpha})^T \Sigma^{-1} (\alpha - \overline{\alpha}) \right\}. \tag{2.2}$$

Frequently, parametric uncertainty is described by upper and lower bounds

$$\alpha \in [\overline{\alpha} - \Delta\alpha,\ \overline{\alpha} + \Delta\alpha], \tag{2.3}$$

where $\Delta\alpha \in \mathbb{R}^{n_\alpha}$. In our robust design approach the rectangular uncertainty region has to be approximated by a smooth hull, e.g., by an ellipse (2.2) with $\Sigma = \mathrm{diag}(\Delta\alpha^2)$ and $\gamma = n_\alpha$.

A simple example of a parameterized disturbance $d(\alpha, t)$ is a step disturbance triggered at t_0,

$$d(\alpha, t) = \begin{cases} 0, & t \leq t_0, \\ \alpha, & t > t_0, \end{cases} \tag{2.4}$$

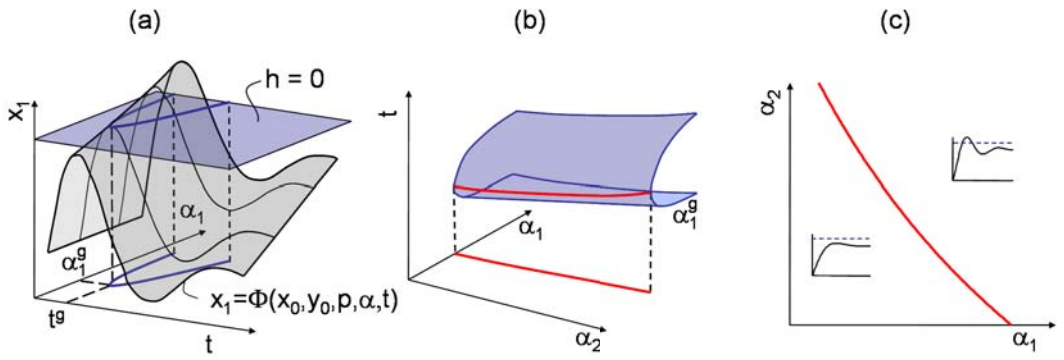


Figure 3. (a) *Trajectories of x_1 after step disturbance parameterized by α_1; a parabola-like curve connects those points where trajectories cross the constraint $h = 0$. The grazing point is located at the extremum of the curve.* (b) *The parabola unfolds into a surface of points in case of two parameters α_1 and α_2.* (c) *Projection of the manifold of grazing points separates the parameter plane (α_1, α_2). Trajectory sketches show on which side the bound h is not violated.*

with α parameterizing the magnitude of the disturbance. More general types of disturbances can be defined by a parameterized time-dependent function or by an additional set of parameterized differential equations. In the following we will replace disturbances $d(\alpha, t)$ by their parameterization α.

The flow of the nonlinear system (1.1) [44], including the dynamic and algebraic states, is denoted by Φ, i.e.,

$$\begin{aligned}[x(x_0, t_0, p, \alpha, t)^T, y(x_0, t_0, p, \alpha, t)^T]^T &= \Phi(x_0, t_0, p, \alpha, t), \\ [x_0^T, y_0^T]^T &= \Phi(x_0, t_0, p, \alpha, t_0).\end{aligned} \tag{2.5}$$

For nonlinear systems, the flow Φ is generally not available in an analytical form but has to be evaluated by numerical integration. In the following sections, we abbreviate the list of arguments of $x(x_0, t_0, p, \alpha, t)$ and $y(x_0, t_0, p, \alpha, t)$ for ease of notation and write $x(t)$, $y(t)$. Likewise, we assume for simplicity that there is only a single constraint (1.2), i.e., $n_h = 1$. This assumption will be dropped in section 4.

2.1. Critical manifold of grazing points. A manifold of grazing points is characterized by the set of trajectories that tangentially touch the hypersurface spanned by an active constraint $h = 0$. Before introducing the mathematical definitions for a grazing point, we want to show qualitatively the relevance of a grazing point for the transient behavior of a dynamic system (1.1) with constraints (1.2). Figure 3a shows several trajectories of the state x_1 after a step disturbance with its magnitude parameterized by α_1. The plane denoted by $h := x_{1\max} - x_1 = 0$ in Figure 3a is an upper bound on x_1 that must not be violated despite disturbances. The grazing trajectory for $\alpha_1 = \alpha_1^g$ touches the bound tangentially at the grazing time $t = t^g$. For values of $\alpha_1 > \alpha_1^g$ the trajectories cross the bounding plane and violate the upper bound for some time. The set of points where the trajectories cross the bound form a parabola-like curve with its extremum located at the grazing bifurcation t^g, α_1^g. For values of $\alpha_1 < \alpha_1^g$ trajectories of x_1 never cross the upper bound $x_{1\max}$. In the presence of a second disturbance parameter α_2 the parabola-like curve of the crossing points unfolds into a surface and the grazing point

into a curve in the (α_1, α_2, t)-space as depicted in Figure 3b. The projection of this curve onto the parameter plane (α_1, α_2) shown in Figure 3c separates the region where the constraint is not violated from the region where disturbances lead to trajectories that cross the bound. Figure 3c stresses the importance of the grazing bifurcation for the design of a dynamic system that has to be restricted to the region, where the constraint is not violated.

In what follows, we derive necessary conditions for a grazing bifurcation that can be used to describe boundaries of the type shown in Figure 3c. The necessary conditions for a grazing point can be derived by taking into account that the grazing trajectory

(i) must fulfill the constraint (1.2) at the grazing point

$$\begin{aligned} [x(t^g)^T, y(t^g)^T]^T &= \Phi(x_0, t_0, p, \alpha^g, t^g), \\ h(x(t^g), y(t^g), p, \alpha^g, t^g) &= 0, \end{aligned} \tag{2.6}$$

(ii) and must have a common tangent space with the hypersurface spanned by constraint (1.2) at the grazing point in the space $(x(t), y(t), t)$.

Let the tangent space of the hypersurface defined by the constraint (1.2) in the space $(x(t), y(t), t)$ be spanned by vectors $[v_d^T, v_a^T, \tilde{v}]^T$ with $v_d \in \mathbb{R}^{n_x}$, $v_a \in \mathbb{R}^{n_y}$, and $\tilde{v} \in \mathbb{R}$. The tangent space must be orthogonal to the normal space of the hypersurface. The normal space is spanned by the gradient of the constraint h with respect to the dynamic and algebraic state variables and time $[h_x, h_y, h_t]^T$ [20]. We can therefore write

$$\begin{aligned} [h_x, h_y, h_t] \begin{bmatrix} v_d \\ v_a \\ \tilde{v} \end{bmatrix} &= 0, \\ v_d^T v_d + v_a^T v_a + \tilde{v}^2 &\neq 0. \end{aligned} \tag{2.7}$$

A tangential vector of a trajectory is spanned by the derivative of the flow (2.5) with respect to time, $[\Phi_t(x_0, t_0, p, \alpha^g, t^g)^T, 1]^T$ with $\Phi_t = [x_t^T, y_t^T]^T \in \mathbb{R}^{n_x + n_y}$. The time derivative x_t is the right-hand side of (1.1) for the dynamic variables $x_t = f$. The time derivative y_t is implicitly defined by the time derivative of the algebraic equations:

$$0 = g_y y_t + g_x f + g_t. \tag{2.8}$$

Since g_y has full rank, by assumption the time derivative y_t can be calculated by solving the linear system (2.8) for y_t. Note, however, that y_t is usually provided by the numerical integrator solving the nonlinear dynamic system (1.1). At the grazing point $(x(t^g), y(t^g), t^g)$ the tangential vector of the corresponding trajectory must belong to the tangent space of the hypersurface defined by the active constraint (1.2); hence the vector $[f^T, y_t^T, 1]^T$ is a valid choice for $[v_d^T, v_a^T, \tilde{v}]^T$ such that (2.7) is satisfied. The augmented system $M^{(g)}$ for a grazing point can then be written as

$$M^{(g)} = \begin{pmatrix} h(x(t), y(t), p, \alpha, t) \\ h_x f(x(t), y(t), p, \alpha, t) + h_y y_t + h_t \end{pmatrix} = 0, \tag{2.9}$$

with the flow $\Phi = [x(t)^T, y(t)^T]^T$ as defined in (2.5). These two equations determine the time t^g and one parameter α_1^g at which the grazing point occurs. Note that the defining equations

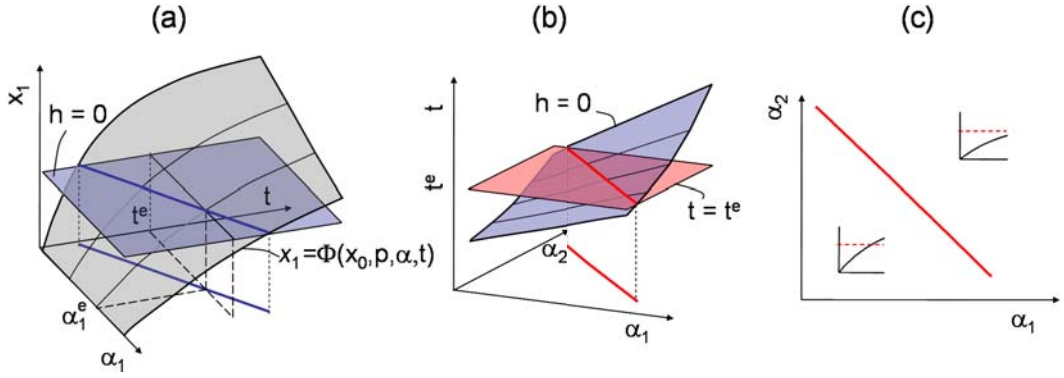


Figure 4. (a) *Trajectories for different values of disturbance parameter* α_1 *and curve connecting points where trajectories cross the boundary* $h = 0$ *with critical point at* t^e, α_1^e. (b) *Critical point unfolds to a curve of points which fulfill constraint* $h = 0$ *at* t^e *in a two-dimensional parameter space spanned by* α_1 *and* α_2. (c) *Projection of the curve splits the parameter plane* (α_1, α_2) *into a region where constraint* h *is violated by corresponding trajectories before* t^e *is reached and a region where the constraint is not crossed for* $t < t^e$.

(2.9) are also valid for those trajectories that have an inflection point at the hypersurface spanned by the active constraint (1.2). These trajectories cross the constraint directly after touching it tangentially. Such inflection points can be excluded by requiring the second order derivative of h with respect to time to be strictly positive,

$$\frac{\partial(h_x f + h_y y_t + h_t)}{\partial t} > 0,$$

as the grazing bifurcation always corresponds to a minimum of h.

In general, more than one critical manifold may exist for a constraint (1.2) if the trajectory touches or crosses the boundary several times.

2.2. Critical manifold for end-point constraints. A second type of critical manifold can be defined for a constraint (1.2), by specifying a final time t^e at which the constraint must be fulfilled exactly. The critical manifold for end-point constraints is defined by the set of parameter values for which the corresponding trajectories reach the constraint $h = 0$ at a specified final time t^e. Before stating the augmented system for the critical manifold we want to discuss the relevance of end-point constraints with the sketches in Figure 4. Several trajectories of a state x_1 after step disturbances triggered at $t = 0$ and parameterized by α_1 are shown in Figure 4a. All the trajectories are monotonically increasing and cross the upper bound $h := x_{1\max} - x_1 = 0$ at some point. For the critical parameter value α^e the upper bound is crossed exactly at the specified final time t^e. For parameter values $\alpha > \alpha^e$ the constraint is crossed for times $t > t^e$ whereas for $\alpha < \alpha^e$ the constraint is violated before t^e is reached. By taking into account a second disturbance parameter α_2, the critical point α^e, t^e unfolds into a curve in the space (t, α_1, α_2) as shown in Figure 4b. The projection of this curve on the parameter plane (α_1, α_2), as shown in Figure 4c, separates a region where the corresponding trajectories do not touch or cross the boundary until the specified time t^e is reached from a region where the constraint is always violated by the trajectories for $t < t^e$.

This type of critical manifold is particularly useful for bounds on monotonically increasing states or outputs for which grazing points cannot occur. End-point constraints can be formulated, e.g., to define an upper bound on the integrated squared error of an output with respect to its specified set-point for a closed-loop control system, as demonstrated in the case study in section 5.2.

The augmented system defining the manifold of end-point constraints is determined by the trajectory that crosses the bound (1.2) at the specified final time t^e. The augmented system is therefore given by

$$M^{(e)} = \begin{pmatrix} h(x(t), y(t), p, \alpha, t) \\ t - t^e \end{pmatrix} = 0. \tag{2.10}$$

These two equations determine the time t^e and one parameter α_1^e.

3. Derivation of the normal vector direction. The normal vector of critical manifolds can be calculated from the defining augmented systems (2.9) and (2.10) following the scheme of derivation developed by Mönnigmann and Marquardt [36]. We assume that initial conditions x_0 and design parameters p are fixed and known. Uncertain initial conditions for state x_i, $i \in \{1, \dots, n_x\}$, can easily be included by introducing a new state variable $\tilde{x}_i$ with initial conditions $\tilde{x}_{i0} = 0$ and time derivative given by $\tilde{x}_{it} = f_i(x, y, p, \alpha)$. Then $x_i = \tilde{x}_i + x_{i0}$, where x_{i0} can be treated as a parameter α. The normal space of the hypersurface defined by an augmented system $M^{(c)}$, $c \in \{g, e\}$, in the (t, α)-space is spanned by the columns of the Jacobian matrix of the partial derivatives $\nabla M^{(c)}$ [20]. In order to simplify the following derivations we collect the variables $x(t)$ and $y(t)$ in the state variable vector

$$z(t) = [x(t)^T, y(t)^T]^T, \quad z \in \mathbb{R}^{n_z}, \quad n_z = n_x + n_y,$$

and the functions f and g of (1.1) in

$$F = [f^T, g^T]^T.$$

Furthermore, we introduce the matrix $A \in \mathbb{R}^{n_z} \times \mathbb{R}^{n_z}$,

$$A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

with identity matrix $I \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_x}$. The dynamic system (1.1) can then be rewritten as

$$Az_t = F(z, p, \alpha, t). \tag{3.1}$$

The normal space of the augmented systems (2.9) and (2.10) is an $(n_\alpha + 1) \times (n_\alpha + 1)$ matrix

$$B = \begin{bmatrix} \nabla_t M_1^{(c)} & \nabla_t M_2^{(c)} \\ \nabla_\alpha M_1^{(c)} & \nabla_\alpha M_2^{(c)} \end{bmatrix}, \quad c \in \{g, e\}. \tag{3.2}$$

For the augmented system of the grazing point (2.9), the entries of B result in

$$\begin{aligned} \nabla_t M_1^{(g)} &= h_z z_t + h_t, \\ \nabla_t M_2^{(g)} &= h_{zz} z_t z_t + h_z z_{tt} + 2h_{zt} z_t + h_{tt}, \\ \nabla_\alpha M_1^{(g)} &= (h_z \Phi_\alpha + h_\alpha)^T, \\ \nabla_\alpha M_2^{(g)} &= (h_{zz} z_t \Phi_\alpha + h_{z\alpha} z_t + h_z \Phi_{t\alpha} + h_{tz} \Phi_\alpha + h_{t\alpha})^T. \end{aligned} \tag{3.3}$$

The entries of B for the end-point system (2.10) are

$$\begin{aligned} \nabla_t M_1^{(e)} &= h_z z_t + h_t, \\ \nabla_t M_2^{(e)} &= 1, \\ \nabla_\alpha M_1^{(e)} &= (h_z \Phi_\alpha + h_\alpha)^T, \\ \nabla_\alpha M_2^{(e)} &= 0. \end{aligned} \tag{3.4}$$

The gradients include sensitivities Φ_α of the flow with respect to the uncertain parameters α. The differential and algebraic equations that define the dynamics of the sensitivities are obtained by differentiating the DAE system (1.1) with respect to the parameters α:

$$A\Phi_{\alpha t} = F_z \Phi_\alpha + F_\alpha. \tag{3.5}$$

The sensitivity system (3.5) consists of one DAE system for each parameter α_i, $i = 1, \ldots, n_\alpha$. These systems are independent of each other but all depend on the solution $z(t)$ of the state system. The sensitivities can be calculated by integrating the sensitivity equations (3.5) together with the state system (3.1). A number of numerical integrators supports efficient evaluation of the sensitivity equations by exploiting their special properties (e.g., [47]).

The parametrically closest point of a critical manifold to a given point is in the direction of the particular vector in the normal space (3.2) that has no contribution along the variable t. According to [36] this vector $r \in \mathbb{R}^{n_\alpha}$ is obtained by choosing $\kappa \in \mathbb{R}^2$ such that

$$B\kappa = \begin{bmatrix} 0 \\ r \end{bmatrix} \in \mathbb{R}^{n_\alpha + 1}, \tag{3.6}$$

where $0 \in \mathbb{R}$. Together with the regularization condition $\kappa^T \zeta - 1 = 0$ with $\zeta \in \mathbb{R}^2$ not orthogonal to κ [36], the two entries of κ are defined by the equations

$$\begin{aligned} \begin{bmatrix} \nabla_t M_1^{(c)} & \nabla_t M_2^{(c)} \end{bmatrix} \kappa = 0, \quad c \in \{g, e\}, \\ \kappa^T \zeta - 1 = 0. \end{aligned} \tag{3.7}$$

For the grazing point $c = g$, this system of equations can be solved by choosing $\kappa = [1, 0]^T$ and $\zeta = \kappa$. The trailing n_α elements of (3.6) then give the n_α equations defining the normal vector r. Substituting this choice of κ into (3.7) and using (3.2) and (3.3), the following system of equations results for the normal vector on grazing point manifolds:

$$G^{(g)} := \begin{pmatrix} z(t) - \Phi(z_0, t_0, p, \alpha, t) \\ h(z(t), p, \alpha, t) \\ h_z z_t + h_t \\ (h_z \Phi_\alpha + h_\alpha)^T - r \end{pmatrix} = 0. \tag{3.8}$$

For the end-point constraint, the system of equations (3.7) is solved by choosing $\kappa = [1, -(h_z z_t + h_t)]$ and $\zeta = [1, 0]$. The augmented system defining the normal direction for the

end-point constraint (2.10) then reads as

$$G^{(e)} := \begin{pmatrix} z(t) - \Phi(z_0, t_0, p, \alpha, t) \\ h(z(t), p, \alpha, t) \\ t - t^e \\ (h_z \Phi_\alpha + h_\alpha)^T - r \end{pmatrix} = 0. \tag{3.9}$$

Note that the required time derivatives $z_t = [x_t^T, y_t^T]$ as well as the sensitivities Φ_α can be provided by numerical simulators that allow for an efficient integration of the sensitivity equations (3.5), such as DDASPK [30], an extended version of LIMEX [47], or DAEPACK [49]. For large nonlinear systems with a Jacobian that has a sparse unordered structure the latter two should be preferred as they use sparse linear algebra packages.

The normal direction of a grazing point manifold is also calculated in a recent work of Donde and Hiskens [16]. The derivation in [16] differs from the calculations presented in this section in that Donde and Hiskens do not make use of the values calculated by the numerical integration for the algebraic states $y(t)$ and their derivatives with respect to time and parameters y_t, y_α. Instead they use the algebraic equations $0 = g$ for the calculation of y and (2.8) for the calculation of y_t. This leads to a more complex normal vector system that includes second order derivatives of the algebraic equations and $2n_x + 4n_y + n_\alpha + 4$ equations instead of $n_x + n_y + n_\alpha + 2$ equations in (3.8). For further illustration the B-matrix and the normal vector system according to [16] are given in Appendix B.

4. Optimization with normal vector constraints.

In this section we adopt the approach for the robust design of nonlinear systems first presented by Mönnigmann and Marquardt [36] that utilizes the normal vector direction of critical manifolds for the robust design of nonlinear systems. In this approach, the normal vector direction r is used to formulate constraints for a nonlinear program (NLP) to ensure that constraints (1.2) hold for the optimal design of nonlinear system (1.1) despite disturbances parameterized by uncertain parameters α. Note that in the following sections we consider again the general case of $n_h \geq 1$ constraints (1.2).

4.1. Normal vector constraints.

The approach enforces a lower bound on the parametric distance between the critical manifold and the boundary of the uncertainty region $\mathcal{A}$ (2.1). We denote the boundary as *robustness manifold* $M^{(\mathrm{rob})}$ defined by

$$M^{(\mathrm{rob})} = \begin{pmatrix} \hat{g}(\tilde{x}^{(\mathrm{rob})}, p, \alpha^{(\mathrm{rob})}) \\ \tilde{g}(\tilde{x}^{(\mathrm{rob})}, p, \alpha^{(\mathrm{rob})}) \end{pmatrix} = 0. \tag{4.1}$$

The normal vector constraint defining the lower bound then reads

$$\begin{aligned} \alpha^{(\mathrm{rob})} &= \alpha^{(c)} + lr, \\ l &\geq 0, \end{aligned} \tag{4.2}$$

with l as the measure for parametric distance and r representing the common normal direction of the critical manifold and the robustness manifold. As shown by Mönnigmann and Marquardt [37] the system of equations defining the normal direction of the robustness manifold

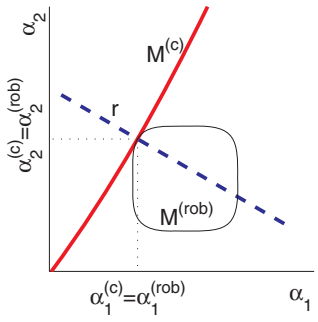


Figure 5. *Normal vector constraint: Uncertainty region is bounded by the robustness manifold $M^{(\mathrm{rob})}$. Distance is measured along common normal direction r of critical and robustness manifold.*

$r^{(\mathrm{rob})}$ can be derived following the scheme described in section 3, resulting in another normal vector system of the form

$$G^{(\mathrm{rob})}\left(\tilde{x}^{(\mathrm{rob})}, p, \alpha^{(\mathrm{rob})}, r^{(\mathrm{rob})}\right) = 0. \tag{4.3}$$

The lower bound on the parametric distance ensures that the complete range of uncertain parameters is at a safe distance from the critical boundary. This is further illustrated with Figure 5.

We are interested in finding a system design that minimizes an objective ϕ for a nominal operating point and, at the same time, satisfies the constraints (1.2). In the presence of fast disturbances parameterized by α bounded within the robustness manifold (4.1) the design problem is addressed by solving the following constrained NLP:[1]

$$\min \phi(z^{(0)}, p, t^{(0)}) \tag{4.4a}$$

$$\begin{aligned} \text{s. t. } z^{(0)} &= \Phi(z_0^{(0)}, t_0^{(0)}, p, t^{(0)}), \\ 0 &\le h_i(z^{(0)}, p, t^{(0)}) \quad \forall i \in \mathcal{I}, \end{aligned} \tag{4.4b}$$

$$0 = G^{(c,i,j)}\left(z^{(c,i,j)}, p, \alpha^{(c,i,j)}, t^{(c,i,j)}, r^{(i,j)}\right) \quad \forall i \in \mathcal{I},\ \forall j \in J_i, \tag{4.4c}$$

$$0 = G^{(\mathrm{rob},i,j)}\left(\tilde{x}^{(\mathrm{rob},i,j)}, p, \alpha^{(\mathrm{rob},i,j)}, \nu^{(i,j)} \cdot r^{(i,j)}\right) \quad \forall i \in \mathcal{I},\ \forall j \in J_i, \tag{4.4d}$$

$$\begin{aligned} 0 &= \alpha^{(c,i,j)} - \alpha^{(\mathrm{rob},i,j)} + l r^{(i,j)} \quad \forall i \in \mathcal{I},\ \forall j \in J_i, \\ 0 &\le l^{(i,j)} \quad \forall i \in \mathcal{I},\ \forall j \in J_i, \end{aligned} \tag{4.4e}$$

$$\begin{aligned} 0 &\le \tilde{h}(z^{(0)}, p, t^{(0)}, \tilde{x}_1), \\ 0 &= \hat{h}(z^{(0)}, p, t^{(0)}, \tilde{x}_2). \end{aligned} \tag{4.4f}$$

Equations (4.4b) define the states $z^{(0)}$ of the nominal system without disturbances. The set $\mathcal{I} = \{1, \ldots, n_h\}$ enumerates the constraints. Equation (4.4c) constitutes the normal vector

[1]For ease of notation, the time argument (t) for states z is omitted in the following sections.

system of the critical points. For each constraint $i \in \mathcal{I}$ several critical points $j \in J_i$ may exist in general. More than one closest point may exist due to nonconvexity of the critical manifolds [37] or in the case that there are several critical manifolds of grazing points for a single constraint. The superscript $c \in \{g, e\}$ denotes the type of critical manifold (grazing point or end-point).

The normal vector of the robustness manifold is defined by (4.4d). The scalar multiplier $\nu^{(i,j)}$ is introduced as the normal vectors of the critical manifold and the robustness manifold are not normalized in their defining equations (3.8), (3.9), and (4.3) and might furthermore point in opposite directions. Equations (4.4e) enforce the minimal distance between a point located on the robustness manifold $\alpha^{(\mathrm{rob},i,j)}$ and the nearest critical point $\alpha^{(c,i,j)}$.

Additional equality and inequality constraints (4.4f) with auxiliary variables $\tilde{x}_1 \in \mathbb{R}^{n_{\tilde{x}_1}}$ and $\tilde{x}_2 \in \mathbb{R}^{n_{\tilde{x}_2}}$ map into $\mathbb{R}^{n_{\tilde{h}}}$ or $\mathbb{R}^{n_{\tilde{h}}}$, respectively. These constraints are included in the formulation to allow for further specifications of the nominal solution that do not have to be robustly enforced. Typically, these are constraints which are not affected by the disturbances, e.g., simple box constraints on the parameters p. The degrees of freedom of the NLP (4.4) are p, $z_0^{(0)}$, $t^{(0)}$, $\tilde{x}_1$, $\tilde{x}_2$, $t^{(c,i,j)}$, $\alpha^{(c,i,j)}$, $\alpha^{(\mathrm{rob},i,j)}$, $\tilde{x}^{(\mathrm{rob},i,j)}$, $r^{(i,j)}$, $\nu^{(i,j)}$, and $l^{(i,j)}$.

4.2. Solution strategy. A number of difficulties has to be tackled to solve the NLP (4.4). Generally, the location of the critical manifolds is not known beforehand. Therefore, an algorithmic solution strategy for critical manifolds of steady states [38] has to be adapted to the new types of critical manifolds.

The algorithm builds up the sets of critical points and corresponding normal vector constraints J_i as the optimization proceeds. Newly detected critical points that violate bounds h_i will in general not satisfy the normal vector constraints (4.2). Therefore, an additional set of points $\mathcal{J}$ is introduced that comprises the newly detected critical points. An initialization step is necessary to move points from $\mathcal{J}$ to the sets of points J_i used for the normal vector constraints. The algorithm also allows for removal of normal vector constraints (i, j) if the parametric distance $l^{(i,j)}$ is larger than a specified threshold $l_{\max}$. The algorithm involves the following four steps, of which the last three may have to be carried out repeatedly:

(i) *Initialization*: Choose an initial nominal operating point for which all constraints h_i are satisfied. If points that violate h_i are known beforehand, put them into the set $\mathcal{J}$.

(ii) *Update of J_i*: Find the points on the critical manifolds $\alpha \in M^{(c,i)}$ that satisfy the normal vector constraints using the points in set $\mathcal{J}$ as initialization. The obtained critical points are moved from set $\mathcal{J}$ to the sets J_i. The last step in the update of J_i is the removal of those critical points (i, j) from the set J_i for which $l^{(i,j)} > l_{\max}$.

(iii) *Optimization*: Solve NLP (4.4) with normal vector constraints to the critical points contained in J_i.

(iv) *Search within the robustness region for constraint violations*: Check if the constraints hold within the region of parametric uncertainty. If critical points are found, update index $\mathcal{J}$ and return to step (ii).

If the test in step (iv) reveals no further critical points, an optimal operating point has been found that is robust with respect to the constraints in the presence of parametric uncertainty. The algorithm is therefore terminated.

In contrast to critical points of steady states, there are no adequate test functions to detect

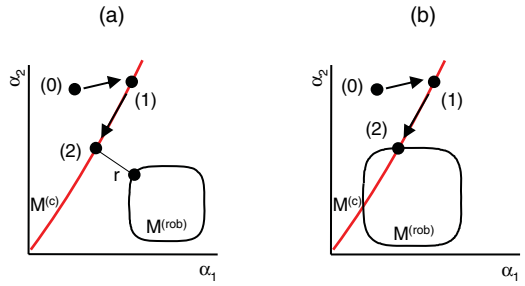


Figure 6. (a) *Initialization steps for points in $\mathcal{J}$ located outside the robustness region.* (b) *Initialization fails if critical manifold crosses the robustness manifold.*

points that directly satisfy all conditions that must hold for a point located on the critical manifold. In the following we therefore elaborate further on the detection of critical points in step (iv) and initialization of normal vector constraints in step (ii), which have to be modified in comparison to [38].

4.2.1. Detection of critical points. Detection of unknown critical points in step (iv) is realized by numerical integration of system (1.1). Since the numerical integration can only be carried out for a finite span of time, the detection is limited to a finite time horizon. The horizon length must be a compromise between computational costs on the one hand and the risk of missing constraint violations on the other hand. The check for constraint violations can be stopped at a finite time if a steady state or a stable limit cycle is reached. The search within the robustness region is carried out in this work on a grid of points of the robustness region, e.g., on the corner or center points of the faces of the hypercube (2.3).

4.2.2. Initialization of normal vector constraints. In general, critical points detected in step (iv) do not satisfy the augmented system (2.9) or (2.10). They will also not be the nearest critical point to the robustness manifold. The solver used to solve NLP (4.4), however, requires a feasible point as initialization. In step (ii), therefore, a strategy has to be employed to obtain a critical point satisfying the normal vector constraints (4.4e). For points in $\mathcal{J}$ located outside the uncertainty region, i.e., $0 \geq \tilde{g}(z, \tilde{x}, p, \alpha, t)$, the two step strategy illustrated by Figure 6a is used:

(i) *Find a point located on the critical manifold*: Here we assume that the second condition holds for the points in $\mathcal{J}$, i.e., the tangential condition for a grazing bifurcation $0 = h_z z_t + h_t$ or the time condition $0 = t - t^e$ for an end-point constraint. Bound (1.2), however, is assumed to be violated, $0 > h$. To obtain a point on the critical manifold the following optimization problem is solved:

$$
\begin{aligned}
&\max_{\alpha,t} \quad h(z,p,\alpha,t) \\
&\text{s. t. } z = \Phi(z_0^{(0)}, t_0^{(0)}, p, \alpha, t), \\
&\qquad 0 \geq h(z,p,\alpha,t), \\
&\qquad 0 = \begin{cases} h_z z_t + h_t & \text{if } c = g, \\ t - t^e & \text{if } c = e. \end{cases}
\end{aligned} \tag{4.5}
$$

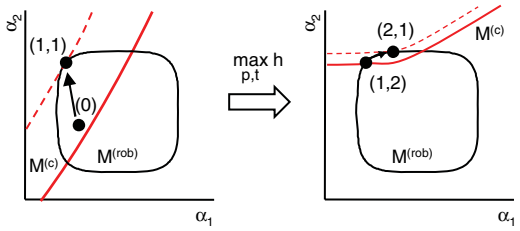


Figure 7. *Two iterations of the initialization steps for points in $\mathcal{J}$ located within the robustness region. Minimization (4.7) ($(0) \to (1,1)$ and $(1,2) \to (2,1)$) and maximization (4.8) ($(1,1) \to (1,2)$). The dashed curve is the level set $h = \text{const} < 0$, which touches the robustness region tangentially.*

This gives a point on the critical manifold if $h = 0$ at the optimum. Note that an equivalent formulation can be used in case that the first condition $h = 0$ of the augmented systems holds at the detected point by using the second condition as objective in (4.5). In Figure 6a transition of point (0) to point (1) illustrates this step.

(ii) *Find the closest critical point*: Points satisfying normal vector constraint (4.4e) are found by minimizing the parametric distance between the critical manifold and the robustness manifold over all $\alpha \in M^{(c,i)}$ and $\alpha^{(\text{rob})} \in M^{(\text{rob})}$ for a fixed nominal point

$$\begin{aligned} \min_{\alpha,t,\alpha^{(\text{rob})},\tilde{x}^{(\text{rob})}} \ & \frac{1}{2}(\alpha - \alpha^{(\text{rob})})^T(\alpha - \alpha^{(\text{rob})}) \\ \text{s. t. } & z = \Phi(z_0^{(0)}, t_0^{(0)}, p, \alpha, t), \\ & 0 = M^{(c,i)}(z, p, \alpha, t), \quad i \in \mathcal{I}, \\ & 0 = M^{(\text{rob})}(\tilde{x}^{(\text{rob})}, p, \alpha^{(\text{rob})}). \end{aligned} \tag{4.6}$$

As pointed out in [37] the Lagrange multipliers required for the definition of the Karush–Kuhn–Tucker conditions of optimality can be used to initialize the normal vector r, the distance variable l, and ν. In Figure 6a solving (4.6) corresponds to the transition from point (1) to point (2). In case $\alpha = \alpha^{(\text{rob})}$ it has to be checked if the critical manifold intersects the robustness region (cf. Figure 6b). In this case the initialization routine fails and the strategy for critical points located within the robustness region has to be pursued.

If a point in $\mathcal{J}$ is located within the uncertainty region, then the critical manifold intersects the robustness region and has to be pushed outside before NLP (4.4) can be solved. This is achieved by the following two iterative steps illustrated by Figure 7:

(i) *Find a worst-case point on the robustness manifold*: The worst-case point with respect to the constraint h on the robustness manifold is found by solving

$$\begin{aligned} \min_{\alpha,t} \quad & h(z, p, \alpha, t) \\ \text{s. t. } & z = \Phi(z_0^{(0)}, t_0^{(0)}, p, \alpha, t), \\ & 0 = \begin{cases} h_z z_t + h_t & \text{if } c = g, \\ t - t^e & \text{if } c = e, \end{cases} \\ & 0 = M^{(\text{rob})}(\tilde{x}, p, \alpha). \end{aligned} \tag{4.7}$$

In Figure 7 this is the transition from point (0) to $(1,1)$ and from point $(1,2)$ to $(2,1)$.

(ii) *Push the critical manifold to the robustness manifold*: By maximizing the constraint for fixed uncertain parameters,

$$
\begin{aligned}
\max_{p,t} \quad & h(z,p,\alpha,t) \\
\text{s. t. } & z = \Phi(z_0^{(0)}, t_0^{(0)}, p, \alpha, t), \\
& 0 \geq h(z,p,\alpha,t), \\
& 0 = \begin{cases} h_z z_t + h_t & \text{if } c = g, \\ t - t^e & \text{if } c = e, \end{cases} \\
& z = \Phi(z_0^{(0)}, t_0^{(0)}, p, t), \\
& 0 \leq h_i(z^{(0)}, p, t^{(0)}) \quad \forall i \in \mathcal{I}, \\
& 0 \leq \tilde{h}(z^{(0)}, p, t^{(0)}, \tilde{x}_1), \\
& 0 = \hat{h}(z^{(0)}, p, t^{(0)}, \tilde{x}_2),
\end{aligned}
\tag{4.8}
$$

the critical manifold crosses the robustness manifold at the former worst-case point ($(1,2)$ in Figure 7) if $h = 0$ (and if the equations defining the critical manifold do not depend on p) at the solution of (4.8).

The two steps have to be repeated iteratively until the constraint violation after step (i) is smaller than a specified threshold. Again, the Lagrangian multipliers of the optimality conditions for (4.7) can be used to initialize the normal vector r and ν of NLP (4.4).

The described initialization steps may fail due to an unsuitable control structure or too tight restrictions (1.2) for the modeled disturbances. In this case the control structure needs to be modified or the requirements have to be relaxed.

4.3. Implementational details. For the solution of the NLP the SQP-solver NPSOL [21] is used. This gradient-based solver requires the derivatives of the constraints of NLP (4.4) with respect to the degrees of freedom of the optimization problem. As the system of equations defining the normal vectors for the critical manifold of grazing points (3.8) and end-point constraints (3.9) involve sensitivities Φ_α of the state variables, the corresponding derivatives of the normal vector equations contain second order sensitivities $\Phi_{\alpha\alpha}$ (and $\Phi_{\alpha p}, \Phi_{\alpha z_0}$). The dynamics of the second order sensitivities are defined by the DAE system that is obtained by differentiating the DAE system for the first order sensitivities (3.5) with respect to α, p, and initial conditions z_0. The DAE system for the second order sensitivities with respect to the uncertain parameters α can be written as

$$
M\Phi_{\alpha\alpha t} = F_{zz}\Phi_\alpha\Phi_\alpha + 2F_{z\alpha}\Phi_\alpha + F_z\Phi_{\alpha\alpha} + F_{\alpha\alpha}. \tag{4.9}
$$

The DAE systems for the mixed second order sensitivities are defined accordingly. Numerical integration of the nonlinear system (1.1) and sensitivity equations (3.5), (4.9) is carried out either by DDASPK [30] or by an extended version of LIMEX [47]. As pointed out in [52] the structure of the equations of the second order sensitivities (4.9) is similar to the structure of the first order sensitivities. They can therefore be integrated in the same efficient fashion as the

first order sensitivities. Furthermore, the symmetry of the second order sensitivities $\Phi_{\alpha_i\alpha_j} = \Phi_{\alpha_j\alpha_i}$ can be easily exploited. For a larger number of parameters a promising alternative based on adjoints has been suggested recently [25] to further reduce the computational effort.

Finally, second order derivatives of the state variables with respect to time $z_{tt} = [x_{tt}^T, y_{tt}^T]^T$ are required to state the derivatives of the grazing bifurcation augmented system (2.9). These derivatives are usually not provided by the numerical integrator and have to be calculated by evaluating the time derivatives of the differential equations defining x_t (1.1) and the equations for y_t (2.8), respectively. x_{tt} is easily calculated by the following expression:

$$x_{tt} = f_x f + f_y y_t + f_t.$$

For y_{tt} the following system of linear equations has to be solved:

$$g_y y_{tt} = -(g_x x_{tt} + g_{xx} f f + 2 g_{xy} f y_t + g_{yy} y_t y_t + g_{tx} f + g_{ty} y_t + g_{tt}).$$

The partial derivatives of the functions f, g, and h that are needed for the constraints and gradients of the NLP (4.4) are evaluated with symbolic differentiation with Maple [40] and automatic differentiation with ADIFOR [3].

5. Illustrative case studies.

In section 5.1 we present two applications of the normal vector constraints to ensure a minimal distance to a manifold of grazing points. In section 5.2 we show an application of the approach to guarantee robustness with respect to a critical manifold for end-point constraints.

5.1. Robustness with respect to grazing points.

We investigate a bioreactor model which has already been robustly optimized with respect to steady-state stability boundaries by Mönnigmann and Marquardt [36]. Here we consider an upper bound on the substrate concentration, which should robustly hold despite fast disturbances. The second case study involves a closed-loop chemical reactor with state feedback control to illustrate the possibility of simultaneous operating point and control design with the presented approach.

5.1.1. Bioreactor.

Consider the continuous bioreactor model with two nonlinear ordinary differential equations [1]:

$$\begin{aligned} x_{1\tau} &= -x_1 + \Lambda(x_2) Da\, x_1 := f_1, \\ x_{2\tau} &= -x_2 + \Sigma(x_2) Da\, x_1 := f_2. \end{aligned} \tag{5.1}$$

Here x_1 denotes the dimensionless biomass concentration, x_2 the dimensionless substrate concentration, and τ is the dimensionless time. The uncertain parameters are the Damköhler number Da and the feed substrate concentration S_q in $\mathrm{kmol\,m^{-3}}$. Λ and Σ are known state functions describing the bioreaction [1, 36]. The bioreactor is optimized with respect to the economic profit function [36]

$$\phi = -\frac{c_\phi \mu(S_q) V}{Da}(c_1(a + bS_q)S_q x_1 - c_2(S_q)), \tag{5.2}$$

with reactor volume V, cost coefficients $c_1, c_2(S_q)$, and kinetic parameters a and b. The substrate feed concentration is limited to values between

$$S_d \le S_q \le S_c, \quad S_d = 0.3\,\mathrm{kmol\,m^{-3}},\ S_c = 1.0\,\mathrm{kmol\,m^{-3}}.$$

Parameter values are taken from [5] and [36].

The optimal operating point is unstable [5] if no additional constraints ensure the stability of the optimal solution. As shown in Figure 8, the bioreactor model exhibits two stability boundaries in the parameter plane (Da, S_q): a manifold of Hopf bifurcations and a manifold of saddle-node bifurcations. These manifolds separate the region with stable stationary points from the region with unstable steady states. In [36] the bioreactor model has been optimized with respect to the profit function (5.2) guaranteeing robust stability by normal vector constraints to the two stability boundaries. In this work we look for an optimal steady-state operating point $(x^{(0)}, p)$ for which we can guarantee that a maximal substrate concentration

$$0 < S_{\max} - S := h, \quad S_{\max} = 0.07\,\mathrm{kmol\,m^{-3}}, \tag{5.3}$$

is not violated at the outflow of the reactor after step disturbances of both the Damköhler number Da and the feed substrate concentration S_q. The disturbances are parameterized by the parameters dDa and dS_q,

$$Da = \begin{cases} \overline{Da}, & t \le t_0, \\ \overline{Da} + dDa, & t > t_0, \end{cases} \tag{5.4a}$$

$$S_q = \begin{cases} \overline{S_q}, & t \le t_0, \\ \overline{S_q} + dS_q, & t > t_0, \end{cases} \tag{5.4b}$$

which are bounded by

$$\begin{aligned} dDa &\in [-\Delta Da, \Delta Da], \quad \Delta Da = 0.05, \\ dS_q &\in [-\Delta S_q, \Delta S_q], \quad \Delta S_q = 0.03\,\mathrm{kmol\,m^{-3}}. \end{aligned}$$

Hence, we are looking for a steady state, defined by $0 = f(x_1^{(0)}, x_2^{(0)}, \overline{Da}, \overline{S_q})$, such that the upper bound on the substrate concentration in the reactor (5.3) is not violated by any of the trajectories $\Phi(x_{10}, x_{20}, t_0, Da, S_q, t)$ with initial conditions $x_{10} = x_1^{(0)}$, $x_{20} = x_2^{(0)}$ and uncertain disturbances defined by (5.4). With regard to the NLP (4.4) we have $\mathcal{I} = \{1\}$, $c = g$, $z = \{x_1, x_2\}$, $p = \{\overline{Da}, \overline{S_q}\}$, and $\alpha = \{dDa, dS_q\}$. We consider the approximation of the uncertainty box by an ellipse $\sum_i (\alpha_i/\Delta\alpha_i)^2 = n_\alpha$.

Optimization starts without robustness constraints. Unknown critical manifolds are detected by numerical integration and repeated optimization steps. Each detected critical point adds normal vector constraints with new variables $t^{(i,j)}$, $l^{(i,j)}$, $\nu^{(i,j)}$, $dDa^{(i,j)}$, $dS_q^{(i,j)}$, $dDa^{(\mathrm{rob},i,j)}$, $dS_q^{(\mathrm{rob},i,j)}$ to the NLP. In this example there is only one constraint $\mathcal{I} = \{1\}$ and one critical point $J_1 = \{1\}$. NLP (4.4) was solved in less than 0.1 seconds on a PC with 2 GHz and 1 GB RAM. The result of the optimization is shown in Figure 8 in the (Da, S_q)-plane. Figure 8 illustrates that the normal vector constraint ensuring the minimal distance between the nominal operating point and the manifold of grazing points $S^{(g)} = S_{\max}$ is active at the optimum. The profit function has a value of $\phi = 0.17$ at $Da^{(0)} = 0.99, S_q^{(0)} = 0.3\,\mathrm{kmol\,m^{-3}}$ compared to $\phi = 1$ at the unstable optimal operating point. The nearest point on the manifold of steady states satisfying the bound (5.3), denoted as $S^{(\mathrm{st})} = S_{\max}$, is located at a greater distance from the nominal point than the manifold of grazing points $S^{(g)}$. This shows that it

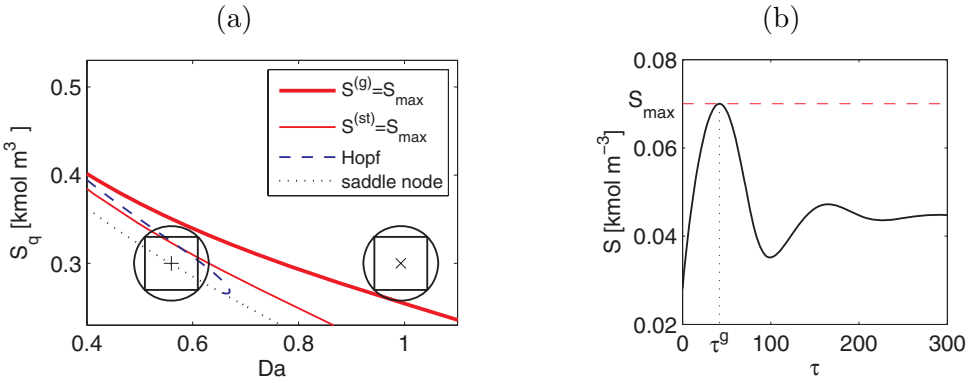


Figure 8. *Optimization of the bioreactor model* (5.1). (a) *The optimal solution without normal vector constraints* (+) *is unstable. At the optimal solution obtained with normal vector constraints* (×) *the constraint robustly holds despite disturbances in Da and* S_q. (b) *Grazing trajectory of critical point nearest to the robust optimum.*

Table 1

Optimal operating point of the fermenter obtained with optimization without normal vector constraints (unstable) (a)*, and with an ellipsoidal robustness region* (b).

	(a)	(b)
$\overline{Da}$	0.56	0.99
$\overline{S_q}\,(\mathrm{kmol\,m^{-3}})$	0.3	0.3
$x_1^{(0)}$	0.275	0.169
$x_2^{(0)}$	0.566	0.891
ϕ	1.0	0.17

is not sufficient to consider robustness with respect to steady-state constraints in the presence of fast disturbances. In this case study the stability boundaries are also located at a greater distance from the nominal point than the manifold of grazing points and do not have to be considered by additional normal vector constraints. In other cases, however, it might be necessary to take into account simultaneously normal vector constraints with respect to critical manifolds of grazing points and end-points, *and* normal vector constraints with respect to steady-state stability boundaries.

The result of the robust optimization is summarized in Table 1 together with the (unstable) optimal operating point obtained by the optimization of the objective (5.2) without normal vector constraints.

5.1.2. Continuous stirred tank reactor with state feedback control. As a second example for the robust optimization with respect to grazing points we consider a continuous stirred tank reactor (CSTR) with the exothermic consecutive reactions $A \to B \to C$ with B being the desired product. The CSTR model consists of nonlinear state equations for material balances of species A and B and energy balances for the reactor and cooling jacket assuming perfect level control [43]

$$c_{At} = \frac{q}{V}(c_{\mathrm{Aq}} - c_A) - r_1, \tag{5.5}$$

$$
\begin{aligned}
c_{Bt} &= \frac{q}{V} - c_B + r_1 - r_2, \\
T_t &= \frac{q}{V}(T_q - T) - \frac{\Delta H_1}{\rho C_p} r_1 - \frac{\Delta H_2}{\rho C_p} r_2 + \frac{UA}{V \rho C_p}(T_c - T), \\
T_{ct} &= \frac{q_c}{V_c}(T_{q_c} - T_c) - \frac{UA}{V_c \rho_c C_{pc}}(T_c - T),
\end{aligned}
$$

with reaction rates r_1 and r_2 defined by

$$
r_1 = k_{10} \exp\left(-\frac{E_1}{RT}\right) c_A^2, \quad r_2 = k_{20} \exp\left(-\frac{E_2}{RT}\right) c_B.
$$

The feed rates of the reactant q and coolant q_c are the manipulated variables; all parameter values are taken from [43]. A PI controller is considered,

$$
u(t) = u_0 + K(x(t) - x_{\mathrm{sp}}) + K_i \int_0^t (x(\tau) - x_{\mathrm{sp}}) d\tau,
$$

with $u = [q, q_c]^T$, $x = [c_A, c_B, T, T_c]^T$, and u_0, x_{sp}, K, and K_i as tunable control parameters. Integral action is used for concentration control of the desired product B and for the reactor temperature T only, i.e.,

$$
K_i = \begin{pmatrix} 0 & K_{i12} & 0 & 0 \\ 0 & 0 & K_{i23} & 0 \end{pmatrix}.
$$

In total there are 10 control parameters $K_{11}, K_{12}, K_{13}, K_{14}, K_{21}, K_{22}, K_{23}, K_{24}, K_{i12}, K_{i23}$. For simplicity, we assume that measurements for all states are available, such that a state estimator is not required. Disturbances are modeled by sinusoidal variations of the feed concentration c_{Aq} and feed temperature T_q:

$$
T_q(t) = \begin{cases} \overline{T_q}, & t \le t_0, \\ \overline{T_q} + dT_q \sin \omega_{T_q}(t - t_0), & t > t_0, \end{cases} \tag{5.6a}
$$

$$
c_{Aq} = \begin{cases} \overline{c_{Aq}}, & t \le t_0, \\ \overline{c_{Aq}} + dc_{Aq} \sin \omega_{c_{Aq}}(t - t_0), & t > t_0. \end{cases} \tag{5.6b}
$$

Disturbances and uncertainties in the feed stream can arise, e.g., if the reactor is part of a larger process plant and the reactant stream is the product of a pretreatment step. If this pretreatment process is not operated at steady state, the resulting feed temperature and concentration may also vary. The disturbance of the feed temperature is parameterized by the amplitude dT_q and frequency ω_{T_q} that are bounded by

$$
\begin{aligned}
dT_q &\in [-\Delta dT_q, \Delta dT_q], \\
\omega_{T_q} &\in [\overline{\omega_{T_q}} - \Delta\omega_{T_q}, \overline{\omega_{T_q}} + \Delta\omega_{T_q}],
\end{aligned} \tag{5.7}
$$

with

$$
\Delta dT_q = 10\,\mathrm{K}, \quad \Delta\omega_{T_q} = 2\,\mathrm{rad\,h^{-1}}, \quad \overline{\omega_{T_q}} = 3\,\mathrm{rad\,h^{-1}}.
$$

Likewise, the disturbance of the feed concentration is parameterized by the amplitude dC_{Aq} and frequency $\omega_{c_{Aq}}$ bounded by

$$\begin{aligned} dc_{Aq} &\in [-\Delta dc_{Aq}, \Delta dc_{Aq}], \\ \omega_{c_{Aq}} &\in [\overline{\omega_{c_{Aq}}} - \Delta\omega_{c_{Aq}}, \overline{\omega_{c_{Aq}}} + \Delta\omega_{c_{Aq}}], \end{aligned} \tag{5.8}$$

with

$$\Delta dc_{Aq} = 0.1\,\mathrm{kmol\,m^{-3}}, \quad \Delta\omega_{c_{Aq}} = 5\,\mathrm{rad\,h^{-1}}, \quad \overline{\omega_{c_{Aq}}} = 8\,\mathrm{rad\,h^{-1}}.$$

Here, we have a four-dimensional uncertainty box with $\alpha = [dT_q, \omega_{T_q}, dc_{Aq}, \omega_{c_{Aq}}]$. A tight approximation of the uncertainty box is important as any overestimation will generally lead to more conservative results or even prevent finding a feasible solution of (4.4). For an increasing number of uncertain parameters n_α the overestimation of the box by the ellipsoidal approximation used in the first example grows exponentially. For $n_\alpha = 2$ the volume ratio between a hypersphere with radius $\sqrt{n_\alpha}$ and a hypercube with side length 2 is approximately 1.6, whereas for $n_\alpha = 10$ the ratio is almost 250. Tighter approximations of the uncertainty box can be realized, e.g., by using higher norms for the approximation

$$\tilde{g} := \sum_{i=1}^{n_\alpha} (\alpha_i - \overline{\alpha_i})^{2j} - n_\alpha \tag{5.9}$$

with $j \in \mathbb{Z}^+$. Alternatively, one can use the approximation introduced by Kreisselmeier and Steinhauser [28]

$$\tilde{g} := \frac{1}{\rho} \ln\left(\frac{1}{2n_\alpha} \sum_{i=1}^{n_\alpha} \exp(\rho(\alpha_i - \overline{\alpha_i} - 1)) + \exp(\rho(-(\alpha_i - \overline{\alpha_i}) - 1)) \right).$$

The approximation of the uncertainty box gets closer for larger values of the parameter ρ. Here we use the less complicated formulation (5.9) with $j = 4$.

We now want to find a controller tuning and process design that guarantee that the outlet concentration of product B stays within specified bounds around the nominal value $c_B^{(0)}$ despite the presence of disturbances, i.e.,

$$0 \le -c_B + c_B^{(0)} + 0.15\,\mathrm{kmol\,m^{-1}} := h_1, \tag{5.10a}$$

$$0 \le c_B - c_B^{(0)} + 0.15\,\mathrm{kmol\,m^{-1}} := h_2. \tag{5.10b}$$

For the nominal operating point (0) steady-state constraints $0 = f^{(0)}$ are employed. Additional constraints $0 = x_{\mathrm{sp}} - x^{(0)}$ ensure that the controller is not active at the nominal point. The initial values for the critical points (i, j) with $\mathcal{I} = \{1, 2\}$, $c = g$ are set to the steady-state values of the nominal system $x_0^{(i,j)} = x^{(0)}$. Normal-vector constraints on critical manifolds of grazing points defined by the upper (5.10a) and lower concentration limit (5.10b) are used to guarantee that the outlet concentration of the product c_B stays within the bounds despite disturbances (5.6). We solve NLP (4.4) with respect to the objective function

$$\phi = q_0(\kappa_A \overline{c_A} - \kappa_B c_B^{(0)}) + q_c\kappa_c$$

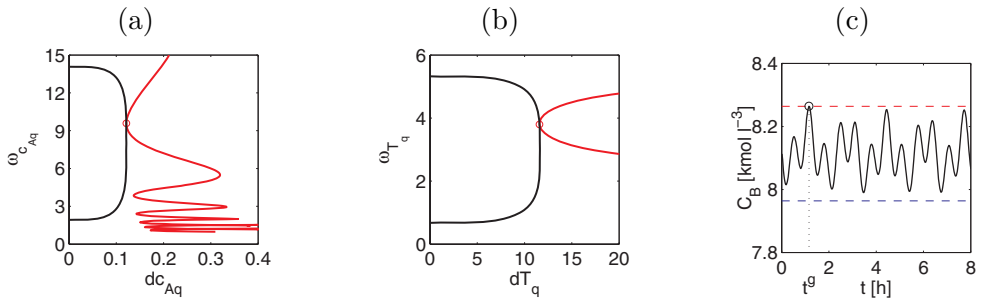


Figure 9. (a) *Robustness manifold and critical manifold for upper bound* (5.10a) *in the plane of the feed concentration disturbance parameters* $(dc_{Aq}, \omega_{c_{Aq}})$. *The normal vector constraint for the closest point* ($\circ$) *on the manifold of grazing points is active.* (b) *Robustness manifold and critical manifold in the plane of the feed temperature disturbance parameters* (dT_q, ω_{T_q}). (c) *Trajectory of the reactor temperature at the critical point.*

Table 2

Results for the optimization of the CSTR (5.5) *with normal vector constraints to manifolds of grazing points. Units of the control parameters:* K_{ij}, $i = 1,2$, $j = 1,2$: $\mathrm{m^6\,kmol^{-1}\,h^{-1}}$; K_{ij}, $i = 1,2$, $j = 3,4$: $\mathrm{m^3\,K^{-1}\,h^{-1}}$; K_{i12}: $\mathrm{m^6\,kmol^{-1}\,h^{-2}}$; K_{i23}: $\mathrm{m^3\,K^{-1}\,h^{-2}}$.

$K_{11} = -210.71$	$K_{12} = -80.06$	$K_{13} = -17.97$	$K_{14} = -10.54$
$K_{21} = -114.16$	$K_{22} = -59.05$	$K_{23} = -10.31$	$K_{24} = -3.97$
$K_{i12} = 21.38$	$K_{i23} = -1.15$	$q_0 = 100$	$q_{c0} = 60$
$\overline{c_{Aq}} = 12$	$\phi = -419.8\kappa_A$		
$dc_{Aq}^{(1,1)} = 0.12$	$\omega_{c_{Aq}}^{(1,1)} = 3.8$	$dT_q^{(1,1)} = 11.58$	$\omega_{T_q}^{(1,1)} = 9.58$

with κ_A, κ_B, and κ_c as cost coefficients of reactant A, product B, and coolant c, respectively, where $\kappa_c = 0.05\kappa_A$, $\kappa_B = 2\kappa_A$ are chosen for this case study. Degrees of freedom for the NLP (4.4) are the 10 control parameters K and K_i, the nominal feed rates q_0, q_{c0}, limited to values between 60 and $100\,\mathrm{m^3\,h^{-1}}$, the nominal feed concentration $\overline{c_{Aq}}$ bounded within 2 and $12\,\mathrm{kmol\,m^{-3}}$, the nominal steady state $c_A^{(0)}, c_B^{(0)}, T^{(0)}, T_c^{(0)}$, and corresponding set points of the controller. Each normal vector constraint involves a set of disturbance parameters on the critical manifold and on the robustness manifold as well as the critical time at which the grazing point occurs.

At the optimal solution there is one active normal vector constraint corresponding to the critical manifold of the upper concentration bound (5.10a). Figures 9a and 9b show the critical manifold corresponding to $h_1 = 0$ and the robustness manifold. The shape of the critical manifold shows that all disturbance parameters have a strong influence on system (5.5). The approach is able to automatically identify the worst-case combination of the disturbances and find an economical optimal design ensuring that the concentration bound holds. The trajectory of the reactor concentration corresponding to the nearest grazing point of the upper bound (5.10a) is shown in Figure 9c. The concentration trajectory touches the upper bound at $t^{(1)} = 1.16\,\mathrm{h}$. The aggregated computational time for the solution of NLP (4.4) and the optimization problems of the initialization routine described in section 4.2 is 18 seconds on a PC with 2 GHz and 1 GB of RAM. The values of parameters p and disturbance parameters α at the optimal operating point are summarized in Table 2.

5.2. Robustness with respect to end-point constraints. Finally we show an application of the normal vector constraints for manifolds of end-point constraints.[2] We consider a cooled continuous stirred tank reactor (CSTR) with an exothermic first order reaction $A \rightarrow B$ with temperature control assuming perfect level control [23]. The CSTR model consists of nonlinear state equations for material and energy balances including reaction kinetics and heat transfer [51]:

$$\begin{aligned} c_{At} &= \frac{q}{V}(c_{\mathrm{Aq}} - c_A) - k_0 \exp\left(-\frac{E}{RT}\right) c_A, \\ T_t &= \frac{q}{V}(T_q - T) - \frac{\Delta H k_0}{\rho C_p} \exp\left(-\frac{E}{RT}\right) c_A + \frac{UA}{V\rho C_p}(T_c - T). \end{aligned} \tag{5.11}$$

The temperature of the cooling fluid T_c is the manipulated variable, the reactor temperature T is the measured and controlled variable, and c_A is the concentration of species A in the reactor. The values of the process parameters are taken from [23]. Control is realized by means of a linearizing feedback including integral action. The control law reads as [23]

$$\begin{aligned} T_c = &\frac{-\frac{q}{V}(T_q - T) + \frac{\Delta H}{\rho C_p} k_0 \exp(-\frac{E}{RT})\tilde{c}_A + \frac{UA}{V\rho C_p} T}{\frac{UA}{V\rho C_p}} \\ &+ \frac{\frac{2}{\varepsilon}(T_{\mathrm{sp}} - T) + \frac{1}{\varepsilon^2}\int_0^t (T_{\mathrm{sp}} - T) d\tau}{\frac{UA}{V\rho C_p}}. \end{aligned} \tag{5.12}$$

The control law comprises one tuning parameter ε, which corresponds to the time constant of the closed-loop dynamics; i.e., the smaller ε is, the faster the dynamics of the closed-loop system are. If the model used for the controller design and the real plant exactly match, the controller stabilizes the process for all possible set-point temperatures T_{sp}. In such a case, ε can be tuned to arbitrary small values resulting in very fast system dynamics. In real applications, however, there will always exist a mismatch between the plant and the model because of disturbances and uncertain parameters. In this case study the measured concentration $\tilde{c}_A$ used in the control law may differ from the true concentration c_A. We investigate a set-point change in the reactor temperature from $T_{\mathrm{sp0}} = 370\,\mathrm{K}$ to $T_{\mathrm{sp}} = 400\,\mathrm{K}$ at $t_0 = 0$. The initial conditions of the nominal case correspond to the steady state at $T_{\mathrm{sp0}} = 370\,\mathrm{K}$, $\overline{T_0} = 370\,\mathrm{K}$, and $\overline{c_{A0}} = 0.206\,\mathrm{mol\,l^{-1}}$. We assume a step disturbance of the feed temperature

$$T_q(t) = \begin{cases} \overline{T_q}, & t \le t_0, \\ \overline{T_q} + dT_q, & t > t_0, \end{cases} \tag{5.13}$$

with $dT_q\,(\mathrm{K}) \in [-15, 15]$. Further, the initial conditions of the reactor temperature and reactor concentration are not known exactly and may vary around the nominal values:

$$\begin{aligned} T_0 &\in [\overline{T_0} - \Delta T_0,\, \overline{T_0} + \Delta T_0],\ \Delta T_0 = 10\,\mathrm{K}, \\ c_{A0} &\in [\overline{c_{A0}} - \Delta c_{A0},\, \overline{c_{A0}} + \Delta c_{A0}],\ \Delta c_{A0} = 0.1\,\mathrm{mol\,l^{-1}}. \end{aligned}$$

[2] In [14] an alternative robust optimization method based on a bilevel formulation is applied to a dynamic optimization problem with an end-point constraint. The close relationship between the bilevel formulation and the normal vector approach is discussed in this paper, too.

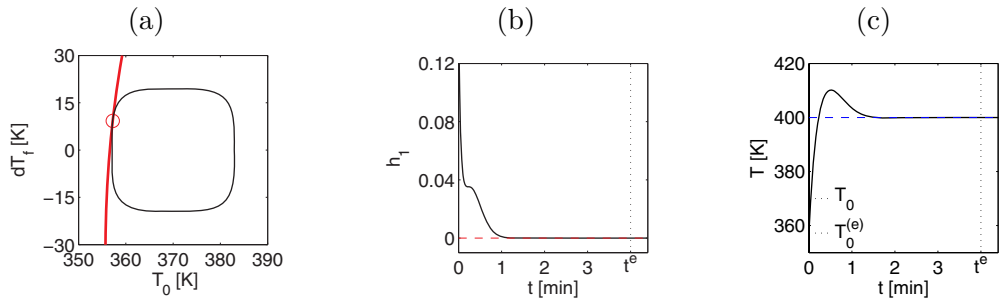


Figure 10. (a) *Robustness and critical manifold for end-point constraint* (5.14) *in the plane of the uncertain initial condition* T_0 *and disturbance parameter* dT_q. *The normal vector constraint for the closest point* ($\circ$) *on the manifold for end-point constraint* (5.14) *is active.* (b) *Trajectory of the value of* h_1 *at the critical point reaches* 0 *after approximately* 2.5 min. (c) *Temperature trajectory of the critical point with initial condition* $T_0^{(e)} = 357.2\,\mathrm{K}$ *and set-point temperature* $T_{\mathrm{sp}} = 400\,\mathrm{K}$.

Finally, the measured concentration $\tilde{c}_A$ used for the control law (5.12) is assumed to be biased from the true concentration $\tilde{c}_A = c_A + dC_A$ with

$$dC_A \in [-\Delta C_A, \Delta C_A], \quad \Delta dC_A = 0.005\,\mathrm{mol\,l^{-1}}.$$

We consider again the tighter approximation (5.9) with $j = 4$. One end-point constraint is imposed that must hold at final time t^e despite the disturbance in T_q, the uncertain initial conditions, and the uncertain measurement:

$$0 < ISE_{\max} - \sqrt{\frac{\int_0^{t_e} (T - T_{\mathrm{sp}})^2 dt}{t_e (T_{\mathrm{sp}} - T_{\mathrm{sp0}})^2}} := h_1. \tag{5.14}$$

Constraint (5.14) sets an upper bound $ISE_{\max} = 0.2$ for the integrated square error (ISE) between the set-point temperature T_{sp} and the reactor temperature T. This constraint can be interpreted as a specification for a desired performance of the closed-loop system that must be achieved despite disturbances and uncertain initial conditions. The final time is set to $t_e = 4\,\mathrm{min}$ for the end-point constraint. The objective is to minimize the control effort required to drive the system to the desired temperature $T_{\mathrm{sp}} = 400\,K$, here represented by the integral

$$\phi = \frac{V}{q(T_{\mathrm{sp}} - T_{\mathrm{sp0}})^2} \int_0^{t_e} (T_c(t) - T_c(T_{\mathrm{sp}}))^2 dt.$$

The cooling water temperature corresponding to the steady state at the new set-point temperature is $T_c(T_{\mathrm{sp}}) = 328.1\,K$.

The results of the optimization are displayed in Figure 10. The projection of the critical manifold on the (dT_q, T_0)-plane in Figure 10a shows that the normal-vector constraint to the end-point constraint h_1 is active. The uncertainty of the initial concentration has no influence on the integral (5.14) as the temperature dependence on the concentration is eliminated by the linearizing feedback control (5.12). The uncertainty of the concentration measurement has a large influence as it directly enters the control law (5.12). The transient behavior of

the value of h_1 at the critical point in Figure 10b shows that $h_1 = 0$ is attained before t_e is reached. The value stays at $h_1 = 0$ for $t > 2.5\,\text{min}$, as the reactor temperature has reached the new set-point temperature $T_{\text{sp}} = 400\,\text{K}$, (cf. Figure 10c). The value of the control parameter at the optimal solution is $\varepsilon = 0.262\,\text{min}$ and the value of the objective function is $\phi = 0.332$. The computational time was less than 0.1 seconds for this case study.

6. Conclusions. In the presence of fast disturbances critical points of trajectories of nonlinear systems can be defined. Specifically, we have investigated critical manifolds in the space of disturbance parameters α corresponding to trajectories with grazing points or trajectories which exactly fulfill an end-point constraint. These critical manifolds separate the parameter space into regions with qualitatively different transient system behavior. A manifold of grazing bifurcations separates those trajectories which do not violate an inequality constraint from those which violate the constraint for some time. A manifold defined for trajectories satisfying end-point constraints separates those trajectories which do not violate the constraint until a specified final time is reached from those that cross the constraint before the final time is reached.

These critical manifolds are used in a constructive manner for the robust design of nonlinear systems by extending a recently presented method for robust optimization. The method is based on the parametric distance between the uncertainty region of the disturbance parameters and the nearest point on the critical manifold. Normal-vector constraints guarantee that specifications for nonlinear systems are not violated in the presence of disturbances. Previous work considered critical points of steady states, e.g., Hopf and saddle-node bifurcations. We have extended this approach from steady-state specifications to state and output constraints on trajectories of nonlinear systems. Application of the approach to closed-loop systems allows for the integrated treatment of system and control design. The methodology has been successfully applied to several illustrative case studies from the area of chemical engineering.

Appendix A. Notation. The subscript μ enumerates the first dimension (rows of matrices), ν enumerates the second dimension (columns of matrices), and ρ enumerates the third dimension of three-dimensional arrays. When an index appears twice in a term it indicates summation over the index. In particular, we use

$$(f_x)_{\mu\nu} = \frac{\partial f_\mu}{\partial x_\nu} \in \mathbb{R}^{n_x \times n_x}, \qquad (f_y)_{\mu\nu} = \frac{\partial f_\mu}{\partial y_\nu} \in \mathbb{R}^{n_x \times n_y},$$

$$(f_x f)_\mu = \frac{\partial f_\mu}{\partial x_\nu} f_\nu \in \mathbb{R}^{n_x}, \qquad (f_y y_t)_\mu = \frac{\partial f_\mu}{\partial y_\nu} \frac{\partial y_\nu}{\partial t} \in \mathbb{R}^{n_y},$$

$$(g_x)_{\mu\nu} = \frac{\partial g_\mu}{\partial x_\nu} \in \mathbb{R}^{n_y \times n_x}, \qquad (g_y)_{\mu\nu} = \frac{\partial g_\mu}{\partial y_\nu} \in \mathbb{R}^{n_y \times n_y},$$

$$(g_x x_{tt})_\mu = \frac{\partial g_\mu}{\partial x_\nu} \frac{\partial^2 x_\nu}{\partial t^2} \in \mathbb{R}^{n_y}, \qquad (g_{xx} f f)_\mu = \frac{\partial^2 g_\mu}{\partial x_\nu \partial x_\rho} f_\nu f_\rho \in \mathbb{R}^{n_y},$$

$$(g_{xy} f y_t)_\mu = \frac{\partial^2 g_\mu}{\partial x_\nu \partial y_\rho} f_\nu \frac{\partial y_\rho}{\partial t} \in \mathbb{R}^{n_y}, \qquad (g_{yy} y_t y_t)_\mu = \frac{\partial^2 g_\mu}{\partial y_\nu \partial y_\rho} \frac{\partial y_\nu}{\partial t} \frac{\partial y_\rho}{\partial t} \in \mathbb{R}^{n_y},$$

$$(h_x)_\mu = \frac{\partial h}{\partial x_\mu} \in \mathbb{R}^{n_x}, \qquad (h_y)_\mu = \frac{\partial h}{\partial y_\mu} \in \mathbb{R}^{n_y},$$

$$(h_z z_t) = \frac{\partial h}{\partial z_\mu}\frac{\partial z_\mu}{\partial t} \in \mathbb{R}, \qquad (h_{zz} z_t z_t) = \frac{\partial^2 h}{\partial z_\mu \partial z_\nu}\frac{\partial z_\mu}{\partial t}\frac{\partial z_\nu}{\partial t} \in \mathbb{R},$$

$$(h_z z_{tt}) = \frac{\partial h}{\partial z_\mu}\frac{\partial^2 z_\mu}{\partial t^2} \in \mathbb{R}, \qquad (h_{zz} z_t z_t) = \frac{\partial^2 h}{\partial z_\mu \partial z_\nu}\frac{\partial z_\mu}{\partial t}\frac{\partial z_\nu}{\partial t} \in \mathbb{R},$$

$$(h_z \Phi_\alpha)_\mu = \frac{\partial h}{\partial z_\nu}\frac{\partial \Phi_\nu}{\partial \alpha_\mu} \in \mathbb{R}^{n_\alpha}, \qquad (h_{zz} z_t \Phi_\alpha)_\mu = \frac{\partial^2 h}{\partial z_\nu \partial z_\rho}\frac{\partial z_\nu}{\partial t}\frac{\partial \Phi_\rho}{\partial \alpha_\mu} \in \mathbb{R}^{n_\alpha},$$

$$(F_z \Phi_\alpha)_{\mu\nu} = \frac{\partial F_\mu}{\partial z_\rho}\frac{\partial \Phi_\rho}{\partial \alpha_\nu} \in \mathbb{R}^{n_z \times n_\alpha}, \qquad (F_{zz} \Phi_\alpha \Phi_\alpha)_{\mu\nu\rho} = \frac{\partial^2 F_\mu}{\partial z_\sigma \partial z_\tau}\frac{\partial \Phi_\sigma}{\partial \alpha_\nu}\frac{\partial \Phi_\tau}{\partial \alpha_\rho} \in \mathbb{R}^{n_z \times n_\alpha \times n_\alpha},$$

$$(F_{z\alpha} \Phi_\alpha)_{\mu\nu\rho} = \frac{\partial^2 F_\mu}{\partial z_\sigma \partial \alpha_\rho}\frac{\partial \Phi_\sigma}{\partial \alpha_\nu} \in \mathbb{R}^{n_z \times n_\alpha \times n_\alpha}, \qquad (F_z \Phi_{\alpha\alpha})_{\mu\nu\rho} = \frac{\partial F_\mu}{\partial z_\sigma}\frac{\partial^2 \Phi_\sigma}{\partial \alpha_\nu \partial \alpha_\rho} \in \mathbb{R}^{n_z \times n_\alpha \times n_\alpha}.$$

Appendix B. Augmented and normal vector system with state variables as independent variables. In this section we present the augmented and normal vector system of the grazing bifurcation if dynamic and algebraic state variables x and y are considered as independent variables as described in [16]. In comparison to (2.9) the flow equation $0 = x - \Phi(x_0, t, p, \alpha)$ is included in the augmented system to define the dynamic variables x, and the algebraic equations $0 = g$ are included to define the algebraic states y. The complete augmented system is then given by

$$M^{(g)} = \begin{pmatrix} x - \Phi(x_0, t, p, \alpha) \\ g(x, y, p, \alpha, t) \\ h(x, y, p, \alpha, t) \\ g_x f + g_y v + g_t \\ h_x f + h_y v + h_t \end{pmatrix} = 0, \tag{B.1}$$

with the time derivative of the algebraic equations denoted by v. The $n_x + 2n_y + 2$ equations define the $n_x + 2n_y + 2$ variables x, y, v, t, α_1. The gradient matrix B of the augemented system with respect to the variables x, y, v, t, α is

$$B = \begin{bmatrix} I & g_x^T & h_x^T & [g_{xx} f + g_x f_x + g_{yx} v + g_{tx}]^T & [h_{xx} f + h_x f_x + h_{yx} v + h_{tx}]^T \\ 0 & g_y^T & h_y^T & [g_{xy} f + g_x f_y + g_{yy} v + g_{ty}]^T & [h_{xy} f + h_x f_y + h_{yy} v + h_{ty}]^T \\ 0 & 0 & 0 & g_y^T & h_y^T \\ -f^T & g_t^T & h_t^T & [g_{xt} f + g_x f_t + g_{yt} v + g_{tt}]^T & [h_{xt} f + h_x f_t + h_{yt} v + h_{tt}]^T \\ -\Phi_\alpha^T & g_\alpha^T & h_\alpha^T & [g_{x\alpha} f + g_x f_\alpha + g_{y\alpha} v + g_{t\alpha}]^T & [h_{x\alpha} f + h_x f_\alpha + h_{y\alpha} v + h_{t\alpha}]^T \end{bmatrix}.$$

According to the scheme presented in section 3 for the derivation of the normal vector system, we are looking for a vector $\kappa \in \mathbb{R}^{n_x+2n_y+2}$ which spans the kernel of the first $n_x + 2n_y + 1$ rows of the matrix B

$$B\kappa = \begin{bmatrix} 0 \\ r \end{bmatrix} \in \mathbb{R}^{n_x+2n_y+1+n_\alpha},$$

with $0 \in \mathbb{R}^{n_x+2n_y+1}$ and $\kappa^T z - 1 = 0$ for some $z \in \mathbb{R}^{n_x+2n_y+2}$ not orthogonal to κ. For the matrix B in (3.3) in section 3 the choice $\kappa = [1, 0]^T$ was obvious and resulted in a drastic simplification of the normal vector system. Here, however, there is no κ such that the second order derivatives disappear in the normal vector system. Instead all entries of κ are also

unknown variables and have to be included in the normal vector system. The normal vector system for augmented system (B.1) with $z = \kappa$ is

$$G^{(g)} = \begin{pmatrix} x - \Phi(x_0, t, p, \alpha) \\ g(x, y, p, \alpha, t) \\ h(x, y, p, \alpha, t) \\ g_x f + g_y v + g_t \\ h_x f + h_y v + h_t \\ B^* \kappa \\ \kappa^T \kappa - 1 \\ \left[-\Phi_\alpha^T \;\; g_\alpha^T \;\; h_\alpha^T \;\; [g_{x\alpha} f + g_x f_\alpha + g_{y\alpha} v + g_{t\alpha}]^T \right. \\ \left. [h_{x\alpha} f + h_x f_\alpha + h_{y\alpha} v + h_{t\alpha}]^T \right] \kappa - r \end{pmatrix} = 0,$$

where $B^* \in \mathbb{R}^{n_x+2n_y+1} \times \mathbb{R}^{n_x+2n_y+2}$ is the submatrix of B with the first n_x+2n_y+1 rows of B. The $2n_x + 4n_y + n_\alpha + 4$ equations define the $2n_x + 4n_y + n_\alpha + 4$ variables $x, y, v, t, \kappa, \alpha_1, r$. In comparison, the normal vector system (3.8) has only $n_x+n_y+n_\alpha+2$ equations without second order derivatives of g. As we use normal vector systems for the formulation of constraints within a nonlinear program, gradient-based solvers would require third order derivatives of the algebraic equations if formulation (B.1) were used as the augmented system for a grazing point.

REFERENCES

[1] P. Agrawal, C. Lee, H. C. Lim, and D. Ramkrishna, *Theoretical investigations of dynamic behavior of isothermal continuous stirred tank biological reactors*, Chem. Eng. Sci., 37 (1982), pp. 453–462.

[2] P. A. Bahri, J. A. Bandoni, and J. A. Romagnoli, *Integrated flexibility and controllability analysis in design of chemical processes*, AIChE J., 43 (1997), pp. 997–1015.

[3] C. Bischof, P. Khademi, A. Mauer, and A. Carle, *Adifor 2.0: Automatic differentiation of Fortran 77 programs*, IEEE Comp. Sci. Eng., 3 (1996), pp. 18–32.

[4] F. Blanchini, *Set invariance in control*, Automatica J. IFAC, 35 (1999), pp. 1747–1767.

[5] D. D. Brengel and W. D. Seider, *Coordinated design and control optimization of nonlinear processes*, Comput. Chem. Eng., 16 (1992), pp. 861–886.

[6] P. J. Campo and M. Morari, *Robust model predictive control*, in Proceedings of the American Control Conference, Vol. 2, Minneapolis, MN, 1987, pp. 1021–1026.

[7] N. Chawankul, L. A. Ricardez Sandoval, H. Budman, and P. L. Douglas, *Integration of design and control: A robust control approach using MPC*, Can. J. Chem. Eng., 85 (2007), pp. 433–446.

[8] W. Chin, E. Ott, H. E. Nusse, and C. Grebogi, *Grazing bifurcations in impact oscillators*, Phys. Rev. E, 50 (1994), pp. 4427–4444.

[9] F. Chu and Z. Zhang, *Bifurcation and chaos in a rub-impact jeffcott rotor system*, J. Sound Vibration, 210 (1998), pp. 1–18.

[10] H. Dankowicz and A. B. Nordmark, *On the origin and bifurcations of stick-slip oscillations*, Phys. D, 136 (2000), pp. 280–302.

[11] M. di Bernardo, C. J. Budd, and A. R. Champneys, *Grazing and border-collision in piecewise-smooth systems: A unified analytical framework*, Phys. Rev. Lett., 86 (2001), pp. 253–256.

[12] M. di Bernardo, F. Garofalo, L. Glielmo, and F. Vasca, *Switchings, bifurcations, and chaos in DC/DC converters*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 45 (1998), pp. 133–141.

[13] M. Diehl, H. G. Bock, and E. Kostina, *An approximation technique for robust nonlinear optimization*, Math. Program., 107 (2006), pp. 213–230.

[14] M. Diehl, J. Gerhard, W. Marquardt, and M. Mönnigmann, *Numerical solution approaches for robust nonlinear optimal control problems*, Comput. Chem. Eng., 32 (2008), pp. 1287–1300.

[15] V. Donde and I. A. Hiskens, *Dynamic performance assessment: Grazing and related phenomena*, IEEE Trans. Power Syst., 20 (2005), pp. 1967–1975.
[16] V. Donde and I. A. Hiskens, *Shooting methods for locating grazing phenomena in hybrid systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 16 (2006), pp. 671–692.
[17] N. H. El-Farra and P. D. Christofides, *Integrating robustness and constraints in control of nonlinear processes*, Chem. Eng. Sci., 56 (2001), pp. 1841–1868.
[18] N. H. El-Farra and P. D. Christofides, *Bounded robust control of constrained multivariable nonlinear processes*, Chem. Eng. Sci., 58 (2003), pp. 3025–3047.
[19] W. F. Feehery and P. I. Barton, *Dynamic optimization with state variable path constraints*, Comput. Chem. Eng., 22 (1998), pp. 1241–1256.
[20] W. Fleming, *Functions of Several Variables*, Undergraduate Texts in Mathematics, Springer Verlag, New York, Heidelberg, Berlin, 1977.
[21] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *User's Guide for NPSOL, Version* 4.0, Tech. report, Systems Optimization Laboratory, Stanford University, Stanford, CA, 1986.
[22] S. Gros, B. Srinivasan, and D. Bonvin, *Robust predictive control based on neighboring extremals*, J. Process Control, 16 (2006), pp. 243–253.
[23] J. Hahn, M. Mönnigmann, and W. Marquardt, *A method for robustness analysis of controlled nonlinear systems*, Chem. Eng. Sci., 59 (2004), pp. 4325–4338.
[24] K. P. Halemane and I. E. Grossmann, *Optimal design under uncertainty*, AIChE J., 29 (1983), pp. 425–433.
[25] R. Hannemann and W. Marquardt, *Fast computation of the Hessian of the Lagrangian in shooting algorithms for dynamic optimization*, in Proceedings of the 8th International IFAC Symposium on Dynamics and Control of Process Systems, Cancun, Mexico, 2007, pp. 105–110.
[26] J. V. Kadam, W. Marquardt, B. Srinivasan, and D. Bonvin, *Optimal grade transition in industrial polymerization processes via NCO tracking*, AIChE J., 53 (2007), pp. 627–639.
[27] A. C. Kokossis and C. A. Floudas, *Stability in optimal design: Synthesis of complex reactor networks*, AIChE J., 40 (1994), pp. 849–861.
[28] G. Kreisselmeier and R. Steinhauser, *Application of vector performance optimization to a robust-control loop design for a fighter aircraft*, Internat. J. Control, 37 (1983), pp. 251–284.
[29] J. H. Lee and Z. H. Yu, *Worst-case formulations of model predictive control for systems with bounded parameters*, Automatica J. IFAC, 33 (1997), pp. 763–781.
[30] S. T. Li and L. Petzold, *Software and algorithms for sensitivity analysis of large-scale differential algebraic systems*, J. Comput. Appl. Math., 125 (2000), pp. 131–145.
[31] C. Loeblein and J. D. Perkins, *Structural design for on-line process optimization:* I. *Dynamic economics of mpc*, AIChE J., 45 (1999), pp. 1018–1029.
[32] L. Magni, G. De Nicolao, R. Scattolini, and F. Allgöwer, *Robust model predictive control for nonlinear discrete-time systems*, Internat. J. Robust Nonlinear Control, 13 (2003), pp. 229–246.
[33] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, *Constrained model predictive control: Stability and optimality*, Automatica J. IFAC, 36 (2000), pp. 789–814.
[34] M. J. Mohideen, J. D. Perkins, and E. N. Pistikopoulos, *Optimal design of dynamic systems under uncertainty*, AIChE J., 42 (1996), pp. 2251–2272.
[35] M. J. Mohideen, J. D. Perkins, and E. N. Pistikopoulos, *Robust stability considerations in optimal design of dynamic systems under uncertainty*, J. Process Control, 7 (1997), pp. 371–385.
[36] M. Mönnigmann and W. Marquardt, *Normal vectors on manifolds of critical points for parametric robustness of equilibrium solutions of ODE systems*, J. Nonlinear Sci., 12 (2002), pp. 85–112.
[37] M. Mönnigmann and W. Marquardt, *Steady state process optimization with guaranteed robust stability and robust feasibility*, AIChE J., 49 (2003), pp. 3110–3126.
[38] M. Mönnigmann and W. Marquardt, *Steady state process optimization with guaranteed robust stability and flexibility: Application to HDA reaction section*, Ind. Eng. Chem. Res., 44 (2005), pp. 2737–2753.
[39] M. Mönnigmann, W. Marquardt, C. H. Bischof, T. Beelitz, B. Lang, and P. Willems, *A hybrid approach for efficient robust design of dynamic systems*, SIAM Rev., 49 (2007), pp. 236–254.
[40] M. B. Monagan, K. O. Geddes, K. M. Heal, G. Labahn, S. M. Vorkoetter, J. McCarron, and P. DeMarco, *Maple* 9 *Advanced Programming Guide*, Waterloo Maple, Waterloo, Canada, 2003.
[41] Z. K. Nagy and R. D. Braatz, *Robust nonlinear model predictive control of batch processes*, AIChE J., 49 (2003), pp. 1776–1786.

[42] A. B. Nordmark, *Nonperiodic motion caused by grazing-incidence in an impact oscillator*, J. Sound Vibration, 145 (1991), pp. 279–297.
[43] C. Panjapornpon, M. Soroush, and W. D. Seider, *Model-based controller design for unstable, nonminimum-phase, nonlinear processes*, Ind. Eng. Chem. Res., 45 (2006), pp. 2758–2768.
[44] L. Perko, *Differential Equations and Dynamical Systems*, 2nd ed., Springer-Verlag, New York, 1996.
[45] V. Sakizlis, J. D. Perkins, and E. N. Pistikopoulos, *Parametric controllers in simultaneous process and control design optimization*, Ind. Eng. Chem. Res., 42 (2003), pp. 4545–4563.
[46] V. Sakizlis, J. D. Perkins, and E. N. Pistikopoulos, *Recent advances in optimization-based simultaneous process and control design*, Comput. Chem. Eng., 28 (2004), pp. 2069–2086.
[47] M. Schlegel, W. Marquardt, R. Ehrig, and U. Nowak, *Sensitivity analysis of linearly implicit differential-algebraic systems by one-step extrapolation*, Appl. Numer. Math., 48 (2004), pp. 83–102.
[48] P. O. M. Scokaert and D. Q. Mayne, *Min-max feedback model predictive control for constrained linear systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 1136–1142.
[49] J. Tolsma and P. I. Barton, *Daepack: An open modeling environment for legacy models*, Ind. Eng. Chem. Res., 39 (2000), pp. 1826–1839.
[50] J. Unger, A. Kroner, and W. Marquardt, *Structural-analysis of differential-algebraic equation systems: Theory and applications*, Comput. Chem. Eng., 19 (1995), pp. 867–882.
[51] A. Uppal, W. H. Ray, and A. B. Poore, *On the dynamic behavior of continuous stirred tank reactors*, Chem. Eng. Sci., 29 (1974), pp. 967–985.
[52] V. S. Vassiliadis, E. B. Canto, and J. R. Banga, *Second-order sensitivities of general dynamic systems with application to optimal control problems*, Chem. Eng. Sci., 54 (1999), pp. 3851–3860.
[53] Y. J. Wang and J. B. Rawlings, *A new robust model predictive control method* I: *Theory and computation*, J. Process Control, 14 (2004), pp. 231–247.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 491–509
© 2008 Society for Industrial and Applied Mathematics

Wave Formation through the Interactions between Clustered States and Local Coupling in Arrays of Neural Oscillators*

Fatma Gürel Kazanci† and Bard Ermentrout‡

Abstract. We study a system of coupled oscillators with global inhibition and local gap junction coupling. The coupling functions are derived from a biological system in the limit of weak coupling. With global inhibition, the system evolves to a clustered state, while with local gap junctions, waves and synchrony are the only attractors. Increasing gap junction strength from zero destroys the clustered state leaving a complex pattern. Decreasing gap junction strength from a high value results in the loss of stability of waves to a Hopf bifurcation and results in periodically modulated waves. We present analytical results along with numerical simulations.

Key words. electrical coupling, inhibition, oscillations, waves, two-cluster solution, stability

AMS subject classifications. 37G05, 92C20

DOI. 10.1137/070699147

1. Introduction.

Globally coupled oscillators have been studied extensively [1, 3, 13, 14, 15]. Such systems arise in natural and physical environments such as groups of flashing fireflies, chirping crickets, coupled multimode lasers, and networks of oscillatory neurons. The behavior of generally coupled limit cycles is a daunting task; thus, simplifications are usually necessary. Reducing each oscillator to a single variable, the phase, is the simplest way to produce an analytically tractable model for systems of coupled oscillators. Through the method of averaging, when the oscillators are nearly identical and the coupling between them is sufficiently small, we can reduce a general system to an equation of the form

$$\theta_j' = \omega_j + H_j(\theta_1 - \theta_j, \ldots, \theta_N - \theta_j), \quad j = 1, \ldots, N. \tag{1.1}$$

This is the most common form for the study of patterns of coupled oscillators. Among the commonly studied classes of patterns are synchrony, in which $\theta_j = \theta_k$ for all k, j, and clustering, in which the oscillators are divided into m groups; within each group, all oscillators are synchronous, but there is no synchrony between groups. In structured networks, waves and many other patterns are possible.

While the analysis for globally coupled identical systems is simpler than that of more structured networks, there are still many complex behaviors which arise. For example, as the parameters characterizing the oscillators change, it is possible to go from synchrony to clustered states and to asynchrony [2, 7, 8, 12].

*Received by the editors August 1, 2007; accepted for publication (in revised form) by T. Kaper January 10, 2008; published electronically May 2, 2008.
http://www.siam.org/journals/siads/7-2/69914.html
†Corresponding author. Department of Biology, Emory University, Atlanta, GA 30322 (fgurelk@emory.edu).
‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (bard@math.pitt.edu).

In a recent paper, we studied a spatially structured network of coupled neural oscillators in which there was local synchronizing coupling (mediated by electrical or gap junction coupling) and long range "desynchronizing" coupling mediated by synaptic inhibition [9]. The motive for this work is the appearance of traveling waves and synchronous oscillations in the olfactory lobe of the garden slug [6]. The neurons which generate these patterns are coupled with both gap junctions and synaptic inhibition. Starting with a synchronous locally coupled network, we showed that the addition of global inhibitory coupling leads to a symmetry-breaking bifurcation and ultimately to traveling waves.

In this paper, we consider the same system (local synchronization and long-range desynchronization) from a different perspective. Starting with a globally coupled network of oscillators, we introduce local synchronizing coupling and ask what kinds of behaviors arise. Our work for the inhibition only case is motivated by Hansel, Mato, and Meunier [8]. In their paper, they show a heteroclinic connection between unstable two-cluster states. The coupling functions they consider are different from the ones we used in our model. We derive the coupling functions via an approximation using the biophysical model. We alter the long range coupling function by changing the sign of one of the Fourier coefficients. This allows the globally coupled network to have stable clustered states. We then add nearest neighbor synchronizing coupling and study the resulting dynamics. Since our main interest lies in pattern formation, we thought it would be a good idea to investigate these stable patterns for the full model. Local coupling (as opposed to all-to-all) requires that we specify a geometry of the network; here we consider the simplest case, a one-dimensional ring of oscillators.

In section 2, we describe the model with only all-to-all coupling and provide a stability analysis for the two-cluster state for our choice of coupling function. In section 3, we add local gap junctions and arrange the oscillators on a ring. We show that for sufficiently strong gap junctions, there are stable traveling waves and that as the gap junction coupling decreases, there is a loss of stability of the traveling waves. In section 4, we start with the clustered states and numerically analyze the bifurcations of the clustered states. In particular, for a structured network, the ordering of the oscillators matters and there are many arrangements for a clustered state. We show that these have different stability behavior when local synchronizing coupling is added and that many new patterns bifurcate. There is a great deal of multistability in such networks.

2. Model and theory of clustering. Our goal in this paper is to study the behavior of a network of neural oscillators which are coupled with long-range inhibition (desynchronizing) and short-range gap junctions (synchronizing). The motive for our work comes from a detailed biophysical model for wave generation in the olfactory lobe of the garden slug [6]. In [9] we used the theory of averaging to reduce the biophysical model to a phase model of the form (1.1). As there are two different types of coupling, it is convenient to write (1.1) as

$$\frac{d\theta_j}{dt} = \omega + g_{syn} \sum_{k=1}^{N} c_{jk} H_{syn}(\theta_k - \theta_j) + g_{gap} \sum_{k=1}^{N} d_{jk} H_{gap}(\theta_k - \theta_j), \tag{2.1}$$

where g_{syn}, g_{gap} are the overall coupling strengths of synaptic inhibition and the gap junction coupling, c_{jk}, d_{jk} are the connectivity matrices, and $H_{syn}(\theta)$, $H_{gap}(\theta)$ are the coupling functions obtained by averaging the dynamics of the biophysical model. As these are computed

numerically, we approximate them by the first few terms of their Fourier series:

$$H_{syn}(x) = 35 + 200\cos(x) + 32\cos(2x) - 95\sin(x) + 5\sin(2x),$$

$$H_{gap}(x) = 87 - 50\cos(x) - 37\cos(2x) + 295\sin(x) - 65\sin(2x).$$

In [9] the actual coefficient of $\sin 2x$ in $H_{syn}(x)$ is -5. The theory in the next subsection requires that at least one sine coefficient of H_{syn} be positive. The small change in $\sin 2x$ results in a new H_{syn} which is nearly indistinguishable from the original but which has a much richer repertoire in its dynamics. As we will need them in subsequent sections, we observe that $H'_{syn}(0) = -85$, $H'_{syn}(\pi) = 105$, $H'_{gap}(0) = 165$, and $H'_{gap}(\pi) = -425$.

Before continuing, it is useful to clarify what we mean by synchronizing and desynchronizing coupling by considering a pair of identical mutually coupled oscillators:

$$\frac{d\theta_1}{dt} = \omega + H(\theta_2 - \theta_1),$$
$$\frac{d\theta_2}{dt} = \omega + H(\theta_1 - \theta_2).$$

Letting $\phi = \theta_2 - \theta_1$, we see that

$$\frac{d\phi}{dt} = H(-\phi) - H(\phi)$$

so that $\phi = 0, \pi$ are always solutions. The synchronous solution is stable if and only if $H'(0) > 0$, while the antiphase solution, π, is stable if $H'(\pi) > 0$. For our choice of models, synchrony is stable and antiphase is unstable for gap junctions, while the opposite is true for the synaptic coupling case. Thus, we say that gap junctions are synchronizing and synaptic inhibition is desynchronizing.

Clustered solutions are generally found with all-to-all coupling so that in this paper we assume that $c_{jk} = 1/N$ and thus synaptic inhibition is global. This is the same assumption made in the previous paper. For gap junction coupling, there are many possible topologies that we could use. However, our motivation comes from a one-dimensional phenomenon, so we will arrange the oscillators in a line. To avoid boundary effects, most of the work is done in a one-dimensional ring of oscillators with nearest neighbor gap junction coupling. Thus, $d_{jk} = 1$ if $k = j \pm 1$ or $j = 1$, $k = N$ or $j = N$, $k = 1$ and $d_{jk} = 0$ otherwise. In the last section, we show that our results do not depend crucially on the periodicity of the boundary condition. However, the analysis in section 2.1 is much easier in this case.

To begin with, we need to study the all-to-all synaptically coupled clustered state. Thus, we now consider the case when $g_{gap} = 0$.

2.1. General theory. We start our analysis by revisiting the work in [8]. Consider a globally coupled network of identical phase oscillators given by the equation

$$\frac{d\theta_j}{dt} = \omega + \frac{g_{syn}}{N}\sum_{k=1}^{N} H_{syn}(\theta_k - \theta_j) \quad \text{for} \quad j = 1, \ldots, N. \tag{2.2}$$

Here, as in the rest of the paper, the frequency of individual oscillators is identical. The case where the oscillators have different intrinsic frequencies has also been studied, but we will not consider it now.

We are interested in clustered solutions to (2.2). We look at the case where m of the oscillators have phase $\theta_A = \Omega t$ and the remaining $N - m$ have phase $\theta_B = \Omega t + \phi$ with $\phi \in [0, 2\pi]$. By letting $p = \frac{m}{N}$, we can rewrite (2.2) as

$$\theta_A' = \omega + g_{syn}(pH_{syn}(0) + (1-p)H_{syn}(\theta_B - \theta_A)),$$
$$\theta_B' = \omega + g_{syn}(pH_{syn}(\theta_A - \theta_B) + (1-p)H_{syn}(0)),$$

which then becomes

$$\Omega = \omega + g_{syn}(pH_{syn}(0) + (1-p)H_{syn}(\phi)),$$
$$\Omega = \omega + g_{syn}((1-p)H_{syn}(0) + pH_{syn}(-\phi)).$$

Subtracting the first equation from the second, we get

$$0 = g_{syn}((1-p)H_{syn}(0) + pH_{syn}(-\phi) - pH_{syn}(0) - (1-p)H_{syn}(\phi)). \tag{2.3}$$

Now, we can solve for p:

$$p = \frac{H_{syn}(0) - H_{syn}(\phi)}{2H_{syn}(0) - H_{syn}(-\phi) - H_{syn}(\phi)} \equiv F(\phi). \tag{2.4}$$

Note that $F(\pi) = 1/2$. Given H_{syn} and ϕ, we can determine the size of the clusters according to (2.4).

So far, we have established the existence of two-cluster states. One could also look at other n-cluster states. Thus, we remark that stable clustered states with more than two clusters will not arise with our particular choice of H_{syn} due to the lack of higher order Fourier terms. We limit the present analysis to two-cluster states and study their stability. To do this, we look at the linearized system for the two-cluster solution. Letting $\theta_j = \Omega t + y_j$, where the y_j's are small perturbations, we get

$$\frac{dy_j}{dt} = \frac{g_{syn}}{N}\sum_{k=1}^{N} H_{syn}'(\theta_k - \theta_j)(y_k - y_j) \quad \text{for} \quad j = 1, \ldots, N, \tag{2.5}$$

which can be written as two separate equations:

$$y_j' = \frac{g_{syn}}{N}\sum_{k=1}^{m} H_{syn}'(0)(y_k - y_j) + \frac{g_{syn}}{N}\sum_{k=m+1}^{N} H'(\phi)(y_k - y_j) \quad \text{for} \quad j = 1, \ldots, m,$$
$$y_j' = \frac{g_{syn}}{N}\sum_{k=1}^{m} H'(-\phi)(y_k - y_j) + \frac{1}{N}\sum_{k=m+1}^{N} H'(0)(y_k - y_j) \quad \text{for} \quad j = m+1, \ldots, N.$$

Letting $\alpha = H'_{syn}(0)$, $\beta = H'_{syn}(\phi)$, $\gamma = H'_{syn}(-\phi)$, $\xi = -(m-1)\alpha - (N-m)\beta$, and $\nu = -m\gamma - (N-m-1)\alpha$, we write

$$A = \begin{bmatrix} \xi & \alpha & \cdots & \alpha \\ \alpha & \xi & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha & \xi \end{bmatrix},$$

$$B = \beta \mathbf{1}_{m\times(N-m)},$$

$$C = \gamma \mathbf{1}_{(N-m)\times m}, \quad \text{and}$$

$$D = \begin{bmatrix} \eta & \alpha & \cdots & \alpha \\ \alpha & \eta & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha & \eta \end{bmatrix}.$$

We can write (2.5) as

$$\mathbf{y}' = g_{syn}\mathbf{M}\mathbf{y}, \tag{2.6}$$

where $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^T$ and $M = \frac{1}{N}\begin{bmatrix} A & B \\ C & D \end{bmatrix}$.

In order to determine the stability of the two-cluster state, we need to find the eigenvalues of the matrix M. A similar analysis was done in [8]. We state the following theorem.

Theorem 2.1. *Let $\alpha = H'_{syn}(0)$, $\beta = H'_{syn}(\phi)$, and $\gamma = H'_{syn}(-\phi)$. The eigenvalues of the matrix M in (2.6) are*

$$\begin{aligned} \lambda_1 &= -p\alpha - (1-p)\beta, \\ \lambda_2 &= -p\gamma - (1-p)\alpha, \\ \lambda_3 &= -p\gamma - (1-p)\beta, \\ \lambda_4 &= 0, \end{aligned}$$

where $p = \frac{m}{N}$. The algebraic multiplicities of λ_1, λ_2, λ_3, and λ_4 are $m-1$, $N-m-1$, 1, and 1, respectively.

Proof. The zero eigenvalue, λ_4, is the eigenvalue corresponding to translation invariance along the limit cycle. The corresponding eigenvector is found as

$$v_4 = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T.$$

λ_1, λ_2 correspond to fluctuations within a cluster, whereas λ_3 corresponds to fluctuations between two clusters. The associated eigenvector for λ_3 has the form

$$v_3 = \begin{bmatrix} 1 & \cdots & 1 & x & \cdots & x \end{bmatrix}^T,$$

where $x \neq 1$. We need to satisfy $Mv_3 = \lambda_3 v_3$,

$$Mv_3 = \begin{bmatrix} (x-1)(1-p)\beta x \\ \vdots \\ (x-1)(1-p)\beta x \\ -(x-1)p\gamma \\ \vdots \\ -(x-1)p\gamma \end{bmatrix} = \lambda_3 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ x \\ \vdots \\ x \end{bmatrix},$$

which implies $(x-1)(1-p)\beta x = \lambda_3$ and $-(x-1)p\gamma = \lambda_3 x$. We solve for x to get $x = \frac{-p\gamma}{(1-p)\beta}$ and $\lambda_3 = -p\gamma - (1-p)\beta$.

We next claim that the eigenvector v_1 associated with λ_1 has the form $[\nu_1, \mathbf{0}]^T$, where ν_1 is in the null-space of C and $A\nu_1 = \lambda_1\nu_1$. We write A as

$$A = \begin{bmatrix} \xi & \alpha & \cdots & \alpha \\ \alpha & \xi & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha & \xi \end{bmatrix} = \begin{bmatrix} \xi-\alpha & 0 & \cdots & 0 \\ 0 & \xi-\alpha & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \xi-\alpha \end{bmatrix} + \underbrace{\begin{bmatrix} \alpha & \cdots & \alpha \\ \vdots & \ddots & \vdots \\ \alpha & \cdots & \alpha \end{bmatrix}}_{\hat{A}}.$$

If $\hat{\lambda} \in \sigma(\hat{A})$, then $\hat{\lambda} + \xi - \alpha \in \sigma(A)$. 0 is an $m-1$-fold eigenvalue for the matrix $\hat{A}$, which implies that the eigenvalues of A are $\xi - \alpha$ with corresponding eigenvector ν_1. Substituting ξ we get $\lambda_1 = -p\alpha - (1-p)\beta$. λ_2 can be found similarly. ■

An immediate corollary of Theorem 2.1 implies that the eigenvalues for the system given in (2.6) are 0, $g_{syn}\lambda_1$, $g_{syn}\lambda_2$, $g_{syn}\lambda_3$. For stability of the two-cluster solution, the real parts of the eigenvalues have to be negative. This translates into the following conditions for stability:

- $p\alpha + (1-p)\beta > 0$,
- $p\gamma + (1-p)\alpha > 0$,
- $p\gamma + (1-p)\beta > 0$.

2.2. Application. We apply the general results of the previous section to our particular choice of H_{syn}. Figure 1 shows the dependence of p on the phase difference for H_{syn}. Since this is a monotonic function, this implies that for each choice of $p \in [0,1]$ there is a unique phase, ϕ.

Next we examine the stability of clusters of different sizes by computing the eigenvalues using Theorem 2.1. Figure 2 shows the nontrivial eigenvalues as a function of ϕ, the phase difference between the two clusters. From the monotonic relationship in Figure 1, we get a unique value of p for each ϕ so that this figure translates into the amount of asymmetry tolerated in the cluster sizes. There is a very narrow window of stable phases centered around π, which means that the cluster size is very close to 1/2; that is, there are equal numbers of oscillators in each cluster. From now on, we restrict our attention to equal size clusters, so $p = 1/2$. For cluster sizes which deviate slightly from equality, we expect the remaining results to be similar.

Using Theorem 2.1, we can simplify the conditions for the stability of the two-cluster state when the clusters are equal since $\phi = \pi$, and so $\beta = \gamma$ and $p = 1/2$. The stability conditions we need to satisfy are

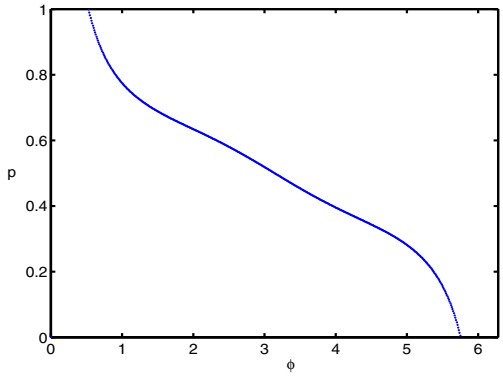


Figure 1. *Possible values for p are graphed as the phase difference, ϕ, varies from 0 to 2π.*

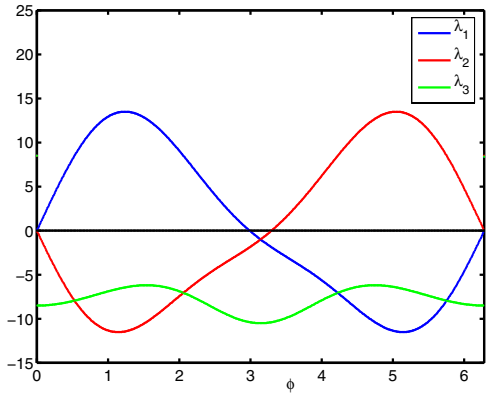


Figure 2. *Eigenvalues for the system in (2.6) are computed for all possible values of (p, ϕ).*

1. $H'_{syn}(0) + H'_{syn}(\pi) > 0$,
2. $H'_{syn}(\pi) > 0$, and
3. $H'_{syn}(0) < 0$.

Note that if the first and third conditions hold, then the second follows automatically. Recall that H_{syn} is supposed to be desynchronizing, which means that $H'_{syn}(0) < 0$ and $H'_{syn}(\pi) > 0$. Thus, desynchronizing long-range coupling is necessary for the formation of stable clusters. Before turning to our specific function H_{syn}, we interpret the stability conditions in terms of the Fourier coefficients of the function H_{syn}. Since the conditions are evaluated at the derivatives of the function at $0, \pi$, only the odd Fourier components come into play. Suppose that the odd components are a_n, $n = 1, 2, \ldots$. Then stability requires

$$\sum_{n>0} n a_n < 0,$$

$$\sum_{n>0} 2n a_{2n} > 0.$$

In particular, there must be at least one positive even Fourier sine coefficient. For our choice of $H_{syn}(\theta)$, $H'_{syn}(0) = -85$ and $H'_{syn}(\pi) = 105$, so that the three conditions hold and the two-cluster state is stable. We remark that the small change in the Fourier coefficient from the one derived from the biophysical model is thus absolutely necessary for the existence of clustered states.

3. Traveling waves in the presence of gap junctions. We now introduce local gap junction coupling and put our oscillators in a one-dimensional ring:

$$\frac{d\theta_j}{dt} = \omega + \frac{g_{syn}}{N}\sum_{k=1}^{N} H_{syn}(\theta_k - \theta_j) + g_{gap}[H_{gap}(\theta_{j+1} - \theta_j) + H_{gap}(\theta_{j-1} - \theta_j)], \tag{3.1}$$

where H_{gap} is the coupling function for the gap junction coupling between nearest neighbors with coupling strength g_{gap}.

In [9], we looked at a network of coupled oscillators with global inhibitory coupling and local (not necessarily nearest neighbor) gap junction coupling. We showed that a traveling wave was a stable solution for all the values of the parameters. In this section, we look at a similar network where the scope of the gap junction coupling extends only to nearest neighbors and the inhibitory coupling function is modified as noted in section 2. The coupling function H_{syn} is approximated from its original form by using the most dominant Fourier components. In order to get stable two-cluster solutions for the all-to-all network, we adjusted one of the coefficients in H_{syn}. Thus, in this section, we examine whether the small change in the coupling function affects the stability of waves and, if so, what kinds of bifurcations occur. A single traveling wave solution, $\theta_j = \Omega t + \frac{2\pi j}{N}$, satisfies (3.1) when

$$\Omega = \omega + \frac{g_{syn}}{N}\sum_{l=1}^{N} H_{syn}(\delta l) + g_{gap}(H_{gap}(\delta) + H_{gap}(-\delta)),$$

where $\delta = \frac{2\pi}{N}$. Given N, H_{syn}, H_{gap}, g_{syn}, and g_{gap}, we can determine Ω uniquely. If there are no cosine components in H_{syn} or H_{gap}, then $\Omega = \omega$. To determine stability, we linearize around the traveling wave. The linearized system is

$$\frac{dy_j}{dt} = \frac{g_{syn}}{N}\sum_{l=1}^{N} H'_{syn}(\delta l)(y_{j+l} - y_j) + g_{gap}[H'_{gap}(\delta)(y_{j+1} - y_j) + H'_{gap}(-\delta)(y_{j-1} - y_j)]. \tag{3.2}$$

The solutions for (3.2) are of the form $y_j = e^{\lambda_m t}e^{i\delta m j}$, where $j = 1, \ldots, N$, $m = 0, \ldots, N-1$. Solving for λ_m gives us

$$\lambda_m = \frac{g_{syn}}{N}\sum_{l=1}^{N} H'_{syn}(\delta l)(e^{i\delta m l} - 1) + g_{gap}[H'_{gap}(\delta)(e^{i\delta m} - 1) + H'_{gap}(-\delta)(e^{-i\delta m} - 1)]. \tag{3.3}$$

To determine the stability, we look at the real parts of λ_m. For $m = 0$, we have $\lambda_0 = 0$, which corresponds to translation invariance. For $m \neq 0$, we need to use the Fourier series expansions of H'_{syn} and H'_{gap}. For δ small, we can get a reasonably good estimate for λ_m when m is small as well. The local coupling contributes a term which is approximately

$$-g_{gap}H'_{gap}(0)\delta^2 m^2,$$

which is negative since $H'_{gap}(0) > 0$ by assumption. Writing

$$H'_{syn}(x) = \sum_n na_n \cos nx - b_n n \sin nx$$

for fixed m, we have to evaluate

$$S_m = \frac{1}{N}\sum_{l=1}^{N}\sum_n [na_n \cos(n\delta l) - nb_n \sin(n\delta l)][(\cos \delta ml - 1) + i \sin \delta ml].$$

The outer sum vanishes except when $n = m$ so that

$$S_m = \frac{1}{2}m(a_m - ib_m).$$

Thus we obtain an approximation for λ_m:

$$\lambda_m \approx \frac{g_{syn}}{2}n(a_m - ib_m) - g_{gap}H'_{gap}(0)\frac{4\pi^2 m^2}{N^2}. \tag{3.4}$$

Once again, the crucial players in the stability of waves are the sine coefficients, a_m. Recalling that stable clusters require that $a_{2n} > 0$ for some n, we see that it is always possible to destabilize traveling waves. We remark that traveling waves with higher wave numbers, $k > 1$, are the same as traveling waves with a wave number $k = 1$ for a ring of size N/k. For our choice of H_{syn}, $a_2 > 0$ so that the mode $m = 2$ destroys stability when either g_{syn} is large enough or g_{gap} is sufficiently small. The critical gap junction strength below which the wave is unstable is approximately

$$g_{gap}^{TW} \approx \frac{a_m}{8\pi^2 m H'_{gap}(0)}N^2 g_{syn}. \tag{3.5}$$

Since b_m is generally nonzero, the loss of stability generically occurs at a pair of imaginary eigenvalues implying a Hopf bifurcation should occur as the gap junction coupling decreases below the critical value. For our model $m = 2$, $a_2 = 5$, $H'_{gap}(0) = 165$, so for $N = 20$, $g_{gap}^{TW} \approx 0.0768 g_{syn}$, which compares well with the exact value $0.0746 g_{syn}$ obtained from (3.3). Large networks are less able to support traveling waves when there are clustering instabilities. Figure 3 shows the critical strength of g_{gap} above which the traveling wave is stable as a function of the length of the ring of oscillators. For this graph, $g_{syn} = 0.1$, so that our approximation is $g_{gap}^{TW} \approx 1.91 \times 10^{-5} N^2$.

In addition to traveling waves, an obvious solution is perfect synchrony, in which $\theta_j = \Omega t$ with

$$\Omega = \omega + g_{syn}H_{syn}(0).$$

(Note that for gap junction coupling, $H_{gap}(0) = 0$.) Stability is determined by linearizing about the synchronous solution; the methods used for traveling waves lead to the following expression for the eigenvalues:

$$\nu_m = \frac{g_{syn}H'_{syn}(0)}{N}\sum_{l=1}^{N}(e^{i\delta ml} - 1) + 2g_{gap}H'_{gap}(0)[\cos(m\delta) - 1],$$

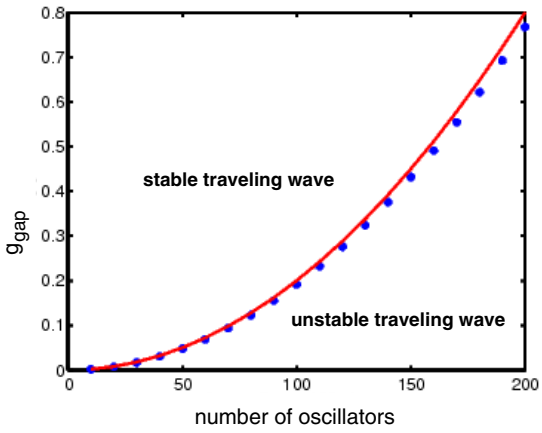


Figure 3. *Relation between g_{gap} and N is shown by the blue dots. The red line is the curve that fits the data with $y = cN^2$ and $c = 1.91 \times 10^{-5}$. g_{syn} is fixed at 0.1 throughout the computation. A Hopf bifurcation is observed at the marked g_{gap} values as $N =$ "number of oscillators" changes. The emerging traveling wave is stable in the region above the curve and unstable below it.*

which leads to $\nu_0 = 0$ and

$$\nu_m = -g_{syn}H'_{syn}(0) + 2g_{gap}H'_{gap}(0)[\cos(2\pi m/N) - 1].$$

Recall that $H'_{syn}(0) < 0$ and $H'_{gap}(0) > 0$. Thus, the first term is always positive. The second term is always negative and is least negative when $m = 1$, so this determines the minimal value of g_{gap} for stable synchrony:

$$g^S_{gap} = \frac{-H'_{syn}(0)}{H'_{gap}(0)} \frac{g_{syn}}{2[1 - \cos(2\pi/N)]}.$$

For N large, we can rewrite this expression as

$$g^S_{gap} \approx \frac{-H'_{syn}(0)}{H'_{gap}(0)} \frac{g_{syn}}{4\pi^2} N^2 \tag{3.6}$$

so that, like the traveling wave, the critical coupling strength grows as N^2. Unlike the traveling wave, however, instability of synchrony does not depend on the details of the Fourier coefficients of $H_{syn}(x)$; in particular, the ability to form clusters is irrelevant. For our model, with $g_{syn} = 0.1$, we find $g^S_{gap} \approx 1.31 \times 10^{-3}N^2$. Note that the eigenvalues are real so that as g_{gap} falls below the critical value a zero eigenvalue is crossed. By symmetry, a pitchfork bifurcation will occur. In [9], we computed this branch of solutions by calculating the normal form. We remark that the traveling wave is "more stable" than synchrony in the sense that it is able to tolerate smaller gap junction coupling than synchrony is.

4. Numerical results. In section 2, we showed that there were stable two-cluster solutions for $g_{gap} = 0$. Since they are asymptotically stable, we expect the two-cluster states to stably persist for sufficiently small values of g_{gap} by the implicit function theorem. In section 3, we

Table 1

Table of all possible two-cluster configurations for a network of 10 oscillators in a ring. The values g_{gap}^C indicate the numerical value of the bifurcation parameter g_{gap}. Two-cluster solution becomes unstable for $g_{gap} > g_{gap}^C$.

Case number	Configuration	$g_{gap}^C \times 10^3$
Case 1	0000011111	2.759
Case 2	0000101111	1.172
Case 3	0000110111	1.099
Case 4	0000111011	1.099
Case 5	0001011101	0.7612
Case 6	0001011011	0.6654
Case 7	0001101101	0.6656
Case 8	0001101011	0.6300
Case 9	0001110101	0.7551
Case 10	0001110011	1.358
Case 11	0011001011	0.7279
Case 12	0011001101	0.7266
Case 13	0011011001	0.7729
Case 14	0011010101	0.4997
Case 15	0010101101	0.5197
Case 16	0101010101	0.6535

demonstrated that there were stable traveling waves for sufficiently large values of g_{gap} and that synchrony was also stable with strong enough gap junction coupling. Thus, our goal here is to use numerical methods to fill in the details as to how the clustered state changes as g_{gap} increases and how traveling waves and synchrony evolve as g_{gap} decreases.

4.1. Bifurcations from clustered states. Traveling waves and synchrony in a one-dimensional ring are unambiguously defined. That is, we can assign phases to each of the N-oscillators in only one way, up to translation around the ring. However, the arrangement of a clustered solution on a ring is much more complicated. Suppose that there are $2m$ oscillators with m at phase 0 and m at phase π. Then, we can ask how many different (up to rotations, reflections, and translations) configurations there are. This question is equivalent to asking how many $2m$ black and white bead necklaces with m black beads there are, the answer to which can be found in The Online Encyclopedia of Integer Sequences (http://www.research.att.com/~njas/sequences/A005648). For $m = 1, \ldots, 10$, the number of such sequences is 1, 2, 3, 8, 16, 50, 133, 440, 1387, and 4752, respectively. Thus, for a twenty oscillator ring, there are 4752 ways to have a two-clustered solution with equal numbers in each cluster. For ten oscillators, it is a more reasonable number of 16. Thus, to analyze the evolution of clustered states as g_{gap} increases, it is necessary to compute the bifurcation for every possible configuration. Table 1 shows the critical values of the gap junction coupling for each of the 16 configurations in a ring of 10 oscillators when $g_{syn} = 0.1$. We use AUTO to compute the bifurcation diagram starting at a given configuration [4]. For example, starting with the configuration 0000011111 (Case 1) corresponding to the first 5 oscillators at zero phase and the next at π, we find that the pattern stably exists for $g_{gap} < 2.759 \times 10^{-3}$. Similarly, the pattern 0000110111 (Case 3) exists up to $g_{gap} = 1.099 \times 10^{-3}$. Figure 4 shows the two

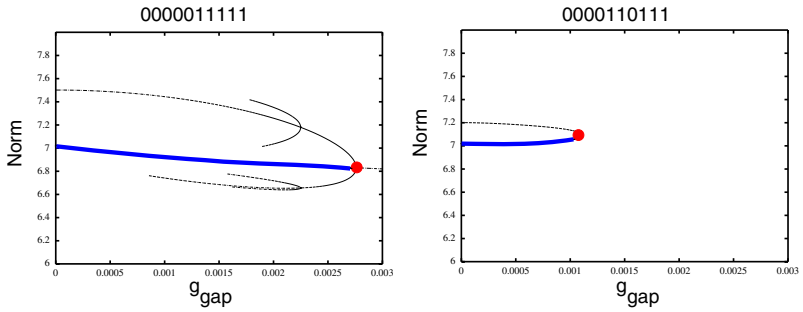


Figure 4. *Bifurcation diagram of two different clustered states as g_{gap} increases with $g_{syn} = 0.1$. Cluster* 0000011111 *loses stability at a subcritical pitchfork (red circle on left) while cluster* 0000110111 *loses stability at a saddle node (red circle on right). Stable clustered states are in blue.*

bifurcation diagrams as g_{gap} increases (stable clustered states are shown in blue). These patterns have quite different local bifurcation diagrams. The pattern in Case 1 undergoes a subcritical pitchfork bifurcation which spawns additional unstable branches. The pattern in Case 3 undergoes a saddle-node bifurcation. Thus, while the local bifurcations are different and while they happen at different values of g_{gap}, in all cases the clusters are disrupted by a small (but not infinitesimally small) amount of gap junction coupling. As can be seen in these two examples, the new branches that arise at the bifurcations do not result in any new stable solutions, so we have to look elsewhere to find out the behavior beyond the critical gap junction strength. One trend that we note is that there is a rough correlation between the minimal strength for disrupting clusters and the number of switches between the 0 and π phases. This is not a strict rule, as, for example, Case 16, which has the maximal number of 10 switches, is more resistant to disruption than is Case 14, which has 8 switches.

In the analysis of waves and synchrony, the critical coupling strength for stability depended strongly on N (cf. Figure 3). We can ask the same question for clustered states. The number of possible cases for clusters is of course rather large, and the possible clustered patterns vary tremendously. However, one pattern which appears for all even N is the pattern of m oscillators at 0 followed by m oscillators at π. For this pattern, we have computed the critical gap junction coupling for $N = 10$, 20, 40 to be, respectively, 0.002759, 0.003490, 0.003771. While there are slight differences, there are not the order of magnitude differences that are observed in the traveling wave case. Another simple pattern common to all networks with size $2m$ is the alternating one, where every other oscillator has phase 0 and the remainder have phase π. For this case, we find that the critical gap junction coupling for $N = 10$, 20, 40 is, respectively, 6.535, 6.064, 5.975×10^{-4}. Thus, there is virtually no difference. Using an eigenvalue perturbation argument, it may be possible to estimate the strength of gap junctions needed to destabilize the clustered state, but due to the multiplicity of the eigenvalues (see section 2), we could at best hope to reduce the stability issue to an $(N/2 - 1)$-dimensional system. However, we can determine the stability of the clustered state in the presence of gap junctions when both gap and synaptic coupling are all-to-all:

$$\theta_j' = \omega + \frac{1}{N}\sum_{k=1}^{N} \left(g_{syn} H_{syn}(\theta_k - \theta_j) + g_{gap} H_{gap}(\theta_k - \theta_j)\right).$$

We can replace H_{syn} in Theorem 2.1 with the combined H:

$$H(x) = g_{syn}H_{syn}(x) + g_{gap}H_{gap}(x).$$

This leads to

$$g_{gap}^{C} = -g_{syn} \min\left\{ \frac{H'_{syn}(0)}{H'_{gap}(0)}, \frac{H'_{syn}(\pi)}{H'_{gap}(\pi)}, \frac{H'_{syn}(0)+H'_{syn}(\pi)}{H'_{gap}(0)+H'_{gap}(\pi)} \right\}. \tag{4.1}$$

For our model, $g_{gap}^{C} = 0.07692\, g_{syn}$. This overestimates the values in the table, but it is at least in the right ballpark and does not depend on N as do waves and synchrony.

4.2. Bifurcations from traveling wave states. As we showed in section 3, as g_{gap} decreases, the traveling wave solution loses stability at a Hopf bifurcation, so that we expect to see oscillations below a critical gap junction strength. Our estimates in this and the previous sections show that the critical gap junction strength for the break-up of clusters is largely independent of N but that the loss of stability of the wave depends strongly on N. Thus, we expect to see qualitative changes in the nature of the bifurcations as N changes. We start with our smallest example, $N = 10$. In this and subsequent figures, we plot the solutions relative to oscillator 1. Thus, clusters, synchrony, and traveling waves are all fixed points; any periodic solutions to the relative phases represent quasi-periodic solutions to the full equations. Figure 5A superimposes the bifurcation diagram for cluster #1 in Table 1 along with the diagram for the traveling wave. As predicted by the analysis in the previous section, the traveling wave (green) loses stability at a Hopf bifurcation (red dot). For this particular model, the bifurcation is supercritical and leads to a branch of periodic solutions (red). Figure 5B shows that along these branches (labeled B) the waves are periodically modulated. As g_{gap} decreases, the modulation becomes deeper and the period longer. At a critical value of g_{gap}, the primary branch of periodic solutions undergoes a pitchfork and two new periodic solutions arise (magenta). Time series from oscillators 5, 10 (Figure 5C) show that there are slight differences in the two periodic orbits. As $g_{gap} \to 0^+$, these solutions appear to terminate on an unstable fixed point. Starting with initial data very close to this fixed point, solutions evolve to cluster #1 in our table. The final state is shown in panel E. The cluster state (shown in blue in Figure 5A) is bistable with the pure traveling wave as well as the periodically modulated traveling wave.

Figure 6 shows the bifurcation diagram for $N = 20$. Like the $N = 10$ case, the clustered state (blue) loses stability at roughly the same value of g_{gap}, while the wave (green) loses stability for a much larger value of g_{gap} as predicted by the theory. Thus, bistability is restricted to the cluster and the modulated waves (red), in contrast to $N = 10$, where there is bistability with respect to the traveling wave as well as to the modulated wave. For $N = 20$ and, likewise, for $N = 40$, the branch of modulated waves appears to terminate at a finite value of g_{gap} on a homoclinic. Initial conditions, near the modulated wave below the critical gap junction strength seem to be attracted to the cluster solution corresponding to the analogue of case #1 for $N = 10$, the "simplest" arrangement of clusters.

We close this section by remarking that, in addition to a single traveling wave, there are waves which go through multiple cycles. For example, when $N = 20$, there is a stable double wave, $\theta_j = \Omega t + 4\pi j/20$, and these waves go through similar sequences of bifurcations as the single wave with $N = 10$.

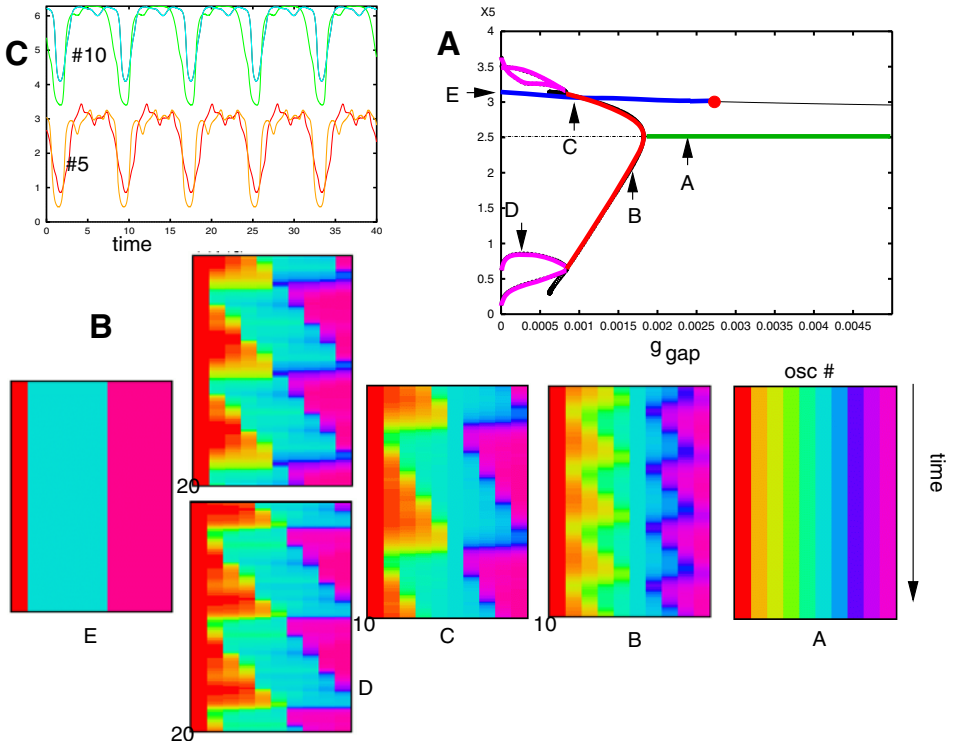


Figure 5. (A) *Bifurcation diagram for $N = 10$ oscillators as g_{gap} decreases. Line starting at E (in blue) is the curve of solutions starting at cluster #1. The green curve through A is the branch of traveling wave solutions. This loses stability at a Hopf bifurcation leading to a branch of periodic solutions (red). This undergoes a pitchfork bifurcation spawning a pair of limit cycles (magenta) which extend to g_{gap} close to 0.* (B) *Sample solutions along the diagram. Oscillators are arranged across and time increases downward.* (C) *Trajectories of $\theta_{5,10}(t)$ on the two different branches of periodic solutions. Green and orange represent one solution and red and blue the other.*

4.3. Bifurcations from synchrony. In our previous paper [9] we showed that as the ratio between the synaptic and gap junction coupling increased, the synchronous state underwent a bifurcation to a patterned state which with further increases disappeared resulting in traveling waves. Here, because of a small change in the synaptic coupling function (needed to stabilize the clustered state), the bifurcations are more complicated. As we showed above, synchrony is stable for g_{gap} sufficiently large. Due to the symmetry of the ring, it is difficult to use numerical continuation to follow bifurcations from the synchronous state. However, we can pick one of the patterned states as a starting point and follow this. Figure 7 shows the result of such a calculation. Rather than draw the full bifurcation diagram as determined via numerical continuation, we divide the behavior into two parts. Figure 7A shows our starting point, a patterned state with $g_{gap} = 0.001$. Figure 7C shows the behavior of this state as g_{gap} is decreased. The patterned state (in blue) loses stability at a Hopf bifurcation (HB) and results in a supercritical branch of periodic solutions (red). These lose stability at a pitchfork bifurcation. Unlike Figure 5, the pitchfork is subcritical and the new solutions are unstable.

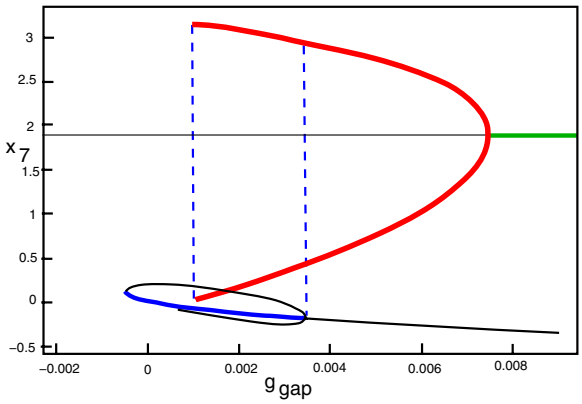


Figure 6. *Bifurcation diagram as g_{gap} varies for a system with 20 oscillators. The two-cluster state (solid blue curve) becomes unstable at $g_{gap}^{C} = 0.003491$ at a subcritical pitchfork (thin black lines). Traveling waves (solid green curve) become unstable through a Hopf bifurcation at $g_{gap}^{TW} = 0.007463$. The periodic orbits (solid red curves) are stable and do not appear to persist to $g_{gap} = 0$. There is a region where both the two-cluster state and periodic orbits are stable.*

Thus, we start near this bifurcating branch and decrease g_{gap} below the point of instability. The new solutions settle on a different branch of periodic solutions (magenta), one of which is shown in Figure 7B. The magenta branch terminates at an infinite period bifurcation (IP) as g_{gap} decreases and at a torus bifurcation (TR) as g_{gap} increases. There is a small region between TR and HB where there is bistability between periodic modulated solutions and the steady patterned solution. That is, patterns like those in Figure 7A, B can stably coexist. An attempt to track the branch of solutions starting at Figure 7A for increasing g_{gap} led to extremely complex branching diagrams that were difficult to understand. Thus, for increasing g_{gap} we took a different tack. We fix g_{gap} and integrate the equations until a steady state is reached, then we increment g_{gap} again, and so on. The result of this is plotted in Figure 7D; the phase of each oscillator is shown as g_{gap} ranges from 0.001 to 0.201 in two hundred steps. The phases all merge to the synchronous solution at $g_{gap} \approx 0.14$. Every solution on this branch looks qualitatively like Figure 7A. A similar picture is obtained with $N = 20$.

4.4. Pattern selection. Suppose that the gap junctions are sufficiently strong to support traveling waves and synchrony. A natural question is which of these patterns are selected. Given that synchrony requires stronger gap junction coupling than traveling waves, we expect traveling waves are more likely. There are many types of traveling waves: waves can travel in both the positive and negative directions or there can be higher wave number traveling waves, e.g., two cycles per ring. In order to get some idea of the basin of attraction, we have set $g_{syn} = 0.1$ and $g_{gap} = 0.6$ for a ring of 20 oscillators. (Note that synchrony is stable for g_{gap} larger than about 0.54.) For these values, synchrony and several types of traveling waves are all stable. We repeated 300 simulations with random initial data and found that 109 of the simulations or roughly one third of these initial conditions led to synchrony. The remainder of steady solutions were traveling waves of one type or another. Thus, it appears that waves are favored over the synchronous solutions but that synchrony still has a reasonably robust basin of attraction.

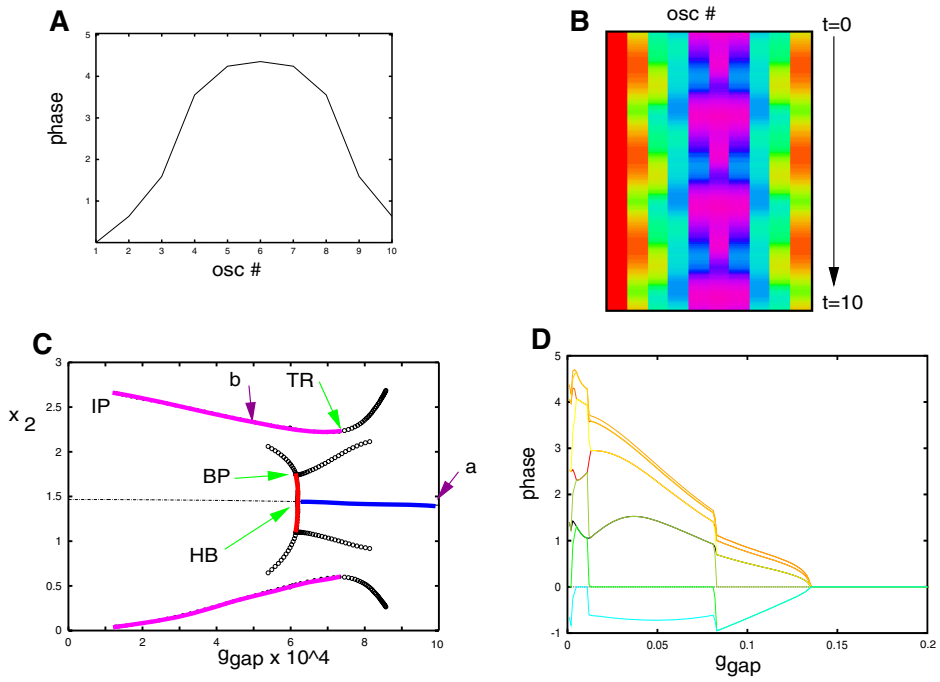


Figure 7. *Behavior of the patterned states as g_{gap} varies.* (A) *Patterned state with $g_{gap} = 0.001$.* (B) *Oscillatory state with $g_{gap} = 0.0005$.* (C) *Starting with the patterned state (blue) shown in* (A), *g_{gap} is decreased leading to a Hopf bifurcation (HB) and a branch of periodic solutions (red) which loses stability at a pitchfork (BP). In addition, there is another branch of periodic solutions (upper and lower magenta curves) that disappears for low values of g_{gap} at an infinite period bifurcation (IP) and at a torus bifurcation (TR) as g_{gap} increases. Points labeled a, b correspond to solutions in panes A, B, respectively.* (D) *Increasing g_{gap} beyond* 0.001 *leads to a cascade of steady states (all oscillator phases are plotted), terminating with the synchronous solution at $g^S_{gap} \approx 0.14$.*

4.5. Nonperiodic domains. The reader might rightfully ask how much of the behavior shown in this section is a consequence of the ring structure. To answer this, we can numerically analyze the bifurcations in a line of cells without periodic boundary conditions. That is, we replace the condition $\theta_0 = \theta_N$ with the condition $\theta_0 = \theta_2$ and the condition $\theta_{N+1} = \theta_1$ with $\theta_{N+1} = \theta_{N-1}$. Biologically, the array configuration is more realistic than the periodic domain considered earlier. Also, in this geometry, the dynamics is remarkably simple; cf. Figure 8. The synchronous solution is lost when $g_{gap} \approx 1.89$ (larger than the value on a periodic domain) via a pitchfork bifurcation. Two stable branches emerge leading to patterned states. Both of the states emerging result in monotonic (as a function of oscillator) relative phases—partial traveling waves. Examples are shown in Figure 8B. As g_{gap} decreases, these come to resemble traveling waves on the ring and nearly cover the entire range between 0 and 2π. Note that the two branches correspond to rightward and leftward waves. At roughly the same value of g_{gap}, the two branches lose stability via a Hopf bifurcation resulting in patterns similar to those in Figure 5, shown here in Figure 8C. These periodic patterns persist down to $g_{gap} \approx 0.0009$, beyond which we could no longer follow them. The period gets quite large and much time is spent near clustered solutions. For smaller g_{gap}, only clustered solutions remain.

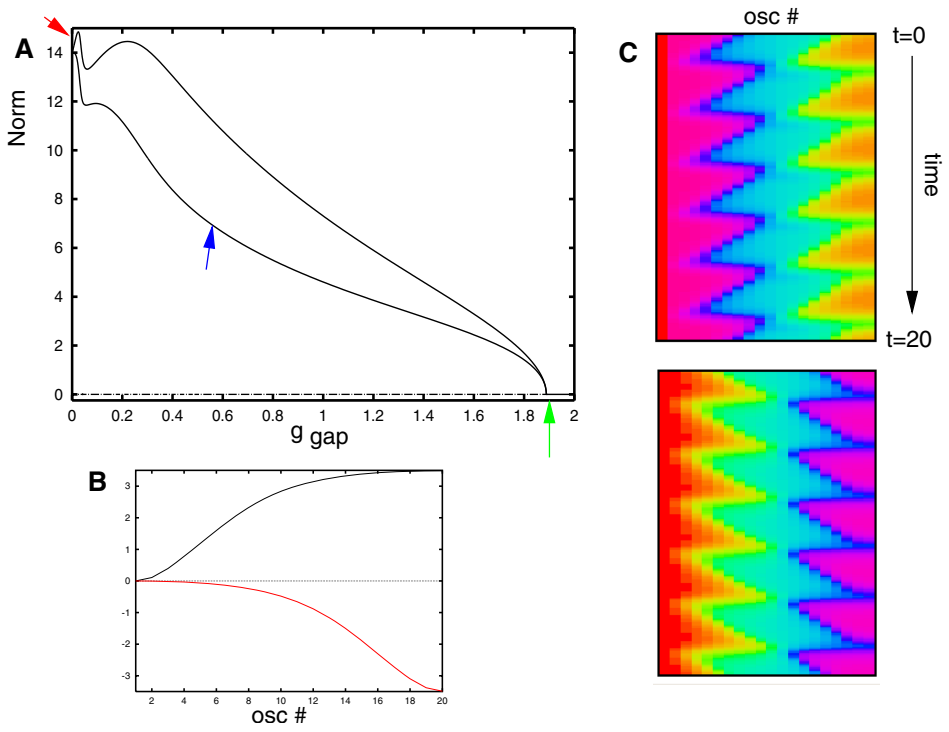


Figure 8. (A) *Bifurcation for* 20 *oscillators arranged in a line as* g_{gap} *decreases. A pair of solutions bifurcates from synchrony (green arrow) at* $g_{gap} \approx 1.89$*, persists until* $g_{gap} \approx 0.008$*, where it loses stability to a Hopf bifurcation. The periodic orbits disappear at* $g_{gap} \approx 0.0009$ *leaving clustered states.* (B) *Relative phases when* $g_{gap} = 0.5$ *(blue arrow).* (C) *Periodic solution* $g_{gap} = 0.006$ *(red arrow).*

If one returns to the diagram in Figure 7, we can see that the patterned states bifurcating from the ring resemble a pair of leftward and rightward moving waves glued together in the middle. Furthermore, the triangle-like periodic patterns are analogous to patterns formed by "gluing" together the pair of periodic patterns shown in Figure 8C.

5. Discussion. This paper continues the analysis of networks of oscillators coupled via a combination of gap junctions and synaptic inhibition. Here, our choice of coupling functions was motivated by a reduction using averaging of a biophysically based model for oscillations in the slug brain. We altered one parameter in the synaptic coupling such that a purely synaptically coupled network was able to produce clustered (rather than asynchronous) solutions. Thus, the present work addresses an interesting mathematical as well as biologically relevant problem: how do clustered states interact with local synchronizing coupling? Surprisingly, the behavior is considerably more complex than if the long-range synaptic coupling is only desynchronizing (as shown in [9]). Lewis and Rinzel have studied the interactions between gap junctions and synaptic inhibition in the case where both types of coupling encourage synchrony [11]. Similarly, Kopell and Ermentrout showed that gap junctions could prevent the suppression of oscillators which were coupled by inhibition which also synchronized [10]. Since synchronizing inhibition precludes the existence of clusters (Theorem 2.1 requires that

$H'_{syn}(0) < 0$, which implies synchrony is unstable), the results of these earlier studies cannot be applied to our model. Inhibition can be synchronizing or desynchronizing depending on the details of the model and the time course of the synapses; thus, this work could apply to systems more general than the slug brain. For example, we have computed the interaction functions for a recent biophysical model of cortical inhibitory neurons when the synaptic time course is short (2.5 msec) and the neuron is firing at 40 Hz [5]. Under reasonable physiological conditions, we compute synaptic interaction functions such that the symmetric two-cluster state is stable.

Interactions between local synchrony and clustered states are complex. We attempt to summarize our findings. For large enough gap junctions, synchrony and traveling waves are always stable, while for sufficiently small gap junctions, clustered states are stable. As gap junctions decrease, the synchronous solution bifurcates to a stationary patterned state, which then bifurcates to a modulated state. This quasiperiodic behavior disappears at sufficiently low gap junction coupling leaving only the clustered state. Conversely, starting with a patterned state, as the gap junctions increase, different patterned states emerge which are neither waves nor synchrony and result ultimately in synchrony. Traveling waves lose stability as gap junction coupling decreases, proceeding first to modulated waves and then to clusters. Thus, there are many regimes of multistability, and the waves and synchronous branches are largely independent.

We have restricted our attention to two-cluster states which have equal numbers of oscillators in each cluster and to networks on a line. It would be interesting to study other topologies with less symmetric arrays of clustering. Bifurcations starting from g_{gap} large (e.g., from synchrony or waves) are independent of the size of clusters and from our observations tend to terminate on simple symmetric clusters in which the oscillators are mainly segregated into a small number of groups. Thus, asymmetry would be expected only for very small gap junctions. Other networks, such as two-dimensional arrays, remain to be studied; these would be more relevant in cortical networks as opposed to the slug brain, which is effectively a one-dimensional network. One other possible extension of our work is to consider more complex coupling functions. This will introduce n-cluster solutions with $n \geq 2$. This study might be motivated by experimental and theoretical explanations of clustered solutions and their possible function.

REFERENCES

[1] P. Ashwin and J. W. Swift, *The dynamics of n weakly coupled oscillators*, J. Nonlinear Sci., 2 (1992), pp. 69–108.
[2] M. Banaji, *Clustering in globally coupled oscillators*, Dyn. Syst., 17 (2002), pp. 263–285.
[3] E. Brown, P. Holmes, and J. Moehlis, *Perspectives and Problems in Nonlinear Science: A Celebratory Volume in Honor of Larry Sirovich*, Springer-Verlag, New York, 2003, Ch. 5.
[4] E. Doedel, *AUTO: A program for the automatic bifurcation analysis of autonomous systems*, Congr. Numer., 30 (1981), pp. 265–284.
[5] A. Erisir, D. Lau, and C. S. Leonard, *Function of specific k(+) channels in sustained high-frequency firing of fast-spiking neocortical interneurons*, J. Neurophysiol., 82 (1999), pp. 2476–2489.
[6] B. Ermentrout, J. W. Wang, J. Flores, and A. Gelperin, *Model for transition from waves to synchrony in the olfactory lobe of limax*, J. Comput. Neurosci., 17 (2004), pp. 365–383.

[7] D. Golomb, D. Hansel, B. Shraiman, and H. Sompolinsky, *Clustering in globally coupled phase oscillators*, Phys. Rev. A, 45 (1992), pp. 3516–3530.
[8] D. Hansel, G. Mato, and C. Meunier, *Clustering and slow switching in globally coupled phase oscillators*, Phys. Rev. E, 48 (1993), pp. 3470–3477.
[9] F. G. Kazanci and B. Ermentrout, *Pattern formation in an array of oscillators with electrical and chemical coupling*, SIAM J. Appl. Math., 67 (2007), pp. 512–529.
[10] N. Kopell and B. Ermentrout, *Chemical and electrical synapses perform complementary roles in the synchronization of interneuronal networks*, Proc. Natl. Acad. Sci., 101 (2004), pp. 15482–15487.
[11] T. J. Lewis and J. Rinzel, *Dynamics of spiking neurons connected by both inhibitory and electrical coupling*, J. Comput. Neurosci., 14 (2003), pp. 283–309.
[12] K. Okuda, *Variety and generality of clustering in globally coupled oscillators*, Phys. D, 63 (1993), pp. 424–436.
[13] H. Sakaguchi, *Cooperative phenomena in coupled oscillator systems under external fields*, Progr. Theoret. Phys., 79 (1988), pp. 39–46.
[14] H. Sakaguchi and Y. Kuramoto, *A soluble active rotator model showing phase transitions via mutual entrainment*, Progr. Theoret. Phys., 76 (1986), pp. 576–581.
[15] S. H. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators*, Phys. D, 143 (2000), pp. 1–20.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 510–531
© 2008 Society for Industrial and Applied Mathematics

A Tracking Algorithm for Car Paths on Road Networks*

Gabriella Bretti† and Benedetto Piccoli†

Abstract. In this paper we introduce a computation algorithm to trace car paths on road networks, whose load evolution is modeled by conservation laws. This algorithm is composed of two parts: computation of solutions to conservation equations on each road and localization of car position resulting by interactions with waves produced on roads. Some applications and examples to describe the behavior of a driver traveling in a road network are shown. Moreover, a convergence result for wave front tracking approximate solutions, with BV initial data on a single road, is established.

Key words. conservation laws, discontinuous ordinary differential equations, finite difference schemes, fluid-dynamic models, traffic flow

AMS subject classifications. 35L65, 65L05, 34B45, 90B10, 90B20

DOI. 10.1137/070697768

1. Introduction.

Consider the Lighthill–Whitham and Richards traffic flow model [18, 19]:

$$\begin{cases} \partial_t \rho + \partial_x f(\rho) = 0, \\ \rho(0,x) = \bar{\rho}(x), \end{cases} \tag{1.1}$$

where $\rho = \rho(t,x)$ is the car density, with $\rho \in [0, \rho_{max}]$, $(t,x) \in \mathbb{R}^+ \times \mathbb{R}$, and $\bar{\rho}$ a suitable initial datum. The flux $f(\rho)$, assumed to be strictly concave, can be written as $f(\rho) = \rho v$, where the average velocity of cars v is assumed to be a smooth strictly decreasing function of the density ρ.

Suppose that a driver travels along a road, whose load is modeled by (1.1), being influenced by traffic along the road but without influencing it significantly. Then, the driver's position $x = x(t)$ can be obtained by solving the following Cauchy problem:

$$\begin{cases} \dot{x} = v(\rho(t,x)), \\ x(\bar{t}) = \bar{x}, \end{cases} \tag{1.2}$$

with $\bar{x}$ the position at initial time $\bar{t} > 0$. Notice that ρ is, in general, a discontinuous function. We look for numerical methods to find solutions to the problem (1.1)–(1.2), which was recently studied from a theoretical point of view by Colombo and Marson in [6, 7]. Note that the speed of the driver v in (1.2) need not be identical to the overall mean traffic speed in the flow in (1.1). Therefore, the assumptions could be significantly relaxed and generalized as in [6].

Fluid-dynamic models can describe macroscopic phenomena as shock formation and propagation. Since they can develop discontinuities in finite time even starting from smooth

*Received by the editors July 19, 2007; accepted for publication (in revised form) by M. Pugh January 18, 2008; published electronically May 2, 2008.
http://www.siam.org/journals/siads/7-2/69776.html
†Istituto per le Applicazioni del Calcolo "M.Picone," Rome, Italy, 00161 (g.bretti@iac.cnr.it, b.piccoli@iac.cnr.it).

initial data, the study of the analytical and numerical aspects is fundamental. In the papers [4, 5, 8, 13, 14], some models for flows on networks based on the conservation law formulation (1.1) were proposed, and existence of solutions to Cauchy problems was proved. In particular, in [5] some rules to uniquely solve Riemann problems at junctions, where interactions between roads in the network occur, were introduced. In [3] some numerical approximations of the traffic models described in [4, 5] were provided. Moreover, a simulation algorithm using the Godunov scheme with boundary conditions at junctions was implemented and tested.

Here we present a simulation algorithm to trace the trajectory of a car traveling in a road network. The approximation procedure is composed of the following two parts:

- the solutions to (1.1) on each road of the network are computed by Wave Front Tracking (WFT) [1, 15] or by the Godunov scheme [10, 17];
- the problem (1.2) is solved by tracing car position through a procedure which takes into consideration interactions between the car trajectory and the (shock or rarefaction) waves on each road.

Focusing on bounded variation data, we establish a convergence result for the approximate solution on a single road, obtained by WFT technique, toward the solution of (1.1)–(1.2). While theoretical approaches do not provide a convergence rate (see, for instance, [6]), we are able to give an explicit linear convergence rate expressed in terms of total variation of initial datum of density, namely, $TV(\bar{\rho})$; see Theorem 4.4, formula (4.15).

Then we apply the algorithm to measure the efficiency of a traffic circle. Simulations are run to test how the total travel time of a driver is influenced by the right of way parameters. It is shown, as intuition may suggest, that the best choice corresponds to traffic inside the circle having priority with respect to incoming traffic. In the opposite situation, the circle may even come to a complete stop.

The case of car accidents on highways is also considered. We can give accurate estimates of the traveling times, assuming that we know the accident removal time. Such information is particularly interesting since static estimates are of low quality.

The paper is organized as follows. Section 2 is devoted to the description of the problem. The approximation algorithm is described in section 3, and in section 4 theoretical results of convergence for WFT are presented. In section 5 we propose an application of the approximation algorithm to determine the trajectory of a car moving into a portion of urban network represented by a traffic circle and into a single road when a car accident occurs. Related animations are reported on the web page [2].

2. Background.

To construct solutions to Cauchy problems like (1.1) it is important to solve Riemann problems, which are Cauchy problems with initial data of Heaviside type. If f is convex or concave, then there exist centered solutions (i.e., constant along rays $\frac{x}{t}$) consisting of a single wave, either a shock or a rarefaction. For instance, if f is concave and the initial datum is

$$\bar{\rho}(x) = \begin{cases} \rho_l, & x < 0, \\ \rho_r, & x > 0, \end{cases} \tag{2.1}$$

with ρ_l and ρ_r fixed constants, then the solution is a shock (if $\rho_l \leq \rho_r$),

$$\rho(t,x) = \begin{cases} \rho_l & \text{if } x \le \frac{f(\rho_r)-f(\rho_l)}{\rho_r-\rho_l} t, \\ \rho_r & \text{if } x > \frac{f(\rho_r)-f(\rho_l)}{\rho_r-\rho_l} t, \end{cases} \tag{2.2}$$

or a rarefaction (if $\rho_l > \rho_r$),

$$\rho(t,x) = \begin{cases} \rho_l & \text{if } x \le f'(\rho_l)t, \\ (f')^{-1}\left(\frac{x}{t}\right), & f'(\rho_l)t \le x \le f'(\rho_r)t, \\ \rho_r & \text{if } x > f'(\rho_r)t. \end{cases} \tag{2.3}$$

For (1.1), the velocity and the flux function are required to satisfy the following assumptions:

(H) $v'(\rho) < 0$ and f is smooth and strictly concave.

A road network is given by a finite number of roads modeled by intervals $[a_i, b_i]$, $i = 1, \dots, N$, that meet at some junctions. We call the Riemann problem for a road network a Cauchy problem with constant initial datum on each road. For road endpoints not linked to any junction, boundary data are required and the corresponding boundary problem is solved.

In treating networks, the main difficulty is the fact that the system at a junction is underdetermined, even imposing the conservation of cars. The latter can be expressed by the Rankine–Hugoniot condition at the junction

$$\sum_{i=1}^{n} f(\rho_i(t, b_i)) = \sum_{j=n+1}^{n+m} f(\rho_j(t, a_j)),$$

where ρ_i, $i = 1, \dots, n$, and ρ_j, $j = n+1, \dots, n+m$, are the car densities, respectively, on incoming and outgoing roads. To uniquely solve Riemann problems at junctions, as in [5], we make the following assumptions:

(A) there are some fixed coefficients, which depend on drivers preferences, expressing the distribution of traffic from incoming to outgoing roads;

(B) respecting (A), drivers behave in order to maximize the flow through junctions.

To deal with rule (A) we fix a matrix, called the *traffic distribution* matrix,

$$A = \{\alpha_{ji}\}_{j=n+1,\dots,n+m,\, i=1,\dots,n} \in \mathbb{R}^{m\times n}, \text{ with } 0 < \alpha_{ji} < 1, \quad \sum_{j=n+1}^{n+m} \alpha_{ji} = 1,$$

for $i = 1, \dots, n$ and $j = n+1, \dots, n+m$, where α_{ji} represents the percentage of drivers arriving from the ith incoming road who take the jth outgoing road. In [14] an approach based only on the maximization of a function, e.g., flux, was proposed.

In [5] the existence of solutions to Cauchy problems respecting rules (A) and (B) was proved. In the case $m < n$ it is necessary to introduce a further rule; see [4]. If, for example, $m = 1$, $n = 2$, we fix a *right of way* parameter $q \in]0, 1[$ and assume the following rule:

(C) Assume that not all cars can enter the outgoing road and C is the quantity that can do so. Then qC cars come from first incoming road and $(1-q)C$ cars from the second.

The rule (C) allows us to uniquely solve Riemann problems.

Let us now briefly describe how solutions to Riemann problems are computed; for the details the reader is referred to [5]. We look for solutions to problem (1.1) with a single

wave on each road. Rules (A)–(B) give rise to a linear programming problem. In particular, rule (B) consists in the maximization of a linear functional on a convex region determined by rule (A). More precisely, initial data of roads linked on the right (incoming roads) or on the left (outgoing roads) and constraints on the sign of wave speed determine the region where incoming fluxes are maximized.

Fix constant initial data $\rho_{i,0}$ on each incoming road and $\rho_{j,0}$ on each outgoing road. The densities of cars on the incoming roads are indicated by $\rho_i(t,x) : \mathbb{R}^+ \times I_i \to [0,1]$, $i \in \{1,\dots,n\}$, and on the outgoing roads by $\rho_j(t,x) : \mathbb{R}^+ \times I_j \to [0,1]$, $j \in \{1,\dots,m\}$. Let $\tau : [0,1] \mapsto [0,1]$ be the continuous map such that

$$f(\tau(\rho)) = f(\rho),$$

and $\tau(\rho) \neq \rho$ for each $\rho \neq \sigma$. We define the densities $\hat\rho_i, \hat\rho_j$ (and the corresponding fluxes $f(\hat\rho_i) = \hat\gamma_i$, $f(\hat\rho_j) = \hat\gamma_j$) as the new states at the junction. The unique admissible weak solution at a junction is given by the solution to the Riemann problem with data $(\rho_{i,0}, \hat\rho_i)$ for incoming roads and $(\hat\rho_j, \rho_{j,0})$ for outgoing roads. For instance, for incoming roads with $\rho_{i,0} \le \hat\rho_i$, the solution (centered in b_i) is a shock, and, for a sufficiently small time, can be expressed as

$$\rho_i(t,x) = \begin{cases} \rho_{i,0} & \text{if } x \le b_i + \frac{f(\hat\rho_i) - f(\rho_{i,0})}{\hat\rho_i - \rho_{i,0}} t, \\ \hat\rho_i & \text{otherwise,} \end{cases} \tag{2.4}$$

and the velocity is given by $\lambda = \frac{f(\hat\rho_i) - f(\rho_{i,0})}{\hat\rho_i - \rho_{i,0}}$ (namely, the Rankine–Hugoniot relation), or, if $\rho_{i,0} > \hat\rho_i$, a rarefaction that, for a sufficiently small time, reads as

$$\rho_i(t,x) = \begin{cases} \rho_{i,0} & \text{if } x \le b_i + f'(\rho_{i,0})t, \\ (f')^{-1}\left(\frac{x}{t}\right), & b_i + f'(\rho_{i,0})t < x < b_i + f'(\hat\rho_i)t, \\ \hat\rho_i & \text{if } x \ge b_i + f'(\hat\rho_i)t. \end{cases} \tag{2.5}$$

Analogously, the waves produced by the solutions to Riemann problems for the outgoing roads are centered in the left endpoint a_i.

Since we look for waves emerging out of junctions, admissible solutions are obtained by solving Riemann problems by waves of negative speed on incoming roads and by waves of positive speed on outgoing roads, as indicated by conditions (2.6)–(2.7):

$$\hat\rho_i \in \begin{cases} \{\rho_{i,0}\} \cup \,]\tau(\rho_{i,0}), 1] & \text{if } 0 \le \rho_{i,0} \le \sigma, \\ [\sigma, 1] & \text{if } \sigma \le \rho_{i,0} \le 1, \end{cases} \tag{2.6}$$

and

$$\hat\rho_j \in \begin{cases} [0, \sigma] & \text{if } 0 \le \rho_{j,0} \le \sigma, \\ \{\rho_{j,0}\} \cup [0, \tau(\rho_{j,0})[& \text{if } \sigma \le \rho_{j,0} \le 1. \end{cases} \tag{2.7}$$

The new states at junctions, namely, $\hat\rho_i$ on incoming roads and $\hat\rho_j$ on outgoing roads, are uniquely obtained by inverting the relations

$$f(\hat\rho_i) = \hat\gamma_i, \qquad f(\hat\rho_j) = \hat\gamma_j \tag{2.8}$$

on the sets given by (2.6) and (2.7).

Due to the rule (A),

$$\hat{\gamma}_j \doteq \sum_{i=1}^{n} \alpha_{ji}\hat{\gamma}_i, \qquad j = n+1, \dots, n+m,$$

it suffices to determine the incoming fluxes $\hat{\gamma}_i$, which solve the following LP problem: Define

$$\begin{aligned} \Omega_i &:= [0, \gamma_i^{max}(\rho_{i,0})], \quad i = 1, \dots, n, \\ \Omega_j &:= \left[0, \gamma_j^{max}(\rho_{j,0})\right], \quad j = n+1, \dots, n+m, \\ \Omega &:= \left\{(\gamma_1, \dots, \gamma_n) \in \Omega_1 \times \cdots \times \Omega_n \mid A \cdot (\gamma_1, \dots, \gamma_n)^T \in \Omega_{n+1} \times \cdots \times \Omega_{n+m}\right\}. \end{aligned} \tag{2.9}$$

The set Ω is compact, convex, and not empty. Then, $\hat{\gamma} \in \Omega$ is the solution to

$$\max_{\gamma \in \Omega} \gamma \cdot \mathbf{1}, \tag{2.10}$$

where $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^n$.

For definiteness, the flux function is chosen to be the Greenshields one (see [12]):

$$f(\rho) = v_{max}\, \rho \left(1 - \frac{\rho}{\rho_{max}}\right). \tag{2.11}$$

We set for simplicity $\rho_{max} = 1 = v_{max}$, so that the velocity is $v = 1 - \rho$ and the flux $f = \rho(1-\rho)$. In particular, $f(0) = 0 = f(1)$ and f has a unique maximum in $\sigma = 1/2$.

3. Approximation schemes.

In this section we describe the simulation algorithm, which is composed of the following two steps:

Step 1: The density values satisfying (1.1) are computed on each road solving Riemann problems. The numerical scheme can be indifferently WFT or Godunov scheme endowed with boundary conditions at junctions.

Step 2: The driver's position is determined solving problem (1.2) by means of an algorithm which, given the densities obtained at the previous step by WFT or the Godunov scheme, determine the car position on the network.

The choice of WFT is due to the possibility of obtaining theoretical results. On the other side, the Godunov scheme is easy to implement and gives good insight into vehicular traffic problems; see [9, 16]. Also, both schemes are based on the solution to Riemann problems, thus permitting a convenient treatment of the car trajectory approximation.

3.1. The Wave Front Tracking algorithm (WFT).

Here we recall briefly the technique of Wave Front Tracking; for a detailed description see [1, 9].

The WFT is a semidiscrete scheme which can be summarized by the following steps:

- approximate initial datum $\bar{\rho} = (\bar{\rho}_1, \dots, \bar{\rho}_N)$ by piecewise constant functions $\bar{\rho}_\nu$;
- construct solutions to Riemann problems of $\bar{\rho}_\nu$ and approximate rarefactions by a set of small shocks;
- piece approximate solutions to Riemann problems together to get a solution for t small;
- prolong waves up to first interaction time. Then one gets a new Riemann problem, solves it approximately, and goes on up to the next interaction time.

Next we detail the construction of WFT approximate solutions. Given a general initial datum $\bar\rho$, we approximate it by a sequence of piecewise constant functions $\bar\rho_\nu$ and we construct the corresponding approximate solutions. If they converge in L^1_{loc}, then the limit is a weak entropy solution on each road; see [1] for the proof.

3.1.1. Step 1: Numerical algorithm for (1.1). Let $\bar\rho = (\bar\rho_1, \dots, \bar\rho_N)$ be a map defined on the road network, $\bar\rho_i : I_i \to \mathbb{R}$, $i = 1, \dots, N$. It is possible to choose a sequence of piecewise constant functions $\bar\rho_\nu$ such that

$$Tot.Var.\{\bar\rho_\nu\} \le Tot.Var.\{\bar\rho\}, \tag{3.1}$$

$$|\bar\rho_\nu|_{L^\infty} \le |\bar\rho|_{L^\infty}, \tag{3.2}$$

$$|\bar\rho_\nu - \bar\rho|_{L^1} < \frac{1}{\nu}. \tag{3.3}$$

By (3.1), $\bar\rho_\nu$ has a finite number of discontinuities, say, $y_1^\nu < \cdots < y_K^\nu$. We approximately solve the Riemann problem generated by the jump $(\bar\rho_\nu(y_k^\nu-), \bar\rho_\nu(y_k^\nu+))$ for each $k = 1, \dots, K$ using piecewise constant functions of the type $\varphi\left(\frac{x-y_k^\nu}{t}\right)$, where $\varphi : \mathbb{R} \to \mathbb{R}$. In particular, if the solution to the Riemann problem generated by $(\bar\rho_\nu(y_k^\nu-), \bar\rho_\nu(y_k^\nu+))$ is a shock, then $\varphi\left(\frac{x-y_k^\nu}{t}\right)$ is the exact solution. On the contrary, if a rarefaction wave appears, then we split it into a centered rarefaction fan formed by a sequence of jumps of size at most $1/\nu$, traveling with the characteristic speed of the left state.

Proceeding in this way, we are able to construct an approximate solution $\rho_\nu(t, x)$ until a time t_1, where either two wave fronts interact together or a wave interacts with a junction. When a wave interacts with another one, we simply solve the new Riemann problem; instead, when a wave reaches a junction, we solve the corresponding Riemann problem at the junction.

We always split rarefaction waves, inserting the value σ (if it is in the range of the rarefaction). Moreover, we let any rarefaction shock with endpoint σ have velocity zero. In order to prove existence of a wave front tracking approximate solution for every $t \in [0, T]$, where T may be also $+\infty$, we need to estimate

1. the number of waves; and
2. the number of interactions between waves.

We call the obtained function *a wave front tracking approximate solution.*

Fix $\nu \in \mathbb{N}$ and $t > 0$. We let $K(\nu, t)$ be the time dependent set of discontinuities and for every $k \in K(\nu, t)$ we let $\rho_-^{\nu,k}(t)$, $\rho_+^{\nu,k}(t)$ be, respectively, the left and the right states of the discontinuity. In other words,

$$\rho_-^{\nu,k}(t) = \rho_\nu(t, y_k^\nu-),$$
$$\rho_+^{\nu,k}(t) = \rho_\nu(t, y_k^\nu+),$$

where y_k^ν is the position of the kth discontinuity at time t in ρ_ν. We also indicate by λ_k the velocity of the kth discontinuity. For simplicity from now on we may eventually drop the index ν.

If we get BV estimates on the WFT approximate solutions, we can pass to the limit, obtaining a weak entropy solution.

For a single conservation law on $\mathbb{R}$, the number of waves decreases; thus the number of interactions is bounded by the number of waves, and the total variation diminishes. For networks the situation is more complex. In particular, the number of waves may increase for interactions of waves at junctions. Still the necessary estimates can be carried out as described in [9].

3.1.2. Step 2: Car path. For every $\nu \in \mathbb{N}$, we call $x_\nu(t)$ the position of the car at time t if the load is given by the approximate solution ρ_ν. Then

$$\dot{x}_\nu(t) = v(\rho_\nu(t, x_\nu(t))). \tag{3.4}$$

In the following lemma we show that at interaction times with waves, the velocity of the car is greater than that of the wave in front of it.

Lemma 3.1. *If $x_\nu(t) \in (y^\nu_k, y^\nu_{k+1})$, then*

$$\dot{x}_\nu(t) = v(\rho^{\nu,k}_+(t)) > \lambda_{k+1}(t) = \frac{f(\rho^{\nu,k+1}_+(t)) - f(\rho^{\nu,k+1}_-(t))}{\rho^{\nu,k+1}_+(t) - \rho^{\nu,k+1}_-(t)}. \tag{3.5}$$

Proof. Using the Rankine–Hugoniot relation we want to prove that

$$\begin{aligned}\frac{f(\rho^{\nu,k+1}_+(t)) - f(\rho^{\nu,k+1}_-(t))}{\rho^{\nu,k+1}_+(t) - \rho^{\nu,k+1}_-(t)} &= \frac{v(\rho^{\nu,k+1}_+(t))\rho^{\nu,k+1}_+(t) - v(\rho^{\nu,k+1}_-(t))\rho^{\nu,k+1}_-(t)}{\rho^{\nu,k+1}_+(t) - \rho^{\nu,k+1}_-(t)} \\ &< v(\rho^{\nu,k+1}_-(t)).\end{aligned}$$

If $\rho^{\nu,k+1}_+(t) > \rho^{\nu,k+1}_-(t)$, then (3.5) is equivalent to

$$v(\rho^{\nu,k+1}_+(t))\rho^{\nu,k+1}_+(t) - v(\rho^{\nu,k+1}_-(t))\rho^{\nu,k+1}_-(t) < v(\rho^{\nu,k+1}_-(t))\rho^{\nu,k+1}_+(t) - v(\rho^{\nu,k+1}_-(t))\rho^{\nu,k+1}_-(t),$$

i.e., to

$$v(\rho^{\nu,k+1}_+(t)) < v(\rho^{\nu,k+1}_-(t)).$$

By the hypothesis (H), v is a strictly decreasing function; hence the last inequality is verified. On the other hand, if $\rho^{\nu,k+1}_+(t) < \rho^{\nu,k+1}_-(t)$, then we obtain the inequality $v(\rho^{\nu,k+1}_+(t)) > v(\rho^{\nu,k+1}_-(t))$, which is again verified due to the decreasing behavior of v. ∎

Therefore, $x_\nu(t)$ interacts with the waves located at y^ν_k in increasing order.

A numerical algorithm is readily obtained by setting the following:

- If $x_\nu(t) \in (y^\nu_k, y^\nu_{k+1})$, then $\dot{x}_\nu(t) = v(\rho^{\nu,k}_+(t))$. In other words, the velocity is constant as long as no interaction with waves occurs.
- If $x_\nu(t)$ interacts with the kth wave of ρ_ν, then the velocity changes from $v(\rho^{\nu,k-1}_+(t)) = v(\rho^{\nu,k}_-(t))$ to $v(\rho^{\nu,k}_+(t))$.

3.2. Godunov scheme on a road network. Now, in order to describe the Godunov scheme we need to introduce a *numerical grid* with the following notation:

- Δx is the space grid size on each road I_i;
- Δt is the time grid size on the time interval $[0, T]$;
- $(t_l, x_m) = (l\Delta t, m\Delta x)$, for $l = 0, 1, \ldots, L$ and $m = 0, 1, \ldots, M$, are the grid points, with L and M, respectively, the number of time and space nodes of the grid.

For a function v defined on the grid we write $v_m^l = v(t_l, x_m)$. Notice that for our simulations we assume to have a constant space increment since all intervals I_i are equal, but, in general, Δx may vary depending on the length of each road.

3.2.1. Step 1: Numerical algorithm for (1.1). The Godunov scheme is based on the local resolution to Riemann problems and it proceeds as follows; for further details see [11, 10].

Piecewise constant approximations of the initial data are used as the initial data of Riemann problems. Waves in two neighbor cells do not interact before time Δt under the CFL condition $\Delta t \leq \frac{1}{2v_{max}}\Delta x$, which, setting $v_{max} = 1$, reads as $\Delta t \leq \frac{1}{2}\Delta x$. It is then possible to define a unique solution in the strip $(t_l, t_{l+1}) \times \mathbb{R}$ by piecing the solutions obtained in each cell together. The exact solution is projected on a piecewise constant function $v_m^{l+1} = \frac{1}{\Delta x}\int_{x_m}^{x_{m+1}} v^{\Delta}(x, t_{l+1})dx$; then the mean is obtained by the Gauss–Green formula, and this procedure is repeated recursively. The Godunov scheme can be expressed in the conservative form as

$$v_m^{l+1} = v_m^l - \frac{\Delta t}{\Delta x}\big(g^G(v_m^l, v_{m+1}^l) - g^G(v_{m-1}^l, v_m^l)\big), \quad l = 0, 1, \ldots, L-1, \quad m = 0, 1, \ldots, M, \tag{3.6}$$

with $g^G(u, v)$ numerical flux. Boundary conditions of the scheme are imposed for any incoming road not linked on the left and for any outgoing road not linked on the right.

Conditions at a junction. For roads connected at the right endpoint, the interaction at a junction is taken into account as follows:

$$v_M^{l+1} = v_M^l - \frac{\Delta t}{\Delta x}\big(\hat{\gamma}_i - g^G(v_{M-1}^l, v_M^l)\big),$$

while for roads connected at the left endpoint we have

$$v_0^{l+1} = v_0^l - \frac{\Delta t}{\Delta x}\big(g^G(v_0^l, v_1^l) - \hat{\gamma}_j\big),$$

where $\hat{\gamma}_i$, $\hat{\gamma}_j$ are the maximized fluxes, respectively, on incoming and outgoing roads, computed solving the linear programming problem (2.10).

3.2.2. Step 2: Car path. Let us consider now a single driver moving on a road network. We develop a numerical scheme to describe car trajectory on each road composing the network. A road is parametrized as an interval and, according to the discretization previously defined, is divided into subintervals or cells of length Δx.

At each time t_l, we determine the position x^l of the driver by studying interactions between the car trajectory and the density waves within a fixed cell of the numerical grid $C_{m_l} = [(m_l - 1)\Delta x, m_j\,\Delta x[$. We distinguish the following two cases:

Case 1. $x^l \in [(m_l - 1)\Delta x, (m_l - \frac{1}{2})\Delta x[$.

Case 2. $x^l \in [(m_l - \frac{1}{2})\Delta x, m_l\,\Delta x[$.

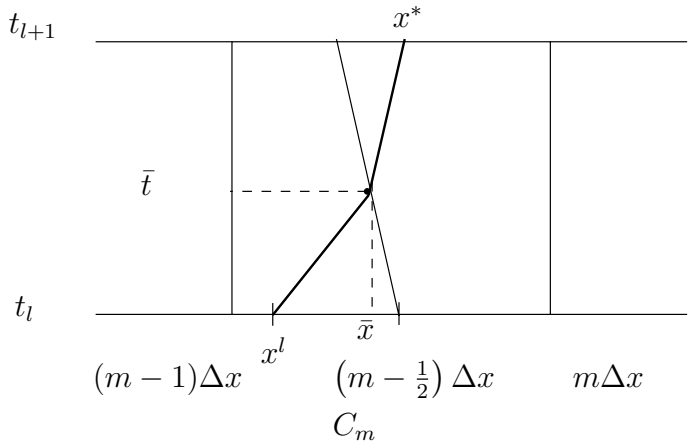


Figure 1. *Interaction with shock.*

In order to describe the car trajectory, we compute the new position x^{l+1} and possibly update the cell index m_{l+1}. The approximate value of the density on the numerical grid is denoted by $\rho_m^l = \rho(t_l, x_m)$.

Let us now detail the algorithm: for notational convenience we drop the index l from m.

3.2.3. Case 1. Two cases are distinguished:

(a) the wave starting from the space node $(m-\frac{1}{2})\Delta x$ is a shock;

(b) the wave starting from the space node $(m-\frac{1}{2})\Delta x$ is a rarefaction.

Let us first study case (a). The velocity of the wave, starting from the point $(m-\frac{1}{2})\Delta x$ at time t_l, is given by

$$\lambda_m = \frac{f(\rho_m^l) - f(\rho_{m-1}^l)}{\rho_m^l - \rho_{m-1}^l}.$$

Then the car and the wave interact at the point $(\bar{t}, \bar{x})$ given by

$$\begin{aligned} \bar{t} &= \frac{\left(m-\frac{1}{2}\right)\Delta x - x^l}{v(\rho_{m-1}^l) - \lambda_m}, \\ \bar{x} &= x^l + v(\rho_{m-1}^l)\bar{t}. \end{aligned} \tag{3.7}$$

We have to further consider the following cases:

(i) $\bar{t} \geq \Delta t$, which means no interaction on the time interval $[t_l, t_{l+1}]$. Then we have

$$x^{l+1} = x^l + v(\rho_{m-1}^l)\Delta t.$$

(ii) $\bar{t} < \Delta t$. Then, after the interaction (see Figure 1), the new position of the car is

$$x^{l+1} = \bar{x} + (\Delta t - \bar{t})v(\rho_m^l).$$

Let us now turn to case (b). Recalling that for the flux function considered we have $v(\rho(t,x)) = 1-\rho$, the driver's position is obtained by solving the ordinary differential equation

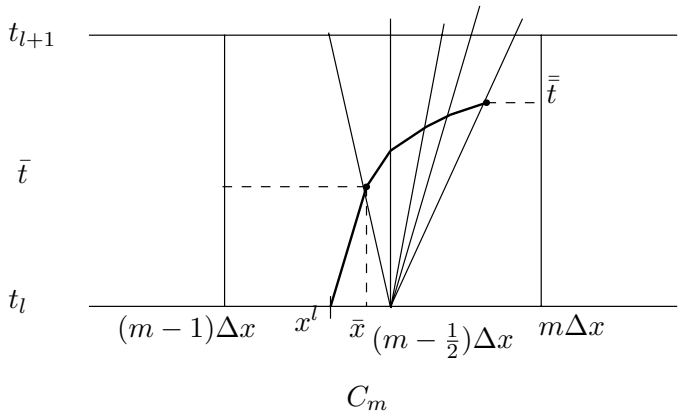


Figure 2. *Interaction with rarefaction.*

$$\dot{x} = \begin{cases} 1-\rho_m^l & \text{if } x \le (m-\frac{1}{2})\Delta x + f'(\rho_m^l)t, \\ 1-\left(\frac{1}{2} - \frac{x-(m-\frac{1}{2})\Delta x}{2\,(t-t_l)}\right) & \text{if } f'(\rho_m^l)t \le x-(m-\frac{1}{2})\Delta x \le f'(\rho_{m+1}^l)t, \\ 1-\rho_{m+1}^l & \text{if } x > (m-\frac{1}{2})\Delta x + f'(\rho_{m+1}^l)t. \end{cases} \tag{3.8}$$

The velocity of the wave, starting from the point $(m-\frac{1}{2})\Delta x$ at time t_l, is given by

$$\lambda_m = f'(\rho_m^l).$$

The first interaction point with the rarefaction wave is again expressed as in (3.7).

Then, the solution to (3.8) is

$$x(t) = \left(m-\frac{1}{2}\right)\Delta x + (t-t_l) - \sqrt{t-t_l}\left(\frac{\bar{t}-t_l+(m-\frac{1}{2})\Delta x-\bar{x}}{\sqrt{\bar{t}-t_l}}\right). \tag{3.9}$$

Now, let $\bar{\bar{t}}$ be the time coordinate of the final intersection point between the rarefaction wave and the car trajectory; see Figure 2. Then

$$\frac{x(\bar{\bar{t}}) - \left(m-\frac{1}{2}\right)\Delta x}{\bar{\bar{t}}} = f'(\rho_m^l),$$

with $x(t)$ given by (3.9).

Again we distinguish two possible cases:

Case (a) $\bar{t} \ge \Delta t$, i.e., no interaction between the car and the rarefaction wave occur. Since the car travels at constant speed, the new position is obtained by

$$x^{l+1} = x^l + v(\rho_{m-1}^l)\Delta t.$$

Case (b) $\bar{t} < \Delta t$. For case (b), we need to further distinguish the following:

Case (b1) $\bar{\bar{t}} \ge \Delta t$. Then the new position is

$$x^{l+1} = x(\Delta t),$$

with $x(t)$ given by (3.9).

Case (b2) $\bar{\bar{t}} < \Delta t$. Then

$$x^{l+1} = x(\bar{\bar{t}}) + (\Delta t - \bar{\bar{t}})v(\rho^l_m),$$

with $x(t)$ given by (3.9).

Since the trajectory remains inside the cell C_{m_l} even after interactions, in all subcases of case 1 the updated cell index is

$$m_{l+1} = m_l.$$

3.2.4. Case 2. Two cases are distinguished:

(a) The wave starting from the space node $(m + \frac{1}{2})\Delta x$ is a shock.

(b) The wave starting from the space node $(m + \frac{1}{2})\Delta x$ is a rarefaction.

We first consider case (a). The velocity of the wave, starting from the point $(m + \frac{1}{2})\Delta x$ at time t_l, is given by

$$\lambda_m = \frac{f(\rho^l_{m+1}) - f(\rho^l_m)}{\rho^l_{m+1} - \rho^l_m}.$$

The interaction point $(\bar{t}, \bar{x})$ of the wave with the car is given by

$$\begin{aligned} \bar{t} &= \frac{\left(m + \frac{1}{2}\right)\Delta x - x^l}{v(\rho^l_{m+1}) - \lambda_m}, \\ \bar{x} &= x^l + v(\rho^l_m)\bar{t}. \end{aligned} \tag{3.10}$$

As before, we have two different cases:

(i) $\bar{t} \geq \Delta t$, which means no interaction on the time interval $[t_l, t_{l+1}]$. Then we have

$$x^{l+1} = x^l + v(\rho^l_m)\Delta t.$$

(ii) $\bar{t} < \Delta t$. Then, after the interaction, the new position of the car is

$$x^{l+1} = \bar{x} + (\Delta t - \bar{t})v(\rho^l_{m+1}).$$

Let us now turn to case (b). We have the same equation as in (3.8) with $(m - \frac{1}{2})\Delta x$ replaced by $(m + \frac{1}{2})\Delta x$ and $f'(\rho^l_m)$ by $f'(\rho^l_{m+1})$. The velocity of the wave, starting from the point $(m + \frac{1}{2})\Delta x$ at time t_l, is given by

$$\lambda_m = f'(\rho^l_{m+1}).$$

The first interaction point with the rarefaction wave is again expressed as in (3.10).

Then, the car position after $\bar{t}$ and before exiting the rarefaction is given by

$$x(t) = \left(m + \frac{1}{2}\right)\Delta x + (t - t_l) - \sqrt{t - t_l}\left(\frac{\bar{t} - t_l + (m + \frac{1}{2})\Delta x - \bar{x}}{\sqrt{\bar{t} - t_l}}\right). \tag{3.11}$$

Now, the final interaction time $\bar{\bar{t}}$ of the car with the rarefaction solves

$$\frac{x(\bar{\bar{t}}) - \left(m + \frac{1}{2}\right)\Delta x}{\bar{\bar{t}}} = f'(\rho^l_{m+1}),$$

with $x(t)$ given by (3.11).

The distinction in subcases is as before:

Case (a) $\bar{t} \geq \Delta t$, i.e., no interaction between the car and the rarefaction wave occurs. Since the car travels at constant speed, the new position is obtained by

$$x^{l+1} = x^l + v(\rho_m^l)\Delta t.$$

Case (b) $\bar{t} < \Delta t$.

Case (b1) $\bar{\bar{t}} \geq \Delta t$. Then the new position is

$$x^{l+1} = x(\Delta t),$$

with $x(t)$ given by (3.11).

Case (b2) $\bar{\bar{t}} < \Delta t$. Then

$$x^{l+1} = x(\bar{\bar{t}}) + (\Delta t - \bar{\bar{t}})v(\rho_{m+1}^l),$$

with $x(t)$ given by (3.11).

Finally, the new cell index is determined as follows. If $x^{l+1} < m\Delta x$, then $m_{l+1} = m_l$; otherwise $m_{l+1} = m_l + 1$.

4. Convergence of car trajectory. This section is devoted to the analysis of the convergence of the car trajectory x_ν in case of the WFT algorithm. We consider the case of a single road $[a, b]$.

Our point of view to estimate the car position as a function of ν is the following. We think of $\rho_{\nu+1}$ as ρ_ν with shifts applied to the initial position of waves. This can be done by fixing an approximation procedure, which is simply based on sampling at points of a grid with mesh size $2^{-\nu}$. Then the position of $x_{\nu+1}$ is also thought of as the position of x_ν plus a shift. The latter changes only at interactions with waves of ρ_ν and $\rho_{\nu+1}$. Thus finally the problem reduces to estimate the increase of the shifts (both of waves and car position) at every interaction.

4.1. Wave and car shifts. Assume we have a BV initial datum $\bar{\rho}(x)$. Then we define $\bar{\rho}_\nu$ to be the sequence approximating $\bar{\rho}$ given by

$$\bar{\rho}_\nu(x) = \bar{\rho}\left(2^\nu \left[\frac{x}{2^\nu}\right]\right), \quad \nu \in \mathbb{N}. \tag{4.1}$$

Then it is easy to notice that every wave of $\rho_{\nu+1}$ corresponds to a wave of ρ_ν with a shift ξ of at most $2^{-(\nu+1)}$; see Figure 3.

Call $\rho_\nu(t, x)$ the WFT solution for initial datum $\bar{\rho}_\nu$. Then it is possible to determine the evolution of shifts of waves $\xi_k^\nu(t)$, $k \in K(\nu, t)$, using the following lemma, proved in [9].

Lemma 4.1. *Consider two waves with speeds λ_1 and λ_2, respectively, that interact together producing a wave with speed λ_3. If the first wave is shifted by ξ_1 and the second wave by ξ_2, then the shift of the resulting wave is given by*

$$\xi_3 = \frac{\lambda_3 - \lambda_2}{\lambda_1 - \lambda_2}\xi_1 + \frac{\lambda_1 - \lambda_3}{\lambda_1 - \lambda_2}\xi_2. \tag{4.2}$$

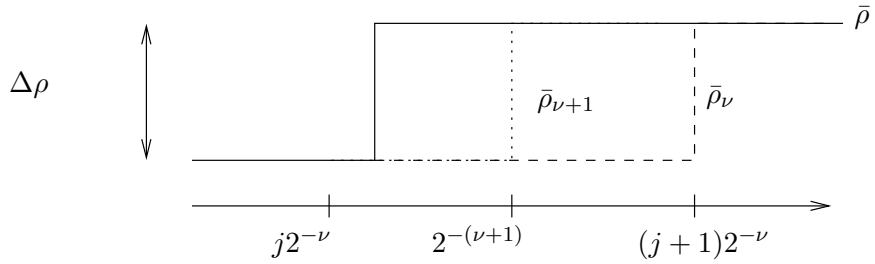


Figure 3. *Shift of discontinuities at time* 0.

Moreover, we have

$$\Delta\rho_3\xi_3 = \Delta\rho_1\xi_1 + \Delta\rho_2\xi_2, \tag{4.3}$$

where $\Delta\rho_i$ *are the signed strengths of the corresponding waves.*

Lemma 4.1 permits us to determine the evolution of shifts of waves. Moreover, even if the shift sizes are not conserved, this happens for the quantity $\xi \cdot \Delta\rho$, which represents the L^1 shift of the approximate solution. Recalling that $(\rho_-^{\nu,k}(t), \rho_+^{\nu,k}(t))$ is the value to the left and, respectively, to the right of the kth jump of $\rho_\nu(t)$, we define $\Delta\rho_k^\nu(t) = \rho_+^{\nu,k}(t) - \rho_-^{\nu,k}(t)$.

To study the convergence of car trajectories, we need to estimate the quantity

$$\|x_\nu(t) - x_{\nu+1}(t)\|. \tag{4.4}$$

As for the waves of ρ_ν, we define the shift of the car position as

$$\eta_\nu(t) = x_{\nu+1}(t) - x_\nu(t).$$

Since the initial car position does not depend on the approximate solution, we have $\eta_\nu(0) = 0$. Notice that our problem reduces to estimate $|\eta(t)|$. Such a quantity varies only at interaction times with waves of ρ_ν and $\rho_{\nu+1}$. Since the latter are obtained by shifting the formers, we can consider generic interactions of a car with waves, both presenting shifts. Thus we let η_k^ν be the value of η_ν after the interaction with the kth wave of ρ_ν and the kth wave of $\rho_{\nu+1}$. In the case of Neumann boundary conditions, and thus no wave from boundaries, the number of waves is decreasing. Hence the number of interactions is bounded by the number of waves in the initial datum, which in turn is at most $(b-a)/2^{-(\nu+1)}$. In the case of Dirichlet boundary conditions, then we approximate also the boundary data with piecewise constant functions, thus getting again a finite number of possible waves and thus of interactions.

4.2. Shifts evolution. Here we consider a generic interaction between a wave front and a car trajectory. In particular, we assume that both have an initial shift and we want to estimate the value of the car position shift after the interaction, since the wave shift does not change. Say x_0 is the initial position of the car and s_0 is the initial position of the wave front with speed $\lambda = \frac{\Delta f}{\Delta\rho}$; then the interaction point is

$$(\bar x, \bar t) = \left(x_0 + \bar t v(\rho_-), \frac{s_0 - x_0}{v(\rho_-) - \lambda}\right), \tag{4.5}$$

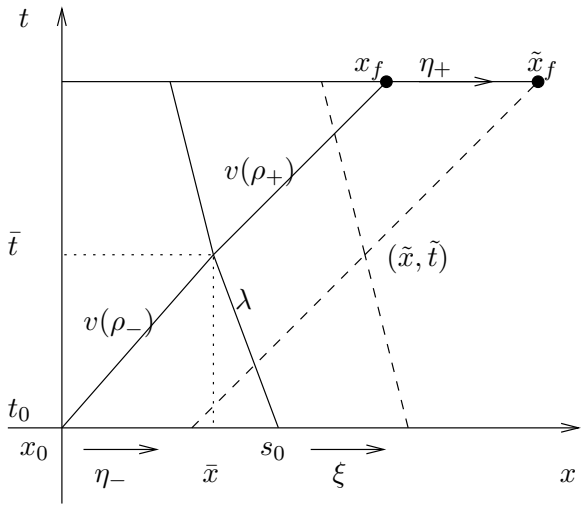


Figure 4. *The shift for the wave front, ξ, and the shifts for the car trajectory, η_-, η_+.*

and, after interaction, the (final) position x_f at time T is given by

$$x_f = \bar{x} + (T - \bar{t})v(\rho_+). \tag{4.6}$$

See Figure 4. Call $\xi \in \mathbb{R}$ the wave shift and $\eta_-, \eta_+ \in \mathbb{R}$ the shifts of the car trajectory, respectively, before and after interacting with the wave. The point of interaction for the shifted wave and car position is

$$(\tilde{x},\tilde{t}) = \left(x_0 + \eta_- + \tilde{t}v(\rho_-), \frac{(s_0+\xi) - (x_0+\eta_-)}{v(\rho_-) - \lambda}\right) \tag{4.7}$$

and the final position of the car $\tilde{x}_f$ is given by

$$\tilde{x}_f = x_f + \eta_+, \tag{4.8}$$

where η_+ is given by

$$\eta_+ = \tilde{x} + v(\rho_+)(\bar{t} - \tilde{t}) - \bar{x}. \tag{4.9}$$

Now, using (4.5)–(4.7), we can express η_+ in terms of η_-:

$$\begin{aligned}\eta_+ &= \eta_- + v(\rho_-)\frac{(s_0+\xi) - (x_0+\eta_-)}{v(\rho_-) - \lambda} + v(\rho_+)\frac{\eta_- - \xi}{v(\rho_-) - \lambda} - v(\rho_-)\frac{s_0 - x_0}{v(\rho_-) - \lambda} \\ &= \eta_-\frac{v(\rho_+) - \lambda}{v(\rho_-) - \lambda} + \xi\frac{v(\rho_-) - v(\rho_+)}{v(\rho_-) - \lambda}. \end{aligned} \tag{4.10}$$

Let us set $\beta = \frac{v(\rho_+)-\lambda}{v(\rho_-)-\lambda}$ and $\gamma = \frac{v(\rho_-)-v(\rho_+)}{v(\rho_-)-\lambda}$. Then, recalling that $\lambda = \Delta f/\Delta\rho$, it follows that

$$\begin{aligned}\beta &= \frac{v(\rho_+)(\rho_+ - \rho_-) - \rho_+ v(\rho_+) + \rho_- v(\rho_-)}{v(\rho_-)(\rho_+ - \rho_-) - \rho_+ v(\rho_+) + \rho_- v(\rho_-)} \\ &= \frac{\rho_-}{\rho_+}, \end{aligned} \tag{4.11}$$

and, similarly,

$$\gamma = \frac{(v(\rho_+) - v(\rho_-))(\rho_+ - \rho_-)}{v(\rho_-)(\rho_+ - \rho_-) - \rho_+ v(\rho_+) + \rho_- v(\rho_-)}$$
$$= \frac{\rho_+ - \rho_-}{\rho_+}. \tag{4.12}$$

Therefore, (4.10) becomes

$$\eta_+ = \eta_- \frac{\rho_-}{\rho_+} + \xi \frac{\rho_+ - \rho_-}{\rho_+}. \tag{4.13}$$

4.3. Estimates for car trajectory. The sequence of interactions between waves and with the car may happen with different orders. However, we know that the car is faster than waves (Lemma 3.1); thus the car always interacts with waves in increasing order. Moreover, the next lemma shows that the worst case, for the car shift increase, happens when the car interacts with waves before wave interactions occur.

Lemma 4.2. *Assume that $\rho_\nu(0)$ is strictly positive and presents only two waves (ρ_-^1, ρ_+^1) and (ρ_-^2, ρ_+^2). Let ξ_1 and ξ_2 be the shifts of the waves in ρ_ν, to get the position of the waves in $\rho_{\nu+1}(0)$. We define η to be the shift generated if x_ν interacts separately with the two waves and $\tilde\eta$ to be the shift if the two waves meet before interacting with x_ν. Then*

$$|\tilde\eta| \le |\eta|.$$

Proof. Using (4.13) and $\eta_0 = 0$, we get

$$\eta_1 = \frac{\rho_-^1}{\rho_+^1}\eta_0 + \frac{|\Delta\rho_1\xi_1|}{\rho_+^1} = \frac{|\Delta\rho_1\xi_1|}{\rho_+^1}, \quad \eta_2 = \frac{\rho_-^2}{\rho_+^2}\eta_1 + \frac{|\Delta\rho_2\xi_2|}{\rho_+^2};$$

then, since $\rho_+^1 = \rho_-^2$, one gets

$$\eta = \frac{1}{\rho_+^2}(|\Delta\rho_1\xi_1| + |\Delta\rho_2\xi_2|),$$

while

$$\tilde\eta = \frac{\rho_-^1}{\rho_+^1}\eta_0 + \frac{\tilde\Delta\rho\tilde\xi}{\rho_+^2} = \frac{\tilde\Delta\rho\tilde\xi}{\rho_+^2},$$

where $\tilde\Delta\rho$ and $\tilde\xi$ are, respectively, the jump and the shift coming from the interaction of the two waves. The conclusion follows by Lemma 4.1. ∎

Now we can state our first estimate for the car shift.

Lemma 4.3. *For every $N \in \mathbb{N}$, provided that $\rho_\nu(0) > 0$, the following recurrence relation holds:*

$$|\eta_N^\nu| \le \frac{1}{\rho_+^{\nu,N-1}} \sum_{k=1}^{N-1} |\Delta\rho_k^\nu(0)\xi_k^\nu(0)|. \tag{4.14}$$

Proof. By Lemmas 3.1 and 4.2 we can assume that x_ν interacts with the waves of $\rho_\nu(0)$ in increasing order before wave interactions occur. In fact, this represents the worst case for the car shift increase.

We proceed by induction. Recall that $\eta_0^\nu = \eta_\nu(0) = 0$. Supposing that (4.14) is true for N, we prove the relation for $N+1$ using (4.13):

$$\begin{aligned}
|\eta_{N+1}^\nu| &\le \frac{\rho_-^{\nu,N}}{\rho_+^{\nu,N}}|\eta_N| \;+\; \frac{|\Delta\rho_N^\nu \xi_N^\nu|}{\rho_+^{\nu,N}} \\
&\le \frac{\rho_-^{\nu,N}}{\rho_+^{\nu,N}} \frac{1}{\rho_+^{\nu,N-1}} \sum_{k=1}^{N-1} |\Delta\rho_k^\nu \xi_k^\nu| \;+\; \frac{|\Delta\rho_N^\nu \xi_N^\nu|}{\rho_+^{\nu,N}} \\
&= \frac{1}{\rho_+^{\nu,N}} \sum_{k=1}^{N} |\Delta\rho_k^\nu \xi_k^\nu| \le \frac{1}{\rho_+^{\nu,N}} \sum_{k=1}^{N} |\Delta\rho_k^\nu(0) \xi_k^\nu(0)|,
\end{aligned}$$

where the latter inequality is obtained by Lemma 4.1. ∎

We can finally state our main result.

Theorem 4.4. *Let $\bar\rho \in BV$ and assume that WFT approximate solutions are constructed taking the initial datum $\bar\rho_\nu$ as in (4.1). Assume that $\bar\rho \ge \tilde\rho > 0$; then*

$$|x_{\nu+1}(t) - x_\nu(t)| \le \frac{2^{-(\nu+1)}}{\tilde\rho} TV(\bar\rho). \tag{4.15}$$

In particular, $x_\nu(t)$ converges uniformly to some $x(t)$ solution of (1.2) when $\nu \to +\infty$. Since the grid mesh parameter is $\Delta x = 2^{-\nu}$ as shown by (4.1), the convergence speed estimate is linear in Δx.

Proof. Recall that $\eta_\nu(t) = x_{\nu+1}(t) - x_\nu(t)$. By Lemma 4.3, for every t it holds that

$$|\eta_\nu(t)| \le \frac{1}{\tilde\rho} \sum |\Delta\rho_k^\nu(0) \xi_k^\nu(0)|. \tag{4.16}$$

As noted above, the shifts at time 0 between ρ_ν and $\rho_{\nu+1}$ are bounded by $2^{-(\nu+1)}$; see Figure 3. Then, we get

$$|\eta_\nu(t)| \le \frac{2^{-(\nu+1)}}{\tilde\rho} \sum |\Delta\rho_k^\nu(0)| = \frac{2^{-(\nu+1)}}{\tilde\rho} TV(\bar\rho).$$

Thus $x_\nu(t)$ converges uniformly exponentially to some function $x(t)$. By the results of Colombo and Marson [6], x is a solution of (1.2). ∎

A possible extension of Theorem 4.4 if initial data vanish is given below.

Theorem 4.5. *Consider a single road $[a,b]$. Then, $\forall \alpha \in (0,1)$ there exist $\rho_\nu \to \rho$ and $x_{\nu+1} \to x$ such that*

$$|\rho - \rho_\nu| < 2^{-\nu\alpha}(b-a), \tag{4.17}$$

$$|x_{\nu+1} - x_\nu| \le 2^{-\nu(1-\alpha)} TV(\bar\rho). \tag{4.18}$$

Proof. It is well known that in the scalar case the following holds:

$$|\rho(t) - \rho_\nu(t)|_{L^1} \le |\rho(0) - \rho_\nu(0)|_{L^1}.$$

Let us take $\rho_\nu(0) \to \bar\rho$ such that $\rho_\nu(0) \ge 2^{-\nu\alpha}$. Then, applying Theorem 4.4, we get (4.17). ∎

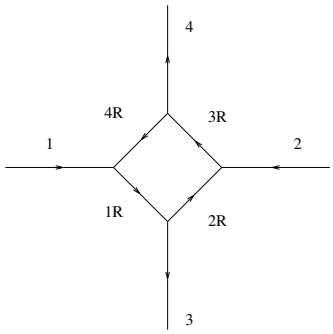


Figure 5. *Traffic circle.*

5. Numerical tests. Simulation results and animations for numerical tests presented in this section are available on the web page [2].

Here we present a study of the trajectory of a car moving into a traffic circle: the evolution of densities was already discussed in [3]. More precisely, by reproducing the evolution in time of traffic, we compute the time needed by a car for covering a fixed path within a traffic circle composed by 8 roads and 4 junctions, as depicted in Figure 5. There are two junctions with one incoming and two outgoing roads, precisely the junction $(1R, 2R, 3)$ and the junction $(3R, 4R, 4)$. Thus we need to assign the corresponding distribution coefficients, namely, $(\alpha_{1R,3}, \alpha_{1R,2R})$ and $(\alpha_{3R,4}, \alpha_{3R,4R})$. We assume all of them to be equal to $\alpha = 0.5$. For junctions with one outgoing road, namely, $(1, 4R, 1R)$ and $(2R, 2, 3R)$, we need to fix a right of way parameter between the two incoming roads in order to describe the priority to pass through the junction, as prescribed by rule (C) in section 2.

We consider the following three cases for the priorities of the roads 1 and 2 bringing traffic to the circle:

(1) $q_1 = q_1(1, 4R, 1R) = q_2 = q_2(2, 2R, 3R) = 0.25$;
(2) $q_1 = q_1(1, 4R, 1R) = q_2 = q_2(2, 2R, 3R) = 0.5$;
(3) $q_1 = q_1(1, 4R, 1R) = q_2 = q_2(2, 2R, 3R) = 0.75$.

Setting parameters as in (1), $4R$ is the through street with respect to road 1 and road $2R$ is the through street with respect to 2. This means that the traffic inside the circle has the priority with respect to the entering traffic. However, if we fix priorities as in (3), the situation is exactly opposite. Case (2), instead, represents the case of the same priority for the traffic inside the circle and that incoming.

As in [3], we consider the initial data

$$\begin{aligned} &\rho_1(0,x) = 0.25,\ \rho_2(0,x) = 0.4,\ \rho_3(0,x) = 0.5,\ \rho_4(0,x) = 0.5, \\ &\rho_{1R}(0,x) = 0.5,\ \rho_{2R}(0,x) = 0.5,\ \rho_{3R}(0,x) = 0.5,\ \rho_{4R}(0,x) = 0.5, \end{aligned} \tag{5.1}$$

and, for roads entering the circle, we impose the following boundary conditions:

$$\rho_{1,b}(t) = 0.25, \quad \rho_{2,b}(t) = 0.4. \tag{5.2}$$

Starting from the configuration given by initial data (5.1), but setting alternatively the right of way parameters according to the three mentioned cases, the behavior of traffic is different.

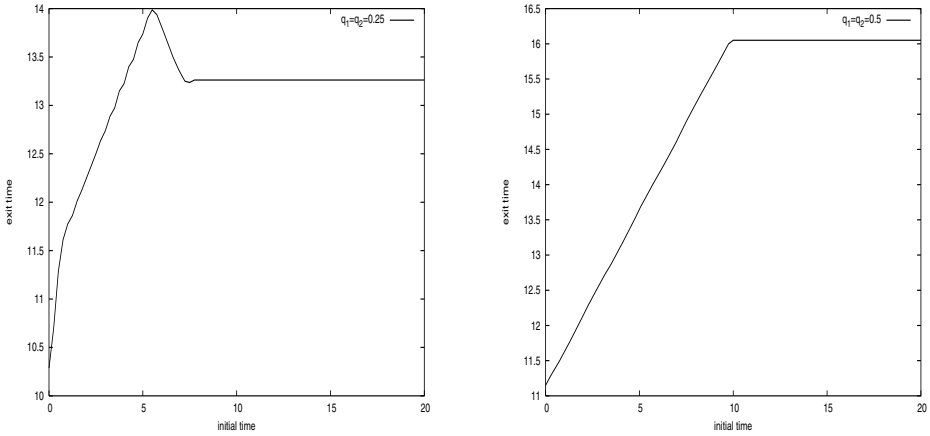


Figure 6. *On the left, exit time for $q_1 = q_2 = 0.25$; on the right, exit time for $q_1 = q_2 = 0.5$ for boundary data* (5.2).

In particular, with priorities set as in (1), shock waves causing an increase in the value of density on incoming roads 1 and 2 are rapidly produced. On the other hand, the density on roads within the circle is kept at a low level.

In case (2), the evolution of the density is similar to the previous case, except for the situation on roads $2R$ and $4R$, where a shock with zero speed is produced.

In case (3) after a short period of time, shocks propagating backward along roads $2R$ and $4R$ provoke an increase in the density inside the circle until traffic becomes completely blocked.

From the analysis of the three considered cases, we want to see how the behavior of a driver moving into the traffic circle can be influenced by the regulation of priority parameters.

Assume that the driver follows the route $(1, 1R, 2R, 3R, 4)$. This means that the car enters the circle from road 1, turns around for 3/4 of the circle, and finally exits to road 4. In the next figures we show that right of way parameters affect the time for covering the path. In particular, in Figure 6 the curve of the exit time as a function of the initial time t_0 varying in $[0, 20]$ is depicted for $q_1 = q_2 = 0.25$ and $q_1 = q_2 = 0.5$. The graph on the left underlines that, after a certain value of the initial time ($t_0 \sim 9$), the exit time becomes stable and corresponds to $T = 13.3$. Similarly, the graph on the right shows that for $t_0 \geq 12$ exit time takes asymptotically the value $T = 16.1$.

On the other hand, setting $q_1 = q_2 = 0.75$, independently by the initial time t_0, the car cannot exit the circle, since traffic within the circle is blocked.

The graph on the left side of Figure 7 represents a comparison between the cases (1) and (2) as regards the space covered by the car when starting time is $t_0 = 8$. As we can observe, taking priority parameters $q_1 = q_2 = 0.25$, the time to exit the circle is lower, even if the first part of the trip is faster for the choice (2). Starting again at time $t_0 = 8$ and setting right of way parameters according to (3), the car cannot exit by the second road of the path, namely, road $1R$, and, consequently, it stops; see the graph in Figure 7 on the right.

Then, keeping the initial data (5.1) as before, we impose at the left endpoint of roads 1

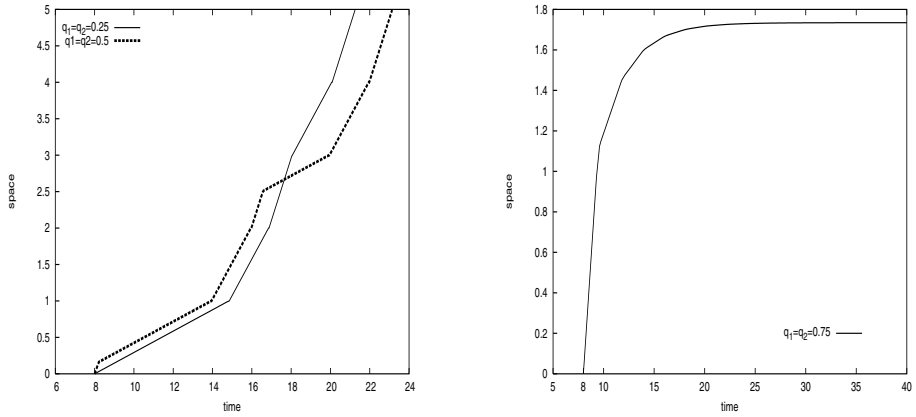


Figure 7. *Time-space diagram of the car trajectory for $t_0 = 8$. On the left, comparison between the cases* (1) *and* (2)*; on the right, priorities set as in* (3).

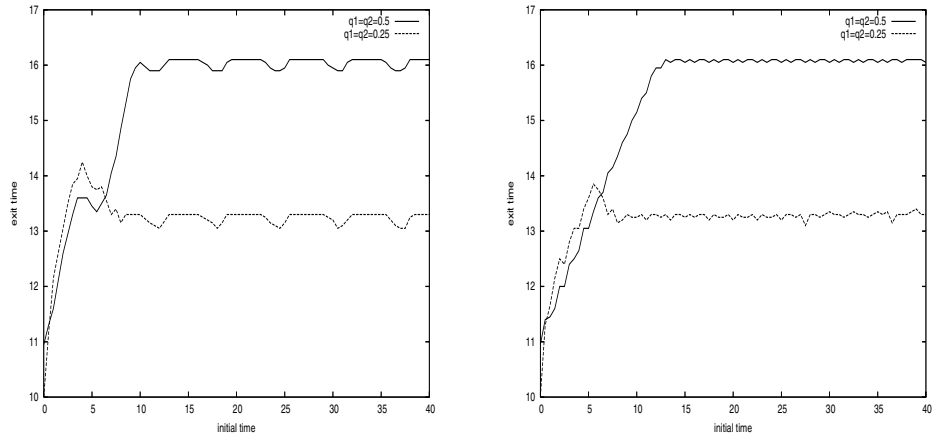


Figure 8. *Exit time for $q_1 = q_2 = 0.25$ (-) and $q_1 = q_2 = 0.5$ $(\cdots)$ when t_0 varies, with $A = 1$ (on the left) and $A = 5$ (on the right) for periodic boundary data* (5.3).

and 2 a periodic data bounded in $[0.1, 0.4]$, namely,

$$\rho_{1,b}(t) = \rho_{2,b}(t) = \frac{1}{4} + \frac{3}{20}\sin(At), \quad t \geq 0, \tag{5.3}$$

where $A > 0$. In Figure 8 the graphs of the time for covering the path for $A = 1$ (on the left) and $A = 5$ (on the right) are depicted, with the initial time t_0 varying in $[0, 40]$.

Observing Figure 8, obtained for $q_1 = q_2 = 0.25$, we can see that after a certain value of the initial time ($t_0 \sim 9$), the exit time becomes more stable, until it reaches the value $T = 13.3$, around which oscillations occur. Similarly, setting $q_1 = q_2 = 0.5$ for $t_0 \geq 12$, the time for covering the path shows oscillations around $T = 16.1$.

This last simulation shows how choice (1) is preferable with respect to choice (2) to reduce the travel time of drivers. Moreover, such convenience is stable with respect to time-varying

Table 1
CPU time and computation of $A_{\Delta x}$ at $T = 8$ for varying Δx.

$T = 8$		
Δx	CPU time	$A_{\Delta x}$
0.1	0.51 s	0.005423
0.05	1.23 s	0.002119
0.025	2.61 s	0.001588
0.0125	5.18 s	0.000635
0.00625	10.13 s	0.000319

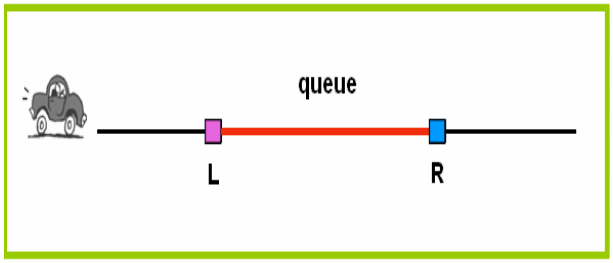


Figure 9. *Car moving on a highway segment where a car accident occurred.*

incoming traffic. In fact, even introducing oscillations with different widths, the two cases of priority parameters still produce quite different travel times.

5.0.1. Test of efficiency and of convergence of the algorithm based on the Godunov scheme. Let us consider a car traveling on a single road parametrized by the interval $[0, 2]$, where the initial and boundary data are, respectively, given by

$$\rho(0,x) = \begin{cases} 0.3 & \text{if } 0 \le x \le 0.2, \\ 0.1 & \text{if } 0.2 < x \le 0.5, \\ 0.8 & \text{if } 0.5 < x \le 2, \end{cases} \qquad \rho(t,0) = 0.3. \tag{5.4}$$

If the car starts traveling at time $t = 0$ with initial position $x^0 = 0$ the time needed to exit the road is $T = 8$. In Table 1 we present the results obtained by applying the simulation algorithm with the Godunov scheme to the proposed example. On one hand we are interested in evaluating the CPU time occupied by the algorithm and on the other hand we estimate the difference between the two positions at final time T, namely, $A_{\Delta x} = |x_{\Delta x}(T) - x_{\frac{\Delta x}{2}}(T)| = |x_\nu(T) - x_{\nu+1}(T)|$, computed numerically when the space mesh parameter $\Delta x = 2^{-\nu}$ decreases. The results in Table 1 represent numerical evidence of the convergence (with a linear rate) of the algorithm, where the evolution in time of the density is computed by the Godunov scheme, which ensures stability of solutions and thus represents a good compromise between numerical accuracy and occupation of CPU time.

All the simulations have been performed by a personal computer, processor AMD Athlon XP 2400 Mhz, RAM 512 Mb.

5.1. Application of the tracking algorithm to highway accidents. We consider the problem of exactly determining the traveling time of a car in case of an accident on a highway.

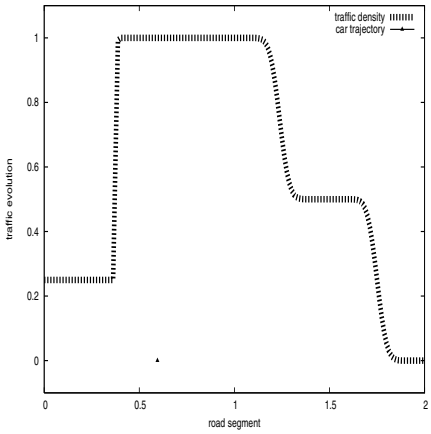


Figure 10. *Evolution of traffic density (line) and car position (point).*

More precisely, we consider a bounded highway segment and a car entering it at time $t = 0$. On the same segment an accident is present at position R and a consequent queue extending backward up to position L; see Figure 9, where (left and right) endpoints of the queue are depicted. We assume to know the inflow of the highway segment and the time t_r at which cars involved in the accident will be removed, thus permitting a free flow of traffic. The question we want to answer is: How much time is needed by the car to cross the congested road?

Using our algorithm we can compute such traveling time. The presence of the accident is simulated as a red light at point R up to time t_r, when the light turns green. If t_r is big enough, the car will reach the back of the queue at L and stop there. After time t_r, the flow will start again at R and the corresponding rarefaction wave (accelerating cars) will eventually reach L, letting the car move.

In Figure 10 a time shot of the evolution of the traffic density on the highway and of the car trajectory is given. Related animations can be found on the web page [2].

6. Conclusions. A new approximation algorithm for tracking the position of a car traveling on a road network is here developed. First the density on the network is simulated using a Godunov or a Wave Front Tracking (WFT) scheme. Then the position of the car is reconstructed determining the effects of interactions with density waves.

The theoretical framework of problem (1.2) was set up in [6].

The algorithm is tested on a portion of an urban network, i.e., a traffic circle. In particular, the time for covering the path of a single driver is measured, showing the convenience of setting the right of way parameters so as to give priority to traffic inside the circle with respect to the entering traffic. Furthermore, a possible application of the algorithm is presented.

As a theoretical result of the present work, the exponential uniform convergence of car trajectories is proved using the WFT algorithm and assuming BV initial data.

Acknowledgment. The authors would like to thank the anonymous referees for many helpful suggestions which improved the paper.

REFERENCES

[1] A. Bressan, *Hyperbolic Systems of Conservation Laws: The One-Dimensional Cauchy Problem*, Oxford University Press, Oxford, UK, 2000.

[2] G. Bretti and B. Piccoli, http://www.iac.rm.cnr.it/~bretti/Car.html.

[3] G. Bretti, R. Natalini, and B. Piccoli, *Numerical approximations of a traffic flow model on networks*, Networks and Heterogeneous Media, 1 (2006), pp. 57–84.

[4] Y. Chitour and B. Piccoli, *Traffic circles and timing of traffic lights for cars flow*, Discrete and Continuous Dynamical Systems Series B, 5 (2005), pp. 599–630.

[5] G. M. Coclite, M. Garavello, and B. Piccoli, *Traffic flow on a road network*, SIAM J. Math. Anal., 36 (2005), pp. 1862–1886.

[6] R. M. Colombo and A. Marson, *Conservation laws and O.D.E.s. A traffic problem*, in Hyperbolic Problems: Theory, Numerics, Applications, Springer-Verlag, Berlin, 2003, pp. 455–461.

[7] R. M. Colombo and A. Marson, *A Hölder continuous O.D.E. related to traffic flow*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 759–772.

[8] M. Garavello, R. Natalini, B. Piccoli, and A. Terracina, *Conservation laws with discontinuous flux*, Netw. Heterog. Media, 2 (2007), pp. 159–179.

[9] M. Garavello and B. Piccoli, *Traffic Flow on Networks*, AIMS Series on Applied Mathematics 1, AIMS, Springfield, MO, 2006.

[10] E. Godlewski and P. A. Raviart, *Hyperbolic Systems of Conservation Laws*, Math. Appl. (Paris) 3/4, Ellipses, Paris, 1991.

[11] S. K. Godunov, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), pp. 271–290.

[12] B. D. Greenshields, *A study in highway capacity*, Highway Research Board Proceedings, 14 (1935), pp. 448–477.

[13] D. Helbing, J. Siegmeier, and S. Lammer, *Self-organized network flows*, Netw. Heterog. Media, 2 (2007), pp. 193–210.

[14] H. Holden and N. H. Risebro, *A mathematical model of traffic flow on a network of unidirectional roads*, SIAM J. Math. Anal., 26 (1995), pp. 999–1017.

[15] H. Holden and N. H. Risebro, *Front Tracking for Hyperbolic Conservation Laws*, Applied Mathematical Sciences 152, Springer-Verlag, New York, 2002.

[16] J. P. Lebacque, *The Godunov scheme and what it means for first order flow models*, in Transportation and Traffic Theory, Proceedings of the 13th ISTT 1996, J. B. Lesort, ed., Pergamon, Oxford, UK, 1996, pp. 647–677.

[17] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 1992.

[18] M. J. Lighthill and G. B. Whitham, *On kinetic waves.* II. *Theory of traffic flows on long crowded roads*, Proc. Roy. Soc. London Ser. A, 229 (1955), pp. 317–345.

[19] P. I. Richards, *Shock waves on the highway*, Operations Res., 4 (1956), pp. 42–51.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 532–560
© 2008 Society for Industrial and Applied Mathematics

Macroscopic Dynamics of Complex Metastable Systems: Theory, Algorithms, and Application to B-DNA*

Illia Horenko†, Evelyn Dittmer†, Filip Lankas‡, John Maddocks‡, Philipp Metzner†, and Christof Schütte†

Abstract. This article is a survey of the present state of the transfer operator approach to the effective dynamics of metastable complex systems, and the variety of algorithms associated with it. We introduce new methods, and we emphasize both the conceptional foundations and the concrete application to the conformation dynamics of a biomolecular system. The algorithmic aspects are illustrated by means of several examples of various degrees of complexity, culminating in their application to a full-scale molecular dynamics simulation of a B-DNA oligomer.

Key words. metastable states, biomolecular conformations, hidden Markov model, maximum likelihood principle, free energy, stochastic differential equations, molecular dynamics, B-DNA, inter base pair parameters

AMS subject classifications. 15A18, 15A51, 60J10, 60J22

DOI. 10.1137/050630064

1. Introduction.

With the increasing availability of ever more powerful computational resources, there is current interest in performing long numerical simulations of large nonlinear dynamical systems, for example, biomolecules, and in examining rather detailed properties of the results. For example, there is a current effort to understand the sequence-dependent physical properties of B-form DNA via the construction and analysis of a self-consistent data base of 39 compatible simulations, each of a 15 base pair fragment or oligomer, with the oligomers constructed in such a way that each of the 136 possible independent tetramer sequences is present at least twice [5]. The time series generated in this particular project comprise more than half a terabyte of data. It is accordingly evident that there is an ever increasing need to analyze such time series efficiently, with mathematical algorithms that are practical for data sets of this order of magnitude. In particular, many nonlinear dynamical systems, including biomolecules and specifically DNA, exhibit the phenomenon of *metastability*; i.e., the trajectory is localized in one subregion of phase space for comparatively long time scales, before undergoing a rapid and rare transition to another region, where it then stays for a comparatively long residency time before eventually undergoing another rapid transition, and so on. Figure 10 illustrates this phenomenon via a plot of a single, scalar-dependent variable, in this

*Received by the editors April 27, 2005; accepted for publication (in revised form) by L.-S. Young September 21, 2007; published electronically May 23, 2008. This work was supported by the DFG research center "Mathematics for key technologies" (FZT 86) in Berlin, and by the Swiss National Science Foundation.
http://www.siam.org/journals/siads/7-2/63006.html
†Institute of Mathematics II, Free University Berlin, Arnimallee 2–6, 14195 Berlin, Germany (horenko@math.fu-berlin.de, dittmer@mi.fu-berlin.de, metzner@mi.fu-berlin.de, schuette@math.fu-berlin.de).
‡Mathematics Institute B, Ecole Polytechnique Federale de Lausanne, CH-1015 Lausanne, Switzerland (filip.lankas@epfl.ch, john.maddocks@epfl.ch).

case a certain torsional angle in one of the DNA backbones between two particular base pairs in a simulation of a poly(AT) oligomer, where A and T denote the base pair of adenine (A) and thymine (T). The time series is of length 10^5, being a sampling of a 100 nanosecond simulation at every picosecond. The time series is evidently multiwell and exhibits metastability with essentially instantaneous sharp transitions between wells that are separated by 180° or so, with rapid oscillations within each well during the long residency times.

The purpose of this article is to survey existing, and introduce new, methods for the identification of the metastable substates that are exhibited in a particular time series, and to estimate the transition probabilities between these sets. The primary messages of the article are the following. First, the notion of the number of metastable states is a hierarchical concept—the appropriate number of metastable sets to be identified in a given time series depends upon the phenomena to be modeled. Second, and nevertheless, the numbers of metastable states are not arbitrary; rather, appropriate choices of the numbers of metastable sets can be identified via various clustering techniques. For example, if the dimension of the time series is not too large, a certain transfer operator can be explicitly computed and the appropriate possible numbers of metastable states can be associated with gaps in the spectrum of the operator via a Perron cluster analysis, while the metastable sets themselves can be identified from the associated eigenfunctions. Third, the usual methods for the computation of the transfer operator suffer from the *curse of dimensionality*, which means that the methods are not practicable for large systems. However, when the dimensionality of the time series is too high, one may be able to use hidden Markov models (HMMs) or the new method of HMM stochastic differential equations (HMMSDEs) to identify metastable substates and to make good estimates of the associated transition probabilities.

The theory developed in the article is illustrated with two examples. First, there is an entirely tutorial and two-dimensional (2D) example involving the high friction or overdamped Brownian dynamics of a particle in a multiwell potential, in which all the conclusions are entirely explicit. Second, the theory is applied to the DNA simulation already mentioned above. That series is of length 10^5 and is of high dimension (for details see section 2). In this context the metastability analysis plays an important role in identifying basins within which the base pair level, structural shape, and stiffness parameters of DNA can be approximated. The DNA example lies within the class of problems that are too large for an explicit computation of the transfer operator for the full system. However, we demonstrate that the metastable sets can be captured via an HMMSDE analysis. We apply the HMMSDE analysis to a description of the system in which the coordinates are backbone angles. In these coordinates the transitions are very rapid, and the states can also be identified via a more standard, but appropriately aggregated, HMM model. However, the HMMSDE analysis additionally yields insight about nanomechanical properties of the molecule within each metastable well, e.g., local stiffness matrices that may allow us to study the changes in elastic properties between conformations.

2. Metastability.

The evolution of a single microscopic system is assumed to be given by a *homogeneous Markov process* $X_t = \{X_t\}_{t\in\mathbf{T}}$ in either continuous or discrete time with state space $\mathbf{X}$. We write $X_0 \sim \mu$ if the Markov process X_t is initially distributed according to the probability measure μ. The motion of X_t is given in terms of the *stochastic transition function*

$$p(t, x, A) \;=\; \mathbf{P}[X_{t+s} \in A \;|X_s = x] \tag{1}$$

for every $t, s \in \mathbf{T}$, $x \in \mathbf{X}$, and $A \subset \mathbf{X}$ that satisfies the well-known Chapman–Kolmogoroff equation $p(t+s, x, A) = \int_{\mathbf{X}} p(t, x, \mathrm{d}z)\, p(s, z, A)$ [15].

We say that the Markov process X_t admits an *invariant probability measure* μ, or that μ is invariant w.r.t. X_t if $\int_{\mathbf{X}} p(t, x, A)\mu(\mathrm{d}x) = \mu(A)$. In the following we always assume that the invariant measure of the process under investigation exists and is unique. A Markov process is called *reversible* w.r.t. an invariant probability measure μ if $\int_A p(t, x, B)\mu(\mathrm{d}x) = \int_B p(t, x, A)\mu(\mathrm{d}x)$ for every $t \in \mathbf{T}$ and $A, B \subset \mathbf{X}$.

2.1. Transition probabilities and transfer operators. Metastability of some subset of the state space is characterized by the property that the dynamical system is likely to remain within the subset for a long period of time, until it exits and a transition to some other region of the state space occurs. There are, in fact, several related but different definitions of metastability in literature (see, e.g., [8, 11, 41, 42]); we will focus on the so-called ensemble concept introduced in (2); for a comparison with, e.g., the exit time concept, see [40].

The objective is an identification of a *decomposition of the state space into metastable subsets* and the corresponding "flipping dynamics" between these substates. In general, a *decomposition* $\mathbf{D} = \{D_1, \ldots, D_m\}$ of the state space $\mathbf{X}$ is a collection of subsets $D_k \subset \mathbf{X}$ with the following properties: (1) positivity $\mu(D_k) > 0$ for every k, (2) disjointness up to null sets, and (3) the covering property $\cup_{k=1}^{m} \overline{D_k} = \mathbf{X}$. In particular, the appropriate number m of metastable subsets must be identified. Within a transfer operator approach this can be achieved via spectral analysis (see *key idea* below).

We define the *transition probability* $p(t, B, C)$ from $B \subset \mathbf{X}$ to $C \subset \mathbf{X}$ within the time span t as the conditional probability

$$p(t, B, C) \;=\; \mathbf{P}_\mu[X_t \in C \;|X_0 \in B] \;=\; \frac{\mathbf{P}_\mu[X_t \in C \text{ and } X_0 \in B]}{\mathbf{P}_\mu[X_0 \in B]}, \tag{2}$$

where $\mathbf{P}_\mu$ indicates that initially $X_0 \sim \mu$. Then (2) may be rewritten as

$$p(t, B, C) \;=\; \frac{1}{\mu(B)} \int_B p(t, x, C)\, \mu(\mathrm{d}x). \tag{3}$$

In other words, the transition probability quantifies the dynamical fluctuations within the stationary ensemble μ. Concomitant with our ensemble dynamics approach to metastability, we call a subset $B \subset \mathbf{X}$ *metastable* on the time scale $\tau > 0$ if

$$p(\tau, B, B^c) \approx 0 \;\text{ or, equivalently, }\; p(\tau, B, B) \;\approx\; 1,$$

where $B^c = \mathbf{X} \setminus B$ denotes the complement of B.

Transfer operator. We define the *semigroup of propagators* or forward transfer operators $P^t : L^r(\mu) \to L^r(\mu)$ with $t \in \mathbf{T}$ and $1 \le r < \infty$ as follows:

$$\int_A P^t v(y)\; \mu(\mathrm{d}y) \;=\; \int_{\mathbf{X}} v(x) p(t, x, A) \mu(\mathrm{d}x) \tag{4}$$

for $A \subset \mathbf{X}$. As a consequence of the invariance of μ, the characteristic function $\mathbf{1}_{\mathbf{X}}$ of the entire state space is an invariant density of P^t; i.e., $P^t \mathbf{1}_{\mathbf{X}} = \mathbf{1}_{\mathbf{X}}$. Furthermore, P^t is a Markov

operator; i.e., P^t conserves both norm $\|P^t v\|_1 = \|v\|_1$ and positivity $P^t v \geq 0$ if $v \geq 0$, which is a simple consequence of the definition. Due to (4), the semigroup of propagators mathematically models the evolution of subensembles in time.

The *key idea of the transfer operator approach* w.r.t. the identification of metastable decompositions can be described as follows:

> Metastable subsets can be detected via eigenvalues of the propagator P close to its maximal eigenvalue $\lambda = 1$; moreover, they can be identified by exploiting the corresponding eigenfunctions. In doing so, the number of metastable subsets is equal to the number of eigenvalues close to 1, including $\lambda = 1$ and counting multiplicity.

This strategy was first proposed by Dellnitz and Junge [12] for discrete dynamical systems with weak random perturbations and has been successfully applied to molecular dynamics in different contexts [38, 39, 40]. Its justification is given below. The key idea requires the following two *conditions on the propagator* P:

- (C1) The essential spectral radius of P is less than 1; i.e., $r_{\text{ess}}(P) < 1$.
- (C2) The eigenvalue $\lambda = 1$ of P is simple and dominant; i.e., $\eta \in \sigma(P)$ with $|\eta| = 1$ implies $\eta = 1$.

In this article, two types of Markov processes will be considered: (1) high-friction Langevin processes and (2) (Nosé–Hoover) constant temperature molecular dynamics. For both cases the dynamics is reversible, and the transfer operator is self-adjoint. For type (1) examples, conditions (C1) and (C2) are known to be satisfied under rather weak conditions on the potential [40]. For type (2) examples, it is unknown whether or not the conditions are satisfied; however, it is normally assumed in molecular dynamics that they are valid for realistically complex systems in solution.

We define the *metastability of a decomposition* $\mathbf{D}$ as the sum of the metastabilities of its subsets. That is, suppose that the time scale τ of interest is fixed. Then, for each arbitrary decomposition $\mathcal{D}_m = \{A_1, \ldots, A_m\}$ of the state space $\mathbf{X}$ into m sets, we define its metastability measure by

$$\text{meta}(\mathbf{D}_m) = \sum_{j=1}^{m} p(\tau, A_j, A_j)/m.$$

For given m the optimal metastable decomposition into m sets can then be defined as that decomposition into m sets which maximizes the functional meta.

The next result [29] justifies the above key idea.

Theorem 2.1. *Let $P^\tau : L^2(\mu) \to L^2(\mu)$ denote a reversible propagator satisfying* (C1) *and* (C2). *Then P^τ is self-adjoint with spectrum of the form*

$$\sigma(P^\tau) \subset [a, b] \cup \{\lambda_m\} \cup \cdots \cup \{\lambda_2\} \cup \{1\}$$

with $-1 < a \leq b < \lambda_m \leq \cdots \leq \lambda_1 = 1$ and λ_i isolated eigenvalues that are counted according to their finite multiplicities. Denote by $v_m, \ldots, v_1$ the corresponding eigenfunctions, normalized to $\|v_k\|_2 = 1$. Let Q be the orthogonal projection of $L^2(\mu)$ onto $\text{span}\{\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_m}\}$. *Then the metastability of an arbitrary decomposition $\mathbf{D}_m = \{A_1, \ldots, A_m\}$ of the state space $\mathbf{X}$ can be bounded from above by*

$$p(\tau, A_1, A_1) + \cdots + p(\tau, A_m, A_m) \leq 1 + \lambda_2 + \cdots + \lambda_m,$$

while it is bounded from below according to

$$1 + \kappa_2\lambda_2 + \cdots + \kappa_m\lambda_m + c \leq p(\tau, A_1, A_1) + \cdots + p(\tau, A_m, A_m),$$

where $\kappa_j = \|Qv_j\|^2_{L^2(\mu)}$ *and* $c = a\,((1-\kappa_2) + \cdots + (1-\kappa_n))$.

Theorem 2.1 highlights the strong relation between a decomposition of the state space into metastable subsets and a *Perron cluster* of dominant eigenvalues close to 1. It states that the metastability of an arbitrary decomposition $\mathbf{D}_m$ cannot be larger than $1 + \lambda_2 + \cdots + \lambda_m$, while it is at least $1 + \kappa_2\lambda_2 + \cdots + \kappa_m\lambda_m + c$, which is close to the upper bound whenever the dominant eigenfunctions $v_2, \ldots, v_m$ are almost constant on the metastable subsets $A_1, \ldots, A_m$ implying $\kappa_j \approx 1$ and $c \approx 0$. The term c can be interpreted as a correction that is small whenever $a \approx 0$ or $\kappa_j \approx 1$. It is demonstrated in [29] that the lower and upper bounds are sharp and asymptotically exact.

2.2. Metastability analysis is hierarchical. The last theorem and the illustrations and examples below contain one main message about metastability analysis: *it has to be hierarchical*. Whenever we approximate the optimal metastable decomposition $\mathbf{D}_2$ of state space into, say, two sets, we should always be aware that there could be a decomposition $\mathbf{D}_3$ into three sets for which meta($\mathbf{D}_3$) is almost as large as meta($\mathbf{D}_2$). For example, one or both of the two subsets in $\mathbf{D}_2$ could decompose into two or several metastable subsets from which exit is comparably difficult for the system under investigation.

However, whenever there is a gap in the spectrum of the transfer operator after m dominant eigenvalues, then the results of, e.g., [28, 8], tell us that any decomposition into more than m sets will be associated with a significantly larger drop in metastability as measured by the function meta. In the context of applications to molecular dynamics, however, one should always be aware that particular aspects of interest may make it desirable to explore the hierarchy of metastable decompositions up to a certain depth that is not necessarily selected only on the values of the functional meta.

2.3. Discretization and PCCA. Let $\chi = \{\chi_1, \ldots, \chi_n\} \subset L^2(\mu)$ denote a set of *nonnegative* functions that are a partition of unity, i.e., $\sum_{k=1}^n \chi_k = \mathbf{1}_\mathbf{X}$. The *Galerkin projection* $\Pi_n : L^2(\mu) \to \mathcal{S}_n$ onto the associated finite-dimensional ansatz space $\mathcal{S}_n = \text{span}\{\chi_1, \ldots, \chi_n\}$ is defined by

$$\Pi_n v \;=\; \sum_{k=1}^{n} \frac{\langle v, \chi_k\rangle_\mu}{\langle \chi_k, \chi_k\rangle_\mu}\, \chi_k.$$

Application of the Galerkin projection to $P^\tau v = \lambda v$ yields an eigenvalue problem for the discretized propagator $\Pi_n P^\tau \Pi_n$ acting on the finite-dimensional space $\mathcal{S}_n$. The matrix representation of this finite-dimensional operator is given by the $n \times n$ *transition matrix* $\top = (\top_{kl})$, whose entries are given by

$$\top_{kl} \;=\; \frac{\langle P^\tau \chi_k, \chi_l\rangle_\mu}{\langle \chi_k, \chi_k\rangle_\mu}. \tag{5}$$

The transition matrix inherits the main properties of the transfer operator: it is a stochastic matrix with invariant measure given by the invariant measure μ of P^τ, it is reversible if P^τ

is self-adjoint, and (if the discretization is fine enough) it also exhibits a Perron cluster of eigenvalues that approximates the corresponding Perron cluster of P^τ, and with eigenvectors that approximate the dominant eigenvectors of P^τ [40]. It thus allows us to compute the metastable sets of interest by computation of the dominant eigenvectors of $\top$ and by realization of the identification strategy described in section 2.1 based on these (discrete) eigenvectors. This has led to the construction of an aggregation technique called "Perron cluster cluster analysis" (PCCA) [13, 14].

The entries of $\top$ can be computed from realizations of the underlying Markov process X_t. We have

$$\top_{kl} \;=\; \frac{1}{\langle \chi_k, \chi_k \rangle_\mu} \int_{\mathbf{X}} \chi_k(x) \mathbf{E}_x[\chi_l(X_\tau)]\, \mu(\mathrm{d}x).$$

If $x_0, \ldots, x_N$ denote a time series obtained from a realization of the Markov process with time stepping τ, then the entries of $\top$ can be approximated from the relative transition rates computed by means of this time series:

$$\top_{kl} \;\approx\; \top_{kl}^{(N)} \;=\; \frac{\sum_{j=1}^{N} \chi_k(x_j) \cdot \chi_l(x_{j+1})}{\sum_{j=1}^{N} \chi_k(x_j)^2}. \tag{6}$$

For a time series of whatever length, but with a high-dimensional configuration variable, practical evaluation of the formula (6) may become problematic. There are two main reasons for potential difficulties.

Trapping problem. The *rate of convergence* of $\top_{kl}^{(N)} \to \top_{kl}$ depends on the smoothness of the partition functions χ_k as well as on the mixing properties of the Markov process [31]. The latter property is crucial here: the convergence is geometric with a rate constant $\lambda_1 - \lambda_2 = 1 - \lambda_2$, where λ_2 denotes the second largest eigenvalue (in modulus). In the case of metastability with λ_2 being very close to $\lambda_1 = 1$, we will have dramatically slow convergence. This is of no surprise because closeness of dominant eigenvalues is typically the main difficulty in all approaches to biomolecular dynamics and statistics, and it is also a bottleneck of the transfer operator approach. Much of the literature aims to tackle this *trapping problem* [4, 17]. In our largest examples, evaluation of (6) is not practical. However, we will *not* go into the depth of the discussion on overcoming the trapping problem, because we propose alternative approaches. We will simply assume in all of the following that we have already generated or can directly generate a time series that is "long enough" in the sense that it contains statistically significant information about more than one—if not all—interesting metastable states of the system under consideration. We will discuss later whether this is the case for our poly(AT) DNA time series.

Curse of dimension. Any discretization of the transfer operator will suffer from the curse of dimension whenever it is based on a uniform partition of all of the hundreds or thousands of degrees of freedom in a typical biomolecular system. Fortunately, chemical observations reveal that—even for larger biomolecules—the curse of dimensionality can be circumvented by exploiting the hierarchical structure of the dynamical and statistical properties of biomolecular systems: only relatively few *essential degrees of freedom* are needed to describe the conformational transitions (see next section); furthermore, the canonical density has a rich spatial multiscale structure induced by the rich structure of the potential energy landscape. This

structure induces a hierarchical cluster structure of the sampling data that can be identified and used to define a multilevel discretization adapted to the structures of the statistical data (see [40] or subsequent examples).

2.4. Illustrative example. For simplicity we consider the Markov process given by the so-called high-friction Langevin equation which is the limit of high friction of the famous Langevin equation; see [36, 39]. The high-friction Langevin equation is stated in the position space only and is given by the equation

$$\dot{x} = -\nabla_x V(x) + \sigma \dot{W}_t, \tag{7}$$

with $x(t) \in \mathbf{R}^d$ being the position vector of the system, W_t denoting d-dimensional standard Brownian motion, and σ denoting the noise intensity parameter. The SDE (7) defines a continuous time Markov process X_t with invariant probability measure $\mu(\mathrm{d}x) \propto \exp(-\beta V(x))\mathrm{d}x$ with $\beta = 2/\sigma^2$ [36]. There is a long history of using it as a simple toolkit for investigation of dynamical behavior in complicated energy landscapes [10]. It is known that under weak conditions on the potential function V the Markov process is reversible [26].

The associated semigroup (P^t) of propagators admits a strong generator $\mathcal{A}$ such that the semigroup can be written as $P^t = \exp(t\mathcal{A})$. For twice continuously differentiable $u \in L^2(\mu)$ we have the identity

$$\mathcal{A}\,u \;=\; \left(\frac{\sigma^2}{2}\Delta_x - \nabla_x V(x)\cdot\nabla_x\right) u.$$

For details on $\mathcal{A}$ see the theory of Fokker–Planck equations and Kolmogoroff forward and backward equations [36, 41]. Under appropriate conditions (the Perron cluster is a discrete part of the spectrum) we can compute the dominant eigenvectors of P^t via those of $\mathcal{A}$.

For illustrative means we use the potential V illustrated in Figure 1 (thus setting $d = 2$). Figure 2 shows typical realizations of the high-friction Langevin Markov process associated with this potential (setting $\sigma = 0.131$). We observe that the vicinity of the wells in the potential energy landscape can be approximately identified with the metastable sets of the process; it is well known from large deviation theory that, in fact, for small enough noise intensity, the vicinity of the wells of the potential energy landscape are the metastable sets of high-friction Langevin processes (at least such wells that are separated from each other by significant energy barriers) [28].

Next, we discretized the transfer operator of the process (fine grid with 100×100 discretization boxes in discretization domain $[-3, 3] \times [-3, 3]$) for different values of τ, which results in the dominant eigenvalues listed in Table 1.

While the eigenvector of the largest eigenvalue is constant, the corresponding second and third eigenvectors of P^τ in $L^2(\mu)$ are shown in Figure 3 (they are identical for all values of τ because of the semigroup property).

Having computed the dominant eigenvectors, we can determine the optimal metastable decomposition by means of PCCA as introduced above. The results on the spectrum (see $\tau = 0.1$, for example) exhibit a hierarchy of metastability that is in perfect agreement with the general insight on metastability of high friction Langevin motion: we can apply PCCA to the first *two* eigenvectors of the transfer operator; this results in the metastable decomposition

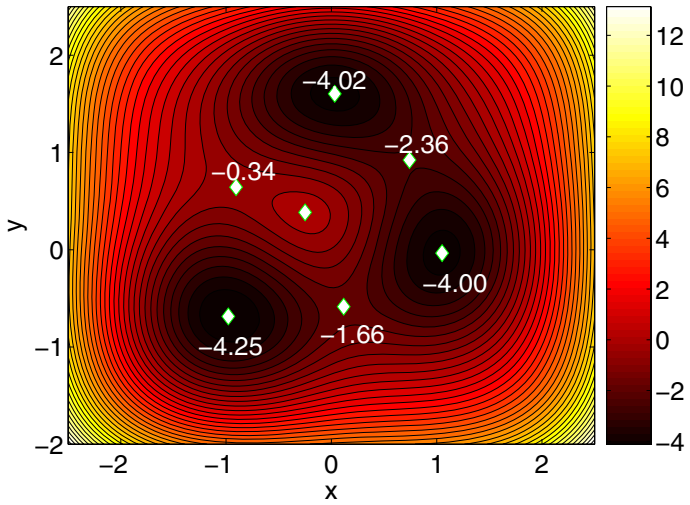


Figure 1. *Potential V used for illustrative example. We observe three wells in the potential landscape (see colorbar). The tags indicate the minima and saddle points of the potential; the numbers give the value of the potential at these points. We observe that the leftmost minimum is the deepest well separated by the most pronounced energy barrier from the other two.*

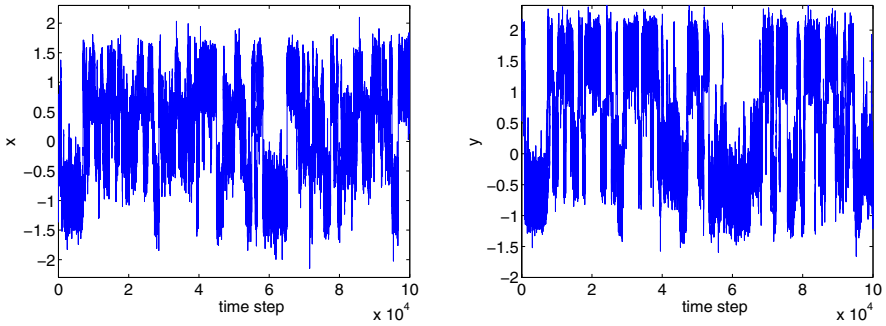


Figure 2. *Typical realization of the high-friction Langevin dynamics (both components (left/right) of the state versus time) in the potential energy landscape V shown in Figure 1 for $\sigma = 0.131$.*

Table 1

Leading four eigenvalues of transfer operator P^τ for different values of τ for high-friction Langevin motion with potential and parameters as described in the text.

$\sigma(P^\tau)$	λ_1	λ_2	λ_3	λ_4
$\tau = 0.01$	1.000	0.999	0.997	0.959
$\tau = 0.10$	1.000	0.994	0.975	0.656
$\tau = 1.00$	1.000	0.937	0.776	0.015

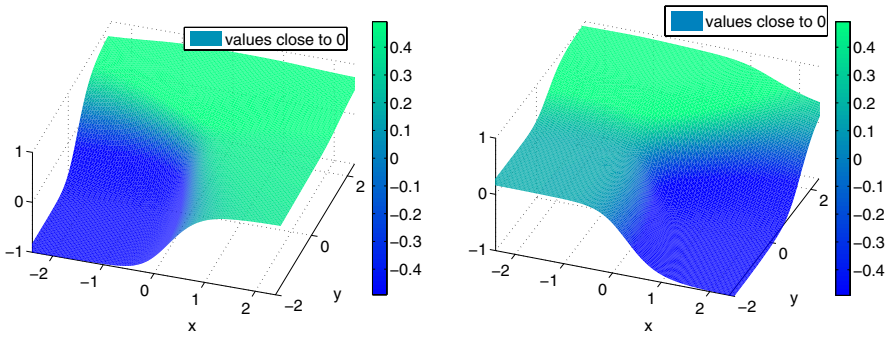


Figure 3. *Second and third eigenvectors of the transfer operator for high-friction Langevin process in potential of Figure 1 (details see text).*

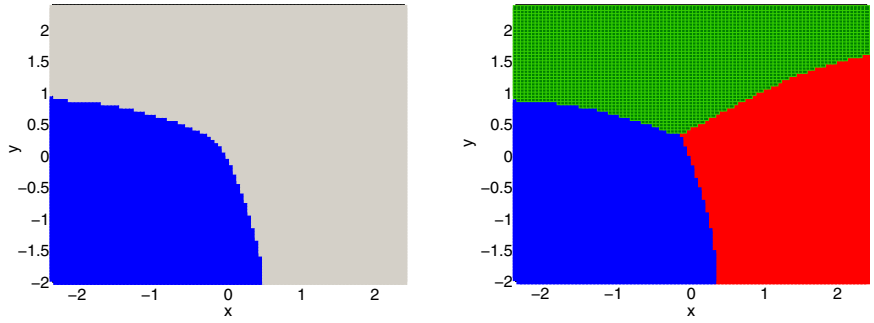


Figure 4. *Optimal metastable decomposition resulting from PCCA based on the first two (left) and first three (right) eigenvectors of the transfer operator.*

that distinguishes between the vicinity of the deepest well and the remaining state space (see Figure 4, left). When applying PCCA to the first *three* eigenvectors, however, the resulting metastable decomposition identifies the vicinities of all three wells as the metastable regions of the system (see Figure 4, right). This outcome is desirable and typical: metastable decomposition via spectral properties of the transfer operator is hierarchical in the sense that the process of including more and more leading eigenvalues uncovers finer and finer details of metastability within the system; see [40, 28].

So, what happens if we take the first four eigenvectors? This we can immediately understand by comparing the values of the functional meta for the optimal metastable decompositions $\mathbf{D}_m$ into $m = 1, 2, 3, 4$ sets ($\tau = 1$) as given in Table 2: Between $m = 3$ and $m = 4$ there is a significant drop in metastability, indicating that it makes no real sense to speak of four metastable sets for the system under consideration.

3. Algorithms. As already mentioned we assume herein that some "long enough" time series $(x_t)_{t=t_0,\ldots,t_N}$ of states (i.e., atomic positions and or momenta) of the system under consideration is already given. We are mainly interested in the case that the states x_{t_l} are from some high-dimensional state space $\mathbf{R}^d$. In this section we will often consider time series of some observables $(Z_t)_{t=t_0,\ldots,t_N}$ computed from the time series (x_t), e.g., the time series of torsion angles or inter base pair parameters; however, we could also consider the case $Z = x$.

Table 2

Metastabilities of the optimal metastable decomposition $\mathbf{D}_m$ into $m = 1, 2, 3, 4$ sets (as computed by PCCA from the dominant eigenvectors) and its theoretical upper bound as in Theorem 2.1.

m	1	2	3	4
$\mathrm{meta}(\mathbf{D}_m)$	1.000	0.967	0.899	0.613
$\frac{1}{k}\sum_{k=1}^{m}\lambda_k$	1.000	0.969	0.904	0.682

We will subsequently follow the following basic idea: from statistical analysis of the time series given we (1) construct a finite-state Markov jump process that models the hops *between* the metastable conformations, and (2) for each conformation we parameterize an appropriate stochastic model that allows us to approximate the dynamics of the systems as long as it is *within* the respective conformation. In [23, 24, 25] appropriate algorithms have been derived that combine hidden Markov models (for the construction of the unobserved jump process) and optimal, likelihood-based parameterization of the local stochastic models. We will review this framework here.

3.1. Algorithms based on hidden Markov models. Suppose the system under consideration has a metastable decomposition. Then, at any time t the system will be in exactly one of the associated metastable sets $B_q \subset \mathbf{X}$, $q = 1, \ldots, m$. Therefore, at each time t we have some "metastable state" $q(t)$ being represented by the number of the presently visited metastable set. Whenever a time series of values of observables (or "observations") $Z = (Z_t)_{t=t_0,\ldots,t_N}$ is given, we want to identify the time series of metastable states $q = (q_t)_{t=t_0,\ldots,t_N}$ associated with it. However, while the time series (Z_t) is observed, i.e., known, the series (q_t) is *hidden* within the data.

Suppose that the observed data (Z_t) is given with constant time stepping τ; i.e., $t_k = t_{k-1} + \tau$ for all $k = 1, \ldots, N$. Setting $t_0 = 0$, we have $t_k = k\tau$ and especially $T = t_N = N\tau$. For the sake of simplicity of notation we thus may simply write $t = 0, \ldots, T$.

The probability to go from one metastable/hidden state q to another, q', is given by $\top_{qq'} = p(\tau, B_q, B_{q'})$. That is, the sequence (q_t) should be seen as the realization of a Markov chain with M states with transition matrix $\top$.

The observations Z_t somehow result from the respective hidden state q_t by a priori unknown rules. For given time series of observations $Z = (Z_t)_{t=t_0,\ldots,t_N}$ one is interested in finding the most probable series of metastable/hidden states.

Models like the one coarsely described above are well known as *hidden Markov models* (HMMs). An HMM is a stochastic process with hidden and observable states; the hidden states of an HMM form a Markov chain, while the observable states are understood as output that is distributed according to a certain conditional distribution (conditioned to the hidden state).

To describe the whole system, we need to know the number M of hidden states, the transition matrix $\top$ between them, an initial distribution, and, for each state, a certain rule governing the probability distribution for the observation.

Stationary output. In standard HMMs the output distributions result from independent and identically distributed random variables; i.e., consecutive output states are statistically independent. That is, conditioned on the hidden state, the output is simply randomly chosen

from a stationary distribution. In application to the analysis of data produced in the context of molecular dynamics, this means that the system reaches the thermodynamical equilibrium distribution immediately after each transition between metastable states; this then is related to abrupt jumps of some physical observables.

The most popular choice for such stationary distributions are (multivariate) normal distributions. However, in the case of circular data (like torsion angle positions in a molecular dynamics simulation) the use of normal distributions often induces crucial problems due to periodicity, and thus they have to be replaced by von Mises distributions [32].

The problem of the statistical analysis of the time series in this case will be reduced to the identification of the Markov transition matrix and equilibrium statistical distributions (often specified in a parameterized form) [3, 34, 35]. This approach was recently successfully applied to analysis of torsion angle dynamics of a trialanine molecule [19].

SDE output. In contrast to the standard HMM approach with stationary output, molecular systems typically do not reach local equilibrium immediately after each jump of the Markov chain but relax into local equilibrium after some characteristic *relaxation time* (see Figure 11 in comparison to Figure 10). Moreover, when replacing stationary output distributions by stochastic models of the local dynamics we can hope to get more insight into the dynamical flexibilities of the molecular system within each conformation and perhaps of the mechanical processes governing this flexibility.

In order to incorporate these aspects, we couple (1) the finite-state Markov jump process that models the hops *between* the hidden states (= metastable conformations), and (2) SDE dynamics *within* the respective conformation. Putting these types of local SDEs and the jump process between conformations together, we get models of the form

$$\begin{aligned} &\dot{z}(t) = F^{(q(t))}(z - \mu^{(q(t))}) + \Sigma^{(q(t))}\,\dot{W}(t), \\ &q(t) = \text{Markov jump process with states } 1, \ldots, L, \end{aligned} \tag{8}$$

where $W(t)$ is denoting standard d'-dimensional Brownian motion (where d' is the dimension of the observations Z_t), $(\Sigma^{(1)}, \ldots, \Sigma^{(L)})$ noise intensity matrices, $(\mu^{(1)}, \ldots, \mu^{(L)})$ equilibrium positions, and $(F^{(1)}, \ldots, F^{(L)})$ appropriate stiffness matrices. The general aim of the HMMSDE extension of the HMM approach is to find the optimal model of the above form for a given time series (Z_t) (in a maximum likelihood sense).

The formal solution to the local SDE $\dot{z} = F(z - \mu) + \Sigma\dot{W}$ on the time interval $[t, t + \tau]$ is given by

$$z(t + \tau) = \mu + e^{\tau F}\,(z(t) - \mu) + \int_0^\tau e^{(\tau - s)F}\Sigma dW(s). \tag{9}$$

Thus, the probability density $\rho(Z_{k+1}|Z_k)$ of observation of Z_{k+1} at time $k + 1$ under the condition of observation of Z_k at k is proportional to

$$\exp\left[-\frac{1}{2}\xi_k^\top R^{-1}(\tau)\,\xi_k\right],$$

where

$$\xi_k = Z_{k+1} - \bar{\mu} - e^{\tau F}\,(Z_k - \bar{\mu})\,, \tag{10}$$

$$R(\tau) = \int_0^\tau e^{sF}\Sigma\Sigma^\top e^{sF^\top}\,ds. \tag{11}$$

We will now exploit this observation to construct the likelihood function of HMMSDEs.

Optimal parametrization. Both types of HMMs, whether with stationary output distribution or with SDE output, contain sets of parameters (the entries of the transition matrix $\top$, the initial output distribution v, as well as the parameters of SDEs or output distributions), herein denoted by θ. We now want to identify optimal parameters for *given* observation data $Z = (Z_t)_{t=1,\dots,T}$. We have to define the likelihood functional w.r.t. which we then will have to determine the optimal parameters θ and the sequence of hidden states $q = (q_t)_{t=1,\dots,T}$.

For given parameters θ, the likelihood $\mathcal{L}(\theta|Z_t, q_t)$ has to be the probability of output Z_t under the condition of being in metastable state q_{t_j} for given parameters θ:

$$\mathcal{L}(\theta|Z_t, q_t) = p(Z, q|\theta) = v(q_0)\rho(Z_0|q_0)\prod_{t=1}^{T}\top(q_{t-1}, q_t)\rho(Z_t|q_t, Z_{t-1}), \tag{12}$$

where $\rho(\cdot|q, Z_{t-1})$ denotes the output distribution at time t under the condition that the system is in hidden state q_t. According to our observations above we have

$$\rho(Z_t|q_t, Z_{t-1}) \propto \exp\left[-\frac{1}{2}\xi_t^\top R^{-1}(\tau)\,\xi_t\right],$$

with ξ_t and R as given by (10) and (11) with $k = t$, $F = F^{(q_t)}$, $\Sigma = \Sigma^{(q_t)}$, and $\mu = \mu^{(q_t)}$.

The next task now will be to construct algorithms that

(1) determine the optimal parameters $(\top, \mu^{(q)}, F^{(q)}, \Sigma^{(q)})_{q=1,\dots,M}$ by maximizing the likelihood $\mathcal{L}(\theta|Z, q)$ (this is a nonlinear global optimization problem),
(2) determine the optimal sequence of hidden metastable states (q_t) for given optimal parameters, and
(3) determine the number of important metastable states. Up to now we also simply assumed that the number M of hidden states is a priori given, but how can we determine the appropriate number?

To solve problem (1) we will use the *expectation-maximization* (EM) algorithm. The EM algorithm is a learning algorithm: it alternately iterates two steps, the expectation step and the maximization step. Starting with some initial parameter set θ_0, the steps iteratively refine the parameter set; i.e., in step k the present parameter set θ_k is refined to θ_{k+1}. We will work out the details of the EM algorithm for the problem under investigation by following the general framework given in [6].

The key object of the EM algorithm is the expectation

$$Q(\theta, \theta_k) \;=\; \mathbf{E}\Big(\log p(Z, q|\theta) \mid Z, \theta_k\Big) \tag{13}$$

of the complete-data likelihood $\mathcal{L}(\theta|Z, q) = p(Z, q|\theta)$ (in our case given by (12)) w.r.t. the hidden sequence q given the observation sequence and the current parameter estimate θ_k. One step of the EM algorithm then realizes the following two steps:

- Expectation step (E-step). This step evaluates the expectation value Q based on the given parameter estimate θ_k.
- Maximization step (M-step). This step determines the refined parameter set θ_{k+1} by maximizing the expectation:

$$\theta_{k+1} = \operatorname*{argmax}_{\theta} Q(\theta, \theta_k). \tag{14}$$

The maximization guarantees that $\mathcal{L}(\theta_{k+1}) \geq \mathcal{L}(\theta_k)$.

Algorithmic realizations of these two steps are standard for stationary Gaussian and von Mises output [19]; for SDEs the results of the maximization step are given below; further algorithmic details can be found in [23, 24, 25]. In both cases the necessary computational effort for one step of the EM algorithm scales linearly in the length of the observation sequence and quadratically in the number of hidden states.

Results of M-step in HMMSDE. In addition to the observation sequence $(Z_k)_{k=1,2,\ldots,}$ the E-step in each iteration yields occupation probabilities $\nu_k(q)$, $q = 1, \ldots, L$, of the hidden state q at time k (i.e., $\nu_k(q)$ denotes the probability to be in hidden state q at time k according to the present state of the EM-iteration). Then, denote the mean and covariance of the time series $(Z_1, \ldots, Z_T)$ in the state q by

$$\bar{Z}_T^{(q)} = \frac{1}{\sum_{k=1}^{T-1} \nu_{k+1}(q)} \sum_{k=1}^{T-1} \nu_{k+1}(q) Z_k,$$

$$\mathrm{Cov}_T^{(q)}(Z) = \frac{1}{\sum_{k=1}^{T-1} \nu_{k+1}(q)} \sum_{k=1}^{T-1} \nu_{k+1}(q)(Z_k - \bar{Z}_T^{(q)})(Z_k - \bar{Z}_T^{(q)})^\top.$$

Furthermore, let the one-step correlation be defined as

$$\mathrm{Cor}_T^{(q)}(Z) = \frac{1}{\sum_{k=1}^{T-1} \nu_{k+1}(q)} \sum_{k=1}^{T-1} \nu_{k+1}(q)(Z_{k+1} - \bar{Z}_T^{(q)})(Z_k - \bar{Z}_T^{(q)})^\top \cdot \mathrm{Cov}_T^{(q)}(Z)^{-1}.$$

Finally, for the sake of convenience let $\delta_T^{(q)}$ denote

$$\delta_T^{(q)} = \frac{1}{\sum_{k=1}^{T-1} \nu_{k+1}(q)} \sum_{k=1}^{T-1} \nu_{k+1}(Z_{k+1} - Z_k).$$

Then, the optimal estimators $\hat{F}^{(q)}$ and $\hat{\mu}^{(q)}$ for the parameters $F^{(q)}$ and $\mu^{(q)}$ of the local SDEs (i.e., the unique maximizers of the expectation of the likelihood) are given by the following statement.

Theorem 3.1. *Let $Cov_T^{(q)}$ be positive definite for all i. Then the optimal estimator satisfies*

$$\exp(\tau \hat{F}^{(q)}) = Cor_T^{(q)}, \tag{15}$$

$$\hat{\mu}^{(q)} = \bar{Z}_T^{(q)} + (\mathrm{Id} - Cor_T^{(q)})^{-1} \delta_T^{(q)}. \tag{16}$$

The second equation is valid whenever the typical case $\|Cor_T^{(q)}\| < 1$ *applies; we then have that the spectrum of the optimal estimator for the stiffness satisfies* $\sigma(\hat{F}^{(q)}) \subset \mathbf{C}^-$.

Furthermore, we get a linear matrix equation for the optimal noise intensity matrix estimator $\hat{\Sigma}^{(q)}\hat{\Sigma}^{(q)^\top}$:

$$e^{-\tau\hat{F}^{(q)}}W^{(q)} = \hat{\Sigma}^{(q)}\hat{\Sigma}^{(q)^\top}e^{\tau\hat{F}^{(q)^\top}} - e^{-\tau\hat{F}^{(q)}}\hat{\Sigma}^{(q)}\hat{\Sigma}^{(q)^\top}, \tag{17}$$

where

$$\begin{aligned} W^{(q)} &= \Omega^{(q)}\hat{F}^{(q)^\top} + \hat{F}^{(q)}\Omega^{(q)}, \\ \Omega^{(q)} &= \left(\frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(q)}\sum_{k=1}^{T-1}\nu_{k+1}(q)\hat{d}_k^{(q)}\hat{d}_k^{(q)^\top}\right), \\ \hat{d}_k^{(q)} &= \left(Z_{k+1} - \hat{\mu}^{(q)} - e^{\tau\hat{F}^{(q)}}\left(Z_k - \hat{\mu}^{(q)}\right)\right). \end{aligned}$$

The matrix equation (17) has a unique solution only if $\sigma(\hat{F}^{(q)}) \subset \mathbf{C}^-$ (see [25]).

This result from [25] gives us the opportunity to realize the required maximization explicitly simply by computing the autocorrelation matrices. There should be a warning: the computation of $\hat{F}^{(q)}$ from $\exp(\tau\hat{F}^{(q)})$ is far from straightforward due to the nonuniqueness of the matrix logarithm (see [25] for details and for appropriate procedures for the computation of $\hat{F}^{(q)}$).

Optimal sequence of hidden states. Based on the results of the EM algorithm, problem (2) can be solved by applying the standard Viterbi algorithm [44]. For given θ and Z this algorithm computes the most probable hidden path $q^* = (q_0^*, \ldots, q_T^*)$. This path is called the *Viterbi path*. For an efficient computation we define the highest probability along a single path, for the first t observations, ending in the hidden state S_i at time t:

$$\delta_t(q) = \max_{q_0,q_1,\ldots,q_{t-1}} P(q_0, q_1, \ldots, q_t = S_i, Z_0, Z_1, \ldots, Z_t|\theta).$$

This quantity is given by induction as

$$\delta_t(q') = \max_{1\le i\le M}[\delta_{t-1}(q)\top_{qq'}]\rho(Z_t|q_t, Z_{t-1}). \tag{18}$$

In addition, the argument q that maximizes (18) is stored in ψ in order to actually retrieve the hidden state sequence. These quantities are calculated for each t and q', and then the Viterbi path will be given by the sequence of the arguments in ψ obtained from backtracking. For more details see [19].

Number of metastable states. In the setup of all HMM techniques for a given observation sequence, one is confronted with the task of selecting *in advance* the number M of hidden states. There are no general solutions to this problem, and the best way to handle this problem often is a mixture of insight and preliminary analysis. However, since our goal is to identify metastable states we can proceed as suggested in [19]: start the EM algorithm with some sufficient number of hidden states, say M, that should be greater than the expected number of metastable states. After termination of the EM algorithm, take the resulting transition

matrix A and aggregate the M hidden states into $\mathcal{M} \leq M$ metastable states by means of PCCA. The resulting conformation states will then allow an interpretation of the results in terms of metastable states.

In all numerical experiments in the following, the initial parameter guesses are based on the same procedure: the initial $M \times M$ transition matrix is chosen to be a stochastic matrix with offdiagonal entries 0.001 and identical diagonal entries. The initial values of the model parameters were obtained by the respective re-estimation formulas of the EM algorithm based on randomized determination of the probabilities $P(Z_t|q_t, Z_{t-1})$ (they were chosen uniformly distributed on $[0, 1]$).

Combination of results from different projections (aggregated HMM). Assume that we already applied HMMSDE or some HMM-based method to several low-dimensional observation time series of the system under consideration, but to each one independently. Suppose that the different time series simply are resulting from different projections of the full time series in state space; for example, think of the different time series given by each single torsion angle of the system, or of the time series given by each single of the leading proper orthogonal decomposition (POD) modes. In this situation one may be interested in combining the hidden states from each of the single projections into "higher-dimensional" metastable states of the system. This can be done by analyzing the Viterbi paths derived from the single low-dimensional observation time series: Suppose we are concerned with J low-dimensional time series and therefore J Viterbi paths. The J Viterbi paths can be understood as a J-dimensional discrete time series. Every state of this time series lies in the discrete state space consisting of all possible combinations of the metastable states of the single low-dimensional time series. We obviously can take this time series, compute its transfer matrix by counting transitions between its discrete states, determine the dominant eigenmodes of this transfer matrix, and again apply PCCA to identify metastable decompositions of the discrete state space. The sets in such a metastable decomposition have to be interpreted as aggregates of the metastable states from the low-dimensional time series where the aggregation is done based on additional insight coming from the combination of all of the low-dimensional information. This concept leads to the following algorithm:

1. Determine model parameters and Viterbi paths for each low-dimensional observation time series.
2. Combine the Viterbi paths and compute the transfer matrix in the discrete state space of combined metastable states.
3. Determine metastable decompositions via PCCA.

3.2. Illustrative example revisited. We now assume a time series $(x(t))_{t=t_0,\ldots,t_N}$ with $t_k - t_{k-1} = \tau = 0.01$ and $N = 10^5$ being given in the test system introduced in section 2.4. For given $t = t_0, \ldots, t_N$ let $x(t) \in \mathbf{R}^2$ be the full state of the system.

For this choice of τ the transfer operator $P^\tau = \exp(\tau\mathcal{A})$ of the high-friction Langevin motion considered in section 2.4 has the following dominant eigenvalues:

$$\sigma(P^t) = \{1.000, 0.999, 0.997, 0.959, \ldots\}.$$

Let us consider the two observation time series $(Z_t^{(j)})_{t=t_0,\ldots,t_N}$, $j = 1, 2$, with $Z_t^{(j)} = x_j(t)$ (the first and second components of the state of the system).

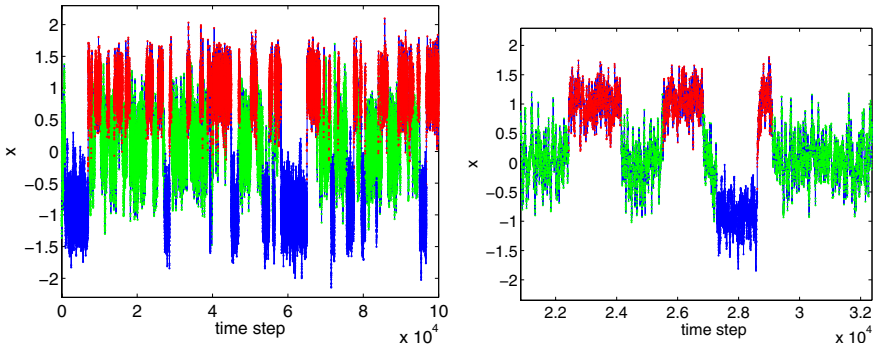


Figure 5. *Observation time series* $(Z_t^{(1)})$. *Left: entire time axis. Right: magnification clearly exhibiting metastability and overlapping. Color scale due to Viterbi path (see text below).*

We first apply HMMSDE to observation time series $(Z_t^{(1)})$ (see Figure 5 for illustration) and set $M = 3$. Eleven iterations of the EM algorithm result in the transition matrix

$$\top = \begin{pmatrix} 0.9983 & 0.0013 & 0.0004 \\ 0.0017 & 0.9983 & 0.0000 \\ 0.0008 & 0.0000 & 0.9992 \end{pmatrix}$$

that has the spectrum

$$\sigma(\top) = \{1.000, 0.999, 0.997\},$$

which agrees perfectly with the results of the transfer operator approach (that is based on the full 2D information instead of on the reduced observation time series). The HMMSDE results for the parameters of the potential and the noise intensities are given in Table 3 and are in very good agreement with the results to be expected.

Table 3
Parameters of HMMSDE *for training with* $(Z_t^{(1)})$.

Parameter	$j = 1$	$j = 2$	$j = 3$
$\mu^{(j)}$	0.0552	1.0169	-0.9584
$\sigma^{(j)^2}$	0.1325	0.1321	0.1302
$D^{(j)}$	0.5589	1.0507	0.9324

Next we apply HMMSDE to observation time series $(Z_t^{(2)})$ (see Figure 5) and set $M = 3$. Nine iterations of the EM algorithm result in the transition matrix

$$\top = \begin{pmatrix} 0.9987 & 0.0013 & 0.0000 \\ 0.0014 & 0.9981 & 0.0005 \\ 0.0000 & 0.0007 & 0.9993 \end{pmatrix}$$

with spectrum

$$\sigma(P^t) = \{1.000, 0.999, 0.997\}.$$

The HMMSDE results now are again in good agreement with the results to be expected (see Table 4).

Table 4
Parameters of HMMSDE *for training with* $(Z_t^{(2)})$.

Parameter	$j=1$	$j=2$	$j=3$
$\mu^{(j)}$	1.5526	-0.0084	-0.6693
${\sigma^{(j)}}^2$	0.1318	0.1347	0.1343
$D^{(j)}$	1.0607	0.5018	1.1037

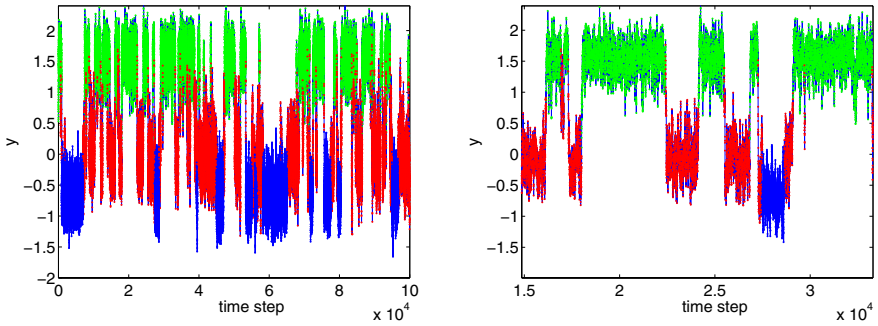


Figure 6. *Observation time series* $(Z_t^{(2)})$. *Left: entire time axis. Right: magnification clearly exhibiting metastability and overlapping. Color scale due to Viterbi path (see text below).*

Next we compute the Viterbi paths for the two HMMSDE results based on $(O_t^{(1)})$ and $(O_t^{(2)})$, respectively. This yields the assignment to metastable states as illustrated in Figures 5 and 6, and in a 2D representation in Figure 7. We observe that the agreement of the assignment with the metastable states resulting from the transfer operator approach (see Figure 3) is good. However, as can be seen from the picture, the assignment of the points in the transition regions gets ambiguous. The algorithm for combining the results of our two different projections (see *aggregated HMM algorithm* described in section 3.1) yields the results shown in Figure 8, where all points which are not clearly assigned to any of the metastable states are identified as belonging to some "transition state."

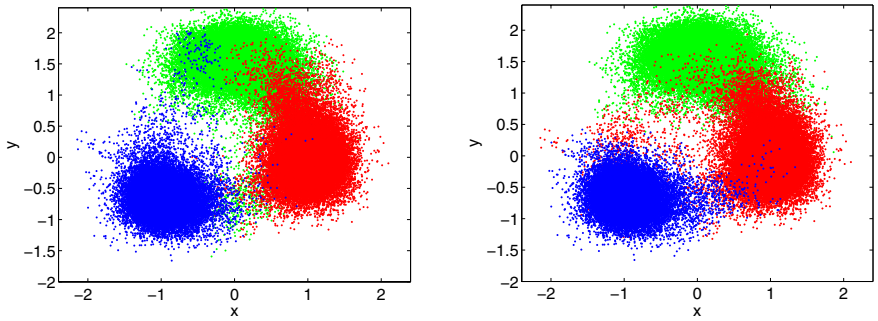


Figure 7. *Visualization of the assignment of states to the three metastable states as resulting from the Viterbi paths computed via* HMMSDE *based on* $(Z_t^{(1)})$ *(left) and* $(Z_t^{(2)})$ *(right).*

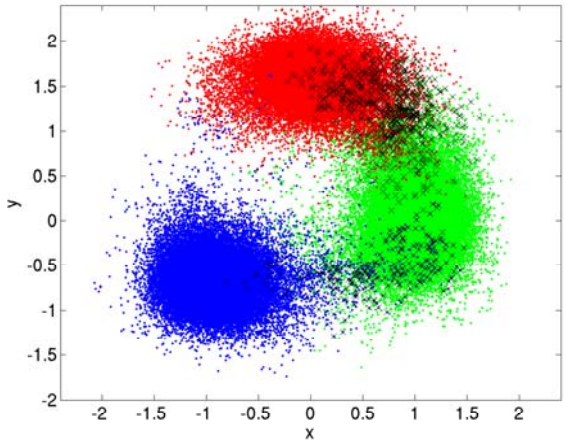


Figure 8. *Visualization of the assignment of states to the three metastable states (points of three different color tones) and transition states (black crosses) as resulting from the clustering of both one-dimensional (1D) Viterbi paths computed according to the transfer operator approach.*

4. Application to molecular dynamics (MD) simulations of a DNA fragment.

4.1. MD and statistics. In classical MD atoms are described as mass points subject to forces that are generated by specified classical interaction potentials V. The dynamical behavior is described by a deterministic Hamiltonian system of the form

$$\dot{q} \;=\; M^{-1}\xi, \qquad \dot{\xi} \;=\; -\nabla_q V(q), \tag{19}$$

defined on the state space $\mathbf{X} = \mathbf{R}^{3N} \times \mathbf{R}^{3N}$ with M denoting the diagonal mass matrix. Equation (19) models an energetically closed system, whose total energy, given by the Hamiltonian

$$H(q,\xi) \;=\; \frac{1}{2}\,\xi^T M^{-1}\xi \;+\; V(q), \tag{20}$$

is preserved under the dynamics.

It is well known that for every smooth function $\mathcal{F} : \mathbf{R} \to \mathbf{R}$ the probability measure $\mu(\mathrm{d}x) \propto \mathcal{F}(H)(x)\mathrm{d}x$ is invariant w.r.t. the Markov process X_t given by the solution of the Hamiltonian system (19). The most frequent choice is the canonical density or *canonical ensemble*

$$f(x) \;\propto\; \exp(-\beta H(x))$$

for some constant $\beta > 0$ that can be interpreted as inverse temperature. The associated measure $\mu(\mathrm{d}x) \propto f(x)\mathrm{d}x$ is called the *canonical measure*. The canonical ensemble is often used in modeling experiments on molecular systems that are performed under the conditions of constant volume and temperature $\top = \frac{1}{k_B\beta}$, where k_B is Boltzmann's constant. Obviously, a single solution of the Hamiltonian system (19) can never be ergodic w.r.t. the canonical measure, since it conserves the internal energy H, as defined in (20). One traditional aspect of MD is the construction of (stochastic) dynamical systems that allow sampling of the canonical ensemble by means of long-term simulation. Several approaches have been discussed, most

of them reducing to the construction of a Hamiltonian system in some slightly extended state space $\hat{\mathbf{X}}$, whose projection onto the lower-dimensional state space $\mathbf{X}$ of positions and momenta generates a sampling according to (4.1). One of the most prominent examples is the Nosé–Hoover thermostat [7].

4.2. 100 ns MD time series of $GT(AT)_6C$ DNA. The time series of interest in the following was generated via an MD simulation of a 15 base pair fragment, or oligomer, using the AMBER package [9]. The detailed protocol was that of the ABC project as described in detail in [5]. That simulation provided a 100 ns time series of the oligomer with the sequence $GT(AT)_6C$ with explicit water- and counterions (see Figure 9). The MD delivers a time series of the Cartesian coordinates of all atoms (about 23000 atoms, including solvent). The MD trajectory was sampled every picosecond to obtain a series of length 10^5.

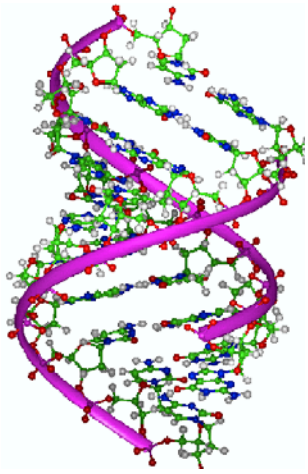


Figure 9. *Illustration of the* 15-*AT B-DNA oligonucleotide in atomic resolution. The attached violet strings indicate the backbones. The molecular dynamics simulation referred to herein includes solvent (water- and counterions), which is not shown here.*

The variables in the time series that we work with are the physically motivated projections onto two different sets of coarse-grained internal coordinates: either the torsion angle series of the backbone data [37] or the inter base pair step parameters [20]. In either case the dimension of the time series is 84, arising from the six degrees of freedom at each of the 14 junctions between 15 base pairs. At each sampling time the coarse-grain variables were extracted from the full set of Cartesian coordinates following standard conventions. Depending on the projection chosen, two basic temporal patterns of dynamics were found: abrupt, almost instantaneous change of the backbone torsion angles (see Figure 10), and slow relaxations of the inter base pair parameters with a relaxation time on the order of 1–2 ns (cf. Figure 11).

Both the backbone angle and base pair parameter descriptions of the DNA oligomer take standard, sequence-independent values on the idealized B-form Watson–Crick double helix. One of the motivations for the development of the time-series analysis developed here is to extract and understand deviations from these standard values both as a function of sequence and as a function of time.

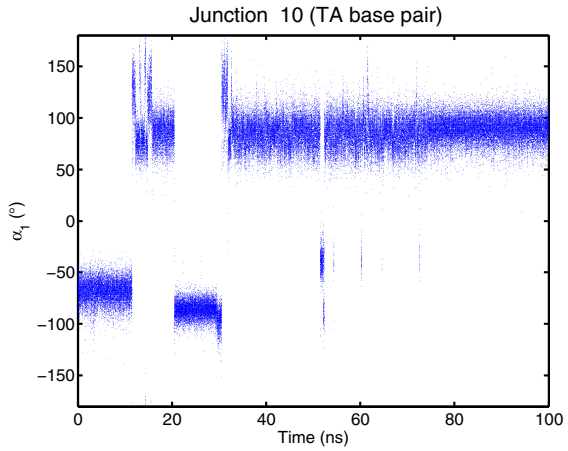


Figure 10. *Time series of the first strand α-torsion angle for the junction* 10. *The dynamics exhibits sharp transitions between the metastable sets.*

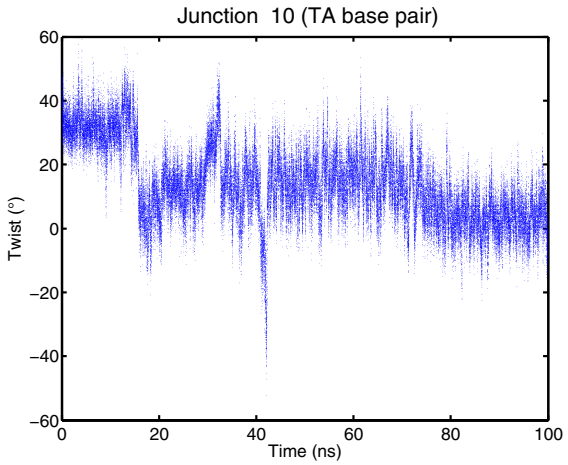


Figure 11. *Time series of the inter base pair coordinate for the junction* 10. *The dynamics exhibits slow relaxational transitions (in a nanosecond region) between the metastable sets.*

In the following we will apply the techniques of section 3 to time series of all backbone torsion angles that result from the 100 ns MD simulation described earlier in this section. This time series of the backbone angles will be denoted by $(Z_t)_{t=1,\ldots,100000}$ with $Z_t \in \mathbf{R}^{84}$ arising from the six backbone torsion angles (α, β, γ-angles at each of the strands) at each of the 14 junctions between 15 base pairs.

We will apply aggregated HMM (based on HMM with stationary van Mises output distribution for each single of the 84 torsion angle time series) and HMMSDE to the full 84-dimensional time series. After a detailed separate analysis of these two algorithms we will finally compare the outcome, especially the Viterbi paths. The section is tailored to demonstrate the application and performance of the respective algorithms; herein discussion of possible physical implications of the results is *not* the focus of our consideration.

4.3. Analysis via aggregated HMM. As can be seen from Figure 10, the torsion angles exhibit a clearly metastable behavior with sharp transitions between the metastable states. We start with the decomposition of the 84-dimensional torsion angle space into the 84 1D spaces defined by each single angle. Then, we apply our HMM analysis of the dynamics to each resulting single torsion angle time series. As a result, we get a set of 84 aggregated Viterbi paths describing the conformational change between the metastable sets in the observed 100 ns time series as identified from each single angle (for examples, see Figure 12). Of course, the aggregation depends on the strictness of the metastability criteria, i.e., on the threshold value above which eigenvalues of the associated transition matrices indicate metastability. To demonstrate that the aggregation does not destroy important dynamical information about metastability, we will in the following present two different results: one computed with a strict threshold (0.97) and another computed with a less strict threshold (0.93).

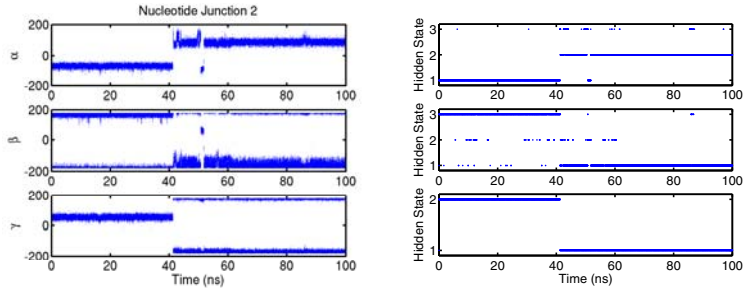


Figure 12. *Time series of the first strand α, β, γ-backbone torsion angles for junction 2 (right-hand panel). Left-hand panel: corresponding Viterbi paths as derived from the independent HMM analysis of these three torsion angle time series with an eigenvalue threshold of 0.97 (this means that all eigenvalues of the corresponding transition matrix ≥ 0.97 define the metastable sets).*

This set of 84 aggregated Viterbi paths is a coarse-grained description of the original time series and encapsulates the dynamical information about the process of the conformation change in the full 84-dimensional torsion angle space. In order to extract this information, we have to construct the transition matrix state space of the combined Viterbi path of the 84 single aggregated Viterbi paths as explained above. The resulting transition matrices have 469 combined states (for metastability threshold 0.97 for the aggregation of the single Viterbi paths) or 12567 combined states (for threshold 0.93). Figure 13 represents the dependence of the resulting eigenvalues of these transition matrices as a function of the lag time τ (lag time $\tau = l\Delta t$ means that the transition matrix counts l-step transitions between instances t and $t + l\Delta t$ along the given time series). Increasing τ means decreasing correlation between successive time steps and therefore a more informative spectrum. As can be seen in both cases, the first eigenvalue gap can always be identified after the first four dominant eigenvalues, which indicates a presence of four metastable sets in 84-dimensional space of torsion angles (for a more detailed level of description a choice of six or eight eigenvalues would be also reasonable).

In Figure 14 we compare the obtained *global* Viterbi path describing the transitions between the resulting four metastable states based on the small transition matrix (469 states) with the results based on the larger transition matrix (12567 states). As can be seen from the figure, the two global Viterbi paths are similar but not identical. The difference between

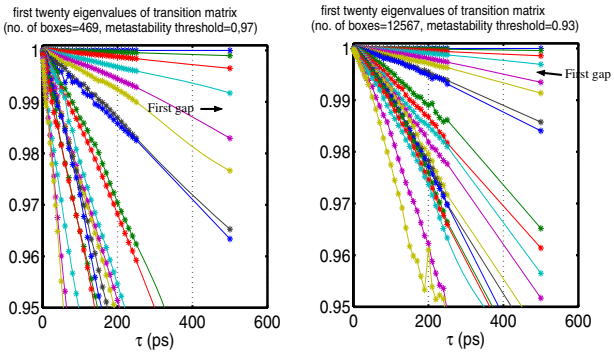


Figure 13. *Eigenvalues of the two global transition matrices as functions of the lag time τ along the respective global Viterbi path. Right-hand panel: metastability threshold* 0.97*; left-hand panel: metastability threshold* 0.93.

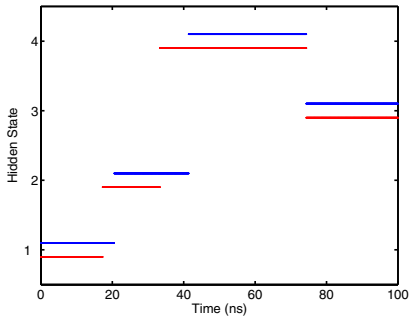


Figure 14. *Different global Viterbi paths, with each resulting from the first four eigenvalues of the global transition matrices based on two different metastability thresholds for the aggregation of the underlying* 84 *single angle Viterbi paths as described in the text. Red: threshold value* 0.93*, resulting in* 12567 *combined states; blue: threshold value of* 0.97*, resulting in* 469 *combined states.*

them can be explained by the presence of *transition states* (i.e., states that lie between the metastable states and are visited by the molecule during the transition from one metastable set to the other). Repetition of the analysis of the two transition matrices based on six dominant eigenvalues results in identification of these transition regions as additional metastable states. Then, the two global Viterbi paths (with six metastable states each) are almost identical.

Global three-dimensional (3D) structures of the DNA molecule as calculated from the mean configurations of the corresponding metastable states are presented in Figure 15.

4.4. Analysis via HMMSDE. We will now apply the HMMSDE procedure to a full 84-dimensional torsion angle time series. Before doing this, one should decide which form of local multidimensional SDE dynamics is adequate to describe the observed time series; the biophysical literature indicates that we should distinguish between Langevin diffusion (SDE in positions and momenta) and overdamped diffusion (SDE in positions only); see [25] for details. The ratio of the decay rates of position and momentum autocorrelation functions

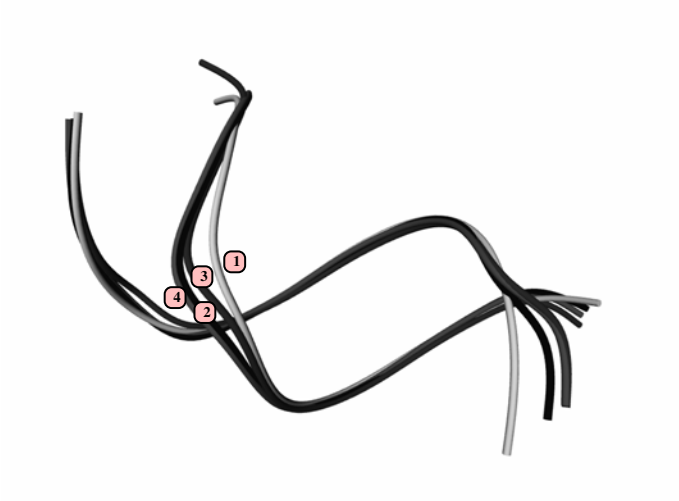


Figure 15. *Average 3D geometry of the four mean configurations defined by the four metastable states shown in Figure 14. The 3D geometry is illustrated by a string-like representation of the backbone. Numbers denote the indices of metastable states as in Figure 14. Visualization is based on AMIRA software [2].*

can indicate which of the models should be applied. This means that if the momentum autocorrelation decays much faster than the position autocorrelation function, then overdamped diffusion (also called Smolochowski dynamics) can be used to describe the observed data. Applying this criterion to the DNA data, we deduce that for a time step of 1 picosecond between observations the overdamped diffusion can be used such that we consequently use local overdamped diffusion models of the form

$$\dot{z}(t) = F^{(q(t))}(z - \mu^{(q(t))}) + \Sigma^{(q(t))}\,\dot{W}(t),$$

where z just denotes the vector of torsion angles. As has been already mentioned above, application of the EM-algorithm results in finding some *local* optimum of the log-likelihood functional. This also means that we should try to select an initial set of parameters that is "close" to the global maximum of (13): If the initial guess for the hidden probabilities $\nu_i(t)$ is not "too far off" from the real probabilities (w.r.t. the global maximum), it will take only a "few" EM-iterations to find the optimal set of parameters. Therefore we start HMMSDE with four hidden states and initialize the EM-iteration and take the hidden probabilities $\nu_i(t)$ as resulting from the Viterbi path computed via aggregated HMM (cf. Figure 14). As expected, it takes just a dozen iterations to get convergence, resulting in a 4×4 transition matrix of the hidden Markov chain; its spectrum is as shown in Figure 16.

Comparison of the resulting Viterbi path associated with the 84-dimensional HMMSDE model to the Viterbi path of the aggregated HMM procedure demonstrates the similarity of the metastable structures identified (see Figure 17).

In addition to the Viterbi path, HMMSDE also yields (optimal) estimators for the parameters of the local overdamped diffusion models (stiffness, equilibrium positions, and noise intensity matrices). Due to the fact that the Viterbi paths in Figure 17 are very similar, the mean configurations of the conformational states (denoted by μ in (8)) are also very close to

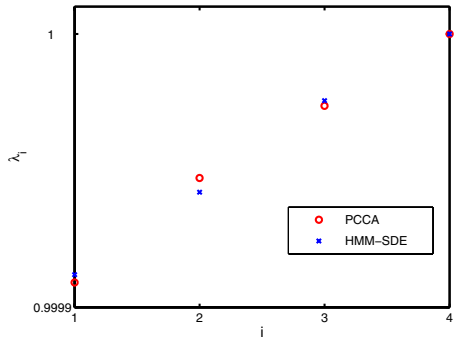


Figure 16. *Four largest eigenvalues of the transition matrix of the hidden Markov process for full-dimension HMMSDE (blue crosses) in comparison to those of aggregated HMM (metastability threshold* 0.93*, red circles).*

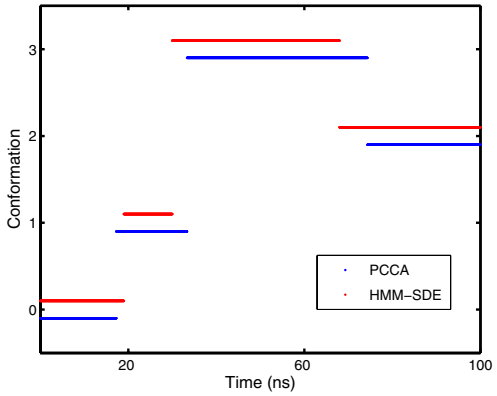


Figure 17. *Viterbi path resulting from full-dimension HMMSDE in comparison to the global Viterbi path resulting from aggregated HMM as of the previous section, with each resulting from the first four eigenvalues of the transition matrices (metastability threshold* 0.93 *for aggregated HMM). Red: full-dimension HMMSDE; blue: aggregated HMM.*

those resulting from aggregated HMM as already shown in Figure 15. Closer inspection of the resulting stiffness matrices allows us to get additional insights into physical properties of the four resulting conformations. The stiffness matrices associated with each of the conformations are shown in Figures 18–21. All four matrices are block-banded: we observe rather strong couplings between neighboring torsion angles in the same strain, weaker coupling between neighboring angles on opposed strains, and almost zero interaction between the angles not directly adjacent to each other. The stiffness matrices are also periodic along the main diagonal and the subdiagonals. This feature can be explained by the periodicity of the underlying DNA sequence. Finally, comparison among the individual stiffness matrices shows that the first conformation is the "most stiff" one, whereas the three other conformations have "loose spots" that correspond to the locations where transitions between the conformations take place.

When repeating the entire HMMSDE analysis with different initial values for the EM and

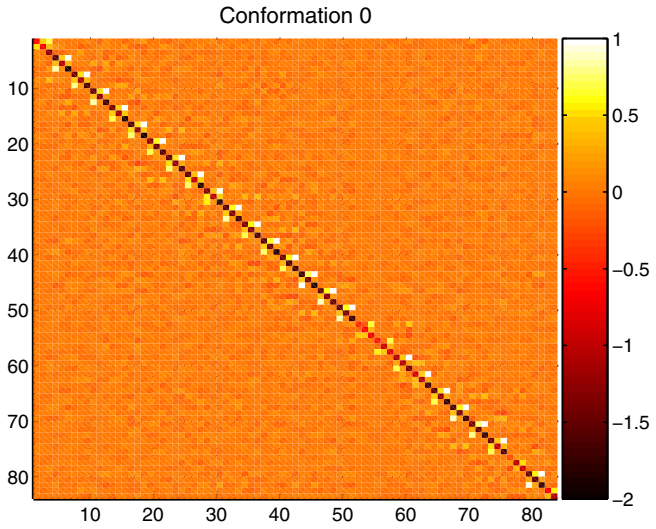


Figure 18. *Optimal estimate for the stiffness matrix F of the first conformation as resulting from full-dimension HMMSDE. For details of its computation, see text.*

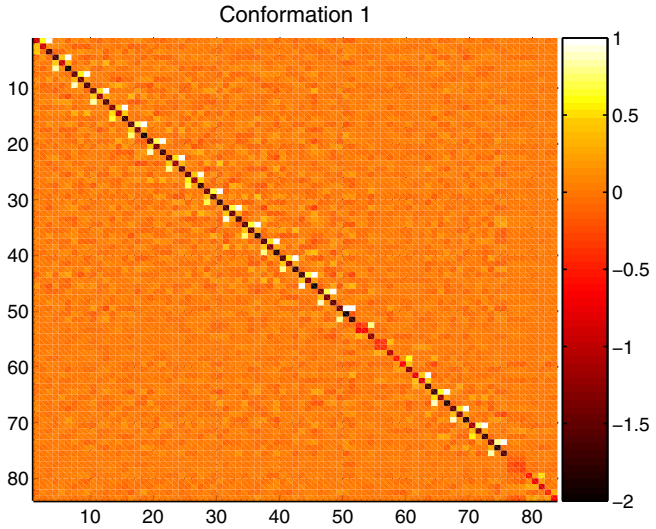


Figure 19. *Optimal estimate for the stiffness matrix F of the second conformation as resulting from full-dimension HMMSDE. For details of its computation, see text.*

Viterbi algorithms, we observe convergence to very similar results, even if the initial values are quite different from the ones taken above (however, in some cases the number of iterations increases significantly).

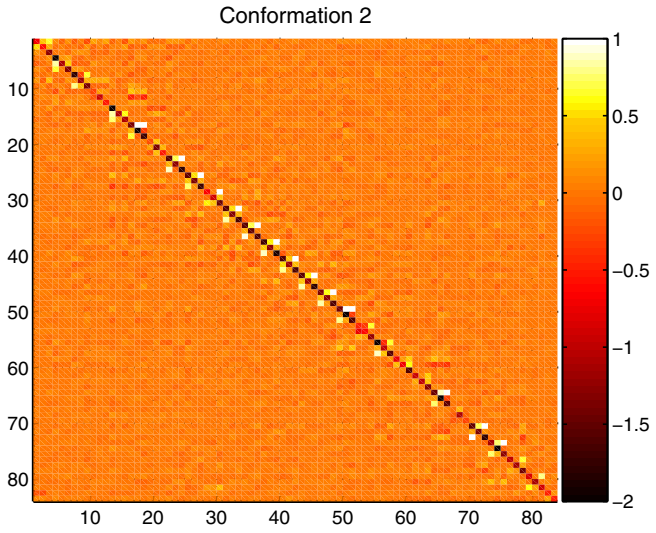


Figure 20. *Optimal estimate for the stiffness matrix F of the third conformation as resulting from full-dimension HMMSDE. For details of its computation, see text.*

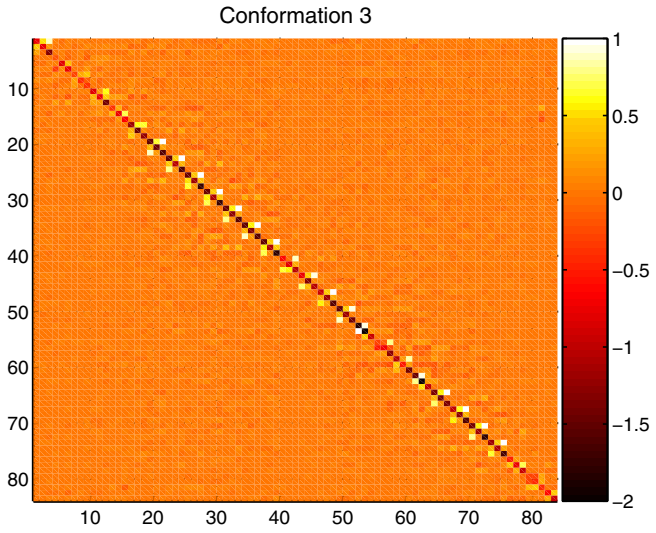


Figure 21. *Optimal estimate for the stiffness matrix F of the fourth conformation as resulting from full-dimension HMMSDE. For details of its computation, see text.*

5. Conclusion. We have presented a variety of algorithmic concepts for the identification of metastable states in dynamical systems. Possible strategies for application to very complex metastable systems, and the performance of the resulting algorithms have been demonstrated by analyzing a full-scale MD simulation of a poly-(AT) B-DNA oligomer.

In regard to the algorithmic aspects, our conclusion is that for realistically large simulations

of biomolecules it is not practical to compute explicitly the full discretized transfer operator in Cartesian coordinates. This is due to the curse of dimensionality. In contrast, aggregated HMM and HMMSDE methods applied to a reduced time series with sharp transitions (the backbone time series in the DNA example) can identify the metastable sets.

Our results show that the backbone dynamics of B-DNA exhibit metastable behavior (visible in both base pair and torsion angles representations of the dynamics) on nano- to microsecond time scales, and that this metastability might be sequence-dependent and of importance for macroscopic modeling of B-DNA elasticity and dynamics. Most specifically the average values of AT and TA base pair parameters are quantified and confirmed to be quite distinct. In addition, the values of these averages are shown to depend upon the particular metastable set of the oligomer.

On a less positive note, it is apparent that the simulation time scale of a few hundred nanoseconds is much too short to compute transition probabilities for the backbone accurately, i.e., to analyze quantitatively the possible sequence dependence of backbone conformation transitions. Most specifically, the trajectory we have computed is demonstrably not ergodic.

REFERENCES

[1] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, *Essential dynamics of proteins*, Proteins, 17 (1993), pp. 412–425.

[2] *Amira—Advanced Visualization, Data Analysis and Geometry Reconstruction, User's Guide and Reference Manual*, Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), Indeed–Visual Concepts GmbH, and TGS Template Graphics Software, 2000.

[3] L. E. Baum, *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes*, Inequalities, 3 (1972), pp. 1–8.

[4] B. J. Berne and J. E. Straub, *Novel methods of sampling phase space in the simulation of biological systems*, Curr. Opinion Struct. Biol., 7 (1997), pp. 181–189.

[5] D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham, III, S. B. Dixit, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, E. Seibert, H. Sklenar, G. Stoll, K. M. Thayer, P. Varnai, and M. A. Young, *Molecular dynamics simulations of the* 136 *unique tetranucleotide sequences of DNA oligonuleotides.* I. *Research design and results on d(CpG) steps*, Biophys. J., 87 (2004), pp. 3799–3813.

[6] J. A. Bilmes, *A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, Technical report, International Computer Science Institute, Berkeley, CA, 1998.

[7] S. D. Bond, B. J. Leimkuhler, and B. B. Laird, *The Nosé–Poincaré method for constant temperature molecular dynamics*, J. Comput. Phys., 151 (1999), pp. 114–134.

[8] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, *Metastability in stochastic dynamics of disordered mean-field models*, Probab. Theory Related Fields, 119 (2001), pp. 99–161.

[9] D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham, III, J. Wang, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R. J. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman, *AMBER* 7, University of California, San Francisco, 2002.

[10] D. Chandler, *Finding transition pathways: Throwing ropes over rough mountain passes, in the dark*, in Classical and Quantum Dynamics in Condensed Phase Simulations, B. Berne, G. Ciccotti, and D. Coker, eds., World Scientific, Singapore, 1998, pp. 51–66.

[11] E. B. Davies, *Metastable states of symmetric Markov semigroups.* I, Proc. London Math. Soc. (3), 45 (1982), pp. 133–150.

[12] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM J. Numer. Anal., 36 (1999), pp. 491–515.
[13] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte, *Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains*, Linear Algebra Appl., 315 (2000), pp. 39–59.
[14] P. Deuflhard and M. Weber, *Robust Perron cluster analysis in conformation dynamics*, Linear Algebra Appl., 398 (2005), pp. 161–184.
[15] J. L. Doob, *Stochastic Processes*, Wiley, New York, 1953.
[16] W. E and E. Vanden-Eijnden, *Metastability, conformation dynamics, and transition pathways in complex systems*, in Multiscale Modelling and Simulation. Part I, Springer, Berlin, 2004, pp. 35–68.
[17] D. M. Ferguson, J. I. Siepmann, and D. G. Truhlar, eds., *Monte Carlo Methods in Chemical Physics*, Advances in Chemical Physics 105, Wiley, New York, 1999.
[18] A. Fischer, F. Cordes, and C. Schütte, *Hybrid Monte Carlo with adaptive temperature in a mixed-canonical ensemble: Efficient conformational analysis of RNA*, J. Comput. Chem., 19 (1998), pp. 1689–1697.
[19] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, and Ch. Schütte, *Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models*, J. Comput. Chem., 28 (2007), pp. 2453–2464.
[20] O. Gonzalez and J. H. Maddocks, *Extracting parameters for base-pair level models of DNA from molecular dynamics simulations*, Theor. Chem. Acc., 106 (2001), pp. 76–82.
[21] C. Hartmann and Ch. Schütte, *A constrained hybrid Monte-Carlo algorithm and the problem of calculating the free energy in several variables*, ZAMM Z. Angew. Math. Mech., 85 (2005), pp. 700–710.
[22] H. Heuser, J. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK, 1996.
[23] I. Horenko, E. Dittmer, A. Fischer, and C. Schütte, *Automated model reduction for complex systems exhibiting metastability*, Multiscale Model. Simul., 5 (2006), pp. 802–827.
[24] I. Horenko, E. Dittmer, and Ch. Schütte, *Reduced stochastic models for complex molecular systems*, Comput. Vis. Sci., 9 (2006), pp. 89–102.
[25] I. Horenko and C. Schütte, *Likelihood-based estimation of Langevin models and its application to biomolecular dynamics*, Multiscale Model. Simul., to appear.
[26] W. Huisinga, *Metastability of Markovian Systems: A Transfer Operator Approach in Application to Molecular Dynamics*, Ph.D. thesis, Free University Berlin, Berlin, 2001.
[27] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, and F. Cordes, *From simulation data to conformational ensembles: Structure and dynamic based methods*, J. Comput. Chem., 20 (1999), pp. 1760–1774.
[28] W. Huisinga, S. Meyn, and Ch. Schütte, *Phase transitions and metastability in Markovian and molecular systems*, Ann. Appl. Probab., 14 (2004), pp. 419–458.
[29] W. Huisinga and B. Schmidt, *Metastability and Dominant Eigenvalues of Transfer Operators*, in preparation.
[30] S. Lall, J. E. Marsden, and S. Glavaski, *A subspace approach to balanced truncation for model reduction of nonlinear control systems*, Internat. J. Robust Nonlinear Control., 12 (2002), pp. 519–535.
[31] P. Lezaud, *Chernoff and Berry–Esséen inequalities for Markov processes*, ESAIM Probab. Statist., 5 (2001), pp. 183–201.
[32] K. V. Mardia, *Statistics of Directional Data*, Academic Press, New York, 1972.
[33] Y. Mu, P. H. Nguen, and G. Stock, *Energy landscape of a small peptide revealed by dihedral angle principal component analysis*, Proteins, 58 (2004), pp. 45–52.
[34] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE, 77 (1989), pp. 257–286.
[35] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice–Hall, Upper Saddle River, NJ, 1993.
[36] H. Risken, *The Fokker-Planck Equation*, 2nd ed., Springer, New York, 1996.
[37] W. Saenger, *Principles of Nucleic Acid Structure*, Springer, New York, 1984.
[38] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys., 151 (1999), pp. 146–168.

[39] Ch. Schütte and W. Huisinga, *On conformational dynamics induced by Langevin processes*, in EQUADIFF 99—International Conference on Differential Equations, Vol. 2, B. Fiedler, K. Gröger, and J. Sprekels, eds., World Scientific, River Edge, NJ, 2000, pp. 1247–1262.

[40] C. Schütte and W. Huisinga, *Biomolecular conformations can be identified as metastable sets of molecular dynamics*, in Handbook of Numerical Analysis, Vol. 10, Handb. Numer. Anal. X, P. G. Ciaret and J.-L. Lions, eds., North–Holland, Amsterdam, 2003, pp. 699–744.

[41] Ch. Schütte, W. Huisinga, and P. Deuflhard, *Transfer operator approach to conformational dynamics in biomolecular systems*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, B. Fiedler, ed., Springer, Berlin, 2001, pp. 191–223.

[42] G. Singleton, *Asymptotically exact estimates for metastable Markov semigroups*, Quart. J. Math. Oxford Ser. (2), 35 (1984), pp. 321–329.

[43] M. Sprik and G. Ciccotti, *Free energy from constrained molecular dynamics*, J. Chem. Phys., 109 (1998), pp. 7737–7744.

[44] A. J. Viterbi, *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*, IEEE Trans. Inform. Theory, 13 (1967), pp. 260–269.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 561–576
© 2008 Society for Industrial and Applied Mathematics

On the Families of Periodic Orbits of the Sitnikov Problem*

Jaume Llibre† and Rafael Ortega‡

Abstract. The main goal of this paper is to study analytically the families of symmetric periodic orbits of the elliptic Sitnikov problem for all values of the eccentricity in the interval $[0, 1)$, providing qualitative and quantitative information on the bifurcation diagram of such families of periodic orbits. The basic tool for proving our results is the global continuation method of the zeros of a function depending on one parameter provided by Leray and Schauder and based in the Brouwer degree.

Key words. 3-body problem, Sitnikov problem, periodic orbits, global continuation

AMS subject classifications. Primary, 70F15; Secondary, 37N05

DOI. 10.1137/070695253

1. Introduction.

The Sitnikov problem is a special case of restricted 3-body problems where the two primaries with equal masses are moving in a circular or an elliptic orbit of the 2-body problem, and the infinitesimal mass is moving on the straight line orthogonal to the plane of motion of the primaries which passes through their center of mass.

When the orbit described by the primaries is circular the Sitnikov problem is known as the *circular Sitnikov problem*. In 1907 Pavanini [33] expressed its solutions by means of Weierstrassian elliptic functions. Four years later MacMillan [26] expressed the solutions in terms of Jacobian elliptic functions (a detailed description of this work can be found in Stumpff [38]). Some other analytical expressions for the solutions of this problem can be found, for instance, in [39], [5], [42], [17], and [20]. The precise definition of the Sitnikov problem is given in section 2.

The *elliptic Sitnikov problem* is the case when the orbit describing the primaries is elliptic. This problem became important in 1960 when Sitnikov [37] used it to show, for the first time, the possibility of the existence of oscillatory motions in the 3-body problem. The existence of this kind of motion was predicted by Chazy [8], [9], [10] in 1922–1932, when he classified the final evolutions of the 3-body problem. Later on Alekseev [2], [3], [4] in 1968–1969 proved that, in the special case of the 3-body problem studied by Sitnikov, all of the possible combinations of final motions in the sense of Chazy were realized. Moser [32] in 1973 gave alternative proofs of the results of Alekseev which are simpler than those in [2], [3], [4]. Since then many other authors have studied the circular or elliptic Sitnikov problem—for instance, Llibre and Simó [24], Perdios and Markellos [34], Liu and Sun [21], Hagel [18], Hagel and Trenkler [19],

*Received by the editors June 23, 2007; accepted for publication (in revised form) by J. Meiss January 3, 2008; published electronically May 23, 2008.
http://www.siam.org/journals/siads/7-2/69525.html

†Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain (jllibre@mat.uab.cat). This author has been partially supported by the grants MCYT/FEDER MTM2005-06098-C02-01 and CIRIT 2005SGR 00550.

‡Departamento de Matemática Aplicada, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain (rortega@ugr.es). This author is partially supported by BFM2002-01038.

Martinez-Alfaro and Chiralt [28], Dvorak [16], Kallrath, Dvorak, and Schlöder [23], Wodnar [40], [41], [42], and Dankowicz and Holmes [15].

The families of symmetric periodic orbits of the elliptic Sitnikov problem have been studied by several authors. These works assume sufficiently small values of the eccentricity and obtain the periodic solution by continuing the known periodic orbit of the circular Sitnikov problem. There are analytic studies by Corbera and Llibre [12], [13], using the analytical continuation method of Poincaré, and Cabral and Xia [7], applying the subharmonic Melnikov method. There are also numerical studies by Belbruno, Llibre, and Ollé [5] and Jiménez-Lara and Escalona-Buendía [22]. In this last paper the authors describe numerically some families of symmetric periodic orbits for almost all values of the eccentricity e in $[0, 1)$.

The main objective of this paper is to study analytically the families of symmetric periodic orbits of the elliptic Sitnikov problem for nonnecessarily small values of the eccentricity e. More precisely, we will show that some periodic orbits for $e = 0$ can be continued to all values of e in $[0, 1)$. In Theorems 3.1, 3.2, and 3.3 are the statements of our main results.

The main tool for proving our results is the global continuation of the zeros of a function depending on one parameter provided by Leray and Schauder and based in the Brouwer degree; see section 4. In section 5 we show that, with convenient formulation, the Sitnikov problem satisfies the basic assumptions of the global continuation theorem. In section 6 we study the dynamics around the unique equilibrium point of the Sitnikov problem. This equilibrium point corresponds to one of the three collinear relative equilibrium solutions of Euler for the general 3-body problem; see section 2 for more details. Finally, in section 7 we provide the last steps in the proofs of Theorems 3.1, 3.2, and 3.3.

The use of the global continuation techniques in the study of nonlinear boundary value problems is classical. We refer the reader to [36] for recent results applicable to general classes of nonlinearities. In another context we should also mention the paper [29] by Mathlouthi. He studies the Sitnikov problem with variational techniques and obtains results about the existence of periodic solutions which are global in the sense that they are valid for arbitrary eccentricity. The use of continuation methods will allow us to obtain many continuous families and to be more precise about the oscillatory properties of the solutions.

2. The Sitnikov problem.

Let $m_1 = m_2$ be two point masses (called *primaries*) describing a circular or an elliptic orbit of the 2-body problem. We consider an infinitesimal mass m_3 that moves on the straight line ρ orthogonal to the plane of motion of the primaries that passes through their center of mass. The *Sitnikov problem* will consist of describing the motion of the infinitesimal mass. In particular, if the primaries are moving in circular (respectively, elliptic) orbits, we have the circular (respectively, elliptic) Sitnikov problem.

We choose the units of mass, length, and time so that $m_1 = m_2 = 1/2$, the gravitational constant $G = 1$, and the period of the orbit described by the primaries is 2π. If z denotes the position of the particle m_3 in a coordinate system on ρ with origin at the center of mass of the primaries (see Figure 1), then the equation of motion of the Sitnikov problem becomes

$$\ddot{z} = -\frac{z}{(z^2 + r^2(t, e))^{3/2}}, \tag{2.1}$$

where $r(t, e)$ is the distance of the primaries to their center of mass and it is given by

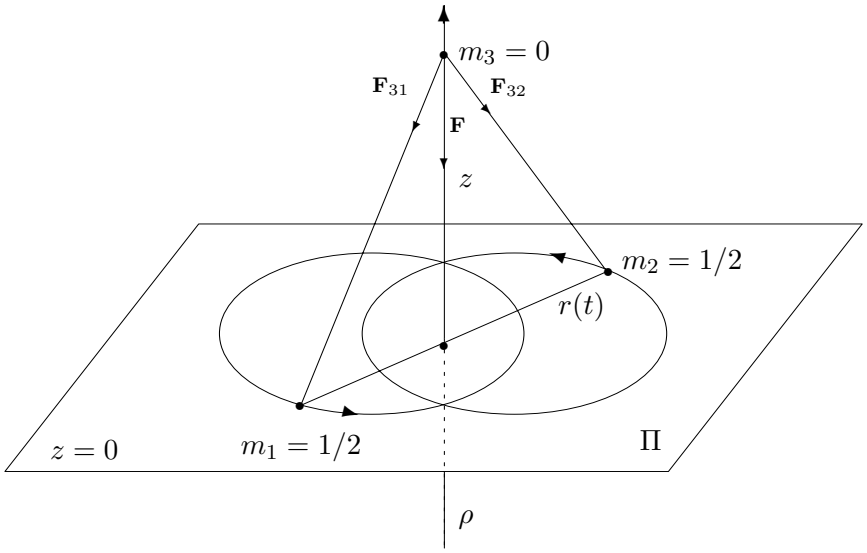


Figure 1. *The Sitnikov problem.*

$$r(t,e) = \frac{1}{2}(1 - e\cos u(t)), \tag{2.2}$$

which is a circular or an elliptic solution of the Kepler problem

$$\ddot{r} = \frac{1-e^2}{16\,r^3} - \frac{1}{8\,r^2}, \tag{2.3}$$

with eccentricity $e = 0$ or $0 < e < 1$, respectively. Here $u(t)$ is the eccentric anomaly which is a function of time via the Kepler equation

$$u - e\sin u = t - \ell, \tag{2.4}$$

with ℓ the time of pericenter passage.

Without loss of generality, when $0 < e < 1$, we usually take the origin of time in such a way that at $t = 0$ the primaries are at the pericenter of the ellipse (i.e., $\ell = 0$).

We note that system (2.1) depends on one parameter, the eccentricity $e \in [0, 1)$.

When the eccentricity e is zero (that is, the primaries move on the circular orbit $r(t) = 1/2$ of the Kepler problem (2.3)), (2.1) becomes the equation of motion

$$\ddot{z} = -\frac{z}{(z^2 + 1/4)^{3/2}} \tag{2.5}$$

for the circular Sitnikov problem. This equation defines an integrable Hamiltonian system of one degree of freedom with Hamiltonian

$$H = \frac{1}{2}v^2 - \left(z^2 + \frac{1}{4}\right)^{-1/2}, \tag{2.6}$$

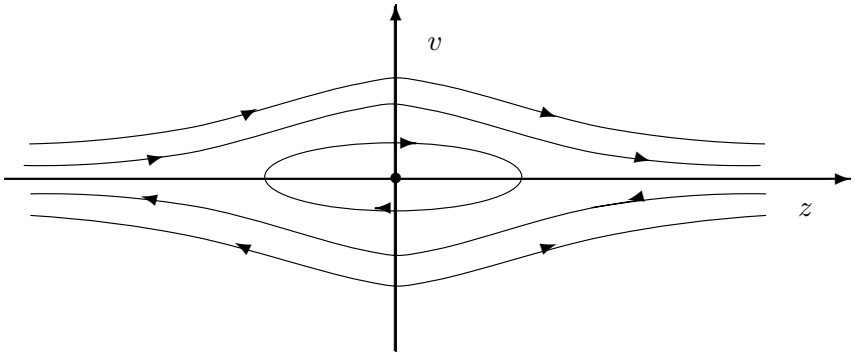


Figure 2. *The circular Sitnikov phase portrait.*

where $v = \dot{z}$.

The orbits for the circular Sitnikov problem in the energy level h are described by the curve $H = h$, where h varies in $[-2, \infty)$. Then, depending on the value of h, we have different types of orbits in the phase space (z, v) (see Figure 2):

(1) when $h < -2$ we have no orbits;
(2) when $h = -2$ we have the equilibrium point ($z = 0$, $v = 0$) or equivalently the trivial solution $z(t) \equiv 0$, which correspond to one of the well-known collinear relative equilibrium solutions of Euler for the 3-body problem (see, for instance, [1]);
(3) when $-2 < h < 0$ we have periodic orbits;
(4) when $h = 0$ we have two parabolic orbits (i.e., two orbits that leave and reach infinity with zero velocity);
(5) when $h > 0$ we have two hyperbolic orbits (i.e., two orbits that leave and reach infinity with positive velocity).

If the eccentricity $e \in (0, 1)$, then differential equation (2.1) corresponds to the elliptic Sitnikov problem. We note that this differential equation is nonautonomous; i.e., the time appears explicitly in the right-hand side of (2.1) through $r(t, e)$. Moreover, $r(t, e)$ is a periodic function in t of minimal period 2π. Consequently all periodic solutions $(z(t), \dot{z}(t))$ of (2.1) with $e \in (0, 1)$ must have period a multiple of 2π. Hence, all periodic orbits of the infinitesimal mass m_3 for the elliptic Sitnikov problem are also periodic orbits involving the three masses. Of course, in general, this was not the case for the circular Sitnikov problem.

3. Statement of the main results.

Given an integer $N \geq 1$, we define

$$\nu = \nu_N = [2\sqrt{2}N],$$

where $[\cdot]$ denotes the integer part function. Our main result is the following.

Theorem 3.1. *For each $p = 1, \ldots, \nu$ and $\varepsilon > 0$ there exists a family (or a branch) of solutions $\{(z_s(t), e_s)\}_{s\in[0,1)}$ of (2.1) satisfying the following:*

(1) *The map $(s, t) \in [0, 1) \times \mathbb{R} \to (z_s(t), \dot{z}_s(t), e_s)$ is continuous.*
(2) *The solutions $z_s(t)$ are even and $2N\pi$-periodic; i.e., for all $s \in [0, 1)$ we have*

$$z_s(-t) = z_s(t), \qquad z_s(t + 2N\pi) = z_s(t).$$

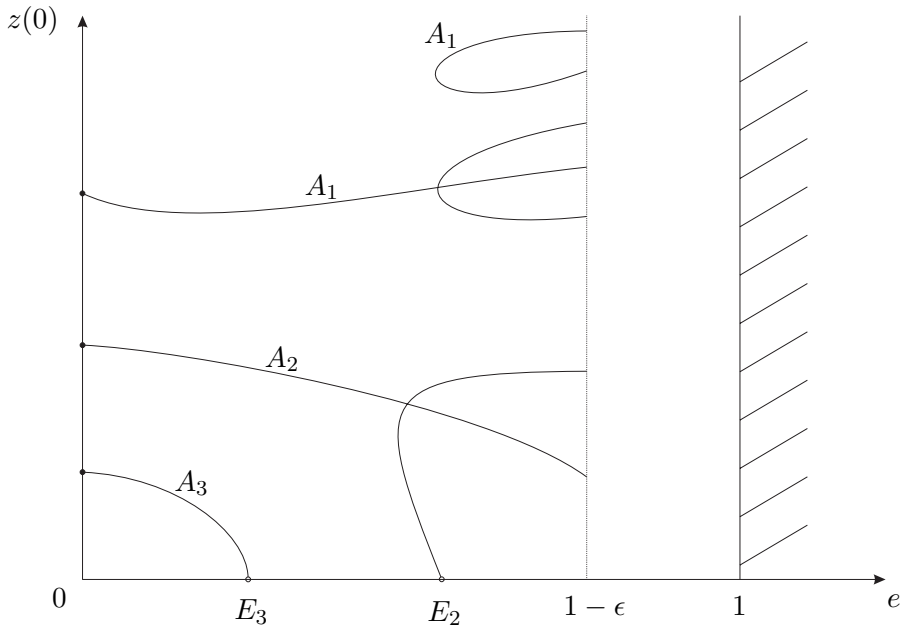


Figure 3. *Families of periodic orbits.*

(3) *For each* $s \in [0,1)$ *we have that* $z_s(0) > 0$ *(hence, the solution* $z_s(t)$ *is nontrivial; i.e.,* $z_s(t) \not\equiv 0$*), and* $z_s(t)$ *has exactly* p *zeros in the interval* $[0, N\pi]$.

(4) $e_0 = 0$, $e_s \in [0, 1-\varepsilon]$ *for each* s *and one of the following alternatives holds:*

(4.a) $e_s \to 1-\varepsilon$ *and* $z_s(0) \to \xi > 0$ *as* $s \nearrow 1$*;*

(4.b) $\lim_{s\nearrow 1} e_s = E$ *exists with* $E < 1-\varepsilon$, $z_s(t)$ *converges to* 0 *as* $s \nearrow 1$, *and the linear differential equation*

$$\ddot{y} + \frac{1}{r(t,E)^3} y = 0$$

has a nontrivial, even, $2N\pi$*-periodic solution with exactly* p *zeros in the interval* $[0, N\pi]$.

Just to illustrate the theorem we sketch a hypothetical situation in Figure 3. For a fixed $N \geq 2$ we take three numbers $1 \leq p_1 < p_2 < p_3 \leq \nu$ and draw the sets A_1, A_2, A_3 defined by

$$A_i = \{(z(0), e) : \ z(t) \text{ is an even } 2N\pi\text{-periodic solution with } p_i \text{ zeros in } [0, N\pi],\ z(0) > 0,\ e \in [0, 1-\varepsilon]\}.$$

For p_1 (respectively, p_3) one finds a family satisfying (4.a) (respectively, (4.b)). For p_2 both alternatives are possible.

The following result provides additional information about Theorem 3.1.

Theorem 3.2. *The following statements hold.*

(1) *If* $p < N$, *then statement* (4.a) *of Theorem 3.1 holds.*

(2) *If* $p \geq N$, $\rho_N < 1-\varepsilon$, *and statement* (4.b) *of Theorem 3.1 holds, then* $E > \rho_N$ *with*

$$\rho_N = \min\left\{2\left(\frac{N}{\nu}\right)^{2/3} - 1,\ 1 - 2\left(\frac{N}{\nu+1}\right)^{2/3}\right\}.$$

We note that statement (1) of Theorem 3.2 provides families which can be globally continued to the whole interval of eccentricities $[0, 1-\varepsilon]$. In the case $N/p \leq 1$, we do not know if the continued family is defined or not in the whole interval $[0, 1)$ but statement (2) of Theorem 3.2 gives us an estimation of the size of the interval of eccentricities where it can be extended. Another consequence of the previous result is the existence of even solutions with minimal period $2N\pi$, $N \geq 2$, for arbitrary eccentricity. It is sufficient to consider the global family associated to $p = 1$. A similar result was obtained in [29]. After adapting Theorem 2 and Corollary 3 of that paper to our notation, one obtains the existence of odd periodic solutions with minimal period $2N\pi$, $N \geq 2$.

Finally, we are going to compare two different families of solutions. Given positive integers M, N, p, q, $1 \leq p \leq \nu_N$, and $1 \leq q \leq \nu_M$, we denote by $\{(z_s, e_s)\}$ and $\{(z_s^*, e_s^*)\}$ the families given by Theorem 3.1 for the couples (p, N) and (q, M), respectively.

Theorem 3.3. *Using the previous notation we assume that $M/q > N/p$.*

(1) *The sets $\{(z_s(0), e_s) : s \in [0, 1)\}$ and $\{(z_s^*(0), e_s^*) : s \in [0, 1)\}$ do not intersect.*
(2) *If statement* (4.a) *of Theorem 3.1 holds for $\{(z_s, e_s)\}$, then the same is true for $\{(z_s^*, e_s^*)\}$.*
(3) *If statement* (4.b) *of Theorem 3.1 holds for $\{(z_s^*, e_s^*)\}$ with $E^* = \lim_{s \nearrow 1} e_s^*$, then the same is true for $\{(z_s, e_s)\}$ with $E = \lim_{s \nearrow 1} e_s \leq E^*$.*

4. Global continuation in the sense of Leray–Schauder. Given an open and bounded subset Ω of $\mathbb{R}^d$ and a function $f : \overline{\Omega} \to \mathbb{R}^d$ which is continuous and does not vanish on the boundary of Ω (i.e., $f(x) \neq 0$ for all $x \in \partial\Omega$), we can define the *Brouwer degree*

$$\deg(f, \Omega),$$

sometimes denoted by $\deg(f, \Omega, 0)$. As usual, $\overline{\Omega}$ denotes the closure of Ω in $\mathbb{R}^d$. There are several equivalent ways to define this degree, and we refer the reader to [25], [30] for more details. For completeness we sketch Nagumo's definition. First we assume that f is of class C^1 and has a finite number of zeros $x_1, \ldots, x_n \in \Omega$ with $\det f'(x_i) \neq 0$ for each i. Then

$$\deg(f, \Omega) = \sum_{i=1}^{n} \operatorname{sign} \det f'(x_i).$$

Now, given an arbitrary continuous function f, we approximate it by functions f_k in the previous conditions and define

$$\deg(f, \Omega) = \lim_{k \to \infty} \deg(f_k, \Omega).$$

Given $x_0 \in \Omega$ a zero of f (i.e., $f(x_0) = 0$), if it is isolated in the set of zeros, then we can define the *Brouwer index of the zero* by

$$\operatorname{ind}(f, x_0) = \deg(f, U),$$

where U is a small neighborhood of x_0. This definition is correct because x_0 is isolated and the degree has the property of excision.

An important property of the degree is its invariance by homotopy. We now state a generalized version where the domain changes with the parameter. It can be found in several papers and books on degree theory, but it is already in [25] (see Lemma 3 of that paper).

Let A be a subset of $\mathbb{R}^d \times [a,b]$. If we denote by (x,λ) the points of $\mathbb{R}^d \times [a,b]$, then we define

$$A_{\lambda_*} = \{\lambda = \lambda_*\} = \{(x,\lambda) \in A : \lambda = \lambda_*\}.$$

Lemma 4.1. *Let U be an open and bounded subset of $\mathbb{R}^d \times [a,b]$ and $f : \overline{U} \to \mathbb{R}^d$ be continuous and such that $f(x,\lambda) \neq 0$ for all $(x,\lambda) \in \partial U$. Then $\deg(f_\lambda, U_\lambda)$ is independent of λ, where $f_\lambda(x) = f(x,\lambda)$.*

We remark that U_λ can be empty for some λ in that case, by Lemma 4.1, the degree is 0.

We conclude these preliminary remarks stating some properties of continua.

Lemma 4.2. *Assume that X is a metric space, $K \subset X$ is a compact set, and $A, B \subset K$ are compact sets such that there is no subcontinuum of K connecting A and B. Then there is an open subset U of X satisfying $A \subset U$, $B \cap \overline{U} = \emptyset$, $K \cap \partial U = \emptyset$.*

Lemma 4.2 or a similar result is employed very often in papers on global bifurcation. See, for instance, [35] and [30].

The next result is stated as a remark after the *Théorème Fondamental* in [25]. Actually, the result in [25] is more general because it works in infinite dimensions for a compact map. We will review the proof of [25]. See also [35] and [30].

Theorem 4.3. *Let $F : \mathbb{R}^d \times [a,b] \to \mathbb{R}^d$ be continuous, and $Z = \{(x,\lambda) : F(x,\lambda) = 0\}$ be the set of zeros of F. Assume that*

(H1) *Z is bounded, and*

(H2) *the set Z_a is finite and there is $(x_0,a) \in Z_a$ with $ind(F_a, x_0) \neq 0$.*

Let C be the connected component of Z containing (x_0,a). Then one of the following alternatives holds:

(a) *$C \cap \{\lambda = b\} \neq \emptyset$.*

(b) *There exists $(x_1,a) \in Z_a$, $x_1 \neq x_0$, such that $(x_1,a) \in C$.*

Proof. Consider the metric space $X = \mathbb{R}^d \times [0,1]$ and the set $K = Z$. The assumption (H1) implies that K is compact. Define

$$A = \{(x_0,a)\}, \quad B = (Z_0 \setminus A) \cup \{(x,b) : |x| \leq M\},$$

where M is a large constant so that Z is included in $|x| < M$. If neither (a) nor (b) holds, then there is no subcontinuum of Z meeting A and B. We find U an open subset of X such that

$$\{(x_0,a)\} = U_a \cap Z, \quad U_b = \emptyset, \quad Z \cap \partial U = \emptyset.$$

By Lemma 4.1, the $\deg(F_\lambda, U_\lambda)$ is independent of λ. Since $U_b = \emptyset$, this degree must be zero. On the other hand, by (H2) we have

$$\deg(F_a, U_a) = \operatorname{ind}(F_a, x_a) \neq 0.$$

This contradiction shows that (a) or (b) must hold. ∎

In general, the continuum C can be rather pathological; however, there is a special case in which one can guarantee that C is arcwise connected. In this case there are arcs joining all points of C, and this corresponds to the usual idea of *continuation*.

Theorem 4.4. *Under the assumptions of Theorem 4.3 suppose that $d = 1$ and F is real and analytic. Then, there is a continuum arc $\alpha : [0,1] \to Z$, $\alpha(s) = (x(s), \lambda(s))$ with $x(0) = x_0$, $\lambda(0) = a$ such that either $\lambda(1) = b$ or $\lambda(1) = a$ and $x(1) \neq x_0$.*

Many results on the effect of analyticity on global continuation can be seen in [14] and [6]. The local structure of the set of zeros of $F(x, \lambda) = 0$ says that C is locally arcwise connected. Since C is, by definition, connected, we conclude that C is arcwise connected.

5. Periodic solutions of the Sitnikov problem. Equation (2.1) is the equation of motion for the Sitnikov problem, where $e \in [0, 1)$ is the eccentricity and $r(t, e)$ is the distance of the primaries to its center of mass. The eccentric anomaly $u(t)$ satisfies

$$u(t + 2\pi) = u(t) + 2\pi, \quad u(-t) = -u(t),$$

and so, by (2.2) and (2.4), $r(t, e)$ is an even and 2π-periodic function.

Given an integer $N \geq 1$, we shall be interested in even, $2N\pi$-periodic solutions of (2.1). They satisfy the boundary conditions

$$\dot{z}(0) = \dot{z}(N\pi) = 0. \tag{5.1}$$

Let $\varphi(t; \xi, e)$ be the solution of (2.1) satisfying

$$z(0) = \xi, \quad \dot{z}(0) = 0.$$

This is a real analytic function in the arguments $(t; \xi, e) \in \mathbb{R} \times \mathbb{R} \times [0, 1)$. Notice that these solutions are globally defined in $(-\infty, +\infty)$ because the nonlinearity in (2.1) is bounded. We define

$$F_N : \mathbb{R} \times [0, 1) \to \mathbb{R}, \qquad F_N(\xi, e) = \dot{\varphi}(N\pi; \xi, e).$$

The research of even $2N\pi$-periodic solutions of (2.1) satisfying (5.1) is equivalent to the study of the equation

$$F_N(\xi, e) = 0.$$

We want to apply Theorem 4.4 to F_N, and so we must verify its assumptions.

Toward assumption (H1). We first consider the circular Sitnikov problem

$$\ddot{z} = -\frac{z}{(z^2 + R^2)^{3/2}}$$

for some $R > 0$. Let $\psi(t; \xi)$ be its solution satisfying

$$z(0) = \xi, \quad \dot{z}(0) = 0, \quad (\xi > 0).$$

Then, $\psi(t; \xi)$ is periodic with minimal period $T(\xi) > 0$, $\lim_{\xi \to +\infty} T(\xi) = +\infty$, and

$$0 < \psi(t; \xi) < \xi, \quad \dot{\psi}(t, \xi) < 0 \text{ if } t \in \left(0, \frac{T(\xi)}{4}\right).$$

Fix $\xi_* > 0$ such that $T(\xi) > 4N\pi$ if $\xi \geq \xi_*$.

Proposition 5.1. *Assume that $r(t,e) \geq R$ for all t, and let $\varphi(t;\xi,e)$ be a solution of (2.1) satisfying (5.1). Then $|\xi| \leq \xi_*$.*

Proof. The equation of motion (2.1) is invariant under the symmetry $(z,t) \to (-z,-t)$. Therefore, we can assume that $\varphi(0;\xi,e) = \xi > 0$. Note that one can deduce from the equation that ξ is a local maximum of $\varphi(t;\xi,e)$. Also, by integrating the equation between 0 and $N\pi$, one deduces that $\varphi(t;\xi,e)$ changes sign in this interval. Let $\tau > 0$ be the first zero of $\varphi(t;\xi,e)$ in $[0,N\pi]$. We must have

$$\dot\varphi(t;\xi,e) < 0 \quad \text{for } t \in (0,\tau).$$

Otherwise, two consecutive critical points of $\varphi(t;\xi,e)$ should be maxima.

Set $\varphi = \varphi(t) = \varphi(t;\xi,e)$. For $t \in (0,\tau)$ we have

$$\frac{d}{dt}\left(\frac{1}{2}\dot\varphi^2 - \frac{1}{(\varphi^2+R^2)^{1/2}}\right) = \dot\varphi\left(-\frac{\varphi}{(\varphi^2+r^2)^{3/2}} + \frac{\varphi}{(\varphi^2+R^2)^{3/2}}\right) \leq 0,$$

and consequently

$$\frac{1}{2}\dot\varphi^2 - \frac{1}{(\varphi^2+R^2)^{1/2}} \leq -\frac{1}{(\xi^2+R^2)^{1/2}}.$$

For $t \in (0,\tau)$, φ is a solution of the differential inequality

$$\dot\varphi \geq -\sqrt{2}\left(\frac{1}{(\varphi^2+R^2)^{1/2}} - \frac{1}{(\xi^2+R^2)^{1/2}}\right)^{1/2}.$$

For $t \in \left(0, \frac{T(\xi)}{4}\right)$, $\psi(t) = \psi(t;\xi)$ satisfies the associated differential equation. In fact, $\psi(t)$ is the minimal solution of

$$\dot x = -\sqrt{2}\left(\frac{1}{(x^2+R^2)^{1/2}} - \frac{1}{(\xi^2+R^2)^{1/2}}\right)^{1/2}, \quad x(0) = \xi.$$

Therefore, the theory of differential inequalities implies that

$$\varphi(t) \geq \psi(t) \quad \text{if } 0 \leq t \leq \min\left\{\tau, \frac{T(\xi)}{4}\right\}.$$

From here we conclude that $\tau \geq T(\xi)/4$. Thus $N\pi > \tau \geq T(\xi)/4$, and so $\xi < \xi_*$. ∎

The previous proposition allows us to verify (H1) in each strip of the type $\mathbb{R}\times[0,E]$ with $E < 1$. In fact, if $e \in [0,E]$, then $r(t,e) \geq (1-E)/2$, and we can apply the previous result with $R = (1-E)/2$.

Toward assumption (H2). We want to study the zeros of $F_N(\cdot,0)$. This is equivalent to studying the solutions of

$$\ddot z = -\frac{z}{(z^2+\frac{1}{4})^{3/2}}, \quad \dot z(0) = \dot z(N\pi) = 0. \tag{5.2}$$

As we already mentioned, the solutions $\varphi(t;\xi,0)$ are periodic with minimal period $T(\xi)$, and $T(\xi)$ is an increasing function in ξ; see [5].

By the symmetry, $\varphi(t;\xi,0)$ with $\xi \neq 0$ is a solution of the boundary problem (5.2) if and only if there is an integer $p \geq 1$ such that $T(\xi)/2 = N\pi/p$.

Since $\inf T(\xi) = \pi/\sqrt{2}$ (see [5]), we have $2N\pi/p > \pi/\sqrt{2}$, and consequently $p < 2\sqrt{2}N$. Define $\nu = \nu_N = [2\sqrt{2}N]$, and let $\xi_1 > \cdots > \xi_\nu > 0$ be the solutions of $T(\xi)/2 = N\pi/p$ with $p = 1, \ldots, \nu$. Then

$$Z_0 = \{-\xi_1, \ldots, -\xi_\nu, 0, \xi_\nu, \ldots, \xi_1\}. \tag{5.3}$$

We compute the indices. Since $T = T(\xi)$ is increasing, we have that

$$\dot\varphi(N\pi;\xi,0) > 0 \quad \text{if } \xi < \xi_1 \text{ close to } \xi_1,$$
$$\dot\varphi(N\pi;\xi,0) < 0 \quad \text{if } \xi > \xi_1 \text{ close to } \xi_1.$$

From here, $\operatorname{ind}(F_N(\cdot,0),\xi_1) = -1$. In general,

$$\operatorname{ind}(F_N(\cdot,0),\xi_p) = (-1)^p. \tag{5.4}$$

The indices for $-\xi_p$ can be computed using the symmetry. We also compute the index at 0, although this information will not be employed in the rest of the paper. We do it by linearization; i.e.,

$$\operatorname{ind}(F_N(\cdot,0),0) = \operatorname{sign}\left(\frac{\partial F_N}{\partial \xi}(0;0)\right).$$

We note that $\frac{\partial F_N}{\partial \xi}(0;0) = \dot y(N\pi)$, where $y(t)$ is the solution of the variational problem

$$\ddot y + 8y = 0, \quad y(0) = 1, \quad \dot y(0) = 0.$$

So, we have

$$\operatorname{ind}(F_N(\cdot,0),0) = \operatorname{sign}\left(-\sin(2\sqrt{2}N\pi)\right) = (-1)^{\nu+1}.$$

To sum up: for $e = 0$ there exist $\nu = \nu_N = [2\sqrt{2}N]$ nontrivial, even, and $2N\pi$-periodic solutions of (5.2) with $z(0) > 0$. They can be labeled by the number of zeros of $z(t)$ in $[0, N\pi]$ for $p = 1, \ldots, \nu$ and

$$\varphi_1(0) = \xi_1 > \cdots > \varphi_\nu(0) = \xi_\nu.$$

Moreover, the index of each of these solutions, $\operatorname{ind}(F_N(\cdot,0),\xi_p)$, is ± 1.

We summarize our knowledge of the set Z. We have the trivial continuum $z = 0$, $e \in [0,1)$. Also, we have the solutions $(\xi_i, 0)$. Since the index is different from zero, there is at least a local branch emanating from them. Finally, we know that Z can only blow up as $e \uparrow 1$. We want to study the possible collisions of the branches emanating from ξ_i with the trivial continuum. To this end we linearize around $z = 0$.

6. Linearization around the equilibrium. The main objective of this section is to prove the next result.

Theorem 6.1. *Consider the boundary value problem*

$$\ddot y + \frac{1}{r(t,e)^3} y = 0, \quad \dot y(0) = \dot y(N\pi) = 0. \tag{6.1}$$

Then, there exists a sequence $\{E_{n,N}\}_{n\geq 1}$ *satisfying* $0 < E_{1,N} < \cdots < E_{n,N} < \cdots < 1$, *converging to* 1, *and such that there is a nontrivial solution of* (6.1) *if and only if* $e = E_{n,N}$. *Moreover,* $E_{1,N} > \rho_N$ *(*ρ_N *was defined in statement* (2) *of Theorem* 3.2*), and the solution* y_n *corresponding to* $E_{n,N}$ *has a number of zeros in* $[0, N\pi]$ *which becomes arbitrarily large as* $n \uparrow \infty$.

The linear differential equation (6.1) has been studied in detail in [28], and most of the statements above follow from their results. There is an alternative way of studying this equation. In fact, the change of the independent variable $u - e\sin u = t$ transforms (6.1) into

$$(1 - e\cos u)\frac{d^2y}{du^2} - e\sin u\frac{dy}{du} + 8y = 0.$$

Here we have used (2.2). This is a particular case of the so-called Ince's equation (see [27]), which has been studied by several authors. In particular, the techniques of [31] can probably be employed for the effective computation of the numbers $E_{n,N}$.

Let $y(t;e)$ be the solution of the equation appearing in (6.1) which satisfies $y(0) = 1$, $\dot{y}(0) = 0$. We study the zeros of $\dot{y}(N\pi, e) = 0$.

Proposition 6.2. *If* $e \leq \rho_N$, *then* $\dot{y}(N\pi, e) \neq 0$.

To prove Proposition 6.2 we need the following result, which is a well-known consequence of Sturm comparison theory (see [11] for more details).

Lemma 6.3. *Assume that* $a(t)$ *is continuous, is* $2N\pi$*-periodic, and for some* $n \geq 0$ *satisfies*

$$\left(\frac{n}{N}\right)^2 \leq a(t) \leq \left(\frac{n+1}{N}\right)^2 \quad \text{for all } t \in \mathbb{R},$$

and that both inequalities are strict somewhere. Then, $\ddot{y} + a(t)y = 0$ *has no* $2N\pi$*-periodic solutions (excepting* $y \equiv 0$*).*

In our case, if $e \leq \rho_N$, since $(1-e)/2 \leq r(t,e) \leq (1+e)/2$, we have

$$\left(\frac{\nu}{N}\right)^2 \leq \frac{8}{(1+e)^3} \leq \frac{1}{r(t,e)^3} \leq \frac{8}{(1-e)^3} \leq \left(\frac{\nu+1}{N}\right)^2.$$

Now, from Lemma 6.3, Proposition 6.2 follows.

Proposition 6.4. *The number of zeros of* $y(t;e)$ *in* $(0, N\pi)$ *tends to infinity as* $e \uparrow 1$.

Proof. First, we notice that $r(t,e)$ can be extended to $e = 1$. This extension is continuous in both variables and, in particular, $r(\cdot, e)$ converges uniformly to $r(\cdot, 1)$ as $e \uparrow 1$. Also, from $u - \sin u = t$ $(e = 1)$, we deduce that $u(t,1) = (6t)^{1/3}a(t)$, where a is continuous and $a(0) = 1$. Next

$$r(t,1) = \frac{1}{2}(1 - \cos u(t,1)) = \sin^2\frac{u(t,1)}{2} = \frac{6^{2/3}}{4}t^{2/3}b(t),$$

with b continuous and $b(0) = 1$. Thus,

$$\frac{1}{r(t,1)^3} = \frac{16}{9t^2b(t)^3}.$$

Let us fix a number γ in the interval $(1/4, 16/9)$. For $\Delta > 0$ small enough,

$$\frac{1}{r(t,1)^3} > \frac{\gamma}{t^2}, \quad t \in (0, \Delta].$$

Consider the Euler equation $\ddot{y}+\frac{\gamma}{t^2}y=0$. The solutions have infinitely many zeros accumulating at $t=0$ because $\gamma>1/4$. This can be checked by direct integration. Now, given an arbitrary $m\geq 1$, we can find $\delta\in(0,\Delta)$ such that the solutions of this Euler equation have at least $m+1$ zeros in the interval (δ,Δ). Since $r(\cdot,e)$ converges uniformly to $r(\cdot,1)$, it is possible to find $e_\star<1$ such that

$$\frac{1}{r(t,e)^3}>\frac{\gamma}{t^2},\quad t\in[\delta,\Delta],\ e\in[e_\star,1].$$

By Sturm comparison theory, $y(t,e)$ will have at least m zeros in $[\delta,\Delta]$. ■

Proof of Theorem 6.1. Changing (6.1) to polar coordinates $y=\rho\cos\theta$, $\dot{y}=\rho\sin\theta$, $\dot{y}(N\pi,e)=0$ becomes equivalent to $\theta(N\pi,e)\in\pi\mathbb{Z}$. The angle $\theta(t,e)$ satisfies

$$\dot{\theta}=-\frac{1}{r(t,e)^3}\cos^2\theta-\sin^2\theta.$$

Thus, $\theta(t,e)$ is decreasing in t. When e is close to 1, $\theta(t,e)\in\pi/2+\mathbb{Z}\pi$ for more and more positive t's in $[0,N\pi]$. This implies that

$$\theta(N\pi,e)=\inf_{t\in[0,N\pi]}\theta(t,e)\to-\infty$$

as $e\uparrow 1$. The function $e\in[0,1)\longmapsto\theta(N\pi,e)$ is analytic and $\lim_{e\uparrow1}\theta(N\pi,e)=-\infty$. The numbers $E_{n,N}$ are the solutions of $\theta(N\pi,e)\in\mathbb{Z}\pi$. ■

7. The conclusion. We need the following two lemmas.

Lemma 7.1. *Let $(z_n(t),e_n)$ be a sequence of solutions of* (2.1) *satisfying $\dot{z}_n(0)=\dot{z}_n(N\pi)=0$, $z_n(0)\to0$, $z_n(0)\neq0$, $e_n\to e_0<1$. Then, the number of zeros of $z_n(t)$ in $[0,N\pi]$ coincides, for large n, with the number of zeros in the same interval of the nontrivial solutions of*

$$\ddot{y}+\frac{1}{r(t,e_0)^3}y=0.$$

In particular, $e_0=E_{n,N}$ for some n.

Proof. By continuous dependence, $z_n(t)\to0$ uniformly in $[0,N\pi]$. Define $v_n(t)=z_n(t)/z_n(0)$. It satisfies

$$\ddot{v}_n+\frac{1}{(z_n(t)^2+r(t,e_n)^2)^{3/2}}v_n=0,\quad v_n(0)=1,\quad \dot{v}_n(0)=0.$$

Again, by continuous dependence $v_n(t)$ converges in $C^1[0,N\pi]$ to the solution $y(t)$ of

$$\ddot{y}+\frac{1}{r(t,e_0)^3}y=0,\quad y(0)=1,\quad \dot{y}(0)=0.$$

Since $\dot{v}_n(0)=\dot{v}_n(N\pi)=0$, we have $\dot{y}(0)=\dot{y}(N\pi)=0$. Thus, all the zeros of $y(t)$ in $[0,N\pi]$ are in its interior. Since $v_n\to y$, $\dot{v}_n\to\dot{y}$ uniformly, and all the zeros of $y(t)$ are nondegenerate (i.e., $y(\tau)=0$ and $\dot{y}(\tau)\neq0$), for large n we deduce that $v_n(t)$ and $y(t)$ have the same number of zeros. ■

Lemma 7.2. *Let $\{x_\lambda(t)\}_{\lambda\in[0,1]}$ be a family of functions in $C^1[0,T]$ satisfying the following:*

(i) $x_\lambda(0) \neq 0$, $x_\lambda(T) \neq 0$ *for all* $\lambda \in [0,1]$.
(ii) *The zeros of* x_λ *are nondegenerate (i.e.,* $x_\lambda(t)^2 + \dot{x}_\lambda(t)^2 > 0$ *everywhere).*
(iii) *The map* $(t,\lambda) \in [0,T] \times [0,1] \longmapsto (x_\lambda(t), \dot{x}_\lambda(t))$ *is continuous.*

Then, the number of zeros of x_λ *in* $[0,T]$ *is independent of* λ.

Proof. The lemma follows from the fact that the number of zeros is locally constant. This is easy because functions which are C^1 close and have nondegenerate zeros have the same number of zeros. ∎

Proof of Theorem 3.1. From the discussions of section 5 we know that searching for $z(t)$, even and $2N\pi$-periodic solution of (2.1), is equivalent to finding a root of the equation

$$F_N(z_0, e) = 0 \quad (z_0 = z(0)).$$

In this way we are lead to the framework of section 4 and we shall apply Theorem 4.4 with $F = F_N$. The parameter λ is now the eccentricity and $[a,b] = [0, 1-\epsilon]$. Let us check that (H1) and (H2) hold. The first condition follows from Proposition 5.1 and the discussion after its proof. To deal with (H2) we recall the information obtained in section 5 about the set Z_0, as given by (5.3). There is a unique $\xi_p \in Z_0 \cap (0,\infty)$ with $T(\xi_p)/2 = N\pi/p$. Here we are using $p \leq \nu$. We know from (5.4) that the index of $F_N(\cdot, 0)$ at ξ_p is different from zero and so (H2) holds for $x_0 = \xi_p$. At this moment it is important to observe that the solution of (2.5) with $z(0) = \xi_p$, $\dot{z}(0) = 0$ is even and $2N\pi$-periodic and has exactly p zeros in $[0, N\pi]$. From Theorem 4.4 we infer the existence of a continuous family $\{(\xi(s), e_s)\}_{s\in[0,1]}$ in $\mathbb{R} \times [0, 1-\epsilon]$ such that

$$F_N(\xi(s), e_s) = 0, \quad \xi(0) = \xi_p, \quad e_0 = 0,$$

and either

$$e_1 = 1 - \epsilon \tag{7.1}$$

or

$$e_1 = 0, \quad \xi(1) \neq \xi_p. \tag{7.2}$$

Let $z_s(t)$ denote the solution of (2.1) for $e = e_s$ which satisfies $z_s(0) = \xi(s)$, $\dot{z}_s(0) = 0$. The family $\{(z_s(t), e_s)\}_{s\in[0,1]}$ satisfies conditions (1) and (2) of Theorem 3.1, but there are no a priori reasons to suppose that it also satisfies (3). We distinguish two cases.

Case 1. $z_s(0) > 0$ for all $s \in [0,1]$.

Lemma 7.2 implies that $z_s(t)$ has p zeros in $[0, N\pi]$ for each $s \in [0,1]$. This proves that (3) also holds. Assuming now that the first alternative (7.1) holds, we arrive at the searched family satisfying (4.a). The second alternative (7.2) cannot occur, for otherwise $e_1 = 0$ and $\xi(1) \neq \xi_p$. Then, since $\xi(1)$ belongs to $Z_0 \cap (0,\infty)$, we deduce that $\xi(1) = \xi_q$ for some $q \neq p$. This would imply that $z_1(t)$ has q zeros in $[0, N\pi]$, a situation which would be incompatible with (3).

Case 2. $z_s(0)$ vanishes for some $s \in [0,1]$.

Let $\sigma \in (0,1]$ be the first zero of $z_s(0)$, so that

$$z_s(0) > 0 \quad \text{if } s \in [0,\sigma),\ z_\sigma(0) = 0.$$

The family $\{(\hat{z}_s, \hat{e}_s)\}_{s\in[0,1)}$, with $\hat{z}_s = z_{s\sigma}$ and $\hat{e}_s = e_{s\sigma}$, satisfies (3). Again Lemma 7.2 has been used. The definition of σ implies that $\lim_{s\to 1^-} \hat{e}_s = e_\sigma \in [0, 1-\epsilon]$ and $\hat{z}_s(t)$ converges to 0 as $s \to 1^-$. Finally, we apply Lemma 7.1 to conclude that $e_\sigma = E_{n,N}$ for some n. Moreover, nontrivial solutions of (6.1) for $e = e_\sigma$ must have p zeros in $[0, N\pi]$. In this way we have constructed a family satisfying (4.b), and the theorem is proven. ■

Proof of Theorem 3.2. Since $r(t, e_s) < 1$ everywhere, we can apply the Sturm comparison theory to the equation appearing in (6.1) and to $\ddot{y} + y = 0$, to deduce that the solutions of (6.1) must have at least N zeros in the interval $[0, N\pi]$. So, if $p < N$, statement (4.b) does not hold. Therefore, statement (1) of Theorem 3.2 is proved.

If $e \le \rho_N$, Proposition 6.2 says that (6.1) has no periodic solutions different from the trivial one. So, statement (2) of Theorem 3.2 follows. ■

Proof of Theorem 3.3. We shall prove statement (1) by contradiction. Assume the existence of $\sigma \in [0, 1)$ such that $z_\sigma(0) = z^*_\sigma(0)$ and $e_\sigma = e^*_\sigma$. Since we know that $\dot{z}_\sigma(0) = \dot{z}^*_\sigma(0) = 0$, by uniqueness, z_σ and z^*_σ are the same periodic solution. This solution has periods $2\pi M$ and $2\pi N$, having q zeros in $[0, M\pi]$ and p zeros in $[0, N\pi]$. Let $2\pi s$ be the minimal period of this solution. Then, there are integers m_1 and m_2 such that $sm_1 = M$ and $sm_2 = N$. Let r be the number of zeros of z_σ in $[0, s\pi]$. Therefore, $m_1 r = q$ and $m_2 r = p$. So, $M/q = N/p$, which is a contradiction. Consequently, statement (1) is proved.

Statements (2) and (3) follow from (1) and the fact that $z^*_0(0) > z_0(0)$. ■

In short, using the global continuation method of the zeros of a function depending on one parameter due to Leray and Schauder and based in the Brouwer degree, we have studied analytically the families of symmetric periodic orbits of the elliptic Sitnikov problem for all values of the eccentricity in the interval $[0, 1)$, providing qualitative and quantitative information on the bifurcation diagram of such families of periodic orbits. The quantitative information mainly is on the periods of the periodic orbits and on some estimations where the different families of periodic orbits can start or end. The precise statements of these results are in Theorems 3.1, 3.2, and 3.3.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, Benjamin/Cummings, Reading, MA, 1978.

[2] V. M. Alekseev, *Quasirandom dynamical systems* I, Math. USSR Sbornik, 5 (1968), pp. 73–128.

[3] V. M. Alekseev, *Quasirandom dynamical systems* II, Math. USSR Sbornik, 6 (1968), pp. 505–560.

[4] V. M. Alekseev, *Quasirandom dynamical systems* III, Math. USSR Sbornik, 7 (1969), pp. 1–43.

[5] E. Belbruno, J. Llibre, and M. Ollé, *On the families of periodic orbits which bifurcate from the circular Sitnikov motions*, Celestial Mech. Dynam. Astronom., 60 (1994), pp. 99–129.

[6] B. Buffoni and J. Toland, *Analytic Theory of Global Bifurcation*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2003.

[7] H. Cabral and Z. Xia, *Subharmonic solutions in the restricted three–body problem*, Discrete Contin. Dynam. Systems, 1 (1995), pp. 463–474.

[8] J. Chazy, *Sur l'allure finale du mouvement dans le problème des trois corps quand le temps croit indefiniment*, Annales de l'Ecole Norm. Sup., 3 sr. 39 (1922), pp. 22–130.

[9] J. Chazy, *Sur l'allure finale du mouvement dans le problème des trois corps*, J. Math. Pures Appl., 8 (1929), pp. 353–380.

[10] J. Chazy, *Sur l'allure finale du mouvement dans le problème des trois corps*, Bull. Astron., 8 (1932), pp. 403–436.

[11] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, Toronto, London, 1955.
[12] M. Corbera and J. Llibre, *Periodic orbits of the Sitnikov problem via a Poincaré map*, Celestial Mech. Dynam. Astronom., 77 (2000), pp. 273–303.
[13] M. Corbera and J. Llibre, *On symmetric periodic orbits of the elliptic Sitnikov problem via the analytic continuation method*, in Celestial Mechanics, Contemp. Math. 292, AMS, Providence, RI, 2002, pp. 91–127.
[14] E. N. Dancer, *Global structure of the solutions of non–linear real analytic eigenvalue problems*, Proc. London Math. Soc., 27 (1973), pp. 747–765.
[15] H. Dankowicz and P. Holmes, *The existence of transverse homoclinic points in the Sitnikov problem*, J. Differential Equations, 116 (1995), pp. 468–483.
[16] R. Dvorak, *Numerical results to the Sitnikov problem*, Celestial Mech. Dynam. Astronom., 56 (1993), pp. 71–80.
[17] S. B. Faruque, *Solution of the Sitnikov problem*, Celestial Mech. Dynam. Astronom., 87 (2003), pp. 353–369.
[18] J. Hagel, *A new analytical approach to the Sitnikov problem*, Celestial Mech. Dynam. Astronom., 53 (1992), pp. 267–292.
[19] J. Hagel and T. Trenkler, *A computer aided analysis of the Sitnikov problem*, Celestial Mech. Dynam. Astronom., 56 (1993), pp. 81–98.
[20] J. Hagel and C. Lhotka, *A high order perturbation analysis of the Sitnikov problem*, Celestial Mech. Dynam. Astronom., 93 (2005), pp. 201–228.
[21] J. Liu and Y.-S. Sun, *On the Sitnikov problem*, Celestial Mech. Dynam. Astronom., 49 (1990), pp. 285–302.
[22] L. Jiménez-Lara and A. Escalona-Buendía, *Symmetries and bifurcations in the Sitnikov problem*, Celestial Mech. Dynam. Astronom., 79 (2001), pp. 97–117.
[23] J. Kallrath, R. Dvorak, and J. Schlöder, *Periodic orbits in the Sitnikov problem*, in The Dynamical Behaviour of Our Planetary System (Ramsau, 1996), Kluwer, Dordrecht, The Netherlands, 1997, pp. 415–428.
[24] J. Llibre and C. Simó, *Estudio cualitativo del problema de Sitnikov*, Publicacions Matemàtiques U.A.B., 18 (1980), pp. 49–71.
[25] J. Leray and J. Schauder, *Topologie et équations fonctionnelles*, Ann. Sci. École Norm. Sup. (3), 51 (1934), pp. 45–78.
[26] W. D. MacMillan, *An integrable case in the restricted problem of three bodies*, Astron. J., 27 (1913), pp. 11–13.
[27] W. Magnus and S. Winkler, *Hill's Equation*, Dover, New York, 1979.
[28] J. Martinez-Alfaro and C. Chiralt, *Invariant rotational curves in Sitnikov's problem*, Celestial Mech. Dynam. Astronom., 55 (1993), pp. 351–367.
[29] S. Mathlouthi, *Periodic orbits of the restricted three-body problem*, Trans. Amer. Math. Soc., 350 (1998), pp. 2265–2276.
[30] J. Mawhin, *Continuation theorems and periodic solutions of ordinary differential equations*, in Topological Methods in Differential Equations and Inclusions, A. Granas and M. Frigon, eds., Kluwer, Dordrecht, The Netherlands, 1995, pp. 291–375.
[31] R. Mennicken, *On Ince's equation*, Arch. Ration. Mech. Anal., 29 (1968), pp. 144–160.
[32] J. Moser, *Stable and Random Motions in Dynamical Systems*, Annals of Math. Studies 77, Princeton University Press, Princeton, NJ, 1973.
[33] G. Pavanini, *Sopra una nuova categoria di soluzioni periodiche nel problema di tre corpi*, Annali di Mathematica, Serie III, Tomo XIII (1907), pp. 179–202.
[34] E. Perdios and V. V. Markellos, *Stability and bifurcations of Sitnikov motions*, Celestial Mech. Dynam. Astronom., 42 (1988), pp. 187–200.
[35] P. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.
[36] C. Rebelo and F. Zanolin, *On the existence and multiplicity of branches of nodal solutions for a class of parameter-dependent Sturm-Liouville problems*, Differential Integral Equations, 13 (2000), pp. 1473–1502.

[37] K. A. Sitnikov, *Existence of oscillating motion for the three–body problem*, Dokl. Akad. Nauk, 133 (1960), pp. 303–306.
[38] K. Stumpff, *Himmelsmeckanik, Band* II, VEB, Berlin, 1965, pp. 73–79.
[39] V. Szebehely, *Theory of Orbits*, Academic Press, New York, 1967.
[40] K. Wodnar, *New formulations of the Sitnikov problem*, in Predictability, Stability, and Chaos in N-Body Dynamical Systems, A. E. Roy, ed., Plenum Press, New York, 1991.
[41] K. Wodnar, *The original Sitnikov article—new insights*, Celestial Mech. Dynam. Astronom., 56 (1993), pp. 99–101.
[42] K. Wodnar, *Analytical approximations for Sitnikov's problem*, in From Newton to Chaos, A. E. Roy and B. A. Steves, ed., Plenum Press, New York, 1995.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 577–608
© 2008 Society for Industrial and Applied Mathematics

Freezing Multipulses and Multifronts*

Wolf-Jürgen Beyn†, Sabrina Selle†, and Vera Thümmler†

Abstract. We consider nonlinear time dependent reaction diffusion systems in one space dimension that exhibit multiple pulses or multiple fronts. In an earlier paper two of the authors developed the freezing method that allows us to compute a moving coordinate frame in which, for example, a traveling wave becomes stationary. In this paper we extend the method to handle multifronts and multipulses traveling at different speeds. The solution of the Cauchy problem is decomposed into a finite number of single waves, each of which has its own moving coordinate system. The single solutions satisfy a system of partial differential algebraic equations coupled by nonlinear and nonlocal terms. Applications are provided to the Nagumo and the FitzHugh–Nagumo systems. We justify the method by showing that finitely many traveling waves, when patched together in an appropriate way, solve the coupled system in an asymptotic sense. The method is generalized to equivariant evolution equations and is illustrated by the complex Ginzburg–Landau equation.

Key words. multipulses, partial differential algebraic equations, unbounded domains, equivariance, Lie groups

AMS subject classifications. 65M99, 35K57

DOI. 10.1137/07070749X

1. Introduction.

Consider a parabolic system for a function $u(x,t) \in \mathbb{R}^m$ on the real line

$$u_t = Au_{xx} + f(u, u_x), \quad x \in \mathbb{R},\ t \geq 0, \quad u(x,0) = u_0(x), \quad x \in \mathbb{R}, \tag{1.1}$$

where $A \in \mathbb{R}^{m,m}$ is assumed to be positive definite and $f : \mathbb{R}^{2m} \to \mathbb{R}^m$ is assumed to be smooth. We are interested in systems that have more than one traveling wave solution

$$u_j(x,t) = w_j(x - c_j t), \quad j = 1, \ldots, N, \tag{1.2}$$

traveling at different speeds c_j and with limiting behavior

$$w_j^- = \lim_{\xi \to -\infty} w_j(\xi), \quad w_j^+ = \lim_{\xi \to \infty} w_j(\xi). \tag{1.3}$$

It is frequently observed that such systems exhibit special solutions that look like a superposition of several waves. In Figure 1 we illustrate the case of two pulses and two fronts that travel in opposite directions ($c_1 < 0 < c_2$) and that can be patched together (i.e., $w_1^+ = w_2^-$); see section 2 for a more precise definition of the meaning of patching. Solutions of this type are usually called multifronts or multipulses depending on whether the limits at $\pm\infty$ agree

*Received by the editors November 6, 2007; accepted for publication (in revised form) by T. Kaper February 16, 2008; published electronically June 13, 2008. This research was supported by CRC 701 'Spectral Analysis and Topological Methods in Mathematics.'
http://www.siam.org/journals/siads/7-2/70749.html
†Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, D-33501 Bielefeld, Germany (beyn@math.uni-bielefeld.de, sselle@math.uni-bielefeld.de, thuemmle@math.uni-bielefeld.de).

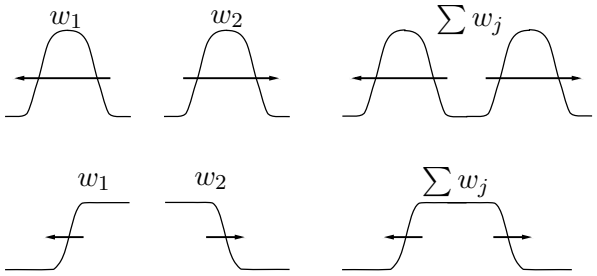


Figure 1. *Single pulses/multipulse and single fronts/multifront.*

or disagree. There is quite an extensive literature that studies existence and stability of multifronts and multipulses close to a fixed pulse configuration with the single pulses far apart and with a common speed; see [15], [13], [14], [16], [22], [7]. More recently, in [23] a center manifold is constructed that contains all types of multipulses (even infinitely many) with large spacings that travel at a slowly varying speed.

In this paper we consider a finite number of pulses, respectively, fronts, that travel at different speeds. We provide a working definition for multifront solutions in an asymptotic sense that will be sufficient for our approach. Note that in the recent paper [17] the authors construct an invariant manifold that contains and attracts such solutions up to a certain time instance prior to collision.

The main goal of this paper is to numerically construct a decomposition of the solution $u(x,t)$ of the Cauchy problem (1.1) of the form

$$u(x,t) = \sum_{j=1}^{N} v_j(x - g_j(t), t). \tag{1.4}$$

The idea is to find functions $v_j : \mathbb{R} \times [0,\infty) \to \mathbb{R}^m$, $(\xi,t) \mapsto v_j(\xi,t)$ that approximate the jth profile in the multifront and that have a rather local support when compared to the overall solution $u(x,t)$. The functions $g_j : \mathbb{R} \to \mathbb{R}$ denote the time-dependent position of the jth profile and will be determined by the numerical process as well. The N-dimensional system that determines the v_j will be set up such that the linear superposition (1.4) is an exact solution of the nonlinear system (1.1) and such that this system can be solved on a much smaller domain than the original equation. Note that, if repelling pulses or fronts appear as in Figure 1, growing spatial domains are needed to compute and represent the solution of (1.1), while our system will be solved on a domain of moderate size that stays constant for all times. Moreover, our method will produce the individual velocities automatically without any a posteriori analysis of simulation data.

We follow the freezing approach for single waves in [4], [5] (see [12] for a related approach) by setting up an appropriate phase condition for each of the single profiles v_j. In section 2 we derive the basic system of N partial differential algebraic equations (PDAEs) for the functions v_j that will be solved numerically. The nonlinearities in this system contain nonlocal terms due to the different positions of the single profiles.

For the numerical computations we truncate this system to a finite interval, use appropriate boundary conditions, and discretize by finite elements in space and BDF methods (based on

backward differentiation formulae) in time. In section 3 we show several applications of our method to multifronts that occur in the Nagumo and in the FitzHugh–Nagumo systems. It may come as a surprise that the method even works in cases for which it was not designed, namely, fronts or pulses that collide and annihilate each other.

In section 4 we give a certain theoretical justification of our method. It is shown that appropriately modified waves (1.2) satisfy the PDAE system in an asymptotic sense.

Finally, in section 5 we generalize our "decompose and freeze" approach to general evolution equations that are equivariant with respect to the action of a (not necessarily compact) Lie group. As an application we discuss the decomposition of solutions of a complex Ginzburg–Landau equation that has a two-dimensional group of equivariances.

2. Decomposition of multifronts. Let us be more precise about the process of patching single waves (1.2). Assume that the left and right limits of the waves match in the sense that

$$w_j^+ = w_{j+1}^-, \quad j = 1, \dots, N-1. \tag{2.1}$$

Then we write down the superposition

$$U(x,t) = \sum_{j=1}^{N} \hat{w}_j(x - c_j t), \quad \hat{w}_j(\xi) = \begin{cases} w_1(\xi), & j = 1, \\ w_j(\xi) - w_j^-, & j \geq 2, \end{cases} \tag{2.2}$$

where we have subtracted the left limits so that the modified profiles $\hat{w}_j$ (cf. Figure 2) fit together upon summation. In particular, this guarantees by (2.1)

$$\lim_{x \to \infty} u(x,t) = \sum_{j=1}^{N} w_j^+ - \sum_{j=2}^{N} w_j^- = w_N^+. \tag{2.3}$$

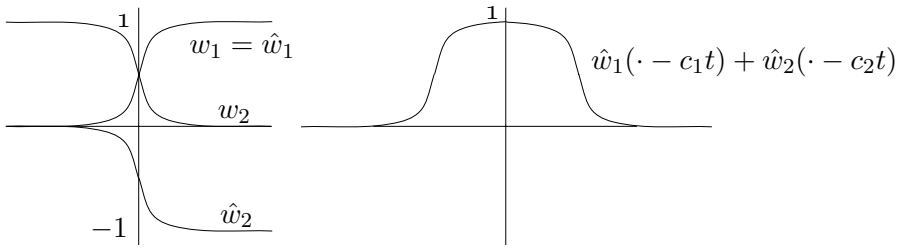


Figure 2. *The modified profiles $\hat{w}_j$, and asymptotic 2-front solution $\hat{w}_1(x - c_1 t) + \hat{w}_2(x - c_2 t)$ at $t > 0$, $c_1 < 0 < c_2$.*

In section 4 we will show that $U(x,t)$ defined by (2.2) satisfies (1.1) in an asymptotic sense, i.e.,

$$\|U_t - (AU_{xx} + f(U, U_x))\| \to 0 \quad \text{as } t \to \infty \tag{2.4}$$

for some suitable norm $\|\cdot\|$, e.g., the $\mathcal{L}_2$-norm.

Our goal is to set up a decomposition (1.4) that approaches the form (2.2) in an asympotic sense.

We use a bump function $\varphi \in C^\infty(\mathbb{R}, \mathbb{R})$ that satisfies

$$0 < \varphi(x) \le C \quad \forall x \in \mathbb{R} \tag{2.5}$$

and has its main mass located near zero. The precise form of φ is not important, but we mention that both numerical computation and the theory in section 4 work for exponential decay of type $\varphi(t) \asymp e^{-\beta|x|^k}$, $\beta > 0$, $k \ge 1$.

We look for a solution of the form (1.4) and insert this into (1.1). We suppress the arguments $(x - g_j(t), t)$ of v_j and find

$$\begin{aligned} u_t &= \sum_{j=1}^N [v_{j,t} - v_{j,\xi} g_{j,t}] \\ &= \sum_{j=1}^N A v_{j,\xi\xi} + f\left(\sum_{k=1}^N v_k, \sum_{k=1}^N v_{k,\xi}\right) \\ &= \sum_{j=1}^N \left[A v_{j,\xi\xi} + \frac{\varphi(\cdot - g_j(t))}{\sum_{k=1}^N \varphi(\cdot - g_k(t))} f\left(\sum_{k=1}^N v_k, \sum_{k=1}^N v_{k,\xi}\right)\right]. \end{aligned} \tag{2.6}$$

Note that the quotients

$$\frac{\varphi(x - g_j(t))}{\sum_{k=1}^N \varphi(x - g_k(t))} \tag{2.7}$$

form a time-dependent partition of unity and that the denominator never vanishes due to (2.5). In (2.6) we have used this partition to localize the nonlinear part of the vector field but not the solutions themselves.

A sufficient condition for (2.6) to hold is that each of the terms in brackets vanishes. Substituting $\xi = x - g_j(t)$ and $\mu_j = g_{j,t}$ leads to the following system of N coupled PDEs for $\xi \in \mathbb{R}$, $t \ge 0$,

$$\begin{aligned} v_{j,t}(\xi, t) = A v_{j,\xi\xi}(\xi, t) + v_{j,\xi}(\xi, t)\mu_j(t) + \frac{\varphi(\xi)}{\sum_{k=1}^N \varphi(\xi - g_k + g_j)} \\ \cdot f\left(\sum_{k=1}^N v_k(\xi - g_k + g_j, t), \sum_{k=1}^N v_{k,\xi}(\xi - g_k + g_j, t)\right), \quad j = 1, \dots, N, \end{aligned} \tag{2.8}$$

and the simple set of ODEs

$$g_{j,t} = \mu_j(t), \quad j = 1, \dots, N. \tag{2.9}$$

In the following it will be convenient to write $v = (v_1, \dots, v_N)$, $g = (g_1, \dots, g_N)$, and $\mu = (\mu_1, \dots, \mu_N)$ and to abbreviate terms in (2.8):

$$\begin{aligned} F_j(v, g)(\xi, t) &= Q_j^g(\xi, t) f\left(\sum_{k=1}^N v_k(\xi_{kj}^g, t), \sum_{k=1}^N v_{k,\xi}(\xi_{kj}^g, t)\right), \\ Q_j^g(\xi, t) &= \frac{\varphi(\xi)}{\sum_{k=1}^N \varphi(\xi_{kj}^g)}, \qquad \xi_{kj}^g = \xi - g_k(t) + g_j(t). \end{aligned} \tag{2.10}$$

Figure 3. *Decomposition of the initial data u_0.*

We note that the nonlinear terms $F_j(v, g)$ couple the single functions v_k, $k = 1, \dots, N$, in a nonlocal fashion. From the derivation we also see that one can allow j-dependent bump functions φ_j that take the size of the jth profile into account. The quotient in (2.7) then reads

$$\frac{\varphi_j(x - g_j(t))}{\sum_{k=1}^N \varphi_k(x - g_k(t))}. \tag{2.11}$$

The system will be completed by initial conditions for v_j, g_j and by phase conditions that compensate for the extra unknowns μ_j.

We impose initial conditions

$$v_j(\xi, 0) = v_j^0(\xi), \quad \xi \in \mathbb{R},\ j = 1, \dots, N, \tag{2.12}$$

$$g_j(0) = g_j^0, \quad j = 1, \dots, N, \tag{2.13}$$

that should satisfy

$$u_0(x) = \sum_{j=1}^N v_j^0(x - g_j^0), \quad x \in \mathbb{R}. \tag{2.14}$$

In most of our applications below we first select v_j^0, g_j^0 and then define u_0 by (2.14). If, on the other hand, u_0 is given, then one has to do some surgery to find appropriate values for v_j^0 and g_j^0. Assume, for example, that a function u_0 is given that forms plateaus near points $x_0 < x_1 < \dots, x_N$ (see Figure 3). Then one may choose $g_j^0 = \frac{1}{2}(x_{j-1} + x_j)$ for $j = 1, \dots, N$ and, similar to (2.2), define

$$v_j^0(\xi) = -u_0(x_{j-1}) + \begin{cases} u_0(x_{j-1}), & \xi + g_j^0 \le x_{j-1}, \\ u_0(\xi + g_j^0), & x_{j-1} \le \xi + g_j^0 \le x_j, \\ u_0(x_j), & x_j \le \xi + g_j^0, \end{cases} \qquad j = 2, \dots, N-1,$$

and

$$v_1^0(\xi) = \begin{cases} u_0(\xi + g_1^0), & \xi + g_1^0 \le x_1, \\ u_0(x_1), & x_1 \le \xi + g_1^0, \end{cases}$$

$$v_N^0(\xi) = \begin{cases} 0, & \xi + g_N^0 \le x_{N-1}, \\ u_0(\xi + g_N^0) - u_0(x_{N-1}), & x_{N-1} \le \xi + g_N^0. \end{cases}$$

Next we discuss the choice of phase condition that will make the solution of the system (2.8), (2.9), (2.13), (2.14) unique. For the case of freezing single waves, two possibilities were suggested in [4], [5].

First, suppose that we have template functions $\hat{v}_j$ (e.g., $\hat{v}_j = v_j^0$) to which we would like the v_j to stay as close as possible. This requires the distance $d_j(g) = \|v_j(\cdot,t) - \hat{v}_j(\cdot - g)\|_{\mathcal{L}^2}$ to achieve its minimum at $g = 0$ for all times. Differentiating with respect to g yields the necessary conditions

$$\langle v_j - \hat{v}_j, \hat{v}_{j,\xi} \rangle_{\mathcal{L}^2} = 0, \quad j = 1, \dots, N. \tag{2.15}$$

In the terminology of differential algebraic equations this constraint leads to an index 2 problem. If we differentiate (2.15) with respect to t and use (2.8), we have

$$\psi_{\mathrm{fix}}(v,\mu) = (\mu_j \langle \hat{v}_{j,\xi}, v_{j,\xi} \rangle_{\mathcal{L}^2} + \langle \hat{v}_{j,\xi}, A v_{j,\xi\xi} + F_j(v,g) \rangle_{\mathcal{L}^2})_{j=1}^N = 0. \tag{2.16}$$

If $\langle \hat{v}_{j,\xi}, v_{j,\xi} \rangle_{\mathcal{L}^2} \neq 0$, then we can determine μ_j from this equation and thus have reduced the problem to index 1.

Second, choose the values μ_j so that $v_{j,t}$ in (2.8) is minimized at each time instance. Geometrically this requires that the time derivative $v_{j,t}(\cdot,t)$ is orthogonal to the group orbit $\{v_j(\cdot - g, t) : g \in \mathbb{R}\}$ at all times. This leads to the phase condition

$$\psi_{\mathrm{orth}}(v,\mu) = (\langle v_{j,\xi}, v_{j,t} \rangle_{\mathcal{L}^2})_{j=1}^N = (\langle v_{j,\xi}, A v_{j,\xi\xi} + \mu_j v_{j,\xi} + F_j(v,g) \rangle_{\mathcal{L}^2})_{j=1}^N = 0, \tag{2.17}$$

which allows us to solve for μ_j whenever $v_{j,\xi}$ is nonconstant. Note that (2.17) can be obtained from (2.16) when replacing $\hat{v}_{j,\xi}$ by $v_{j,\xi}$. The complete system to be solved is now given by the PDAE (2.8), (2.9), (2.12), (2.13) with either (2.16) or (2.17) as phase condition.

The relative merits of both types of conditions have been discussed for the single freezing in [4], [20]. It was shown in [20], [19] that the fixed phase condition leads to a well-posed PDAE in the neighborhood of a relative equilibrium in one space dimension. Moreover, the PDAE as well as its discretization on a finite interval have the wave and its velocity as an asymptotically stable steady state in the classical Lyapunov sense. In [5] we have shown that this pertains on the continuous level to the orthogonality constraint (2.17). Locally near relative equilibria there is not much of a difference between both conditions. It is hard to make a general statement for more global situations, when the initial data are far from any equilibrium. Generally, the orthogonality condition is more flexible globally since it requires no preknowledge of the solution, whereas the fixed phase condition tends to lead to PDAEs with a better conditioning.

We conclude with some remarks concerning the numerical solution of the PDAE system. In section 3 we will discretize the PDAE as a whole by conventional methods. It is clear that the effort of solving the system grows linearly with the number of pulses or fronts present in the solution. On the other hand, in contrast to the original equation, one can solve the PDAE system on a fixed and relatively small spatial domain. So far, the interaction terms

that need values outside this domain were calculated by extrapolating with constant boundary values. One may think of reducing the spatial domain further by solving linearized equations (explicitly) in the outside domain and using this for calculating the interaction. We have not yet pursued the details of such an extension. A method of this type will be reminiscent of the vortex blob method in fluid dynamics (see [2], [8]), which follows moving vortices and then uses the Biot–Savart law for treating interactions.

3. Applications. We illustrate the method on two examples which possess traveling fronts and pulses—the Nagumo and the FitzHugh–Nagumo equations—which both model nerve conduction.

For two components the PDAE (2.8), (2.9) with the phase fixing condition (2.16) reads

$$
\begin{aligned}
v_{1,t} &= Av_{1,\xi\xi} + v_{1,\xi}\mu_1(t) + \frac{\varphi(\cdot)}{\varphi(\cdot)+\varphi(\cdot - g_2 + g_1)} f(v_1(\cdot,t) + v_2(\cdot - g_2 + g_1, t)),\\
v_{2,t} &= Av_{2,\xi\xi} + v_{2,\xi}\mu_2(t) + \frac{\varphi(\cdot)}{\varphi(\cdot)+\varphi(\cdot - g_1 + g_2)} f(v_2(\cdot,t) + v_1(\cdot - g_1 + g_2, t)),\\
0 &= \langle v_1(\cdot,t) - \hat{v}_1, \hat{v}_{1,\xi}\rangle, \qquad 0 = \langle v_2(\cdot,t) - \hat{v}_2, \hat{v}_{2,\xi}\rangle,\\
g_{1,t} &= \mu_1(t), \qquad g_{2,t} = \mu_2(t),
\end{aligned}
\tag{3.1}
$$

with initial conditions (2.12), (2.13) that will be specified below.

To solve (3.1) numerically we restrict to a finite interval $[-L, L]$ and impose Dirichlet or Neumann boundary conditions. Then we use the finite element package Comsol Multiphysics$^{\text{TM}}$ [6] with second order elements in space and a BDF method in time. As a bump function we first set $\varphi(x) = \exp(-x^2/\alpha)$ with suitable α. Later we see that the computations prove to be quite robust with respect to the choice of φ.

Whenever the nonlocal terms in the nonlinearity f have to be evaluated at arguments outside the computational domain, we extrapolate with the boundary values $v_j(\pm L)$. Inside $[-L, L]$ we use linear interpolation. As in the case of freezing single waves, we cannot expect the solutions $v_j(\cdot, t)$ to converge to $\hat{w}_j$ from (3.1) but rather to an approximation $\hat{w}_j^L$ that solves the stationary boundary value problem on $[-L, L]$.

3.1. Nagumo equation. A simple example is the scalar Nagumo equation [3]

$$
u_t = u_{xx} + u(1-u)(u-a), \quad u(x,t) \in \mathbb{R},\ x \in \mathbb{R},\ t > 0,
\tag{3.2}
$$

where $a \in (0, \frac{1}{2})$.

It has explicit traveling waves connecting the stationary points $w_1^- = w_2^+ = 0$ and $w_1^+ = w_2^- = 1$:

$$
\begin{aligned}
w_1(\xi) &= \frac{1}{1+\exp(\frac{-\xi}{\sqrt{2}})}, \quad c_1 = \sqrt{2}\,(a - \tfrac{1}{2}),\\
w_2(\xi) &= \frac{1}{1+\exp(\frac{\xi}{\sqrt{2}})}, \quad c_2 = -\sqrt{2}\,(a - \tfrac{1}{2}).
\end{aligned}
\tag{3.3}
$$

In addition, there exists a multitude of other solutions which can be computed explicitly [1].

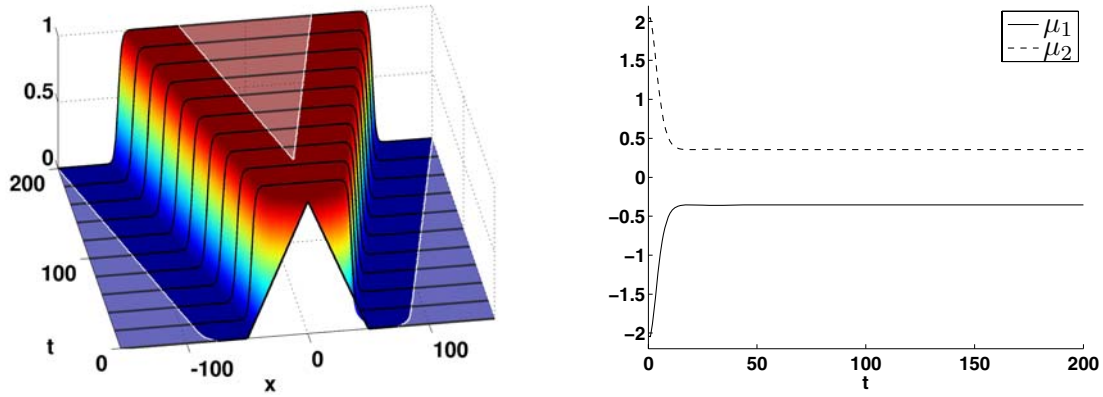


Figure 4. *Fronts moving in opposite directions in the Nagumo equation, and evolution of superposition u_c and velocities μ_1, μ_2.*

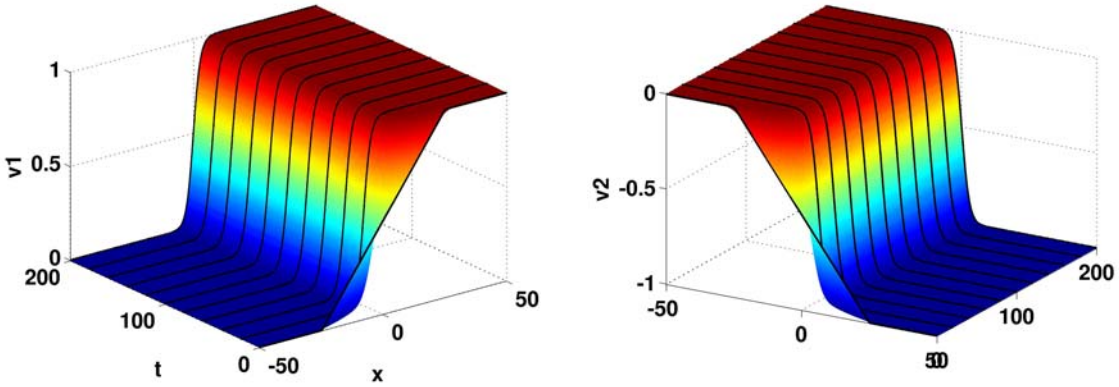


Figure 5. *Fronts moving in opposite directions in the Nagumo equation, and evolution of frozen fronts v_1, v_2.*

We use $a = \frac{1}{4}$, the spatial stepsize $\Delta x = 0.1$, and $\hat{v}_1 = v_1^0$, $\hat{v}_2 = v_2^0$ as template functions for the phase fixing condition. As bump function we take $\varphi(\xi) = \exp(x^2/\alpha)$, where the parameter $\alpha = 20$ is chosen such that the function is localized around the region of interest (see Figure 7). This setting will be used for all computations with the Nagumo equation, unless indicated otherwise.

In Figures 4 and 5 we show the result of a computation starting with initial data v_1^0, v_2^0, g_1^0, g_2^0 that add up to a hat function u^0 via the superposition (2.14). For the numerical solution on the finite interval $[-L, L]$, $L = 50$ we use Dirichlet boundary conditions

$$v_1(-L,t) = w_1^- = 0,\ v_2(-L,t) = 0,\ v_1(L,t) = w_1^+ = 1,\ v_2(L,t) = w_2^+ - w_2^- = -1.$$

Figure 4 displays the sum (1.4),

$$u_c(x,t) = v_1(x - \gamma_1(t), t) + v_2(x - \gamma_2(t), t), \tag{3.4}$$

while Figure 5 shows the components v_1 and v_2. The darker shaded regions indicate the two moving intervals $\gamma_j(t) + [-L, L]$, $j = 1, 2$, where u_c uses the computed values of v_1 or v_2, whereas the lighter shaded regions use exclusively the extrapolated boundary values of v_1 and v_2. After a short transient period the components v_1, v_2 and the velocities μ_1, μ_2 become stationary with opposite values resulting in a broadening plateau for u. The slopes of the plateau travel at speeds $\mu_1 = -\mu_2$ in opposite directions.

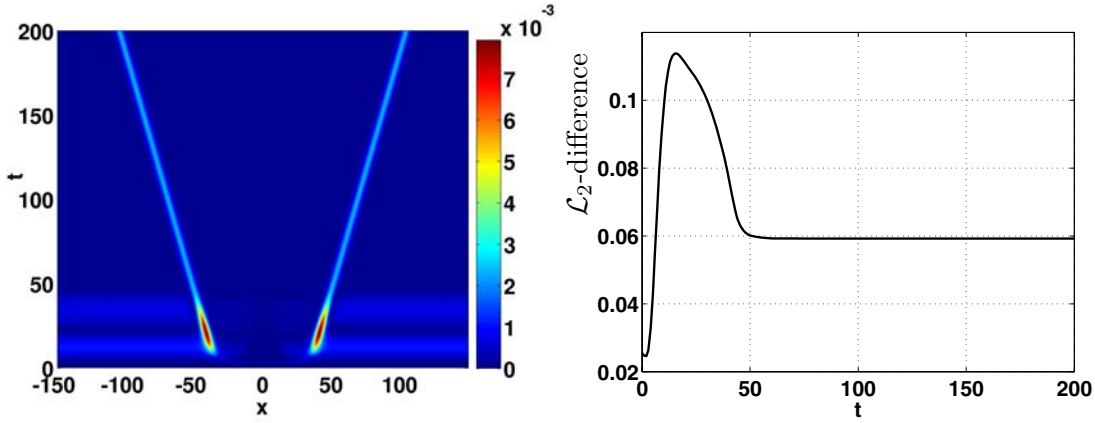


Figure 6. *Nagumo equation: Difference of traveling plateau and superposition of frozen fronts.*

For comparison, we show in Figure 6 the pointwise difference $|u_{\rm dns}(x,t) - u_c(x,t)|$ and the $\mathcal{L}_2$ difference $\|u_{\rm dns}(\cdot,t) - u_c(\cdot,t)\|_{\mathcal{L}_2}$ between the function u_c and a solution $u_{\rm dns}$ that is obtained by solving (3.2) directly on a sufficiently large interval.

There is very good agreement of the two solutions except in two thin layers close to the two fronts. Note that such errors cannot be corrected by a single phase shift, and, therefore, the $\mathcal{L}_2$- error becomes asymptotically constant; see Figure 6 (right) and (3.6) below. For the individual solutions v_j of (3.1) we expect in suitable norms

$$v_j(\cdot,t) - \hat{w}_j(\cdot - \tau_j) \to 0, \quad \mu_j(t) \to c_j \quad \text{as } t \to \infty,\ j = 1, 2. \tag{3.5}$$

Since we do not use the given profiles (3.3) for the phase condition, we can expect only convergence toward $\hat{w}_j(\cdot - \tau_j)$ for some suitable time shift τ_j. This shift is determined by the phase condition in (3.1), i.e., $\langle \hat{w}_j(\cdot - \tau_j) - \hat{v}_j, \hat{v}_{j,\xi} \rangle = 0$. From the last equation in (3.5) we then obtain

$$g_j(t) - (c_j t + g_j^0) \to \tau_j \quad \text{as } t \to \infty.$$

Therefore, our numerical calculation suggests that the exact solution $u_{\rm dns}$ of (3.2) satisfies in a suitable norm

$$u_{\rm dns}(\cdot,t) - \big(\hat{w}_1(\cdot - c_1 t - g_1^0 - \tau_1) + \hat{w}_2(\cdot - c_2 t - g_2^0 - \tau_2)\big) \to 0 \quad \text{as } t \to \infty \tag{3.6}$$

provided the difference of initial positions $g_2^0 - g_1^0$ is sufficiently large and the difference of initial values $u_{\rm dns}(\cdot,0) - \big(\hat{w}_1(\cdot - g_1^0) + \hat{w}_2(\cdot - g_2^0)\big)$ is sufficiently small. A proof of such a result is a work in progress.

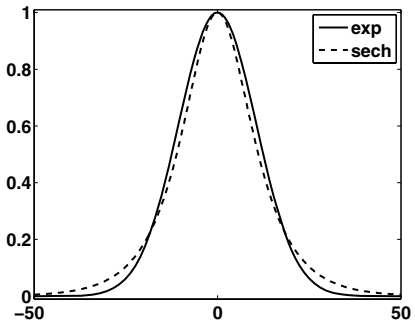


Figure 7. *Comparison of functions* $\varphi(x) = \exp(-x^2/\alpha)$ *and* $\varphi(x) = \operatorname{sech}(-x/\beta)$ *with equal integral.*

In Figure 8 we show the behavior of time derivatives $\|u_t\|$ and $\|\mu_t\|$ (left) including a comparison of $\|u_t\|$ for different bump functions (right). The second function is $\varphi(\xi) = \operatorname{sech}(x/\beta)$, where β is chosen such that the integrals of both functions over $\mathbb{R}$ coincide; see Figure 7. For a certain time interval the rate of decay is the same for both bump functions. From the numerical data one finds the slope 0.25 which coincides with the spectral gap between zero and the smallest negative eigenvalue of the linearization about the single traveling waves w_1 and w_2. This relation was verified in [20], [19] for the case of freezing a single wave. For larger times t the decay is better for the Gaussian $\varphi(x) = \exp(-x^2/\alpha)$ than for the sech-function. The effect vanishes on larger computational domains where both functions are sufficiently localized.

In the second numerical experiment we consider a situation that, in a sense, is opposite to the first case; see Figures 9 and 10. We start with a downward hat function and obtain two fronts traveling toward each other with opposite speeds. Eventually they annihilate each other resulting in a value of zero for the speeds μ_1, μ_2. In Figure 10 one can observe slight disturbances in v_1, v_2 during the strong interaction at collision. Note that after the collision the superposition (1.4) yields a constant, although the components v_1, v_2 cannot become constant due to Dirichlet boundary conditions. Rather, the asymptotic state of our system is formed by two ramp functions that are at rest and add up to a constant; see Figure 10. As before we observe a phase shift difference when comparing this with the solution of (3.2) on a large interval; see

Figure 11. But in this case the difference converges to zero as the constant solution is approached. Similar results are obtained for Neumann boundary conditions.

In Figures 12 and 13 we consider a case where a two-front turns into a single front. This is a case where the number N of components in our ansatz is larger than the number of components which constitute the final solution. This does not create any problems for our method. We use boundary conditions that do not require any a priori knowledge of the limiting stationary points for the components v_1 and v_2. Instead of prescribing $v_1^{\pm}$ and $v_2^{\pm}$ directly, we require $v_1^- + v_2^- = w^-$ and $v_1^+ + v_2^+ = w^+$ and impose Neumann boundary conditions at the remaining ends. If one insists on Dirichlet boundary conditions for every single wave, then additional boundary layers will develop.

In the current example the wave behind has a larger speed and merges with the first wave.

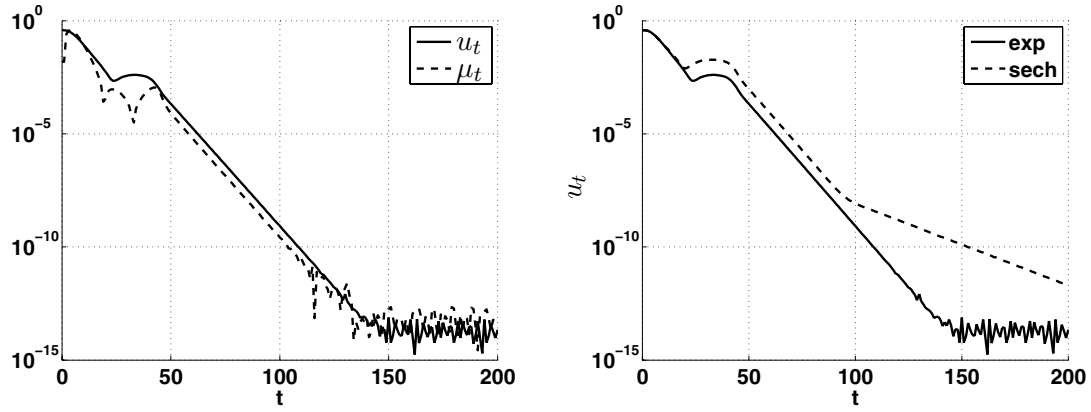


Figure 8. *Nagumo equation:* $\|u_t\|$ *and* $\|\mu_t\|$ *(on a logarithmic scale) versus time and* $\|u_t\|$ *(on a logarithmic scale) for functions* $\varphi(x) = \exp(-x^2/\alpha)$, $\alpha = 20$ *and* $\varphi(x) = \operatorname{sech}(-x/\beta)$, $\beta = 8.5$.

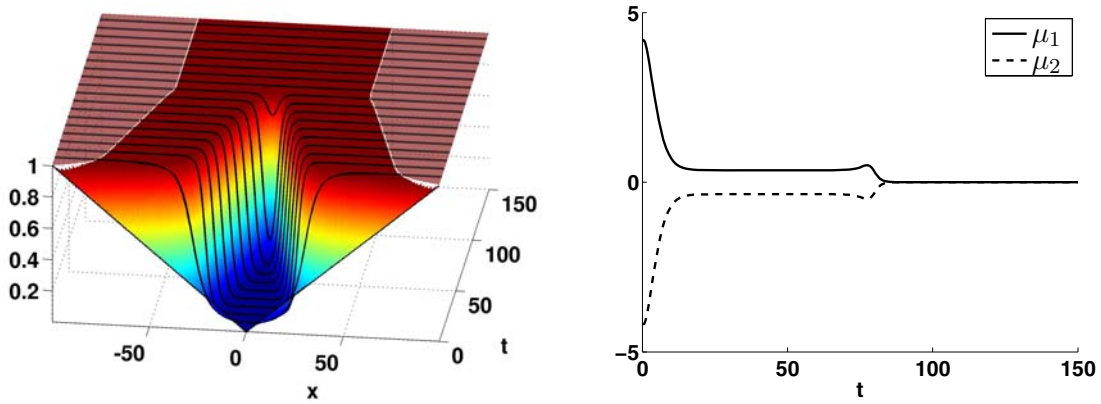


Figure 9. *Collision in the Nagumo equation, and evolution of superposition* u_c *and velocities* μ_1, μ_2.

This is correctly reproduced by our method. The speeds μ_1, μ_2 of both components converge to the same value and the superposition of the profiles v_1, v_2 forms a single front; see Figure 12.

In Figure 14 we show the difference between superposition and solution of the original equation on a large interval. After the strong interaction the rates of decay for $||u_t||_{\mathcal{L}_2}$ and $\|\mu_t\|$ (not shown) turn out to be quite similar to the previous case in Figure 8.

3.2. FitzHugh–Nagumo system.

The two component FitzHugh–Nagumo system (FHN)

$$\begin{aligned} V_t &= V_{xx} + V - \frac{1}{3}V^3 - R, \\ R_t &= \epsilon(V + a - bR) \end{aligned} \tag{3.7}$$

models nerve conduction and possesses different types of traveling wave solutions such as fronts, multifronts [11], and pulses [9], [10]. We consider the same parameters $a = 0.7$, $b = 0.8$, $\epsilon = 0.08$ as in [4] for which traveling pulses exist. Because of reflectional symmetry, with each

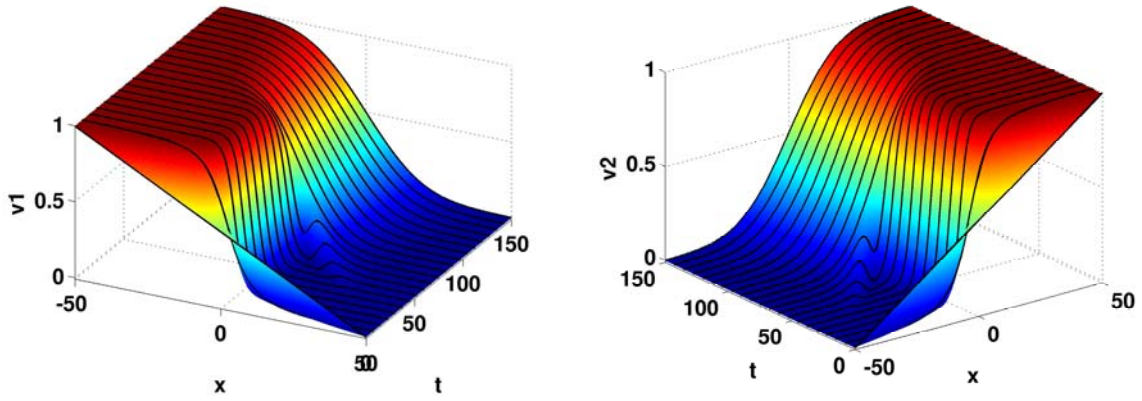


Figure 10. *Collision in the Nagumo equation, and evolution of frozen fronts* v_1, v_2.

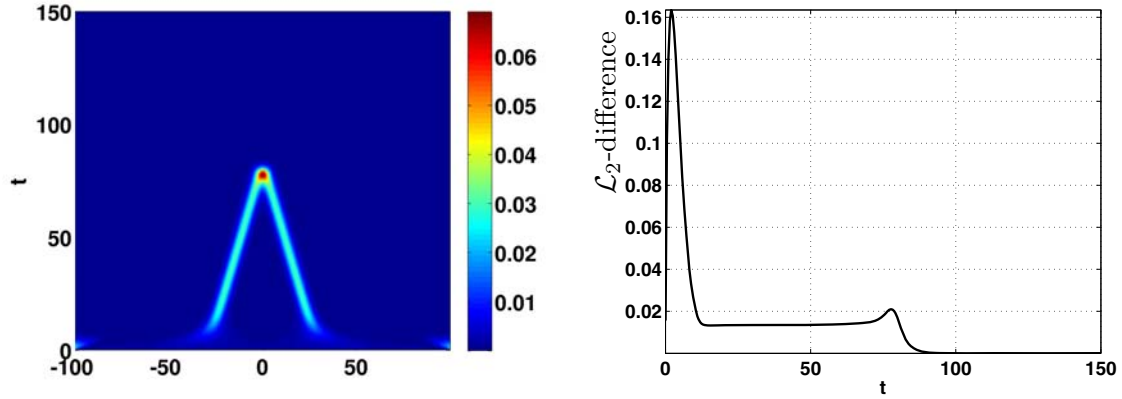


Figure 11. *Collision in the Nagumo equation, and evolution of difference* $|u_{\rm dns}(x,t) - u_{\rm c}(x,t)|$ *and of* $\mathcal{L}_2$ *difference* $\|u_{\rm dns}(\cdot,t) - u_c(\cdot,t)\|_{\mathcal{L}_2}$.

solution its mirror image is also a solution traveling at opposite speed. It is important to note that, due to the lack of diffusion in the second equation of (3.7), the PDAE system (3.1) is of mixed hyperbolic-parabolic type. In the two R-equations the convective terms that allow freezing form the principal part, and this requires some cautionary measure for the numerical solution. With the finite element code Comsol Multiphysics$^{\rm TM}$ we used streamline diffusion ($\delta_{sd} = 0.25$) in order to treat the hyperbolic part correctly. The system is solved on $[-70, 70]$ with Dirichlet boundary conditions, stepsize $\Delta x = 0.1$, and $\varphi(x) = \operatorname{sech}(x/\beta)$, $\beta = 11.3$ (results for a Gaussian φ are similar).

First we consider the formation of two pulses out of a single initial pulse. This situation was already studied in [4] with the single freezing method. There only one of the two forming pulses could be frozen depending on the choice of phase condition. Figure 15 displays the behavior of the superposition (1.4) as obtained by our method, and Figure 17 shows the difference to a solution of (3.7) obtained directly on a large interval. While the difference is small in most of the space time domain, the $\mathcal{L}_2$-difference appears to grow linearly with

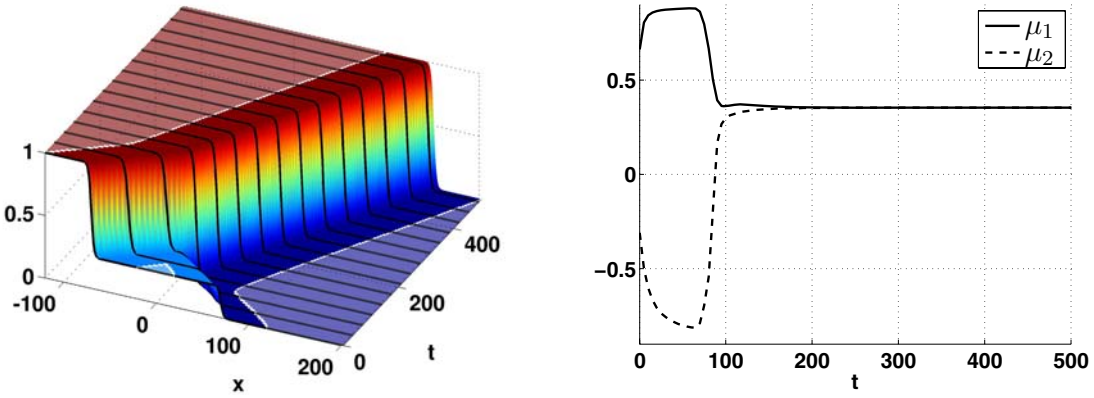


Figure 12. *Merging fronts in the Nagumo equation, and evolution of superposition u_c and velocities μ_1, μ_2.*

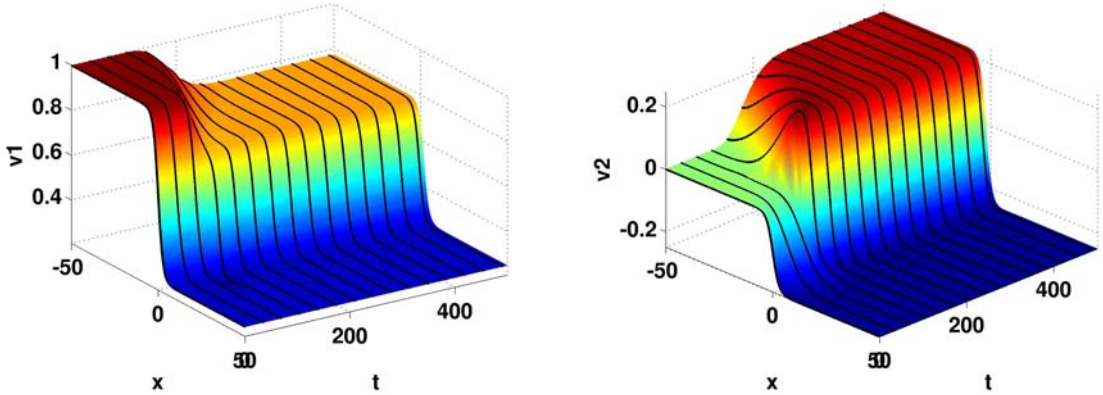


Figure 13. *Merging fronts in the Nagumo equation, and evolution of frozen fronts v_1, v_2.*

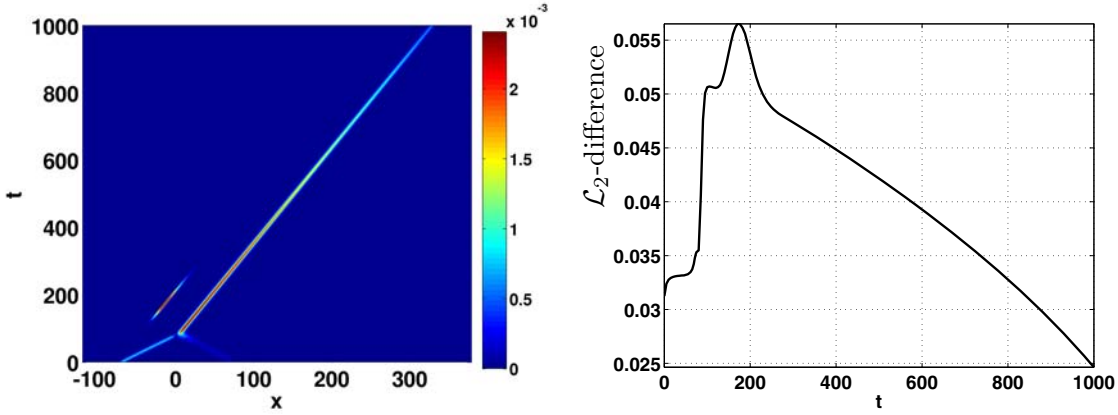


Figure 14. *Nagumo, and evolution of difference $\|u_{\mathrm{dns}}(\cdot,t) - u_c(\cdot,t)\|_{\mathcal{L}_2}$.*

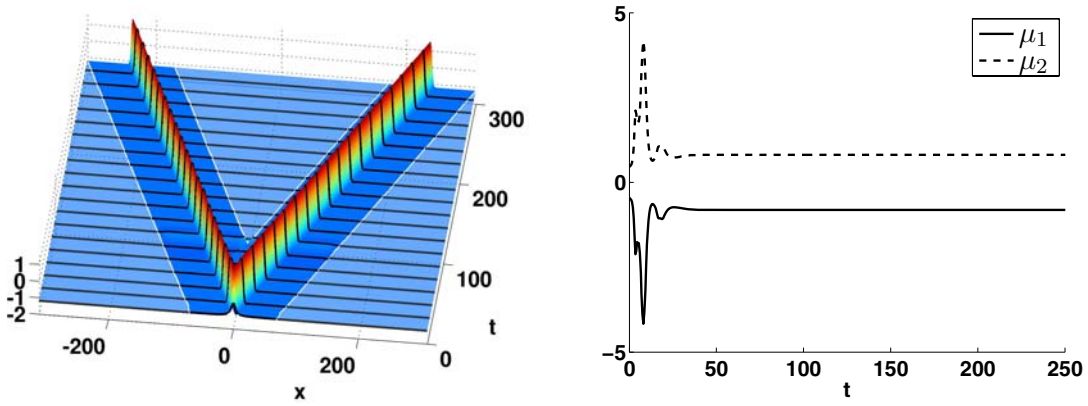


Figure 15. *Splitting of a pulse in the FHN, and evolution of superposition $V = u_c$ and of velocities μ_1, μ_2.*

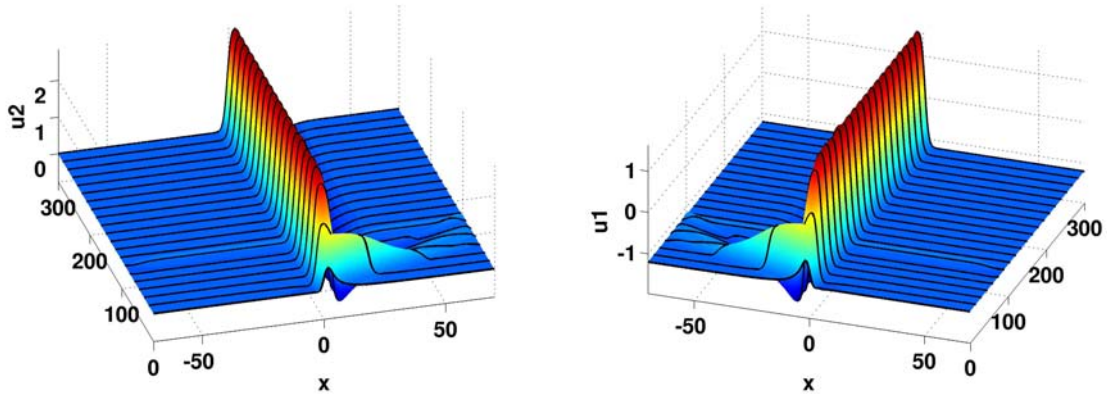


Figure 16. *Splitting of a pulse in the FHN, and evolution of pulses V_1 traveling to the left and V_2 traveling to the right.*

time. This is in contrast to the parabolic case, and we suspect that the effect is due to the hyperbolic part of the frozen equation.

In Figure 16 one observes that the components V_1 and V_2 develop tiny secondary waves that travel toward the boundary. The superposition is still correct until these reach the boundary. When they arrive a slight disturbance of the superposition in the middle of the interval develops (at about $t = 90$; see Figure 17). We expect that these boundary effects can be further reduced by using, e.g., transparent boundary conditions. After the separation phase both pulses quickly reach their asymptotic states (see Figure 16), and $||u_t||$ decays exponentially, as in the Nagumo case; see Figure 18.

Now we take the two traveling pulses which have been computed in this way and interchange their initial positions g_j^0. Then the two pulses start to move toward each other and eventually annihilate, as shown in Figure 19. The difference of the superposition and the solution of the original equation on a large domain behaves in a fashion quite similar to the Nagumo case. At the collision both wave speeds converge to zero. Figure 20 shows that both

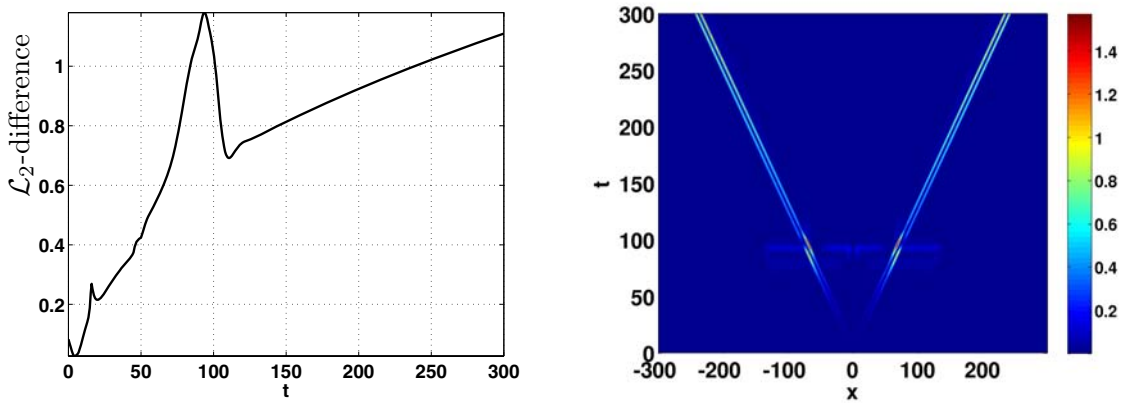


Figure 17. *Splitting of a pulse in the FHN, difference of the two-pulse computed on a large domain, and superposition of frozen single pulses* V_1, V_2.

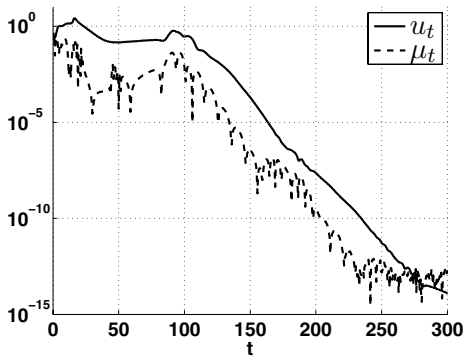


Figure 18. *Splitting of pulse in the FHN and behavior of* $\|u_t\|$, $u = (V_1, R_1, V_2, R_2)$, *and* $\|\mu_t\|$.

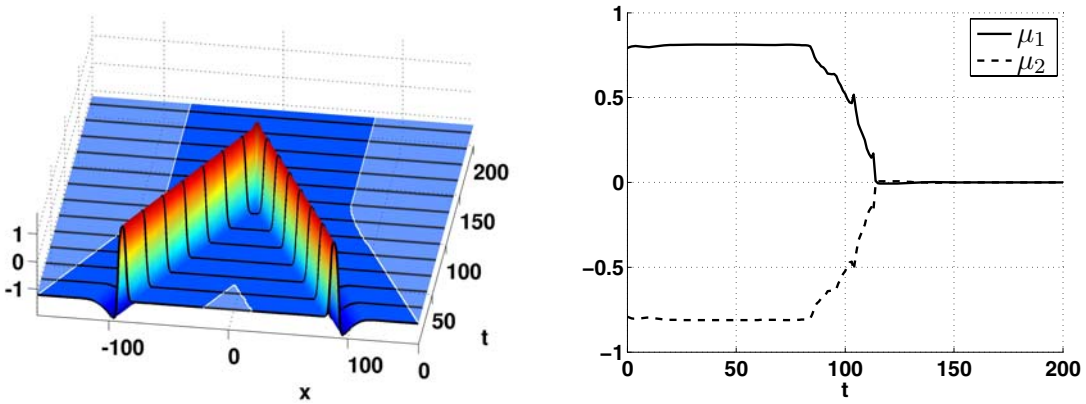


Figure 19. *Collision of pulses in the FHN, and evolution of superposition* $u_c = V$ *and velocities* μ_1, μ_2.

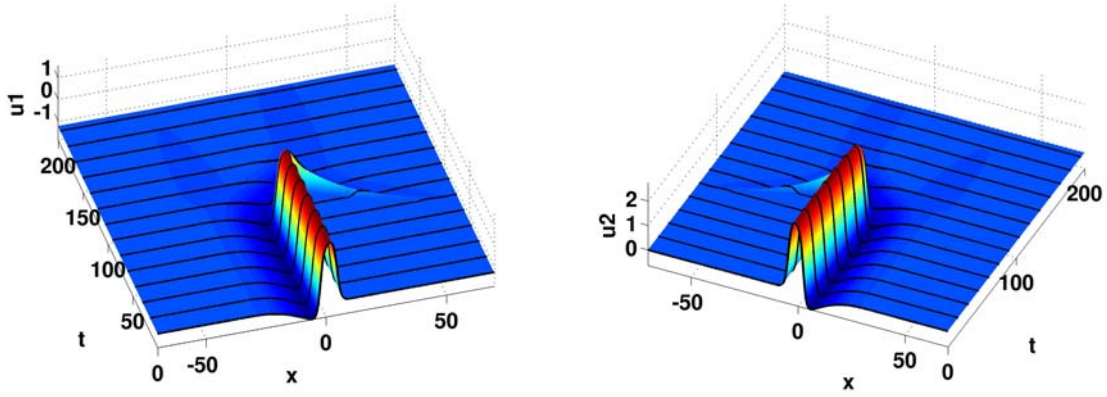


Figure 20. *Collision of pulses in the FHN, and evolution of frozen pulses* V_1, V_2.

components V_1 and V_2 converge to constants.

Contrary to the case of fronts, this can lead to numerical difficulties for large times since our phase conditions are no longer well-posed for constant functions (cf. (2.16), (2.17)). We note that after collision we have again the situation where we use a larger number N than necessary for representing the solution and where our method reproduces the behavior in a consistent manner.

4. Asymptotic properties of multifronts. In this section we show that traveling waves, when shifted as in (2.2), satisfy the PDAE system (2.8) in an asymptotic sense as $t \to \infty$. This will imply a corresponding property of the superposition (2.2) for the original system (1.1).

Definition 4.1. *A smooth function* $V : \mathbb{R}^N \to \mathbb{R}^m$ *is called an* asymptotic N-front solution *of* (1.1) *if there exist constants*

$$c_1, c_2, \ldots, c_N$$

such that

$$u(x,t) = V(x - c_1 t, \ldots, x - c_N t) \tag{4.1}$$

satisfies

$$||(u_t - Au_{xx} - f(u, u_x))(\cdot, t)||_{\mathcal{L}_2} \to 0 \quad \text{as } t \to \infty. \tag{4.2}$$

We look for asymptotic N-front solutions of the type

$$V(x_1, \ldots, x_N) = \sum_{j=1}^{N} \hat{w}_j(x_j), \tag{4.3}$$

where $\hat{w}_j$ is defined in (2.2) and $w_j(x - c_j t)$ are C^2-smooth traveling waves of (1.1). In particular, they satisfy the stationary equation

$$0 = Aw_{j,\xi\xi} + c_j w_{j,\xi} + f(w_j, w_{j,\xi}). \tag{4.4}$$

With (2.10) let us write (2.8) as

$$v_{j,t} = M_j(v,g) = Av_{j,\xi\xi} + v_{j,\xi}g_{j,t} + F_j(v,g), \quad j = 1,\dots,N. \tag{4.5}$$

For the special functions

$$v_j = \hat{w}_j,\ g_j(t) = c_j t,\ j = 1,\dots,N,\ t \in \mathbb{R}_+, \tag{4.6}$$

we show in Theorem 4.2 below that

$$||M_j(\hat{w},g)(\cdot,t)||_{\mathcal{L}_2} \to 0 \quad \text{as } t \to \infty; \tag{4.7}$$

i.e., they are asymptotic steady states of the nonautonomous system (4.5). Using the basic calculation (2.6) we obtain that

$$u(x,t) = \sum_{j=1}^{N} \hat{w}_j(x - c_j t) \tag{4.8}$$

satisfies the estimate

$$\begin{aligned} &||(u_t - Au_{xx} - f(u,u_x))(\cdot,t)||_{\mathcal{L}_2} \\ &\quad = \left|\left| \sum_{j=1}^{N} M_j(\hat{w},g)(\cdot - c_j t, t) \right|\right|_{\mathcal{L}_2} \to 0 \quad \text{as } t \to \infty. \end{aligned}$$

Therefore, the function (4.3) is an asymptotic N-front solution.

Theorem 4.2. *Let $w_j(x - c_j t)$, $j = 1,\dots,N$, be C^2-smooth traveling wave solutions of the system (1.1) that satisfy for some constants $C, \alpha > 0$*

$$c_1 < c_2 < \cdots < c_N, \tag{4.9}$$

$$\|w_j(\xi) - w_j^{\pm}\| \le Ce^{\mp\alpha\xi} \quad \textit{and} \quad \|w_{j,\xi}(\xi)\| \le Ce^{-\alpha|\xi|}, \quad j = 1,\dots,N, \tag{4.10}$$

$$w_j^{+} = w_{j+1}^{-}, \quad j = 1,\dots,N-1. \tag{4.11}$$

Moreover, let $\varphi \in C^{\infty}(\mathbb{R},\mathbb{R})$ be a function for which the exponential estimate

$$C_0 e^{-\beta_0|x|} \le \varphi(x) \le C_1 e^{-\beta_1|x|}, \quad x \in \mathbb{R}, \tag{4.12}$$

holds for some positive constants $C_0 \le C_1$ and $\beta_1 < \beta_0$.

Then the shifted waves $\hat{w}_j$ from (2.2) and $g_j(t) = c_j t$ satisfy for some constants $C, \varepsilon > 0$

$$||M_j(\hat{w},g)(\cdot,t)||_{\mathcal{L}_2} \le Ce^{-\varepsilon t} \quad \forall t \ge 0, \tag{4.13}$$

where M_j denotes the right-hand side in the PDAE system (4.5). In particular, $V(x_1,\dots,x_N) = \sum_{j=1}^{N} \hat{w}_j(x_j)$ is an asymptotic N-front solution of (1.1).

Remark 4.3. *Clearly, this result does not yet prove the behavior of the numerical solutions observed in section 3. For such a result we must show that the PDAE system (4.5) is well-posed and, moreover, that for stable traveling waves the solutions $v_j(\cdot,t)$ converge to $\hat{w}_j$ in a*

suitable way as $t \to \infty$ for sufficiently small initial perturbations. We think of Theorem 4.2 as a first step toward such a result that justifies the overall ansatz in section 2.

Remark 4.4. *The theorem remains valid for more general bump functions that satisfy*

$$C_0 e^{-\beta_0 |x|^k} \le \varphi(x) \le C_1 e^{-\beta_1 |x|^k}, \quad x \in \mathbb{R}, \quad \text{with } 0 < C_0 \le C_1,\ 0 < \beta_1 < \beta_0,\ k \ge 1.$$

The case $k = 2$ was used in some of our simulations above. The following proof will show that the only modifications occur in the estimates on the intervals marked by Q_j^g in Figure 21 and in the condition (4.17) that determines the subdivision of the real line.

Proof. Let us first note that (4.10) and (4.4) imply $\lim_{\xi \to \pm\infty} w_{j,\xi\xi}(\xi) = 0$ and hence

$$f(w_j^{\pm}, 0) = 0. \tag{4.14}$$

As noted above it suffices to prove (4.13). For ease of reading we restrict our attention to the case where f depends on u only, $f(u, u_x) = f(u)$. Using (4.14), (4.4), and (4.10) the details of the general case can be filled in easily. In the following we use C to denote a generic constant. First, the stationary equation (4.4) yields

$$\begin{aligned} ||M_j(\hat{w}, g)(\cdot, t)||_{\mathcal{L}_2}^2 &= \left|\left| Q_j^g(\cdot, t) f\left(\sum_{k=1}^N \hat{w}_k(\xi_{k,j}^g) \right) - f(w_j) \right|\right|_{\mathcal{L}_2}^2 \\ &\le C \left[\int_{\mathbb{R}} Q_j^g(\xi, t)^2 \left| f\left(\sum_{k=1}^N \hat{w}_k(\xi_{k,j}^g) \right) - f(w_j(\xi)) \right|^2 d\xi \right. \\ &\quad \left. + \int_{\mathbb{R}} (1 - Q_j^g(\xi, t))^2 |f(w_j(\xi))|^2 d\xi \right] =: I_1 + I_2. \end{aligned}$$

We estimate the integrals I_1, I_2 separately. When comparing f arguments we frequently use the following equality:

$$\sum_{k=1}^N \hat{w}_k(\xi_{kj}^g) - w_j(\xi) = \sum_{k=1}^{j-1} (w_k(\xi_{kj}^g) - w_k^+) + \sum_{k=j+1}^N (w_k(\xi_{kj}^g) - w_k^-). \tag{4.15}$$

Consider first I_1 and indices $2 \le j \le N-1$. For $q > 0$ sufficiently small we define

$$\gamma_k^{\pm} = (1 \pm q)(c_k - c_j)t, \qquad \gamma_k^0 = (c_k - c_j)t$$

and partition $\mathbb{R}$ into subintervals as follows (cf. Figure 21):

$$\begin{aligned} -\infty < \gamma_1^+ < \gamma_1^0 < \gamma_1^- < \gamma_2^+ < \cdots < \gamma_{j-1}^+ < \gamma_{j-1}^0 < \gamma_{j-1}^- < 0 = \gamma_0^{\pm} \\ < \gamma_{j+1}^- < \gamma_{j+1}^0 < \gamma_{j+1}^+ < \cdots < \gamma_N^- < \gamma_N^0 < \gamma_N^+ < \infty. \end{aligned} \tag{4.16}$$

Note that the relations $\gamma_k^- < \gamma_{k+1}^+$ for $k \le j-1$ and $\gamma_k^+ < \gamma_{k+1}^-$ for $k \ge j+1$ follow if q satisfies

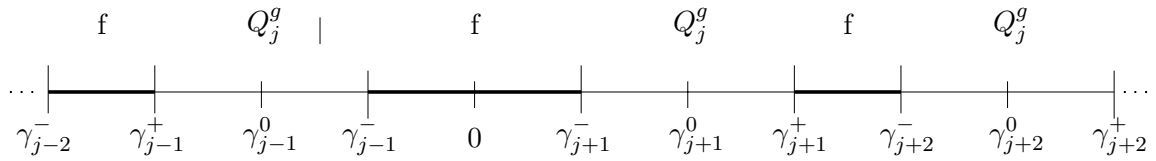


Figure 21. *Decomposition of the interval* $\mathbb{R}$.

$$q < \min\left\{ \frac{c_{k+1}-c_k}{|2c_j - c_k - c_{k+1}|} : \ 1 \le j \le N, \ 1 \le k \le N-1 \right\}.$$

Our second condition on q is

$$q < \frac{\min\big(\min\limits_{j=2,\dots,N}(c_j - c_{j-1}), 1\big)\,\beta_1}{\max\big(\max\limits_{j=2,\dots,N}(c_j - c_{j-1}), 1\big)(\beta_1+\beta_0)}. \tag{4.17}$$

For the estimate of I_1 we use $0 \le Q_j^g(\cdot,\cdot) \le 1$ and the fact that all arguments of f lie in a compact interval. On each subinterval we use the smallness of either f or Q_j^g, as indicated in Figure 21. We obtain

$$\begin{aligned}
I_1 \le C \Bigg[& \int_{-\infty}^{\gamma_1^+} \left| f\left(\sum_{k=1}^N \hat{w}_k(\xi_{k,j}^g)\right) - f(w_j(\xi)) \right|^2 d\xi + \sum_{l=1}^{j-1} \int_{\gamma_l^+}^{\gamma_l^0} Q_j^g(\xi,t)^2 d\xi \\
& + \sum_{l=1}^{j-1} \int_{\gamma_l^0}^{\gamma_l^-} Q_j^g(\xi,t)^2 d\xi + \sum_{l=1}^{j-2} \int_{\gamma_l^-}^{\gamma_{l+1}^+} \left| f\left(\sum_{k=1}^N \hat{w}_k(\xi_{k,j}^g)\right) - f(w_j(\xi)) \right|^2 d\xi \\
& + \int_{\gamma_{j-1}^-}^{\gamma_{j+1}^-} \left| f\left(\sum_{k=1}^N \hat{w}_k(\xi_{k,j}^g)\right) - f(w_j(\xi)) \right|^2 d\xi + \sum_{l=j+1}^{N} \int_{\gamma_l^-}^{\gamma_l^0} Q_j^g(\xi,t)^2 d\xi \\
& + \sum_{l=j+1}^{N} \int_{\gamma_l^0}^{\gamma_l^+} Q_j^g(\xi,t)^2 d\xi + \sum_{l=j+1}^{N-1} \int_{\gamma_l^+}^{\gamma_{l+1}^-} \left| f\left(\sum_{k=1}^N \hat{w}_k(\xi_{k,j}^g)\right) - f(w_j(\xi)) \right|^2 d\xi \\
& + \int_{\gamma_N^+}^{\infty} \left| f\left(\sum_{k=1}^N \hat{w}_k(\xi_{k,j}^g)\right) - f(w_j(\xi)) \right|^2 d\xi \Bigg] \\
=: & \ I_1^b + \sum_{l=1}^{j-1} I_{1,l}^{1-} + \sum_{l=1}^{j-1} I_{1,l}^{2-} + \sum_{l=1}^{j-2} I_{1,l}^{3-} + I_1^c \\
& + \sum_{l=j+1}^{N} I_{1,l}^{1+} + \sum_{l=j+1}^{N} I_{1,l}^{2+} + \sum_{l=j+1}^{N-1} I_{1,l}^{3+} + I_1^e.
\end{aligned}$$

For the convenience of the reader here we give only the estimate for the crucial central term I_1^c and defer the remaining laborious estimates to the appendix. With (4.9), (4.10), (4.11),

and (4.15) we obtain

$$
\begin{aligned}
I_1^c &\le C \int_{\gamma_{j-1}^-}^{\gamma_{j+1}^-} \left| \sum_{k=1}^{N} \hat{w}_k(\xi + (c_j - c_k)t) - w_j(\xi) \right|^2 d\xi \\
&\le C \int_{\gamma_{j-1}^-}^{\gamma_{j+1}^-} \left(\sum_{k=1}^{j-1} |w_k(\xi + (c_j - c_k)t) - w_k^+|^2 + \sum_{k=j+1}^{N} |w_k(\xi + (c_j - c_k)t) - w_k^-|^2 \right) d\xi \\
&\le C \left[\int_{\gamma_{j-1}^-}^{\gamma_{j+1}^-} \sum_{k=1}^{j-1} e^{-2\alpha(\xi + (c_j - c_k)t)} d\xi + \int_{\gamma_{j-1}^-}^{\gamma_{j+1}^-} \sum_{k=j+1}^{N} e^{2\alpha(\xi + (c_j - c_k)t)} d\xi \right] \\
&\le C \left[\int_{\gamma_{j-1}^-}^{\gamma_{j+1}^-} e^{-2\alpha(\xi + (c_j - c_{j-1})t)} d\xi + \int_{\gamma_{j-1}^-}^{\gamma_{j+1}^-} e^{2\alpha(\xi + (c_j - c_{j+1})t)} d\xi \right] \\
&\le C \left[e^{-2\alpha q(c_j - c_{j-1})t} + e^{-2\alpha q(c_{j+1} - c_j)t} \right]. \qquad \blacksquare
\end{aligned}
$$

5. Generalization to equivariant evolution equations. In this section we generalize the idea from section 2 to evolution equations in Banach spaces that are equivariant with respect to the action of a Lie group. The abstract setting follows the approach from [4], [5].

5.1. The abstract formulation. Consider an evolution equation

$$u_t = Au + F(u), \quad u(0) = u_0, \tag{5.1}$$

where $A, F : Y \subset X \to X$ are linear, respectively, nonlinear, operators from a dense subspace Y of some Banach space X into X. We assume equivariance of both A and F with respect to some action of the Lie group G on X,

$$a : G \to GL(X), \ g \mapsto a(g);$$

that is,

$$F(a(g)u) = a(g)F(u), \quad A(a(g)u) = a(g)Au \tag{5.2}$$

holds for all $u \in Y$, $g \in G$. In order to mimic the partition of unity construction we assume that there is a module E (i.e., a real vector space with an Abelian multiplication) acting on X via

$$\bullet : E \times X \to X, \ (\varphi, u) \mapsto \varphi \cdot u,$$

such that both distributive laws and the associative law hold.

Moreover, we assume that the group also acts on E via a possibly different action

$$\alpha : G \to GL(E), \ g \mapsto \alpha(g),$$

such that for all $g \in G$, $\varphi, \psi \in E$, $u \in X$

$$\alpha(g)(\varphi\psi) = (\alpha(g)\varphi)(\alpha(g)\psi), \tag{5.3}$$

$$a(g)(\varphi \cdot u) = (\alpha(g)\varphi) \cdot (a(g)u). \tag{5.4}$$

Furthermore, we assume that the map

$$a(\cdot)u : G \to X, \quad g \mapsto a(g)u$$

is continuous for any $u \in X$ and that it is continuously differentiable for any $u \in Y$ with derivative denoted by

$$d[a(g)u] : T_gG \to X, \quad \lambda \mapsto d[a(g)u]\lambda.$$

Example 5.1. *Consider as an example $X = \mathcal{L}^2(\mathbb{R}, \mathbb{C})$, $G = S^1 \times \mathbb{R}$ with the action given by*

$$[a(\theta, \tau)u](x) = e^{i\theta}u(x - \tau), \quad u \in X, \ (\theta, \tau) \in S^1 \times \mathbb{R}. \tag{5.5}$$

Then with $E = C^0_{unif}(\mathbb{R}, \mathbb{R})$ we find that (5.3), (5.4) hold for the setting $\alpha(\theta, \tau)\varphi(x) = \varphi(x-\tau)$ for $\varphi \in E$. Moreover, with this choice the action is continuous on E. We note, however, that this property will not be needed for the arguments to follow.

In the following we assume that we are given some $\varphi \in E$ such that $\sum_{j=1}^N \alpha(g_j)\varphi$ is invertible for any choice of $g_j \in G$. In section 2 and in Example 5.1 above this property is a consequence of (2.5). For the inverse element of some $\varphi \in E$ we use the notation $\frac{1}{\varphi} = \varphi^{-1}$.

The generalization of (1.4) is to write the solution u as

$$u(t) = \sum_{j=1}^N a(g_j(t))v_j(t), \tag{5.6}$$

with unknowns $g_j(t) \in G$, $v_j \in Y$. Inserting this into (5.1) and using equivariance (5.2) as well as (5.3), (5.4) leads to

$$\begin{aligned}
u_t &= \sum_{j=1}^N \left(a(g_j)v_{j,t} + d[a(g_j)v_j]g_{j,t}\right) \\
&= \sum_{j=1}^N A(a(g_j)v_j) + F\left(\sum_{k=1}^N a(g_k)v_k\right) \\
&= \sum_{j=1}^N a(g_j)Av_j + \sum_{j=1}^N \frac{\alpha(g_j)\varphi}{\sum_{k=1}^N \alpha(g_k)\varphi} \cdot F\left(\sum_{k=1}^N a(g_k)v_k\right) \\
&= \sum_{j=1}^N a(g_j)\left[Av_j + \frac{\varphi}{\sum_{k=1}^N \alpha(g_j^{-1}g_k)\varphi} \cdot F\left(\sum_{k=1}^N a(g_j^{-1}g_k)v_k\right)\right].
\end{aligned}$$

This equation is fulfilled if the v_j, g_j satisfy the system

$$v_{j,t} = Av_j + \frac{\varphi}{\sum_{k=1}^N \alpha(g_j^{-1}g_k)\varphi} \cdot F\left(\sum_{k=1}^N a(g_j^{-1}g_k)v_k\right) - a(g_j^{-1})d[a(g_j)v_j]g_{j,t}. \tag{5.7}$$

We simplify the last term as in [5]. Let $\mathbb{1}$ be the unit element in G; then the tangent space $T_{\mathbb{1}}G$ is the Lie algebra associated with G. By $dg(\mathbb{1}) : T_{\mathbb{1}}G \to T_gG$ we denote the derivative of

the multiplication from the left by g at $\mathbb{1}$. Differentiating the relation $a(g\circ\gamma)v = a(g)(a(\gamma)v)$ for $v \in Y$ at $\gamma = \mathbb{1}$ yields

$$a(g)d[a(\mathbb{1})v]\mu = d[a(g)v](dg(\mathbb{1})\mu) \quad \text{for } \mu \in T_{\mathbb{1}}G,\ v \in Y.$$

Therefore, defining new coordinates $\mu_j(t) \in T_{\mathbb{1}}G$ by $g_{j,t}(t) = dg_j(\mathbb{1})\mu_j(t)$ turns (5.7) into

$$\begin{aligned} v_{j,t} &= Av_j + \frac{\varphi}{\sum_{k=1}^N a(g_j^{-1}g_k)\varphi}\cdot F\left(\sum_{k=1}^N a(g_j^{-1}g_k)v_k\right) - d[a(\mathbb{1})v_j]\mu_j \\ &= Av_j + \mathcal{F}_j(v,g) - d[a(\mathbb{1})v_j]\mu_j \end{aligned} \tag{5.8}$$

and

$$g_{j,t} = dg_j(\mathbb{1})\mu_j. \tag{5.9}$$

As usual, we add initial data

$$v_j(0) = v_{j,0}, \quad g_j(0) = g_{j,0}, \quad j = 1,\dots,N, \tag{5.10}$$

which should satisfy

$$u_0 = \sum_{j=1}^N a(g_{j,0})v_{j,0}. \tag{5.11}$$

Finally, we assume that a continuous inner product $\langle\cdot,\cdot\rangle_2$ on X is available and use this to derive N phase conditions each of dimension $\dim(G)$. Suppose we have template functions $\hat v_j$ and require the distance $\operatorname{dist}(v_j, \mathcal{O}(\hat v_j)) = \inf_{g\in G}\|v_j - a(g)\hat v_j\|_2$ to the group orbit $\mathcal{O}(\hat v_j) = \{a(g)v_j : g \in G\}$ to achieve its minimum at $g = \mathbb{1}$. Then we find the necessary condition

$$\langle v_j - \hat v_j, d[a(\mathbb{1})\hat v_j]\lambda\rangle_2 = 0 \quad \forall\lambda \in T_{\mathbb{1}}G,\ j = 1,\dots,N. \tag{5.12}$$

While this generalizes (2.15), the corresponding generalization of (2.17) is

$$\langle v_{j,t}, d[a(\mathbb{1})v_j]\lambda\rangle_2 = 0 \quad \forall\lambda \in T_{\mathbb{1}}G,\ j = 1,\dots,N. \tag{5.13}$$

Note that this requires $v_{j,t}$ to be orthogonal to the group orbit $\mathcal{O}(v_j)$ at all times. When $v_{j,t}$ from (5.8) is inserted into (5.13), we obtain a linear system of dimension $\dim(G)$ for $\mu_j(t) \in T_{\mathbb{1}}G$ that has a unique solution provided $d[a(\mathbb{1})v_j] : T_{\mathbb{1}}G \to X$ is one to one.

To realize the above abstract equations in $\mathbb{R}^s$, where $s = \dim(G)$, we take a basis $\{e^1,\dots,e^s\}$, in the Lie algebra $T_{\mathbb{1}}G$ and write $\mu_j = \sum_{i=1}^s \mu_{j,i}e^i$. Then the differentiated form of (5.12) and (5.13) reads (cf. (2.16), (2.17))

$$\begin{aligned} \hat B_j\mu_j = \hat r_j,\ \text{where}\ \left((\hat B_j)_{ik}\right)_{i,k=1}^s &= \left(\langle d[a(\mathbb{1})v_j]e^k,\, d[a(\mathbb{1})\hat v_j]e^i\rangle_{\mathcal{L}_2}\right)_{i,k=1}^s \in \mathbb{R}^{s,s}, \\ \hat r_j &= \left(\langle Av_j + \mathcal{F}_j(v,g),\, d[a(\mathbb{1})\hat v_j]e^i\rangle_{\mathcal{L}_2}\right)_{i=1}^s \end{aligned} \tag{5.14}$$

and

$$\begin{aligned} B_j\mu_j = r_j,\ \text{where}\ ((B_j)_{ik})_{i,k=1}^s &= \left(\langle d[a(\mathbb{1})v_j]e^k,\, d[a(\mathbb{1})v_j]e^i\rangle_{\mathcal{L}_2}\right)_{i,k=1}^s \in \mathbb{R}^{s,s}, \\ r_j &= \left(\langle Av_j + \mathcal{F}_j(v,g),\, d[a(\mathbb{1})v_j]e^i\rangle_{\mathcal{L}_2}\right)_{i=1}^s. \end{aligned} \tag{5.15}$$

Altogether we have to solve the differential algebraic system (5.8), (5.10) with phase conditions (5.14) or (5.15).

5.2. An application to the Ginzburg–Landau equation. The cubic quintic Ginzburg–Landau equation [18], [21]

$$u_t = \alpha u_{xx} + \delta u + f(u), \quad f(u) = \beta|u|^2u + \gamma|u|^4u \\ = \alpha u_{xx} + F(u) \tag{5.16}$$

with $\delta \in \mathbb{R}$, $\alpha, \beta, \gamma \in \mathbb{C}$, $u(x,t) \in \mathbb{C}$ shows a variety of coherent structures, like pulses, fronts, sources, and sinks [21]. For certain parameter values this equation exhibits stable rotating pulses [18] as well as fronts that rotate and travel simultaneously. Equation (5.16) has the same equivariance as Example 5.1. Thus we write u as

$$u(x,t) = \sum_{j=1}^{N} e^{-i\theta_j(t)} v_j(x - \tau_j(t), t) \tag{5.17}$$

and define $\mu_j^\theta(t)$ by $\theta_{j,t}(t) = \mu_j^\theta(t)$ and $\mu_j^\tau(t)$ by $\tau_{j,t}(t) = \mu_j^\tau(t)$. The system (5.8) is of the form

$$v_{j,t}(\xi,t) = Av_j(\xi,t) + i\mu_j^\theta(t)v_j(\xi,t) + \mu_j^\tau(t)v_{j,\xi}(\xi,t) \\ + \frac{\varphi(\xi)}{\sum_{k=1}^N \varphi(\xi - \tau_k(t) + \tau_j(t))} F\left(\sum_{k=1}^{N} e^{-i(\theta_k(t)-\theta_j(t))} v_k(\xi - \tau_k(t) + \tau_j(t),\, t)\right).$$

The phase conditions are derived from the $\mathcal{L}_2$-inner product in the corresponding two dimensional real system.

$$0 = \langle \mathrm{Re}(v_j - \hat{v}_j), \mathrm{Re}(\hat{v}_{j,\xi})\rangle_{\mathcal{L}_2} + \langle \mathrm{Im}(v_j - \hat{v}_j), \mathrm{Im}(\hat{v}_{j,\xi})\rangle_{\mathcal{L}_2}, \\ 0 = \langle \mathrm{Re}(v_j - \hat{v}_j), \mathrm{Im}(\hat{v}_j)\rangle_{\mathcal{L}_2} - \langle \mathrm{Im}(v_j - \hat{v}_j), \mathrm{Re}(\hat{v}_j)\rangle_{\mathcal{L}_2}, \qquad j = 1, \ldots, N.$$

For numerical computations we used the parameters $a = 1$, $\delta = -0.1$, $\beta = 3 + i$, $\gamma = -2.75 + i$ for which the fronts and pulses mentioned above exist.

We first look at the case where the solution for the original system consists of two waves that rotate at the same speed but travel in opposite directions. As Figure 22 shows, this is reproduced correctly by our method. The values obtained from extrapolation are shown in grey shades.

The components v_1, v_2 given in Figure 23 become stationary and the parameters μ_i^τ, μ_i^θ, $i = 1, 2$, converge quickly to the correct values for the velocities of rotation and translation. Again the difference to a solution of the QCGL problem (5.16) on a large domain gives similar results as in section 3.1 (not shown), and the decay of the time derivative is exponential as before; see Figure 24 (left). Figures 25 and 26 show another result in a case where the multipulse consists of a rotating stationary pulse and a rotating traveling front with the decay rate displayed in Figure 24 (right).

6. Conclusions. We propose a numerical method for separating drifting motions of interacting pulses and fronts in a nonlinear reaction diffusion system. The method builds on an earlier approach for freezing single pulses and fronts in a comoving frame that is determined by the numerical process. The contribution of this paper is to embed the given equation into

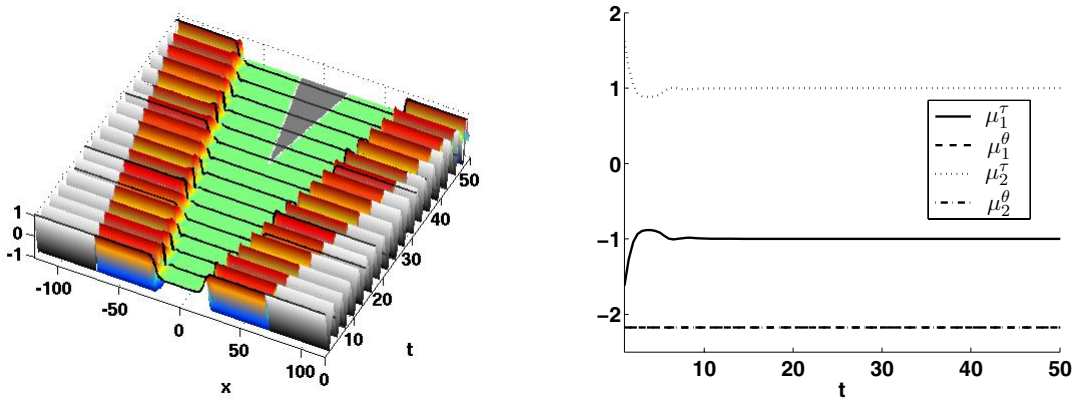


Figure 22. *Fronts moving in opposite directions in the QCGL system: Evolution of superposition* $\mathrm{Re}(u_c)$ *and velocities* $\mu_1^\tau, \mu_1^\theta, \mu_2^\tau, \mu_2^\theta$.

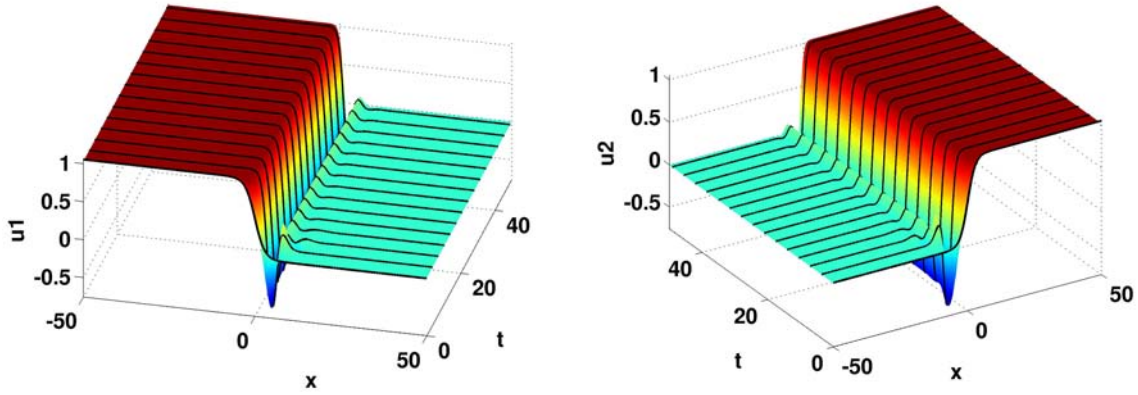


Figure 23. *Fronts moving in opposite directions in the QCGL system: Evolution of frozen fronts* $u_1 = \mathrm{Re}(v_1)$, $u_2 = \mathrm{Re}(v_2)$.

a system of N PDEs where N is at least the number of pulses, respectively, fronts, that is expected for the solution. An essential feature of the approach is to decompose the nonlinear vector field by a time-dependent partition of unity into local parts that decouple when pulses and fronts are far apart. Each subsystem is expected to describe a single front or pulse in its own moving reference frame, and the superposition of these single solutions provides an exact solution of the original system. Except for the nonlinear coupling terms, each subsystem retains a certain shift symmetry that is made use of by imposing appropriate phase conditions. Altogether, a system of partial differential algebraic equations (PDAEs) arises that is solved numerically by restricting to a finite domain and employing suitable time integrators.

There are at least two advantages of our method over solving the original equation on a (potentially very large) domain. Each subsystem can be solved on a relatively small and time-independent domain. The advantage becomes more pronounced the further apart the fronts and pulses are in the original system. Interactions in the far field are treated by extrapolating

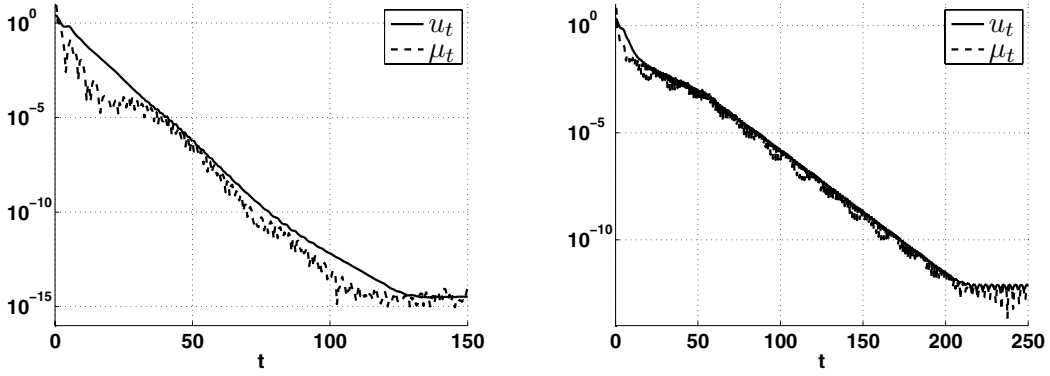


Figure 24. *Fronts moving in opposite directions in the QCGL system: Evolution of temporal change* $\|u_t\|_{\mathcal{L}_2}$ *and* $\|\mu_t\|$ *for two fronts (left) and pulse and front (right).*

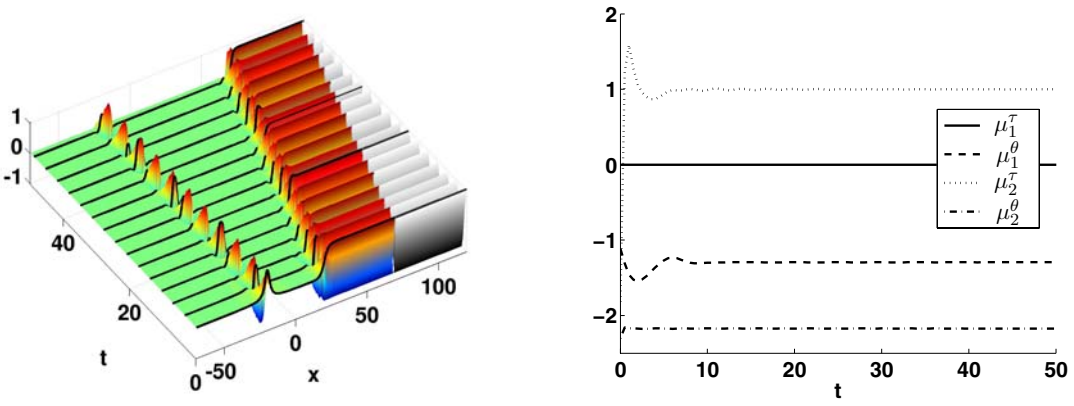


Figure 25. *Pulse and front moving in opposite directions in the QCGL system: Evolution of superposition* $\mathrm{Re}(u_c)$ *and velocities* $\mu_1^\tau, \mu_1^\theta, \mu_2^\tau, \mu_2^\theta$.

the solutions of the subsystems. Second, the approach provides direct access to the shape and velocity of the pulses and fronts present in the original solution. It avoids any a posteriori analysis of the numerical data in order to extract such information. The price to be paid for this advantage is the size of the system to be solved, which grows with the number of pulses occurring.

The method turns out to be quite robust with respect to the choice of bump function which forms the building block of the decomposition. Moreover, several numerical tests confirm that the method is able to handle strong interactions that occur during collision or merging of pulses. Typically, after such collisions the dimension of our system is larger than necessary for decomposing the solution. Then the method still works and provides single components that add up to the correct solution and travel at a common speed or do not travel at all.

The theoretical foundation of the proposed method is still in its beginning phase. We prove that single waves of the given system provide a solution of our PDAE system in an asymptotic sense. Future work will require us to show that the PDAE system is generally

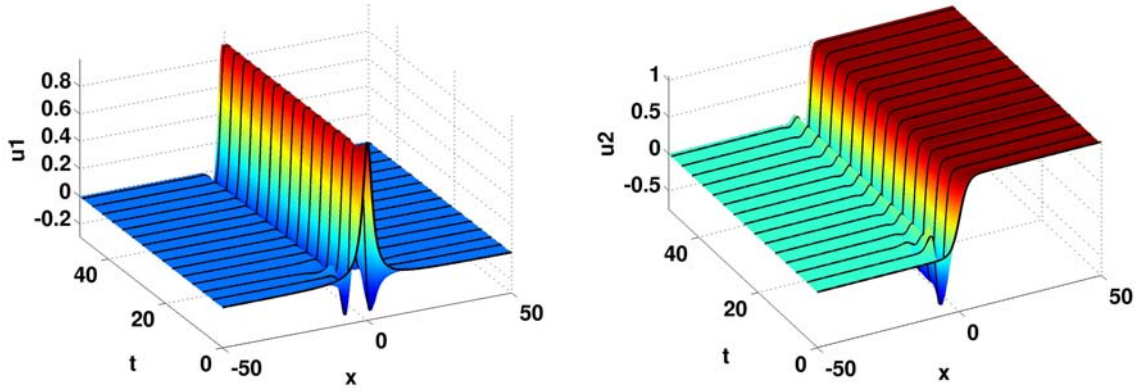


Figure 26. *Pulse and front moving in opposite directions in the QCGL system: Evolution of frozen pulse* $u_1 = \mathrm{Re}(v_1)$ *and front* $u_2 = \mathrm{Re}(v_2)$.

well-posed. Moreover, for the case of stable waves repelling each other, one expects that the set of single waves forms an asymptotically stable equilibrium for the PDAE system.

Finally, the method is formulated in the abstract framework of equivariant evolution equations, which encourages applications to much more general equations than the one-dimensional reaction diffusion systems discussed in this paper. First successful numerical tests are provided for the quintic-cubic Ginzburg–Landau equation with a two-dimensional group of equivariances.

Appendix A. Proof of Theorem 4.2 (continued). From the Lipschitz property of f and (4.12), (4.9), (4.10), (4.11), (4.14), (4.15), (4.17) we obtain the following estimates:

$$\begin{aligned} I_1^b &\le C \int_{-\infty}^{\gamma_1^+} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi + (c_j - c_k)t) \right) - f(w_1^-) + f(w_j^-) - f(w_j(\xi)) \right|^2 d\xi \\ &\le C \int_{-\infty}^{\gamma_1^+} \sum_{k=1}^{N} |w_k(\xi + (c_j - c_l)t) - w_k^-|^2 d\xi \\ &\le C \int_{-\infty}^{\gamma_1^+} \sum_{k=1}^{N} e^{2\alpha(\xi + (c_j - c_k)t)} d\xi \le C \int_{-\infty}^{\gamma_1^+} e^{2\alpha(\xi + (c_j - c_1)t)} d\xi \\ &\le C e^{-2\alpha q (c_j - c_1)t}; \end{aligned}$$

for $l \in \{1, \dots, j-1\}$,

$$\begin{aligned} I_{1,l}^{1-} &\le C \int_{\gamma_l^+}^{\gamma_l^0} \frac{\varphi(\xi)^2}{\varphi(\xi + (c_j - c_l)t)^2} d\xi \\ &\le C \int_{\gamma_l^+}^{\gamma_l^0} e^{2((\beta_1 - \beta_0)\xi - \beta_0(c_j - c_l)t)} d\xi \le C e^{2((\beta_0 - \beta_1)q - \beta_1)(c_j - c_l)t}; \end{aligned}$$

for $l \in \{1, \dots, j-1\}$,

$$I_{1,l}^{2-} \le C \int_{\gamma_l^0}^{\gamma_l^-} \frac{\varphi(\xi)^2}{\varphi(\xi + (c_j - c_l)t)^2} d\xi$$
$$\le C \int_{\gamma_l^0}^{\gamma_l^-} e^{2((\beta_1+\beta_0)\xi + \beta_0(c_j-c_l)t)} d\xi \le C e^{2((\beta_0+\beta_1)q-\beta_1)(c_j-c_l)t};$$

for $l \in \{1, \dots, j-2\}$,

$$I_{1,l}^{3-} \le C \int_{\gamma_l^-}^{\gamma_{l+1}^+} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi + (c_j - c_k)t) \right) - f(w_l^+) + f(w_j^-) - f(w_j(\xi)) \right|^2 d\xi$$
$$\le C \left[\int_{\gamma_l^-}^{\gamma_{l+1}^+} \sum_{k=1}^{l} |w_k(\xi + (c_j - c_k)t) - w_k^+|^2 d\xi \right.$$
$$\left. + \int_{\gamma_l^-}^{\gamma_{l+1}^+} \sum_{k=l+1}^{N} |w_k(\xi + (c_j - c_k)t) - w_k^-|^2 d\xi \right]$$
$$\le C \left[\int_{\gamma_l^-}^{\gamma_{l+1}^+} \sum_{k=1}^{l} e^{-2\alpha(\xi + (c_j-c_k)t)} d\xi + \int_{\gamma_l^-}^{\gamma_{l+1}^+} \sum_{k=l+1}^{N} e^{2\alpha(\xi+(c_j-c_k)t)} d\xi \right]$$
$$\le C \left[\int_{\gamma_l^-}^{\gamma_{l+1}^+} e^{-2\alpha(\xi+(c_j-c_l)t)} + e^{2\alpha(\xi+(c_j-c_{l+1})t)} d\xi \right]$$
$$\le C \left[e^{-2\alpha q(c_j-c_l)t} + e^{-2\alpha q(c_j-c_{l+1})t} \right].$$

We further obtain for $l \in \{j+1, \dots, N\}$

$$I_{1,l}^{1+} \le C \int_{\gamma_l^-}^{\gamma_l^0} \frac{\varphi(\xi)^2}{\varphi(\xi + (c_j - c_l)t)^2} d\xi$$
$$\le C \int_{\gamma_l^-}^{\gamma_l^0} e^{2((-\beta_1-\beta_0)\xi - \beta_0(c_j-c_l)t)} d\xi \le C e^{2((\beta_0+\beta_1)q-\beta_1)\gamma_l^0};$$

for $l \in \{j+1, \dots, N\}$,

$$I_{1,l}^{2+} \le C \int_{\gamma_l^0}^{\gamma_l^+} \frac{\varphi(\xi)^2}{\varphi(\xi + (c_j - c_l)t)^2} d\xi$$
$$\le C \int_{\gamma_l^0}^{\gamma_l^+} e^{2((\beta_0-\beta_1)\xi + \beta_0(c_j-c_l)t)} d\xi \le C e^{2((\beta_0-\beta_1)q-\beta_1)\gamma_l^0};$$

and for $l \in \{j+1, \dots, N-1\}$,

$$
\begin{aligned}
I_{1,l}^{3+} &\le C \int_{\gamma_l^+}^{\gamma_{l+1}^-} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi + (c_j - c_k)t) \right) - f(w_l^+) + f(w_j^+) - f(w_j(\xi)) \right|^2 d\xi \\
&\le C \left[\int_{\gamma_l^+}^{\gamma_{l+1}^-} \sum_{k=1}^{l} |w_k(\xi + (c_j - c_k)t) - w_k^+|^2 d\xi \right. \\
&\quad \left. + \int_{\gamma_l^+}^{\gamma_{l+1}^-} \sum_{k=l+1}^{N} |w_k(\xi + (c_j - c_k)t) - w_k^-|^2 d\xi \right] \\
&\le C \left[\int_{\gamma_l^+}^{\gamma_{l+1}^-} \sum_{k=1}^{l} e^{-2\alpha(\xi + (c_j - c_k)t)} d\xi + \int_{\gamma_l^+}^{\gamma_{l+1}^-} \sum_{k=l+1}^{N} e^{2\alpha(\xi + (c_j - c_k)t)} d\xi \right] \\
&\le C \left[\int_{\gamma_l^+}^{\gamma_{l+1}^-} \sum_{k=1}^{l} e^{-2\alpha(\xi + (c_j - c_l)t)} d\xi + \int_{\gamma_l^+}^{\gamma_{l+1}^-} \sum_{k=l+1}^{N} e^{2\alpha(\xi + (c_j - c_{l+1})t)} d\xi \right] \\
&\le C \left[e^{-2\alpha q \gamma_l^0} + e^{-2\alpha q \gamma_{l+1}^0} \right].
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
I_1^e &\le C \int_{\gamma_N^+}^{\infty} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi + (c_j - c_k)t) \right) - f(w_N^+) + f(w_j^+) - f(w_j(\xi)) \right|^2 d\xi \\
&\le C \sum_{k=1}^{N} |w_k(\xi + (c_j - c_k)t) - w_k^+|^2 d\xi \le C \int_{\gamma_N^+}^{\infty} \sum_{k=1}^{N} e^{-2\alpha(\xi + (c_j - c_k)t)} d\xi \\
&\le C \int_{\gamma_N^+}^{\infty} e^{-2\alpha(\xi + (c_j - c_N)t)} d\xi \le C e^{-2\alpha q (c_N - c_j)t}.
\end{aligned}
$$

For $j = 1$, the estimate of I_1 has fewer terms:

$$
\begin{aligned}
I_1 &\le C \left[\int_{-\infty}^{\gamma_2^-} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi_{k,1}^g) \right) - f(w_1(\xi)) \right|^2 d\xi + \sum_{k=2}^{N} \int_{\gamma_k^-}^{\gamma_k^0} Q_1^g(\xi, t)^2 d\xi \right. \\
&\quad + \sum_{k=2}^{N} \int_{\gamma_k^0}^{\gamma_k^+} Q_1^g(\xi, t)^2 d\xi + \sum_{k=2}^{N-1} \int_{\gamma_k^+}^{\gamma_{k+1}^-} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi_{k,1}^g) \right) - f(w_1(\xi)) \right|^2 d\xi \\
&\quad \left. + \int_{\gamma_N^+}^{\infty} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi_{k,1}^g) \right) - f(w_1(\xi)) \right|^2 d\xi \right] \\
&=: I_1^c + \sum_{k=2}^{N} I_{1,k}^{1+} + \sum_{k=2}^{N} I_{1,k}^{2+} + \sum_{k=2}^{N-1} I_{1,k}^{3+} + I_1^e.
\end{aligned}
$$

$I^{1+}_{1,k}, I^{2+}_{1,k}, I^{3+}_{1,k}, I^e$ are estimated as before, and for I^c_1 we have

$$\begin{aligned} I^c_1 &\le C \int_{-\infty}^{\gamma_2^-} \sum_{k=2}^{N} e^{2\alpha(\xi+(c_1-c_k)t)} d\xi \le C \int_{-\infty}^{\gamma_2^-} e^{2\alpha(\xi+(c_1-c_2)t)} d\xi \\ &\le C e^{-2\alpha q(c_2-c_1)t}. \end{aligned}$$

The estimate of I_1 for $j = N$ proceeds as follows:

$$\begin{aligned} I_1 \le C \Bigg[&\int_{-\infty}^{\gamma_1^+} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi^g_{k,N}) \right) - f(w_N(\xi)) \right|^2 d\xi + \sum_{k=1}^{N-1} \int_{\gamma_k^+}^{\gamma_k^0} Q^g_N(\xi,t)^2 d\xi \\ &+ \sum_{k=1}^{N-1} \int_{\gamma_k^0}^{\gamma_k^-} Q^g_N(\xi,t)^2 d\xi + \sum_{k=1}^{N-2} \int_{\gamma_k^-}^{\gamma_{k+1}^+} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi^g_{k,N}) \right) - f(w_N(\xi)) \right|^2 d\xi \\ &+ \int_{\gamma^-_{N-1}}^{\infty} \left| f\left(\sum_{k=1}^{N} \hat{w}_k(\xi^g_{k,N}) \right) - f(w_N(\xi)) \right|^2 d\xi \Bigg] \\ =: I^b_1 &+ \sum_{k=1}^{j-1} I^{1-}_{1,k} + \sum_{k=1}^{j-1} I^{2-}_{1,k} + \sum_{k=1}^{j-2} I^{3-}_{1,k} + I^c_1. \end{aligned}$$

The terms $I^b_1, I^{1-}_{1,k}, I^{2-}_{1,k}, I^{3-}_{1,k}$ are treated as before, and for I^c_1 we have

$$\begin{aligned} I^c_1 &\le C \int_{\gamma^-_{N-1}}^{\infty} \sum_{k=1}^{N-1} e^{-2\alpha(\xi+(c_N-c_k)t)} d\xi \\ &\le C \int_{\gamma^-_{N-1}}^{\infty} e^{-2\alpha(\xi+(c_N-c_{N-1})t)} d\xi \le C e^{-2\alpha q(c_N-c_{N-1})t}. \end{aligned}$$

Finally, we estimate I_2. For $2 \le j \le N-1$ we partition into four terms:

$$\begin{aligned} I_2 \le C \Bigg[&\int_{-\infty}^{\gamma_{j-1}} |f(w_j(\xi))|^2 d\xi + \int_{\gamma_{j-1}}^{0} (1 - Q^g_j(\xi,t))^2 d\xi \\ &+ \int_{0}^{\gamma_{j+1}} (1 - Q^g_j(\xi,t))^2 d\xi + \int_{\gamma_{j+1}}^{\infty} |f(w_j(\xi))|^2 d\xi \Bigg] \\ =: &\, I_{2,1} + I_{2,2} + I_{2,3} + I_{2,4}, \end{aligned}$$

where

$$\gamma_k = q(c_k - c_j)t.$$

Employing the same estimates as above, we find

$$
\begin{aligned}
I_{2,1} &\le C \int_{-\infty}^{\gamma_{j-1}} |f(w_j(\xi)) - f(w_j^-)|^2 d\xi \\
&\le C \int_{-\infty}^{\gamma_{j-1}} e^{2\alpha\xi} d\xi \le C e^{2\alpha\gamma_{j-1}}, \\
I_{2,2} &\le C \int_{\gamma_{j-1}}^{0} \frac{\left(\sum_{k=1}^{j-1} \varphi(\xi + (c_j - c_k)t) + \sum_{k=j+1}^{N} \varphi(\xi + (c_j - c_k)t)\right)^2}{\varphi(\xi)^2} \, d\xi \\
&\le C \left[\int_{\gamma_{j-1}}^{0} \sum_{k=1}^{j-1} e^{-2(\beta_0+\beta_1)\xi - 2\beta_1(c_j - c_k)t} + \sum_{k=j+1}^{N} e^{2(-\beta_0+\beta_1)\xi + 2\beta_1(c_j-c_k)t} d\xi \right] \\
&\le C \int_{\gamma_{j-1}}^{0} e^{-2(\beta_0+\beta_1)\xi - 2\beta_1(c_j - c_{j-1})t} + e^{2(-\beta_0+\beta_1)\xi + 2\beta_1(c_j-c_{j+1})t} d\xi \\
&\le C(e^{2((\beta_0+\beta_1)q - \beta_1)(c_j - c_{j-1})t} + e^{2(\beta_0-\beta_1)q(c_j-c_{j-1})t - 2\beta_1(c_{j+1}-c_j)t}), \\
I_{2,3} &\le C \int_{0}^{\gamma_{j+1}} \frac{\left(\sum_{k=1}^{j-1} \varphi(\xi + (c_j - c_k)t) + \sum_{k=j+1}^{N} \varphi(\xi + (c_j - c_k)t)\right)^2}{\varphi(\xi)^2} \, d\xi \\
&\le C \left[\int_{0}^{\gamma_{j+1}} \sum_{k=1}^{j-1} e^{2(\beta_0-\beta_1)\xi - 2\beta_1(c_j - c_k)t} + \sum_{k=j+1}^{N} e^{2(\beta_0+\beta_1)\xi + 2\beta_1(c_j-c_k)t} d\xi \right] \\
&\le C \int_{0}^{\gamma_{j+1}} e^{2(\beta_0-\beta_1)\xi - 2\beta_1(c_j - c_{j-1})t} + e^{2(\beta_0+\beta_1)\xi + 2\beta_1(c_j-c_{j+1})t} d\xi \\
&\le C \left[e^{2(\beta_0-\beta_1)\gamma_{j+1} - 2\beta_1(c_j - c_{j-1})t} + e^{2((\beta_0+\beta_1)q - \beta_1)(c_{j+1}-c_j)t} \right], \\
I_{2,4} &\le C \int_{\gamma_{j+1}}^{\infty} |f(w_j(\xi)) - f(w_j^+)|^2 d\xi \le C \int_{\gamma_{j+1}}^{\infty} e^{-2\alpha\xi} d\xi \le C e^{-2\alpha\gamma_{j+1}}.
\end{aligned}
$$

For $j = 1$ we estimate I_2 as follows:

$$
\begin{aligned}
I_2 &\le C \left[\int_{-\infty}^{-qt} |f(w_1(\xi))|^2 d\xi + \int_{-qt}^{0} (1 - Q_1^g(\xi, t))^2 d\xi \right. \\
&\qquad \left. + \int_{0}^{\gamma_2} (1 - Q_1^g(\xi, t))^2 d\xi + \int_{\gamma_2}^{\infty} |f(w_1(\xi))|^2 d\xi \right] \\
&=: I_{2,1} + I_{2,2} + I_{2,3} + I_{2,4}.
\end{aligned}
$$

$I_{2,3}$ and $I_{2,4}$ are estimated as before, and for $I_{2,1}$ and $I_{2,2}$ we obtain

$$
\begin{aligned}
I_{2,1} &\le C \int_{-\infty}^{-qt} |f(w_1(\xi)) - f(w_1^-)|^2 d\xi \le C \int_{-\infty}^{-qt} e^{2\alpha_1\xi} d\xi \le C e^{-2\alpha_1 qt}, \\
I_{2,2} &\le C \int_{-qt}^{0} \frac{\sum_{k=2}^{N} \varphi(\xi + (c_1 - c_k)t)}{\varphi(\xi)^2}^2 d\xi \le C \int_{-qt}^{0} \sum_{k=2}^{N} e^{2(-\beta_0+\beta_1)\xi + 2\beta_1(c_1-c_k)t} d\xi \\
&\le C \int_{-qt}^{0} e^{2(-\beta_0+\beta_1)\xi + 2\beta_1(c_1-c_2)t} d\xi \le C(e^{2(\beta_0-\beta_1)qt - 2\beta_1(c_2-c_1)t}).
\end{aligned}
$$

Similarly, we find for $j = N$

$$\begin{aligned} I_2 &\le C\left[\int_{-\infty}^{\gamma_{N-1}} |f(w_N(\xi))|^2 d\xi + \int_{\gamma_{N-1}}^{0} (1 - Q_N^g(\xi,t))^2 d\xi \right. \\ &\quad \left. + \int_0^{qt} (1 - Q_N^g(\xi,t))^2 d\xi + \int_{qt}^{\infty} |f(w_N(\xi))|^2 d\xi \right] \\ &=: I_{2,1} + I_{2,2} + I_{2,3} + I_{2,4}. \end{aligned}$$

$I_{2,1}$ and $I_{2,2}$ are as before, and $I_{2,3}$, $I_{2,4}$ satisfy

$$\begin{aligned} I_{2,3} &\le C\int_0^{qt} \frac{\sum_{k=1}^{N-1} \varphi(\xi + (c_N - c_k)t)}{\varphi(\xi)^2}^2 d\xi \le C\int_0^{qt} \sum_{k=1}^{N-1} e^{2(\beta_0-\beta_1)\xi - 2\beta_1(c_N-c_k)t} d\xi \\ &\le C\int_0^{qt} e^{2(\beta_0-\beta_1)\xi - 2\beta_1(c_N-c_{N-1})t} d\xi \le Ce^{2(\beta_0-\beta_1)qt - 2\beta_1(c_N-c_{N-1})t}, \\ I_{2,4} &\le C\int_{qt}^{\infty} |f(w_N(\xi)) - f(w_N^+)|^2 d\xi \le C\int_{qt}^{\infty} e^{-2\alpha\xi} d\xi \le Ce^{-2\alpha qt}. \end{aligned}$$

Acknowledgments. The authors thank Clarence Rowley for stimulating discussions that furthered the subject and for pointing out the relations to methods in fluid dynamics. The authors are also grateful for the constructive comments of the anonymous referees that led to improvements of the first version of the paper.

REFERENCES

[1] H. Airault, *Équations asymptotiques pour des cas spéciaux de l'équation de Nagumo*, C. R. Acad. Sci. Paris Sér. I Math., 301 (1985), pp. 295–298.
[2] C. Anderson and C. Greengard, *On vortex methods*, SIAM J. Numer. Anal., 22 (1985), pp. 413–440.
[3] S. Arimoto, J. Nagumo, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proceedings of the IRE, 50 (1962), pp. 2061–2070.
[4] W.-J. Beyn and V. Thümmler, *Freezing solutions of equivariant evolution equations*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 85–116.
[5] W.-J. Beyn and V. Thümmler, *Phase conditions, symmetries, and PDE continuation*, in Numerical Continuation Methods for Dynamical Systems, Underst. Complex Syst., B. Krauskopf, H. Osinga, and J. Galan-Vioque, eds., Springer-Verlag, New York, 2007, pp. 301–330.
[6] *Comsol Multiphysics* 3.3, Comsol, Stockholm, Sweden, 2007, http://www.comsol.com.
[7] J. W. Evans, N. Fenichel, and J. A. Feroe, *Double impulse solutions in nerve axon equations*, SIAM J. Appl. Math., 42 (1982), pp. 219–234.
[8] A. J. Majda and A. L. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge Texts in Applied Mathematics 27, Cambridge University Press, Cambridge, UK, 2002.
[9] R. M. Miura, *Accurate computation of the stable solitary waves for the FitzHugh-Nagumo equations*, J. Math. Biol., 13 (1982), pp. 247–269.
[10] J. D. Murray, *Mathematical Biology*, 2nd ed., Biomathematics 19, Springer-Verlag, Berlin, 1993.
[11] S. Nii, *A topological proof of stability of N-front solutions of the FitzHugh-Nagumo equations*, J. Dynam. Differential Equations, 11 (1999), pp. 515–555.
[12] C. W. Rowley, I. G. Kevrekidis, J. E. Marsden, and K. Lust, *Reduction and reconstruction for self-similar dynamical systems*, Nonlinearity, 16 (2003), pp. 1257–1275.
[13] B. Sandstede, *Stability of multiple-pulse solutions*, Trans. Amer. Math. Soc., 350 (1998), pp. 429–472.

[14] B. Sandstede, *Stability of N-fronts bifurcating from a twisted heteroclinic loop and an application to the FitzHugh-Nagumo equation*, SIAM J. Math. Anal., 29 (1998), pp. 183–207.
[15] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems, Vol. 2, North–Holland, Amsterdam, 2002, pp. 983–1055.
[16] B. Sandstede and A. Scheel, *Gluing unstable fronts and backs together can produce stable pulses*, Nonlinearity, 13 (2000), pp. 1465–1482.
[17] A. Scheel and J. D. Wright, *Colliding dissipative pulses: The shooting manifold*, J. Differential Equations, to appear.
[18] O. Thual and S. Fauve, *Localized structures generated by subcritical instabilities*, J. Phys. France, 49 (1988), pp. 1829–1833.
[19] V. Thümmler, *Numerical Analysis of the Method of Freezing Traveling Waves*, Ph.D. thesis, Dept. of Mathematics, Bielefeld University, Bielefeld, Germany, 2005.
[20] V. Thümmler, *The effect of freezing and discretization to the asymptotic stability of relative equilibria*, J. Dynam. Differential Equations, 20 (2008), pp. 425–477.
[21] W. van Saarloos and P. C. Hohenberg, *Fronts, pulses, sources and sinks in generalized complex Ginzburg-Landau equations*, Phys. D, 56 (1992), pp. 303–367.
[22] E. Yanagida and K. Maginu, *Stability of double-pulse solutions in nerve axon equations*, SIAM J. Appl. Math., 49 (1989), pp. 1158–1173.
[23] S. Zelik and A. Mielke, *Multi-pulse evolution and space-time chaos in dissipative systems*, Mem. Amer. Math. Soc., to appear.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 2, pp. 609–649
© 2008 Society for Industrial and Applied Mathematics

Mechanisms for Frequency Control in Neuronal Competition Models*

Rodica Curtu†, Asya Shpiro‡, Nava Rubin‡, and John Rinzel§

Abstract. We investigate analytically a firing rate model for a two-population network based on mutual inhibition and slow negative feedback in the form of spike frequency adaptation. Both neuronal populations receive external constant input whose strength determines the system's dynamical state—a steady state of identical activity levels or periodic oscillations or a winner-take-all state of bistability. We prove that oscillations appear in the system through supercritical Hopf bifurcations and that they are antiphase. The period of oscillations depends on the input strength in a nonmonotonic fashion, and we show that the increasing branch of the period versus input curve corresponds to a release mechanism and the decreasing branch to an escape mechanism. In the limiting case of infinitely slow feedback we characterize the conditions for release, escape, and occurrence of the winner-take-all behavior. Some extensions of the model are also discussed.

Key words. Hopf bifurcation, antiphase oscillations, slow negative feedback, winner-take-all, release and escape, binocular rivalry, central pattern generators

AMS subject classifications. 34C15, 34C23, 37G05, 92B20

DOI. 10.1137/070705842

1. Introduction.

Competition models have a long tradition in ecology and population biology (see, e.g., [22]). Typically, the competition involves negative interactions in the battle for a common resource. Eventually, one of the participant populations emerges as the winner eliminating the competitors. This framework has appeared in models of neuronal development where the competition is for synapse formation such as the development of neuromuscular connections for innervated muscle fibers and for the formation of ocular dominance columns and topographic maps (as reviewed in [35]). The notion of competition has also been applied in the modeling of various neuronal computational tasks. *Winner-take-all* behavior, when one neural population remains active and the others inactive indefinitely as a result of inhibitory interactions, has been proposed in models for short term memory and attention [13] or for the selection and switching in the striatum of the basal ganglia under both normal and pathological conditions [15, 21].

The winner-take-all steady state may persist for a long time but not indefinitely if some

*Received by the editors October 19, 2007; accepted for publication (in revised form) by D. Terman February 26, 2008; published electronically June 13, 2008.
http://www.siam.org/journals/siads/7-2/70584.html

†Department of Mathematics, The University of Iowa, 14 MacLean Hall, Iowa City, IA 52242, and Transilvania University of Brasov, Romania (rodica-curtu@uiowa.edu). The work of this author was supported by the Romanian grants CNCSIS AT55 and PNCDI-2 11-039/2007.

‡Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003 (asya@cns.nyu.edu, Nava.Rubin@nyu.edu). The work of the second author was supported by NIH grant EY07-158-03. The work of the third author was supported by the Swartz Foundation.

§Courant Institute of Mathematical Sciences and Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003 (rinzel@cns.nyu.edu). The work of this author was supported by the Swartz Foundation.

mechanism for slow fatigue or adaptation is at work. In this case one population may be dominant for a while, then another, and so on. Competition between, say, two neuronal populations, via reciprocal inhibition and slow adaptation underlies models for alternating rhythmic behavior in central pattern generators (CPGs) [11, 32, 6, 33] and in perceptual bistability [17, 37, 24]. CPGs consist of neural circuits that drive alternately contracting muscle groups. Perceptual bistability refers to a class of phenomena in which a deeply ambiguous stimulus gives rise to two different interpretations that alternate over time, only one being perceived at any given moment. Slow adaptation may be implemented via a cellular mechanism (fatigue in the spike generation mechanism) or a negative feedback in the coupling (depression of the synaptic transmission mechanism). In some neuronal competition models the alternations may be irregular and caused primarily by noise, with adaptation playing a secondary role [30, 10, 24].

Both for CPGs and perceptual bistability the issue of oscillations' frequency or period detection (and eventually control) seems to be important. For example, a classical example of perceptual bistability is binocular rivalry whose properties were summarized in the so-called Levelt's propositions [19]. In binocular rivalry, a subject views an ambiguous stimulus in which each eye is presented with a drastically different image. Instead of perceiving a mixture of the two images, the subject reports (over a large range of stimulus conditions) an alternation between the two competing percepts; one image is perceived for a while (a few seconds), then the other, etc. Levelt's proposition IV (LP-IV) states that *increasing the contrast of the rivaling images increases the frequency of percept switching*, or, in other words, that *dominance times of both perceived images decrease with equal increase of stimulus strength*. Since 1968, binocular rivalry has been investigated intensively in other psychophysics experiments [2, 25, 20, 1, 28, 29, 3], in experiments using fMRI techniques [34, 26, 38, 18], and also in modeling studies [17, 37, 10, 24].

In a recent modeling paper, Shpiro et al. [31] show that for a class of competition models, the LP-IV type of dynamics occurs in fact only within a limited range of stimulus strength. Outside this range four other types of behavior were observed: (i) fusion at a very high level of activity, (ii) winner-take-all behavior, (iii) a region where dominance times increase with stimulus strength (as opposed to LP-IV), and then (iv) fusion again for very low levels of activity (see Figure 3F in section 2). These differences between experimental reports and theory have important implications, either predicting new possible dynamics in binocular rivalry or, if future experiments do not confirm them, pointing to the necessity for other types of models. Meanwhile, it is important to understand the sources or mechanisms that lead to the nonmonotonic dependence of oscillation period on the stimulus strength for this class of neuronal competition models. Our paper aims to investigate this issue.

We analyze a firing rate model in which competition between populations is a result of a combination of reciprocal inhibition and a slow negative feedback process. We prove that, as the input strength changes, oscillations appear in the system through a Hopf bifurcation and that they are antiphase. Due to the two time scales involved in the system there is a regime where periodic solutions take the form of relaxation-oscillators. Their period of oscillations depends nonmonotonically on the input strength, say, I: in a range of low values for I, the period increases with I, and we show that the dynamics is due to a *release* mechanism; on the other hand, in a range of higher values for I, the period decreases with I, and *escape* is

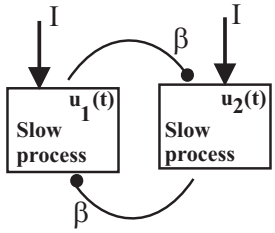


Figure 1. *Network architecture of neuronal competition model.*

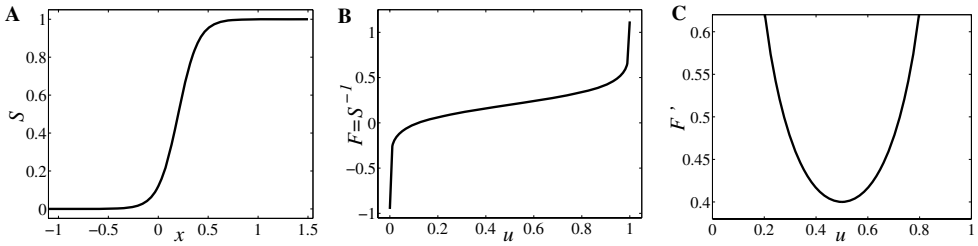


Figure 2. *Graphical representations of generic* (A) *gain function S;* (B) *its inverse F; and* (C) *first derivative F'.*

the underlying mechanism (see section 4; we define release as the case when the switch in dominance during oscillation occurs due to a significant change in the response to an input to the dominant population. On the contrary, escape corresponds to the case of a significant change in the input-output function for the suppressed population). For intermediate values of I, winner-take-all is possible and we explain how it appears. Then in section 5 we present some model modifications that allow for reducing, or even excluding, one of the escape and release regimes, thus leading to a monotonic period versus input curve.

2. The mathematical model. The model we investigate in this paper assumes a network architecture of two populations of neurons (Figure 1) that respond to two competing stimuli of equal strength:

$$\begin{aligned} \dot{u}_1 &= -u_1 + S(I - \beta u_2 - g a_1), \\ \dot{u}_2 &= -u_2 + S(I - \beta u_1 - g a_2), \\ \tau \dot{a}_1 &= -a_1 + u_1, \\ \tau \dot{a}_2 &= -a_2 + u_2. \end{aligned} \tag{2.1}$$

Variables u_j $(j = 1, 2)$ measure short-time and spatially averaged firing rates of the two populations that inhibit each other. The system is nonlinear due to the gain function S; it is the steady input-output function for the population and it has a sigmoid shape as in Figure 2A. The strength of inhibition is modeled by the positive parameter β, while I is the control parameter directly associated to the external stimulus strength (e.g., it grows with growing stimulus strength such as contrast). Each population is subject to a slow negative feedback process a_j such as spike frequency adaptation of positive strength g. Since variables a_j evolve in much slower time than u_j, the parameter τ takes large values, $\tau \gg 1$ (e.g., the

time-scale for u_j is about 10 msec, while for a_j it is about 1000 msec).

Remark 2.1. *In general, firing-rate models like (2.1) include in the equation of u_j a nonlinear term of the form $S(I + \alpha u_j - \beta u_k - g a_j)$. The product αu_j is associated with the intrapopulation recurrent excitation. It is important to note that for neuronal competition model (2.1) we have disallowed recurrent excitation (taken $\alpha = 0$) in order to preclude an isolated population ($\beta = 0$) from oscillating on its own. This is a restriction imposed by experimental observations on binocular rivalry and other perceptual bistable phenomena.*

The nonlinear gain function S that appears in the differential equations for u_j is usually defined in neuronal models by

$$S(x) = \frac{1}{1 + e^{-r(x-\theta)}} \tag{2.2}$$

with positive r and real θ.

Function S is invertible with $F = S^{-1}$ a $\mathcal{C}^\infty(0,1)$-function and $F'(u) = \frac{1}{ru(1-u)}$ (Figure 2B–C). Based on this example, we consider the following *assumptions* for the gain function S.

$S : \boldsymbol{R} \to (0,1)$ is a differentiable, monotonically increasing function with $S(\theta) = u_0 \in (0,1)$ and $\lim_{x\to-\infty} S(x) = 0$, $\lim_{x\to\infty} S(x) = 1$. Moreover, its first and second derivatives satisfy the conditions $\lim_{x\to\pm\infty} S'(x) = 0$, $S''(x) > 0$ for $x < \theta$, $S''(x) < 0$ for $x > \theta$, and $S''(\theta) = 0$, so S' has a maximum at θ.

As a consequence, S is invertible with $F = S^{-1} : (0,1) \to \boldsymbol{R}$ monotonically increasing function such that $\lim_{u\to0} F(u) = -\infty$, $\lim_{u\to1} F(u) = \infty$, $F(u_0) = \theta$, and $\lim_{u\to0} F'(u) = \lim_{u\to1} F'(u) = \infty$, $F''(u) < 0$ for $u \in (0, u_0)$, $F''(u) > 0$ for $u \in (u_0, 1)$, $F''(u_0) = 0$. Obviously, F' has a minimum value at u_0 which is $F'(u_0) = 1/S'(\theta)$.

Additionally we assume that F is a $\mathcal{C}^\infty$-function on $(0,1)$, or at least $\mathcal{C}^2$ on $(0,1)$ and $\mathcal{C}^\infty$ on $(0,1) \setminus \{u_0\}$.

The typical graphs of function S and its corresponding F and F' are drawn in Figure 2A–C. We used the example (2.2) with parameter values $r = 10$ and $\theta = 0.2$; obviously in this case $S(\theta) = u_0 = 0.5$.

All the experiments that motivated our work report oscillatory phenomena with frequencies tightly connected to the stimulus strengths. Moreover, as Levelt [19] pointed out for binocular rivalry, those experiments show large ranges for stimulus strength where the corresponding oscillation periods/frequencies behave monotonically. What kind of possible mechanism is behind this type of dynamics is the question we will focus on in this paper.

Given the neuronal competition model (2.1), the goal is to examine the effect the parameter I has on the existence of oscillations and on their period. The system is a simplified version of an entire class of competition models that, as we found [31], share important dynamical features.

To illustrate those commonalities we draw in Figure 3A–E the timecourses of activity variables $u_1(t)$ and $u_2(t)$ for different values of control parameter I. Then we summarize the result in the bifurcation diagram of the period T versus I (Figure 3F). Other parameter values are fixed to $\beta = 1.1$, $g = 0.5$, $\tau = 100$, and S as in (2.2) with $r = 10$ and $\theta = 0.2$ (here $u_0 = 1/2$). The system (2.1) exhibits five possible types of behavior: for large values of I (region I in Figure 3F) both populations are active at identically high levels (Figure 3A); the timecourses of $u_1(t)$ and $u_2(t)$ tend to a stable steady state larger than u_0. As I decreases (region II,

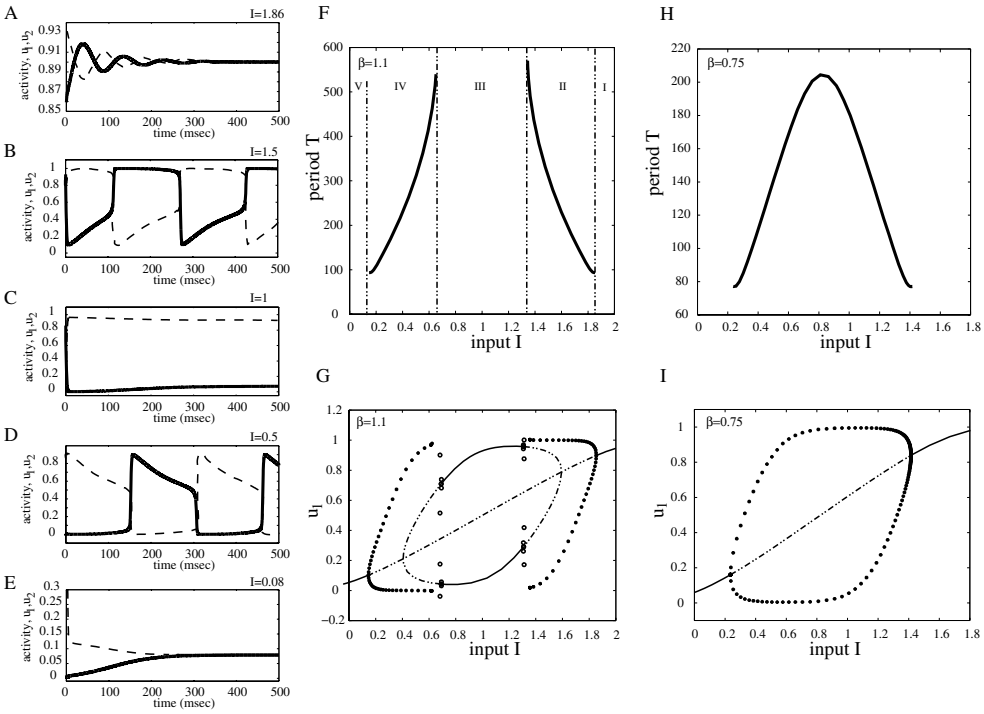


Figure 3. *Bifurcation diagrams and examples of activity timecourses for neuronal competition model* (2.1) *with parameter values* $g = 0.5$, $\tau = 100$, $r = 10$, $\theta = 0.2$, *and* $\beta = 1.1$ (A–G), *respectively,* $\beta = 0.75$ (H–I). *Timecourses of* u_1, u_2 *corresponding to panel F for different values of* I: (A) $I = 1.86$, (B) $I = 1.5$, (C) $I = 1$, (D) $I = 0.5$, *and* (E) $I = 0.08$. *Bifurcation diagrams of period* T *of the network oscillation versus input strength* I *(F and H). Bifurcation diagram of population activity* u_1 *versus* I *(G and I).*

Figure 3F) the system starts oscillating with $u_1(t)$ and $u_2(t)$ alternatively on and off (Figure 3B); in this region the period of oscillation decreases with increasing input strength. At intermediate values of I a winner-take-all kind of behavior is observed (region III, Figure 3F); depending on the choice of initial conditions, one of the two populations is active indefinitely, while the other one remains inactive (Figure 3C). Decreasing I even more (region IV, Figure 3F) the neuronal model becomes oscillatory again (Figure 3D) with u_1 and u_2 competing for the active state; however, for this range of parameter the oscillation period T increases with input value I—an opposite behavior to that observed in region II. Last, for small values of stimulus strength (region V, Figure 3F) both populations remain inactive at identically low level firing rates (Figure 3E); the timecourses of $u_1(t)$ and $u_2(t)$ tend to a stable steady state less than u_0.

To further characterize system (2.1)'s dynamics as the input value I is varied, we also construct the local bifurcation diagram of amplitude response u_1 to I (Figure 3G). For the parameter ranges I and V the trajectories are attracted to a stable fixed point satisfying the $\tilde{u}_1 = \tilde{u}_2$ condition (thick line in Figure 3G). This fixed point becomes unstable (dashed line) in regions II, III, and IV, where the attractor is replaced by either a stable limit cycle (regions II and IV: branched filled-circle curves corresponding to the maximum and minimum amplitudes during rivalry oscillations) or another stable fixed point with $\tilde{u}_1 \neq \tilde{u}_2$ (region III).

Due to the symmetry in the equations of (2.1) whenever $(\tilde{u}_1, \tilde{u}_2, \tilde{a}_1, \tilde{a}_2)$ is an equilibrium point, $(\tilde{u}_2, \tilde{u}_1, \tilde{a}_2, \tilde{a}_1)$ is as well. The local bifurcation diagram suggests the existence of some Hopf and pitchfork bifurcations in the model that we will further investigate in section 3.

There is another common feature of many neuronal competition models based on reciprocal inhibition architecture with slow negative feedback in the form of spike frequency adaptation and/or synaptic depression [31]: the absence of the winner-take-all behavior when inhibition strength β is sufficiently small. As we illustrate in Figure 3H–I for $\beta = 0.75$ (all other parameters are the same as above), the winner-take-all regime at intermediate I disappears. However, the dependency of period T on stimulus strength remains nonmonotonic.

An intriguing question is: What are the neuronal mechanisms underlying the two distinct dynamics—one of increasing period with increase of stimulus strength (for smaller values of input), and one of decreasing period (for larger values)? This issue is discussed in section 4 and then extended in section 5.

Our general goal is to understand and possibly to analytically characterize the numerical results obtained for this specific competition model (2.1). Consequently, as already pointed out, this step will help us understand the typical behavior of an entire class of neuronal competition models.

3. Oscillatory antiphase solutions and local analysis. In this section we use methods from local bifurcation theory [14, 16] to prove the existence of periodic solutions $(u_1(t), u_2(t), a_1(t), a_2(t))$ for the two-population network (2.1). We also show that the main variables u_1 and u_2 oscillate in antiphase, therefore competing for the ON/active state.

The bifurcation diagrams obtained numerically in section 2 suggest the existence of an equilibrium point satisfying $u_1 = u_2$ no matter the value of parameter I. Let us now investigate system (2.1) theoretically.

All equilibria satisfy the conditions $u_1 = S(I - \beta u_2 - g a_1)$, $u_2 = S(I - \beta u_1 - g a_2)$, $a_1 = u_1$, and $a_2 = u_2$ that are equivalent (due to the invertibility of S) to $F(u_1) = I - \beta u_2 - g a_1$, $F(u_2) = I - \beta u_1 - g a_2$, and $a_1 = u_1$, $a_2 = u_2$. Looking for a particular type of equilibrium point, that is, for points with $u_1 = u_2$, we obtain the equation $I = H(u)$ with H defined by

$$H : (0,1) \to \boldsymbol{R}, \quad H(u) \stackrel{\text{def}}{=} F(u) + (\beta + g)u. \tag{3.1}$$

Since F is monotonically increasing on $(0,1)$ with vertical asymptotes $\lim_{u\to 0} F(u) = -\infty$ and $\lim_{u\to 1} F(u) = \infty$, (3.1) has a unique solution $u_I \in (0,1)$ for any real value of the parameter I. Moreover, from the identity $I = F(u_I) + (\beta + g)u_I$ we compute

$$\frac{du_I}{dI} = \frac{1}{\beta + g + F'(u_I)}, \tag{3.2}$$

so a decrease in I leads to a decrease in u_I with $\lim_{I\to\infty} u_I = 1$ and $\lim_{I\to-\infty} u_I = 0$.

The neuronal competition model (2.1) always possesses an equilibrium of the type (u_I, u_I, u_I, u_I). Its stability properties are then defined by the linearized system $dY/dt = \mathcal{A}Y$, $Y = (u_1 - u_I, u_2 - u_I, a_1 - u_I, a_2 - u_I)^{\mathrm{T}}$, where $(\,)^{\mathrm{T}}$ stays for the transpose, and matrix

$$\mathcal{A} = \begin{pmatrix} -1 & -\beta/F'(u_I) & -g/F'(u_I) & 0 \\ -\beta/F'(u_I) & -1 & 0 & -g/F'(u_I) \\ 1/\tau & 0 & -1/\tau & 0 \\ 0 & 1/\tau & 0 & -1/\tau \end{pmatrix}.$$

This form of the matrix relies on the equality $S'(I-\beta u_2-ga_1) = S'(F(u_1)) = S'(S^{-1}(u_1)) = 1/(S^{-1})'(u_1) = 1/F'(u_1)$, which is true at the equilibrium point.

The characteristic equation of $\mathcal{A}$ takes the form

$$\left[(\lambda+1)\left(\lambda+\frac{1}{\tau}\right)+\frac{g}{\tau F'(u_I)}\right]^2-\left[\frac{\beta}{F'(u_I)}\left(\lambda+\frac{1}{\tau}\right)\right]^2=0.$$

As a difference of squares it can be decomposed into two quadratic equations: the first is $\lambda^2+\lambda\left(1+\frac{1}{\tau}+\frac{\beta}{F'(u_I)}\right)+\frac{1}{\tau}\left(1+\frac{g+\beta}{F'(u_I)}\right)=0$, so two eigenvalues of the matrix $\mathcal{A}$, say, λ_1 and λ_2, have negative real part no matter the value of parameter I. The other eigenvalues λ_3 and λ_4 satisfy the second quadratic equation $\lambda^2+\lambda\left(1+\frac{1}{\tau}-\frac{\beta}{F'(u_I)}\right)+\frac{1}{\tau}\left(1+\frac{g-\beta}{F'(u_I)}\right)=0$, and their real part can change sign when I is varied.

For $|I|$ sufficiently large, u_I is close to either zero or one, keeping $F'(u_I)$ larger than both $\beta/(1+\frac{1}{\tau})$ and $\beta-g$ (see Figure 2C); the corresponding equilibrium point (u_I,u_I,u_I,u_I) is asymptotically stable.

There are two ways this equilibrium point can lose stability: either through a pair of purely imaginary eigenvalues $\lambda_{3,4}=\pm i\omega$ at $F'(u_I)=\beta/(1+\frac{1}{\tau})$ or through a zero eigenvalue $\lambda_3=0$, $\lambda_4<0$ at $F'(u_I)=\beta-g$. Which of these two cases occurs first depends on the relationship between $\beta/(1+\frac{1}{\tau})$ and $\beta-g$: if $\beta/(1+\frac{1}{\tau})>\beta-g$, i.e., $\beta/g<\tau+1$, then the eigenvalues λ_3,λ_4 change the sign of their real part from negative to positive by crossing the imaginary axis ($\lambda_{3,4}=\pm i\omega$); if $\beta/g>\tau+1$, then the case $\lambda_3=0$, $\lambda_4<0$ is encountered first.

At this point we remind the reader of our assumption of a large time constant value τ. (The competition between the populations in the network comes from the combination of two important ingredients: reciprocal inhibition and the addition of a slow negative feedback process.) Therefore, it makes sense to situate ourselves in the case of $\beta/g\ll\tau$, which implies

$$\beta<g(\tau+1). \tag{3.3}$$

Inequality (3.3) can be interpreted as a feature of the neuronal competition model to be rather (adaptation) feedback-dominated than (inhibitory) coupling-dominated.

We will assume in the following that parameters g and τ are fixed and that β is chosen such that (3.3) is true.

Another observation is that the graph of F' has a *well-like* shape with positive minimum at $1/S'(\theta)$ as in Figure 2C. Consequently the straight horizontal line $y=\beta/(1+\frac{1}{\tau})$ intersects it twice (if $\beta/(1+\frac{1}{\tau})>1/S'(\theta)$), once (for the equality), or not at all (if $\beta/(1+\frac{1}{\tau})<1/S'(\theta)$). As observed in numerical simulations, in order for the system to oscillate, a sufficiently large inhibition strength has to be considered. Mathematically that reduces to

$$\beta>\frac{1+1/\tau}{S'(\theta)}. \tag{3.4}$$

Thus we are able to characterize the stability of the equilibrium (u_I,u_I,u_I,u_I).

Theorem 3.1. *The dynamical system (2.1) has a unique equilibrium point with $u_1=u_2$, say, $\mathbf{e}_I=(u_I,u_I,u_I,u_I)$, for any real I. The value u_I increases monotonically with I and belongs to the interval $(0,1)$.*

Let us assume that the adaptation-dominance condition (3.3) is true.

(i) *If* $\beta < \frac{1+1/\tau}{S'(\theta)}$, *then* $\mathbf{e}_I$ *is asymptotically stable for all* $I \in \boldsymbol{R}$.

(ii) *If* $\beta > \frac{1+1/\tau}{S'(\theta)}$, *then there exist exactly two values* $u^*_{hb}, u^{**}_{hb} \in (0,1)$ *such that* $u^*_{hb} < u_0 < u^{**}_{hb}$ *and*

$$F'(u^*_{hb}) = F'(u^{**}_{hb}) = \frac{\beta}{1+\frac{1}{\tau}}. \tag{3.5}$$

The equilibrium point $\mathbf{e}_I$ *is asymptotically stable for all* $I \in (-\infty, I^*_{hb}) \cup (I^{**}_{hb}, \infty)$ *and unstable for* $I \in (I^*_{hb}, I^{**}_{hb})$, *where* $I^*_{hb} = H(u^*_{hb})$, $I^{**}_{hb} = H(u^{**}_{hb})$ *defined by* (3.1). *At* I^*_{hb} *and* I^{**}_{hb} *the stability is lost through a pair of purely imaginary eigenvalues.*

Proof. (i) Since $\beta - g < \beta/(1+\frac{1}{\tau}) < 1/S'(\theta) = \min(F')$, all eigenvalues of the linearized system about $\mathbf{e}_I$ have negative real part.

(ii) The conclusion is based on the properties of F', which decreases on interval $(0, u_0)$ and increases on $(u_0, 1)$ with $F'(u_0) = \min(F')$. The sum $\lambda_3 + \lambda_4$ changes sign from negative to positive when I increases through I^*_{hb} and then back from positive to negative when passing through I^{**}_{hb}. For $I \in (I^*_{hb}, I^{**}_{hb})$ at least one real part of λ_3 and λ_4 is positive, so $\mathbf{e}_I$ is unstable. At $I = I^*_{hb}$ and $I = I^{**}_{hb}$ we have $\lambda_{3,4} = \pm i\omega$. For all other values of I we have $\lambda_3 + \lambda_4 < 0$, $\lambda_3\lambda_4 > 0$; so $\mathbf{e}_I$ is asymptotically stable. ∎

Since the equilibrium point $\mathbf{e}_I$ becomes unstable through a pair of purely imaginary eigenvalues as I crosses the values I^*_{hb} and I^{**}_{hb}, we expect to find there two Hopf bifurcation points. Indeed, in section 3.1 we prove the existence of a supercritical Hopf bifurcation at both I^*_{hb} and I^{**}_{hb} and, as a consequence, the existence of stable oscillatory solutions for system (2.1).

3.1. Normal form for the Hopf bifurcation. In the following we assume that both inequalities (3.3) and (3.4) are true; that is, we take the coupling in (2.1) to be sufficiently strong but still adaptation-dominated.

Let us use the notation I^* for any of the critical values I^*_{hb} and I^{**}_{hb} and similarly the notation u^* for $u^* \in \{u^*_{hb}, u^{**}_{hb}\}$. Then the linearization matrix $\mathcal{A}$ at u^* becomes

$$\mathcal{A}_0 = \begin{pmatrix} -1 & -(1+\frac{1}{\tau}) & -\frac{g}{\beta}(1+\frac{1}{\tau}) & 0 \\ -(1+\frac{1}{\tau}) & -1 & 0 & -\frac{g}{\beta}(1+\frac{1}{\tau}) \\ \frac{1}{\tau} & 0 & -\frac{1}{\tau} & 0 \\ 0 & \frac{1}{\tau} & 0 & -\frac{1}{\tau} \end{pmatrix},$$

and it has eigenvalues $\lambda_{1,2}$ with $\mathrm{Re}(\lambda_{1,2}) < 0$ and $\lambda_{3,4} = \pm i\omega$,

$$\omega = \frac{1}{\tau}\sqrt{\frac{g(\tau+1)}{\beta} - 1}. \tag{3.6}$$

The system (2.1) has an equilibrium $\mathbf{e}_I$ for any $I \in \boldsymbol{R}$; we translate $\mathbf{e}_I$ to the origin with the change of variables $v_j = u_j - u_I$, $b_j = a_j - u_I$ $(j = 1, 2)$ and obtain a system topologically equivalent to (2.1),

$$\begin{aligned} \dot{v}_1 &= -v_1 + S(F(u_I) - \beta v_2 - g b_1) - u_I, \\ \dot{v}_2 &= -v_2 + S(F(u_I) - \beta v_1 - g b_2) - u_I, \\ \tau \dot{b}_1 &= -b_1 + v_1, \\ \tau \dot{b}_2 &= -b_2 + v_2. \end{aligned} \tag{3.7}$$

Near u^*, $u^* \neq u_0$, we expand the nonlinear terms in (3.7) with respect to u_I and obtain for v_1 (and similarly for v_2) an equation of the form

$$\dot{v}_1 = -v_1 - S'(F(u_I)) \cdot (\beta v_2 + g b_1) + \frac{1}{2} S''(F(u_I)) \cdot (\beta v_2 + g b_1)^2 - \frac{1}{6} S'''(F(u_I)) \cdot (\beta v_2 + g b_1)^3 + h.o.t. \tag{3.8}$$

Here *h.o.t.* means the *higher order terms*. The parameter value I^* is a possible Hopf bifurcation, so we consider small perturbations about it and about the solution u^*. That is, we take

$$I - I^* = \varepsilon^2 \alpha, \qquad V(t) = \varepsilon V_0(t) + \varepsilon^2 V_1(t) + \varepsilon^3 V_2(t) + \cdots, \tag{3.9}$$

where α is the bifurcation parameter and $V = (v_1, v_2, b_1, b_2)^{\mathrm{T}}$.

The expansions of the coefficients $S^{(k)}(F(u_I))$, $k = 1, 2, 3, \ldots$, with respect to ε take the form $S'(F(u_I)) = (1 + \frac{1}{\tau})/\beta + \alpha A \varepsilon^2 + \mathcal{O}(\varepsilon^4)$, $S''(F(u_I)) = B + \mathcal{O}(\varepsilon^2)$, and $S'''(F(u_I)) = D + \mathcal{O}(\varepsilon^2)$, where A, B, and D are defined by

$$A = -\frac{F''(u^*)}{F'(u^*)^2(\beta + g + F'(u^*))}, \quad B = -\frac{F''(u^*)}{F'(u^*)^3}, \quad D = \frac{3F''(u^*)^2 - F'(u^*) \cdot F'''(u^*)}{F'(u^*)^5} \tag{3.10}$$

(see Appendix A for more details). Let us introduce the following notation: for two vectors $U = (v_1, v_2, b_1, b_2)^{\mathrm{T}}$ and $W = (w_1, w_2, c_1, c_2)^{\mathrm{T}}$ we define first the quantities $\tilde{v}_{ij} = \beta v_j + g b_i$, $\tilde{w}_{ij} = \beta w_j + g c_i$, and then, using the scalars from (3.10), we define the operators

$$L_0 U = \frac{dU}{dt} - \mathcal{A}_0 U, \quad \Lambda U = -\alpha A \, (\tilde{v}_{12}, \tilde{v}_{21}, 0, 0)^{\mathrm{T}},$$
$$\mathcal{B}(U, W) = \frac{B}{2} (\tilde{v}_{12}\tilde{w}_{12}, \, \tilde{v}_{21}\tilde{w}_{21}, \, 0, \, 0)^{\mathrm{T}}, \quad \mathcal{C}(U, U, U) = -\frac{D}{6} (\tilde{v}_{12}^3, \tilde{v}_{21}^3, 0, 0)^{\mathrm{T}}.$$

Then based on (3.8), (3.9), and (3.10), we write system (3.7) as

$$L_0 V_0 = \varepsilon [\mathcal{B}(V_0, V_0) - L_0 V_1] + \varepsilon^2 [\mathcal{C}(V_0, V_0, V_0) + 2\mathcal{B}(V_0, V_1) + \Lambda V_0 - L_0 V_2] + \mathcal{O}(\varepsilon^3).$$

The construction of the normal form relies on an algorithm that we describe in Appendix A (see also [4, 5] for a similar approach) and that involves tedious calculations; we present here only the main result.

Theorem 3.2. *Let us assume that conditions (3.3) and (3.4) are true (sufficiently strong coupling and adaptation-dominated system), and take $I^* \in \{I^*_{hb}, I^{**}_{hb}\}$, $u^* \in \{u^*_{hb}, u^{**}_{hb}\}$ as in Theorem 3.1. Then the system (2.1) has in the neighborhood of I^* the normal form*

$$\dot{z} = A\varphi (I - I^*) \, z - \mathcal{L} z^2 \bar{z}, \tag{3.11}$$

where z is a complex variable and

$$\mathcal{L} = 4\tau^2\omega^2 \frac{|\varphi|^2}{|\psi|^2} \varphi\psi \left(\frac{\beta + g}{2} + i\beta\omega\tau \right) B^2 + 2\tau^2\omega^2 |\varphi|^2 \varphi \left(\frac{2(\beta + g)}{1 + (\frac{g}{\beta} + 1)(1 + \frac{1}{\tau})} B^2 - D \right)$$

with $\varphi = \frac{\beta}{2} + i\frac{g-\beta}{2\tau\omega}$, $\psi = 2 - \frac{3g}{\beta} + \frac{1}{\tau}\left(5 - \frac{3g}{\beta}\right) - 4i\omega(\tau+1)$, *and* A, B, D *as in* (3.10).

In order to show that I^*_{hb} and I^{**}_{hb} are supercritical Hopf bifurcation points, we need to check that the coefficient of $z^2\bar{z}$ has negative real part, i.e., that $\mathrm{Re}(\mathfrak{L}) > 0$.

As we prove in the appendix, for our range of parameters β, g, and τ, the first term in $\mathfrak{L}$ has indeed positive real part (see (A.3)); on the other hand, the real part of the second term (see (A.4)) is larger than

$$\beta\tau^2\omega^2|\varphi|^2B^2F'(u^*)\left(\frac{F'(u^*)F'''(u^*)}{F''(u^*)^2} - 2\right).$$

Remark 3.1. *The inverse of function* S *defined by* (2.2) *satisfies the condition*

$$F'(u^*)F'''(u^*) > 2F''(u^*)^2. \tag{3.12}$$

Moreover, the inequality (3.12) *is true not only for* u^* *but for all* $u \in (0,1)$.

We use this observation to state our next result.

Theorem 3.3. *Let us assume the same hypotheses as in Theorem* 3.2. *Given at* $u^* \in \{u^*_{hb}, u^{**}_{hb}\}$ *the property* (3.12) *for the gain function* S, *the input value* I^* *is a supercritical Hopf bifurcation point for system* (2.1). *The stable limit cycle occurs on the left side of* I^{**}_{hb} *(that is, for sufficiently close* $I < I^{**}_{hb}$*) and on the right side of* I^*_{hb} *(*$I > I^*_{hb}$*).*

Proof. The Hopf bifurcation is supercritical since $\mathrm{Re}(\mathfrak{L}) > 0$ in the normal form; the nondegeneracy condition is $\mathrm{Re}(A\varphi) = A\beta/2 \neq 0$.

The sign of A is opposite to the sign of $F''(u^*)$, so $A > 0$ for u^*_{hb} and $A < 0$ for u^{**}_{hb}. Consequently the sign of $\mathrm{Re}(A\varphi(I - I^*))$ that shows the direction of limit cycle bifurcation is positive for $I > I^*_{hb}$ and $I < I^{**}_{hb}$. ∎

Remark 3.2. *It is possible to obtain supercritical Hopf bifurcation points for other types of gain-function than* (3.12), *as long as* $\mathrm{Re}(\mathfrak{L})$ *has positive value. When* (3.12) *is not valid, the sign of* $\mathrm{Re}(\mathfrak{L})$ *will be computed directly from the definition formula of* $\mathfrak{L}$.

3.2. Antiphase oscillations. The stable limit cycle that exists in the neighborhood of the bifurcation points I^*_{hb} (for $I > I^*_{hb}$) and I^{**}_{hb} (for $I < I^{**}_{hb}$) is a periodic solution $L_1(t) = (u_1(t), u_2(t), a_1(t), a_2(t))$ of period, say, T. Due to the symmetry of the system (2.1) with respect to the group $\Gamma = \{\mathbf{1}_4, \gamma\}$ where $\mathbf{1}_4$ is the unitary 4-by-4 matrix and

$$\gamma = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$L_2(t) = \gamma L_1(t) = (u_2(t), u_1(t), a_2(t), a_1(t))$ is also a periodic solution for (2.1). L_2 results automatically from solution L_1 by relabeling the two network's populations. Moreover, it belongs to the same neighborhood of the equilibrium point as L_1 does.

Since the limit cycle born through the Hopf bifurcation is unique, the corresponding phase space trajectories of L_1 and L_2 coincide. Therefore, there exists a phase shift $\alpha_0 \in [0, T)$ such that $L_2(t) = L_1(t + \alpha_0)$. This implies $u_2(t) = u_1(t + \alpha_0)$ and $u_1(t) = u_2(t + \alpha_0)$, i.e., $u_1(t) = u_1(t + 2\alpha_0)$ for all real t [12]. The phase shift α_0 needs to satisfy $\alpha_0 = kT/2$ with k

an integer, so k is either $k = 0$ or $k = 1$. If $k = 0$, the two populations in the network will oscillate in synchrony. If $k = 1$, we have $u_2(t) = u_1(t + T/2)$ and $a_2(t) = a_1(t + T/2)$, which means an antiphase oscillation.

Let us assume for the moment that $k = 0$. Then $U(t) = (u_1(t), u_1(t), a_1(t), a_1(t))$ is a periodic solution of (2.1), and, as a consequence, $(u_1(t), a_1(t))$ is a periodic solution of the two-dimensional system

$$\dot{u}_1 = -u_1 + S(I - \beta u_1 - g a_1), \quad \tau \dot{a}_1 = -a_1 + u_1.$$

This contradicts Bendixon's criterion: the expression

$$\frac{\partial}{\partial u_1}(-u_1 + S(I - \beta u_1 - g a_1)) + \frac{\partial}{\partial a_1}\left(-\frac{a_1}{\tau} + \frac{u_1}{\tau}\right) = -1 - \beta S'(I - \beta u_1 - g a_1) - \frac{1}{\tau}$$

is always negative, so our initial assumption should be false.

Excluding the first case, the limit cycle has to be an antiphase solution of (2.1): the two populations compete indeed for the active state (see, for example, Figure 3B or 3D).

We state our conclusion in the following theorem.

Theorem 3.4. *Let us assume that conditions (3.3) and (3.4) are true and the coefficient $\mathcal{L}$ in the normal form (3.11) has positive real part. Then the stable limit cycle obtained at the supercritical Hopf bifurcation (as I crosses either I^*_{hb} or I^{**}_{hb}) corresponds to an antiphase oscillation: the limit cycle of period T satisfies $u_2(t) = u_1(t + T/2)$ and $a_2(t) = a_1(t + T/2)$ for any real t.*

3.3. Multiple equilibria for large enough inhibition. Our local analysis shows how stable oscillations occur in system (2.1)—through a Hopf bifurcation. The uniform equilibrium point $\mathbf{e}_I$ has four eigenvalues, λ_1 and λ_2 with negative real part independent of I ($\mathrm{Re}(\lambda_{1,2}) < 0$), and λ_3 and λ_4 that can cross the imaginary axis. Besides Hopf, another type of local bifurcation appears in (2.1) when one of the eigenvalues λ_3, λ_4 takes zero value, that is, when $F'(u_I) = \beta - g$. Because of the system's symmetry we expect it to be a pitchfork bifurcation.

Numerical simulations of system (2.1) reveal indeed the existence of additional equilibrium points. However, they exist for stronger (Figure 3G, $\beta = 1.1$) but not for weaker inhibition (Figure 3I, $\beta = 0.75$). We explain analytically how that happens.

Theorem 3.5. (i) *If $\beta - g < 1/S'(\theta)$, then the dynamical system (2.1) has a unique equilibrium point for all real I and this is $\mathbf{e}_I = (u_I, u_I, u_I, u_I)$.*

(ii) *For strong inhibition,*

$$\beta - g > 1/S'(\theta), \tag{3.13}$$

there are exactly two values, say, $u^*_{pf}, u^{**}_{pf} \in (0, 1)$, such that $u^*_{pf} < u_0 < u^{**}_{pf}$ and

$$F'(u^*_{pf}) = F'(u^{**}_{pf}) = \beta - g. \tag{3.14}$$

At $I^*_{pf} = H(u^*_{pf})$ and $I^{**}_{pf} = H(u^{**}_{pf})$ defined by (3.1), the equilibrium point $\mathbf{e}_I$ has a zero eigenvalue.

Proof. The condition that characterizes the equilibrium points of (2.1) is equivalent to $G(u_1) = G(u_2) = I - \beta(u_1 + u_2)$, where we define G by $G(u) = F(u) + (g - \beta)u$, $u \in (0, 1)$. We

have $\lim_{u\to 0} G(u) = -\infty$, $\lim_{u\to 1} G(u) = \infty$, and $G'(u) = F'(u) + g - \beta \geq F'(u_0) + g - \beta = 1/S'(\theta) + g - \beta$. Obviously, based on hypothesis (i), we have $G'(u) > 0$. Therefore, G is a monotonically increasing function; so it is injective, and the conclusion follows immediately. Statement (ii) results from the shape of F'. ■

By constructing the normal form of the system around the bifurcation point I^*_{pf} or I^{**}_{pf}, we prove the existence of a subcritical pitchfork bifurcation. Therefore, in the neighborhood of I^*_{pf} or I^{**}_{pf} the system (2.1) has multiple (three) equilibria. However, since the pitchfork is subcritical, the two newly born equilibrium points are unstable. In the four-dimensional eigenspace they actually possess two unstable modes (see Remark 3.4).

In some cases these nonuniform equilibria (having $u_1 \neq u_2$) might change their stability for I between I^*_{pf} and I^{**}_{pf}. Depending on the initial condition a trajectory will be attracted either to the fixed point with $u_1 > u_2$ or to that with $u_1 < u_2$, so one population is dominant and the other is suppressed forever. We call this type of dynamics in (2.1) *winner-take-all behavior*. The issue of the existence of the winner-take-all regime will be discussed separately in section 4.2. In this section we focus only on the mechanism that introduces additional equilibrium points to system (2.1).

Theorem 3.6. *Let us assume $\beta/(1+\frac{1}{\tau}) > \beta - g > 1/S'(\theta)$ and take $I^\circ \in \{I^*_{pf}, I^{**}_{pf}\}$, $u^\circ \in \{u^*_{pf}, u^{**}_{pf}\}$ as in (3.14), $I^\circ = F(u^\circ) + (\beta + g)u^\circ$. Then the system (2.1) has in the neighborhood of I° the normal form*

$$\dot{z} = \frac{(I - I^\circ)F''(u^\circ)}{2\beta[g(\tau+1) - \beta]}\, z + \frac{(\beta - g)^2}{6[g(\tau+1) - \beta]}\left(F'(u^\circ)F'''(u^\circ) - \frac{3}{2}F''(u^\circ)^2 + \frac{3g}{2\beta}\right) z^3. \tag{3.15}$$

Moreover, if the gain function S satisfies (3.12) at $u^\circ \in \{u^*_{pf}, u^{**}_{pf}\}$, then I° is a subcritical pitchfork bifurcation point for the system (2.1). Two additional unstable equilibrium points occur on the left side of I^{**}_{pf} ($I < I^{**}_{pf}$) and on the right side of I^*_{pf} ($I > I^*_{pf}$).

Proof. The construction of the normal form is sketched in Appendix B. Since (3.12) is true, the coefficient of z^3 in the normal form is positive and the pitchfork is subcritical. Additional equilibrium points appear for $(I - I^\circ)F''(u^\circ) < 0$ with $F''(u^\circ)$ negative at u^*_{pf} and positive at u^{**}_{pf}. ■

Remark 3.3. *In case of adaptation-dominated systems (when condition (3.3) or, equivalently, $\beta/(1+\frac{1}{\tau}) > \beta - g$ is true) we conclude the following: (1) for weak inhibition ($\beta - g < \beta/(1+\frac{1}{\tau}) < 1/S'(\theta)$), system (2.1) has a unique equilibrium point $\mathbf{e}_I$ which is asymptotically stable for all I; (2) for some intermediate value of inhibition ($\beta - g < 1/S'(\theta) < \beta/(1+\frac{1}{\tau})$), the system still has a unique equilibrium point for all I but this becomes unstable in the interval (I^*_{hb}, I^{**}_{hb}). However in order to obtain this case we need to properly adjust the maximum gain to the adaptation parameters (i.e., we need $S'(\theta) > 1/(\tau g)$); (3) for strong inhibition ($1/S'(\theta) < \beta - g < \beta/(1+\frac{1}{\tau})$), additional equilibrium points occur in system (2.1) for $I > I^*_{pf}$ and $I < I^{**}_{pf}$.*

Remark 3.4. *In case of strong inhibition and adaptation-dominated system we obtain $u^*_{hb} < u^*_{pf} < u_0 < u^{**}_{pf} < u^{**}_{hb}$ and*

$$I^*_{hb} < I^*_{pf} < I_0 < I^{**}_{pf} < I^{**}_{hb}.$$

(Note that $I_0 = H(u_0)$ is independent of β.) At each I between I^*_{hb} and I^{**}_{hb}, the equilibrium point $\mathbf{e}_I$ has at least one eigenvalue of positive real part. In fact for $I \in (I^*_{hb}, I^*_{pf}) \cup (I^{**}_{hb}, I^{**}_{pf})$,

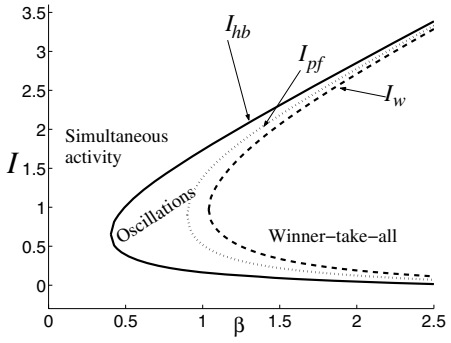


Figure 4. *Dynamical regimes in system (2.1) as inhibition strength β and stimulus strength I vary (other parameters are fixed: $g = 0.5$, $\tau = 100$, $r = 10$, $\theta = 0.2$). To the left of curve I_{hb} (solid line) the system has a unique stable equilibrium, and this satisfies $u_1 = u_2$ (simultaneous activity); in the region between curves I_{hb} and I_w (dashed line) the system oscillates; then to the right of I_w the system has a winner-take-all behavior (two stable and one unstable equilibria). The curve I_{pf} (dotted line) indicates a transition from one equilibrium to multiple equilibria in (2.1). While the attractor's type (limit cycle) does not change between I_{hb} and I_w, the number of equilibria does: we find one unstable equilibrium between I_{hb}, I_{pf} and three unstable equilibria between I_{pf}, I_w. The turning points of curves I_{hb}, I_{pf}, and I_w are obtained at $\beta = \frac{1+1/\tau}{S'(\theta)} = 0.404$, $\beta = g+1/S'(\theta) = 0.9$, and $\beta_{wta} = 1.0387$, respectively.*

it has exactly two eigenvalues with positive real part and for $I \in (I^*_{pf}, I^{**}_{pf})$ it has only one eigenvalue with positive real part. Due to the multidimensionality of the eigenspace, at I^*_{pf} and I^{**}_{pf} the equilibrium $\mathbf{e}_I$ does not actually change its stability (even if an eigenvalue takes zero value). Instead two new (nonuniform) equilibria are born (e.g., Figure 3G). The two nonuniform equilibria inherit the number of unstable modes from their "parent"–fixed point, which means they have exactly two unstable modes.

As we see, condition (3.13) is necessary but not sufficient to obtain a winner-take-all behavior in system (2.1). In section 4.2, equations (4.11) and (4.10), we will determine the minimum value of β for which winner-take-all exists (β_{wta}) and the corresponding values I^*_w, I^{**}_w where transition from oscillation to winner-take-all dynamics takes place. We summarize all these results in Figure 4 by drawing the bifurcation diagram in the parameter plane (I, β).

4. Release, escape, and winner-take-all mechanisms in neuronal competition models.

As we mentioned in section 1, some common features are observed for a large class of neuronal competition models based on mutual inhibition and slow negative feedback process. An important example is the nonmonotonic dependency of the rivalry-oscillation's period T on the stimulus strength I: in a range of small values for I the period increases with input strength; however, there exists another range for I, at larger values, where the period decreases with stimulus strength. These two dynamical regimes are usually separated by another one that is nonoscillatory; it occurs for sufficiently strong inhibition and corresponds to winner-take-all behavior (see Figures 3F and 3H).

The goal of this section is to characterize the underlying mechanisms of the above dynamical scheme. We aim to understand what causes the two opposite rivalry dynamics: as we will see, a *release* kind of mechanism is associated with the increasing branch of the T versus I curve (region IV in Figure 3F); on the other hand, for the decreasing branch of the I-T curve

(region II in Figure 3F), an *escape* mechanim is responsible.

The terms *release* and *escape* were previously introduced by [36] for inhibition-mediated rhythmic patterns in thalamic model neurons and then extended and refined by [32] for Morris–Lecar equations. Most cases include an autocatalytic process either intrinsic by voltage-gated persistent inward currents or synaptic by intrapopulation recurrent excitation. In neuronal competition model (2.1) mutual inhibition plays the role of autocatalysis: one population inhibits the network partner that inhibited it; thus the combination of these two negative factors has a positive effect on its own activity. Rhythmicity is obtained due to a fast positive feedback (disinhibition) and a slow negative feedback process. The slow negative feedback process can be either an intrinsic property of the neuronal populations (e.g., spike frequency adaptation as in (2.1)) or a property of the inhibitory connections between them (e.g., synaptic depression). A simplified model similar to (2.1) but with synaptic depression and a Heaviside step-gain function was analytically investigated in [33]. Numerical results for models with synaptic depression and smooth sigmoid gain functions were also reported in [33] and [31].

In the context of neuronal competition models we define *release* and *escape mechanisms* as follows: The two populations in the network oscillate in antiphase competing for the active state; for the dominant population of variable, say, u_1, the net input $I-\beta u_2-ga_1$ decreases as the slow negative feedback accumulates; on the contrary, for the suppressed population u_2 the feedback recovers (decays) so the net input $I-\beta u_1-ga_2$ increases. However, since the function S is highly nonlinear, equal changes in the net input $I-\beta u_j-ga_k$ of both populations can lead to drastically different changes in the corresponding effective response $S(I-\beta u_j-ga_k)$. This transformation has a direct influence on the variation of u_1 and u_2. The switch in dominance is due either to a significant, more abrupt change (decrease) in the response to an input to the dominant population, or, on the contrary, to a significant change (increase) in the response to an input to the suppressed population. In the first case the dominant population loses control, its activity drops, and it no longer suppresses its competitor, which becomes active. We call this mechanism *release*. In the latter case, when the input-output function S of the suppressed population changes faster, this population regains control, its activity rises, and it forces its competitor into the inhibited state. We call this mechanism *escape*.

Intuitively, escape occurs for higher stimulus ranges than release. Therefore, we expect that an escape (release) mechanism underlies the dynamics in region II (region IV) in Figure 3F with decreasing (increasing) I-T curve. For large values of I the gain function for the dominant population is relatively constant and close to 1 while that for the suppressed population falls in the interval where it is steeper. For example, let us consider the fast plane (u_1, u_2) and assume that u_1 is ON and u_2 is OFF; then the dominance switching point is on the shallow part of the active population nullcline $u_1 = S(I-\beta u_2-ga_1)$ and on the steeper part of the down population nullcline $u_2 = S(I-\beta u_1-ga_2)$ (see the animation 70584_01.gif [3.7MB] in Appendix C). A larger variation in u_2 than in u_1 is expected, and that corresponds to the escape mechanism. For small values of I the gain function for the dominant population is steeper (the steeper part of active population nullcline $u_1 = S(I-\beta u_2-ga_1)$), while the gain function for the suppressed population is relatively constant and close to 0 (the shallow part of the down population nullcline $u_2 = S(I-\beta u_1-ga_2)$)—see the animation 70584_02.gif [3.8MB] in Appendix C. That is what we call release.

For sufficiently large inhibition β, at intermediate I the effective response to an input to

both populations might end up being relatively constant: closer to 1 for the active population and closer to 0 for the down population. The switching does not take place anymore; instead a *winner-take-all* dynamics is obtained.

In the following we investigate analytically how the period of oscillations for system (2.1) depends on the input strength and show that the increasing (decreasing) branch of the I-T curve is associated with the release (escape) mechanism. Our analysis is done in two steps: first, in section 4.1, we consider the limiting case sigmoid function $S(x) = Heav(x-\theta)$ such that $S(x) = 0$ if $x < \theta$ and $S(x) = 1$ if $x > \theta$; the function does not obey the hypotheses in section 2, but it provides a useful example where escape, release, and winner-take-all dynamics are easily characterized. Then in section 4.2 we return to the case of smooth sigmoid function and describe the notions defined above in this more general context. We find a precise mathematical characterization for the minimum value of β and then for the corresponding input values, say, I_w^* and I_w^{**}, where the winner-take-all regime appears. In the absence of a winner-take-all regime we provide a mathematical definition for the transition between escape and release.

4.1. A relevant example: The Heaviside step function. The choice of $S(x) = Heav(x-\theta)$ allows us to solve completely for the intervals of the stimulus strength I where oscillations and winner-take-all dynamics exist. For system (2.1) with a Heaviside step function, there are only four possible equilibrium points: $(1,1,1,1)$, $(1,0,1,0)$, $(0,1,0,1)$, and $(0,0,0,0)$. Here $(1,1,1,1)$ and $(0,0,0,0)$ correspond respectively to the simultaneously high and low activity states observed in numerical simulations in regions I and V of Figure 3F. Oscillations can occur only between the states $(u_1,u_2) = (1,0)$ and $(u_1,u_2) = (0,1)$ in which one population is dominant and the other one is suppressed. Let us now determine the necessary and sufficient conditions for the oscillations to exist. The idea used in our analysis is similar to that in [17] and [33].

In the fast plane, the nullclines of u_1 and u_2 consist in two constant plateaus of zero and unit value discontinuously connected at a "threshold" point $(I-(\theta+ga_1))/\beta$ and $(I-(\theta+ga_2))/\beta$, respectively (Figure 5A). During oscillation, due to the change in slow variables a_1 and a_2 these thresholds move along the vertical and horizontal axes. For example, assuming $u_1 = 1$, $u_2 = 0$, the slow equations become $\tau\dot{a}_1 = -a_1 + 1 > 0$ and $\tau\dot{a}_2 = -a_2 < 0$; thus the u_1-nullcline moves down while the u_2-nullcline moves to the right. If these nullclines slide enough and the thresholds cross either 0 (for the u_1-nullcline) or 1 (for the u_2-nullcline), i.e., either $a_{1J} = (I-\theta)/g$ or $a_{2J} = (I-(\theta+\beta))/g$ are reached, then the equilibrium point $(u_1,u_2) = (1,0)$ disappears and the system will be attracted to $(u_1,u_2) = (0,1)$. The switch takes place and the slow equations change to $\tau\dot{a}_1 = -a_1 < 0$, $\tau\dot{a}_2 = -a_2+1 > 0$, now pushing the nullclines in opposite directions. As we explain below, depending on which of the two *jumping values* a_{1J} or a_{2J} is reached first, a *release* or an *escape* mechanism will underlie the oscillation.

We note that for an oscillatory solution, $u_1+u_2 = 1$ always, and so $\tau(\dot{a}_1+\dot{a}_2) = 1-(a_1+a_2)$. Asymptotically, the slow dynamics will occur along the diagonal $a_1+a_2 = 1$ of the unit square (Figure 5D or 5F). The positions the horizontal line $a_2 = (I-(\theta+\beta))/g$ and the vertical line $a_1 = (I-\theta)/g$ have relative to the unit square is important when the trajectory points to the lower-right corner; on the other hand, when the trajectory points to the upper-left corner, the position of vertical line $a_1 = (I-(\theta+\beta))/g$ and horizontal line $a_2 = (I-\theta)/g$ will matter.

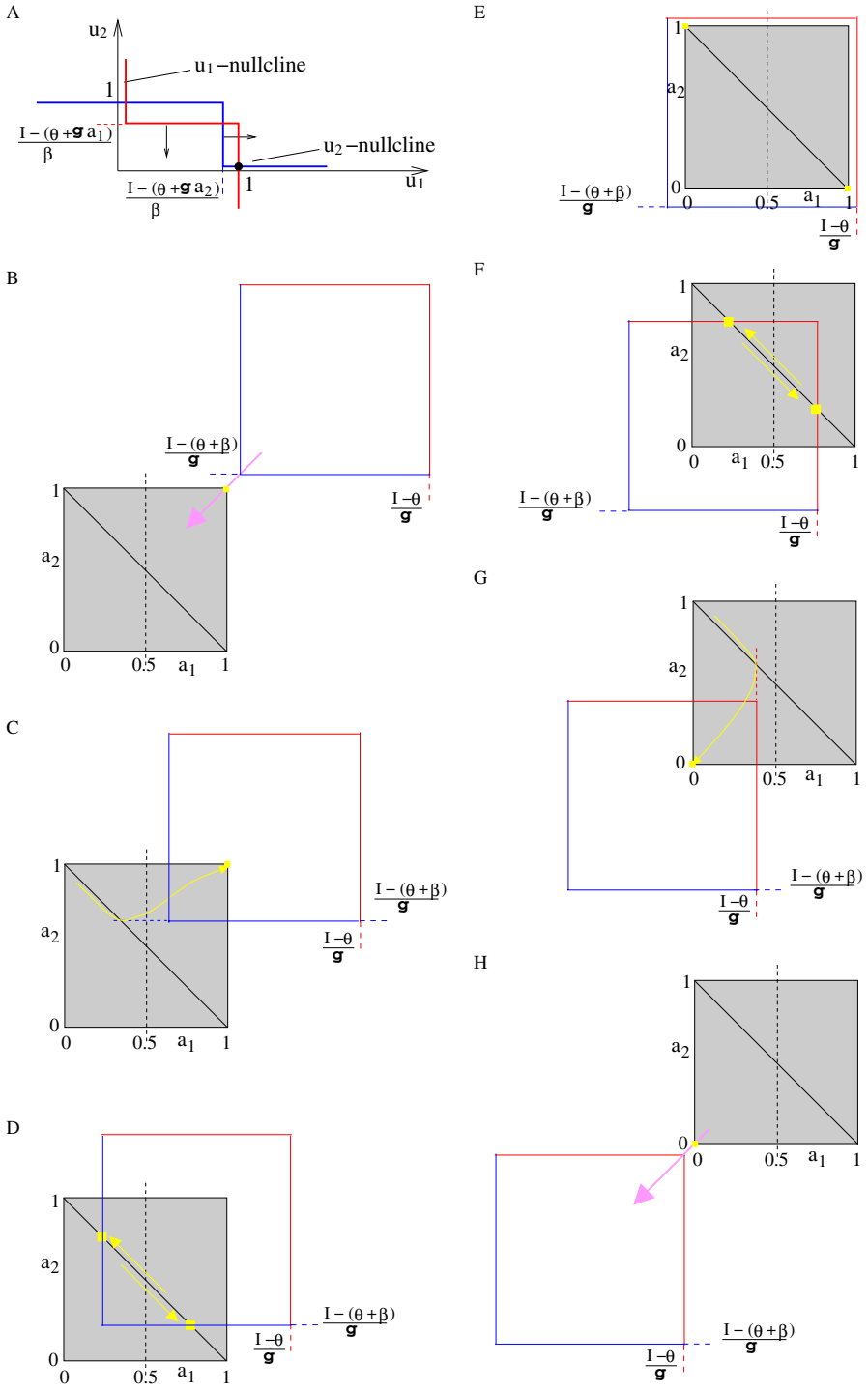


Figure 5. *System (2.1)'s dynamics for large inhibition strength ($\beta/g > 1$) and Heaviside step function.* (A) *Nullclines in the plane of fast variables (u_1, u_2);* (B–H) *System's dynamics in the slow variables plane (a_1, a_2) for:* (B) $I \geq \theta+\beta+g$; (C) $\theta+\beta+g/2 \leq I < \theta+\beta+g$; (D) $\theta+\beta < I < \theta+\beta+g/2$; (E) $\theta+g \leq I \leq \theta+\beta$; (F) $\theta + g/2 < I < \theta + g$; (G) $\theta \leq I \leq \theta + g/2$; (H) $I < \theta$.

Therefore, in the slow plane (a_1, a_2) (see Figure 5B–5H) we are interested in the intersection of the unit square (grey) with the square defined by the possible jumping values $(I-(\theta+\beta))/g$ (blue) and $(I-\theta)/g$ (red). If the red line is reached, then the active cell ($u_j = 1$) becomes suddenly inactive ($u_j = 0$), allowing its competitor to go up. That is why we call this case a release mechanism; otherwise, if the blue line is reached first, then the suppressed cell becomes suddenly active, forcing its competitor to go down. That is the escape mechanism.

As parameter I decreases, the blue-red square slides down along the first diagonal, starts to intersect the grey unit square, and then leaves it (Figure 5B–5H). The way those two squares intersect determines the system's dynamics, so we need to take into account the relative size of their sides: 1 and β/g.

Theorem 4.1 (five modes of behavior for large enough inhibition strength). *Let us assume that $\beta/g > 1$. The following five dynamical regimes exist for the neuronal competition model (2.1) with $S(x) = Heav(x-\theta)$.*

(i) *If $I \geq \theta + \beta + g/2$, then the system's attractor is the simultaneously high activity state $u_1 = u_2 = 1$.*

(ii) *If $\theta+\beta < I < \theta+\beta+g/2$, then the system oscillates between the states $(u_1, u_2) = (1, 0)$ and $(u_1, u_2) = (0, 1)$ due to an escape mechanism. The period of oscillations decreases with I and satisfies*

$$T_{escape} = 2\tau \ln\left(\frac{g}{I-(\theta+\beta)} - 1\right). \tag{4.1}$$

(iii) *If $\theta + g \leq I \leq \theta + \beta$, then the system is in a winner-take-all regime with fast variables either $(1, 0)$ or $(0, 1)$ depending on the initial condition choice.*

(iv) *If $\theta + g/2 < I < \theta + g$, then the system oscillates between the states $(u_1, u_2) = (1, 0)$ and $(u_1, u_2) = (0, 1)$ due to a release mechanism. The period of oscillations increases with I and satisfies*

$$T_{release} = 2\tau \ln\left(\frac{g}{\theta+g-I} - 1\right). \tag{4.2}$$

(v) *If $I \leq \theta + g/2$, then the system's attractor is the simultaneously low activity state $u_1 = u_2 = 0$.*

Proof. Since $\beta/g > 1$, there are exactly seven relative positions of the blue-red square to the unit square that lead to conclusions (i) to (v).

(i) For $1 \leq (I-(\theta+\beta))/g$ (Figure 5B) both a_1, a_2 are smaller than $(I-(\theta+\beta))/g$, so $I - \beta u_i - g a_k \geq I - \beta - g \geq \theta$, and we always have $u_1 = u_2 = 1$.

If $1/2 \leq (I-(\theta+\beta))/g < 1 < (I-\theta)/g$ (Figure 5C), let us assume $u_1 = 1$, $u_2 = 0$, and $a_1 + a_2 = 1$ as initial values. The slow variable a_1 increases while a_2 decreases; in this case only a_2 can cross the horizontal blue line, producing the jump of u_2 from 0 to 1. However, just after the jump, in the equation of u_1 we have $I - \beta u_2 - g a_1 = I - \beta - g(1 - a_{2J}) = 2(I-(\beta+\theta)) - g + \theta \geq \theta$, which keeps u_1 at its value 1. Therefore, the point will be attracted to the corner $(a_1, a_2) = (1, 1)$ and the oscillation dies.

(ii) For $0 < (I-(\theta+\beta))/g < 1/2 < 1 < (I-\theta)/g$ (Figure 5D), a_2 will first cross the blue line and induce a sudden change in u_2. Then in the u_1-equation we obtain $I - \beta u_2 - g a_1 =$

$I-\beta-g(1-a_{2J})=2(I-(\beta+\theta))-g+\theta<\theta$, which forces u_1 to take zero value. The fast system switches from $(1,0)$ to $(0,1)$ and the slow dynamics changes its direction of movement along the upper-left–lower-right diagonal of the unit square. The change $I-\beta u_1-ga_2=I-ga_2\geq I-g\geq\theta$ does not affect the new value $u_2=1$; oscillation exists indeed and its projection on the slow plane is the segment defined by the intersection of the unit square's secondary diagonal with the two blue lines. When I decreases, the length of this segment increases, so it will take longer to go from one endpoint to the other. We expect the period T to increase as I decreases. Indeed, at the jumping point, the slow variable for the suppressed population takes the value $a_f:=a_{2J}=(I-(\theta+\beta))/g$; however, just after the previous jump it was $a_i:=1-a_{1J}=1-(I-(\theta+\beta))/g$. Therefore, from $\tau\dot{a}_2=-a_2$ we compute the solution $a(t)=a_ie^{-t/\tau}$, $t\in(0,T/2)$. The period T of oscillation is $T=2\tau\ln(a_i/a_f)$, i.e., exactly (4.1). Moreover, $dT/dI<0$, so T decreases with I.

(iii) By choosing I such that $(I-(\theta+\beta))/g<0<1\leq(I-\theta)/g$, the unit square falls completely inside the blue-red square (Figure 5E). Starting at $u_1=1$, $u_2=0$ we have $I-\beta u_2-ga_1=I-ga_1\geq I-g\geq\theta$ and $I-\beta u_1-ga_2=I-\beta-ga_2\leq I-\beta<\theta$; the slow variables a_1 and a_2 continue to increase, respectively, decrease, approaching the lower-right corner of the unit square, and a winner-take-all state is achieved. For $u_1=0$, $u_2=1$, the point (a_1,a_2) will be attracted to the upper-left corner.

(iv) Decreasing I even more, we enter the region $(I-(\theta+\beta))/g<0<1/2<(I-\theta)/g<1$ (Figure 5F). Here the threshold is crossed at the red line and (when $u_1=1$, $u_2=0$) u_1 jumps from 1 to 0. In the u_2-equation, the expression $I-\beta u_1-ga_2$ becomes $I-g(1-a_{1J})=2(I-\theta)-g+\theta\geq\theta$, which leads to $u_2=1$. That is the release mechanism. In the slow plane (a_1,a_2) the point changes its direction of movement; the oscillation occurs along the segment defined by the intersection of the unit square's secondary diagonal with the two red lines. The length of this segment decreases as I decreases; it takes less time to go from one endpoint to the other, so we expect the period T to decrease. Indeed, for the release mechanism, $a_f:=a_{1J}=(I-\theta)/g$, with value just after the previous jump $a_i:=1-a_{2J}=1-(I-\theta)/g$. The slow differential equation is now $\tau\dot{a}_1=1-a_1$, that is, $a(t)=1-(1-a_i)e^{-t/\tau}$, $t\in(0,T/2)$. The period T of oscillation satisfies $T=2\tau\ln((1-a_i)/(1-a_f))$, or (4.2). Obviously $dT/dI>0$.

(v) Oscillations cannot exist anymore for $(I-(\theta+\beta))/g<0\leq(I-\theta)/g\leq1/2$ (Figure 5G). Say, again, that we have $u_1=1$, $u_2=0$: only a_1 can reach the threshold (red) line for u_1 to become inactive. However, just after the jump, in the u_2-equation the net input $I-\beta u_1-ga_2=I-g(1-a_{1J})=2(I-\theta)-g+\theta$ is still less than θ, and it keeps u_2 at zero. Both slow variables start to decrease approaching the corner $(a_1,a_2)=(0,0)$. The oscillation dies, and the fast system has $(0,0)$ as a single attractor. For even smaller values of input strength, $(I-(\theta+\beta))/g<(I-\theta)/g<0$ (Figure 5H), it is true that $I-\beta u_i-ga_k\leq I<\theta$ and $u_1=u_2=0$. ■

Remark 4.1. *We note that in the escape regime, $T\to0$ as $I\to\theta+\beta+g/2$ and $T\to\infty$ as $I\to\theta+\beta$. Similarly, in the release regime, $T\to0$ as $I\to\theta+g/2$ and $T\to\infty$ as $I\to\theta+g$.*

Case (iii) in Theorem 4.1 (case (E) in Figure 5) is not possible if the side of the blue-red square is smaller than the unit. Therefore, the winner-take-all dynamics is eliminated. On the other hand, the blue-red square can lie completely inside the unit square. Then we need to distinguish between two cases: the point moving along the secondary diagonal of the unit square may first reach the blue line (Figure 6A) or the red line (Figure 6B).

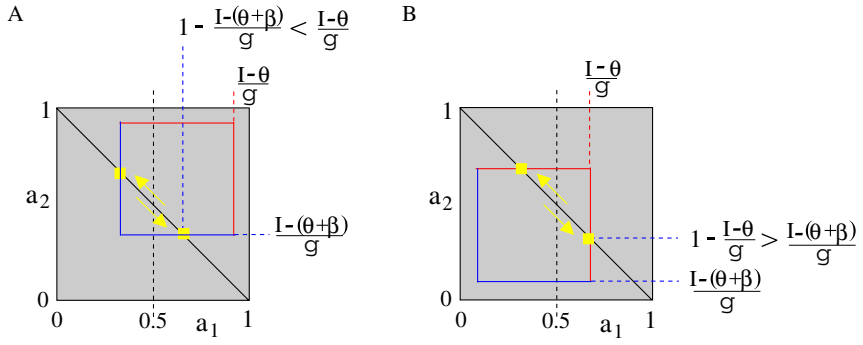


Figure 6. (A) *Escape and* (B) *release mechanisms for low inhibition strength* ($\beta/g < 1$) *in system* (2.1) *with Heaviside step function.*

Theorem 4.2 (no winner-take-all regime for low inhibition strength). *Let us assume that* $\beta/g < 1$. *Then the neuronal competition model* (2.1) *with* $S(x) = Heav(x - \theta)$ *exhibits only four dynamical regimes.*

If $I \geq \theta + \beta + g/2$, *then the system's attractor is the high activity state* $u_1 = u_2 = 1$. *Similarly, if* $I \leq \theta + g/2$, *the system's attractor is the low activity state* $u_1 = u_2 = 0$.

For intermediate values of input strength, the system oscillates between the states $(u_1, u_2) = (1, 0)$ *and* $(u_1, u_2) = (0, 1)$. *If* $\theta + (\beta + g)/2 < I < \theta + \beta + g/2$, *oscillations occur due to an escape mechanism; the period* T *decreases with* I *and satisfies* (4.1). *If* $\theta + g/2 < I < \theta + (\beta + g)/2$, *then a release mechanism underlies the oscillations and* T *increases with* I *according to* (4.2).

Moreover, at $I_{Tmax} = \theta + (\beta + g)/2$ *the system has maximum oscillation period*

$$T_{max} = 2\tau \ln\left(\frac{1 + \beta/g}{1 - \beta/g}\right).$$

Proof. Let us assume again initial conditions $u_1 = 1$, $u_2 = 0$, and $a_1 + a_2 = 1$; therefore, a_1 increases and a_2 decreases. In order to first reach the blue line (Figure 6A), the inequality $1 - (I - (\theta + \beta))/g < (I - \theta)/g$, i.e., $I > I_{Tmax}$, must be true. When the blue line is reached, the down variable u_2 switches from 0 to 1. If $(I - (\theta + \beta))/g < 1/2$, the net input for u_1 becomes $N_e := I - \beta u_2 - g a_1 = I - \beta - g(1 - a_{2J}) = 2(I - \beta - \theta) - g + \theta < \theta$, so u_1 changes to 0 and oscillation occurs due to an escape mechanism. If $(I - (\theta + \beta))/g \geq 1/2$, then $N_e \geq \theta$ and u_1 remains 1; the fast system has $(1, 1)$ as an attractor. Similar arguments are used for the release mechanism; the condition for intersection with the red line (Figure 6B) is equivalent to the inequality $1 - (I - \theta)/g > (I - (\theta + \beta))/g$, i.e., $I < I_{Tmax}$. For both (4.1) and (4.2), $T \to T_{max}$ as $I \to \theta + (\beta + g)/2$. Also, $T_{escape} \to 0$ as $I \to \theta + \beta + g/2$ and $T_{release} \to 0$ as $I \to \theta + g/2$. ■

4.2. Global features of the competition model with smooth sigmoid gain function. Numerical simulations of system (2.1) with smooth gain function S indicate that the limit cycle born through the Hopf bifurcation as in section 3 takes a relaxation-oscillator form just beyond the bifurcation (Figure 3B and 3D). That is, because of the two time-scales involved in the system, variables u_1 and u_2 evolve much faster than a_1 and a_2 ($\tau \gg 1$). We use this

observation to describe for (2.1) the relaxation-oscillator solution in the singular limit $1/\tau = 0$. In the plane (a_1, a_2) of slow variables we construct the curve of "jumping" points from the dominant to the suppressed state of each population, and back. This curve is the equivalent of the blue-red square from the case of Heaviside step function, and similarly it consists of two arcs associated with an escape and release mechanism, respectively. Then we give analytical conditions for the winner-take-all regime to exist.

4.2.1. The singular relaxation-oscillator solution. For large values of τ let us consider the slow time $s = \varepsilon t$ ($\varepsilon = 1/\tau$; $' = d/ds$) and rewrite system (2.1) as $\varepsilon u_1' = -u_1 + S(I - \beta u_2 - g a_1)$, $\varepsilon u_2' = -u_2 + S(I - \beta u_1 - g a_2)$, $a_1' = -a_1 + u_1$, $a_2' = -a_2 + u_2$.

In the singular perturbation limit $\varepsilon = 0$ any solution will belong to the slow manifold Σ defined by $-u_1 + S(I - \beta u_2 - g a_1) = 0$, $-u_2 + S(I - \beta u_1 - g a_2) = 0$ or, based on the inverse property of S ($S^{-1} = F$),

$$\Sigma = \Big\{ (u_1, u_2, a_1, a_2) \ : \ u_1, u_2 \in (0,1),\ a_1, a_2 \in \mathbf{R},\ \text{and} \\ u_2 = S(I - \beta u_1 - g a_2),\ a_1 = \frac{1}{g}[I - F(u_1) - \beta S(I - \beta u_1 - g a_2)] \Big\}. \tag{4.3}$$

The surface Σ is multivalued and can be visualized by plotting a_1 as a function of u_1 and a_2 (Figure 7).

Then the slow dynamics is according to equations $a_1' = -a_1 + \tilde{u}_1(a_1, a_2)$, $a_2' = -a_2 + \tilde{u}_2(a_1, a_2)$, where $(\tilde{u}_1, \tilde{u}_2, a_1, a_2) \in \Sigma$. The "slow" nullclines are characterized by additional conditions: $a_1 = u_1$ for a_1-nullcline ($\mathcal{N}_1$) and $a_2 = u_2$ for a_2-nullcline ($\mathcal{N}_2$); geometrically these are two curves situated on the surface Σ and defined by

$$\mathcal{N}_1(a_1' = 0) = \Big\{ (u_1, u_2, a_1, a_2) \ : \ u_1 = a_1,\ u_2 = S(I - \beta a_1 - g a_2(a_1)), \\ a_2 = \frac{1}{g}\left[I - \beta a_1 - F\left(\frac{I - g a_1 - F(a_1)}{\beta} \right) \right] \ \text{with } a_1 \in (\alpha_1, \alpha_2) \Big\} \tag{4.4}$$

and

$$\mathcal{N}_2(a_2' = 0) = \Big\{ (u_1, u_2, a_1, a_2) \ : \ u_1 = S(I - \beta a_2 - g a_1(a_2)),\ u_2 = a_2, \\ a_1 = \frac{1}{g}\left[I - \beta a_2 - F\left(\frac{I - g a_2 - F(a_2)}{\beta} \right) \right] \ \text{with } a_2 \in (\alpha_1, \alpha_2) \Big\}. \tag{4.5}$$

Here $\alpha_1, \alpha_2 \in (0,1)$ are the unique solutions of the equations $F(\alpha_1) + g\alpha_1 = I - \beta$ and $F(\alpha_2) + g\alpha_2 = I$, respectively.

The equilibrium points are at the nullclines' intersection, which means $a_1 = u_1$ and $a_2 = u_2$ simultaneously on Σ. They are characterized by the conditions

$$\mathcal{N}_1 \cap \mathcal{N}_2 : \ u_1 = a_1,\ u_2 = a_2,\ F(a_1) + g a_1 + \beta a_2 = F(a_2) + g a_2 + \beta a_1 = I. \tag{4.6}$$

We recall from section 3 that if $\beta > (1 + \frac{1}{\tau})/S'(\theta)$ (which in the limit case $\varepsilon = 0$ corresponds to $\beta > 1/S'(\theta)$), there exist two values of input parameter $I_{hb}^* < I_{hb}^{**}$ such that the system has

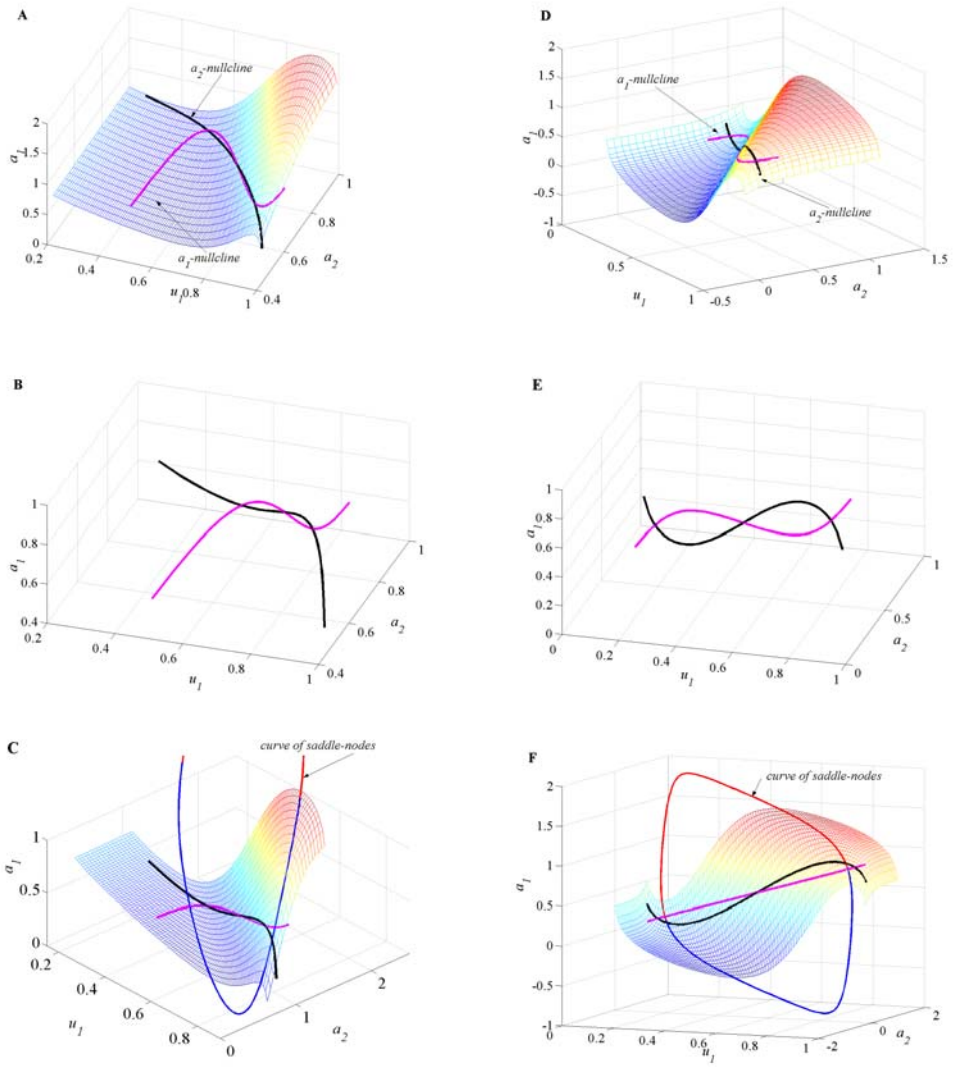


Figure 7. *Projection of the slow manifold Σ on the space (u_1, a_1, a_2). System (2.1)'s parameters are $\beta = 1.1$, $g = 0.5$, $\theta = 0.2$, $r = 10$, and $I = 1.5$ (A–C), respectively, $I = 1$ (D–F). The curve $\mathcal{SN}$ of saddle-nodes (jumping points) is represented in blue-red* (C, F). *Nullclines $\mathcal{N}_1$ (magenta) and $\mathcal{N}_2$ (black curve) intersect in three points for* (B) $I = 1.5$ *and* (E) $I = 1$. *However, either the equilibrium points can all lie on the middle branch of Σ* (A), *that is, inside the curve of saddle-nodes* (C), *or two equilibrium points can move to the lateral branches, outside of $\mathcal{SN}$* (D, F).

a unique stable equilibrium point for $I \in (-\infty, I^*_{hb}) \cup (I^{**}_{hb}, \infty)$. This equilibrium point takes the form $\mathbf{e}_I = (u_I, u_I, u_I, u_I)$ with $u_I \in (0,1)$, $F(u_I) + (\beta + g)u_I = I$, and it becomes unstable at I^*_{hb} and I^{**}_{hb}; a stable limit cycle appears for $I > I^*_{hb}$ and $I < I^{**}_{hb}$. The critical parameter values are defined by $I^*_{hb} = F(u^*_{hb}) + (\beta + g)u^*_{hb}$ and $I^{**}_{hb} = F(u^{**}_{hb}) + (\beta + g)u^{**}_{hb}$ with u^*_{hb}, u^{**}_{hb} according to (3.5). Again, in the limit case $\varepsilon = 0$, condition (3.5) becomes

$$F'(u^*_{hb}) = F'(u^{**}_{hb}) = \beta, \tag{4.7}$$

and obviously we have $F'(u_I) < \beta$ for each $I \in (I^*_{hb}, I^{**}_{hb})$.

Moreover, if $\beta - g > 1/S'(\theta)$, two additional equilibrium points occur for $I > I^*_{pf}$ and $I < I^{**}_{pf}$ through a pitchfork bifurcation. However, at least in the neighborhood of I^*_{pf} or I^{**}_{pf}, these equilibrium points are unstable. Their coordinates (u_{1p}, u_{2p}) satisfy (4.6), so either $u_{1p} < u_I < u_{2p}$ or $u_{1p} > u_I > u_{2p}$. In addition, they are close to u_I for all I sufficiently close to I^*_{pf} or I^{**}_{pf}; in conclusion, at least in a small region, $F'(u_{1p}) < \beta$ and $F'(u_{2p}) < \beta$.

In Figure 7 we plotted the projection on the three-dimensional space (u_1, a_1, a_2) of the slow manifold Σ and the nullclines $\mathcal{N}_1$ and $\mathcal{N}_2$ ($\mathcal{N}_1$ is colored in magenta and $\mathcal{N}_2$ is colored in black; because Σ is defined over the full range $-\infty$ to ∞ with respect to a_1 and a_2, we plot only part of it). All plots are done for S as in (2.2), $\beta = 1.1$, $g = 0.5$, $r = 10$, $\theta = 0.2$, and two values of input strength, $I = 1.5$ (Figure 7A–C) and $I = 1$ (Figure 7D–F). For these values of parameters we have $I^*_{pf} = 0.4064$ and $I^{**}_{pf} = 1.5936$, so at both $I = 1.5$ and $I = 1$ the system has three equilibrium points; nullclines intersect three times and the middle intersection point is exactly $\mathbf{e}_I$. We note that all equilibrium points are unstable at $I = 1.5$ while two of them become stable at $I = 1$ (see also Figure 3G).

On the surface Σ we have $\frac{\partial u_1}{\partial a_1} = g/[\beta^2 S'(I - \beta u_1 - g a_2) - F'(u_1)]$ and $\frac{\partial u_1}{\partial a_2} = -\beta g/[\beta^2 - F'(u_1)/S'(I - \beta u_1 - g a_2)]$, i.e.,

$$\frac{\partial u_1}{\partial a_1} = \frac{g F'(u_2)}{\beta^2 - F'(u_1)F'(u_2)} \quad \text{and} \quad \frac{\partial u_1}{\partial a_2} = -\frac{\beta g}{\beta^2 - F'(u_1)F'(u_2)}.$$

For an initial condition (u_1, u_2, a_1, a_2) with u_1 sufficiently large (close to 1) we have $\frac{\partial u_1}{\partial a_1} < 0$ and $\frac{\partial u_1}{\partial a_2} > 0$; therefore, a simultaneous increase in a_1 and decrease in a_2 lead to a decrease in u_1. Similarly, for a choice of u_1 sufficiently low (close to 0), $\frac{\partial u_1}{\partial a_1} < 0$, $\frac{\partial u_1}{\partial a_2} > 0$, and a simultaneous decrease in a_1 and increase in a_2 lead to an increase in u_1. On the other hand, the limit cycle exists for some $I \in (I^*_{hb}, I^{**}_{hb})$. Since here $\mathbf{e}_I \in \Sigma$ with $F'(u_I) < \beta$, we have $\frac{\partial u_1}{\partial a_1} > 0$ and $\frac{\partial u_1}{\partial a_2} < 0$ in a neighborhood of this equilibrium point; the behavior of u_1 with respect to a_1 and a_2 is opposite that previously described.

Therefore, relative to the plane (u_1, a_1) the surface Σ has a cubic-like shape: its left and right branches decrease with u_1 while the middle branch increases. The curve of lower $(u_1 < u_2)$ and upper $(u_1 > u_2)$ knees on Σ is defined by $F'(u_1)F'(u_2) = \beta^2$ (blue-red curve in Figure 7C and 7F). As we show in the following, when the trajectory on surface Σ reaches the curve of knees on its upper side, the point will jump from the right branch of Σ to its left branch. Similarly, when it reaches the lower side of the curve of knees, the point will jump from the left to the right branch of Σ. In the plane of fast variables (u_1, u_2) the curve of knees corresponds to a saddle-node bifurcation (a node approaching a saddle then merging with it and disappearing—see the animations 70584_01.gif [3.7MB] and 70584_02.gif [3.8MB] in Appendix C). For this reason we also call it the *curve of saddle-nodes* ($\mathcal{SN}$) or the *curve of jumping points*. It is defined by the equations

$$(4.8)\quad \mathcal{SN}:\ F'(u_{1J})F'(u_{2J}) = \beta^2,\ a_{1J} = \frac{1}{g}[I - F(u_{1J}) - \beta u_{2J}],\ a_{2J} = \frac{1}{g}[I - F(u_{2J}) - \beta u_{1J}].$$

We do not prove here the existence of the relaxation-oscillator singular solution. We aim only to provide the reader with the intuition for how oscillations occur in the competition

model (2.1) if a smooth sigmoid is taken as a gain function. Thus, if the oscillations exist and, for example, u_1 is dominant and u_2 is suppressed ($u_1 > u_2$), we have $-a_1 + u_1 > 0$, $-a_2 + u_2 < 0$, so a_1 increases and a_2 decreases. They push u_1 down and the point moves on the trajectory until it reaches $\mathcal{SN}$ at an upper knee U of coordinates (u_{1J}, u_{2J}); here the derivatives $\frac{\partial u_1}{\partial a_1}$ and $\frac{\partial u_1}{\partial a_2}$ become infinite, so u_1 jumps from the upper to the lower branch of Σ (Figure 8A–8B). On the lower branch ($u_1 < u_2$) we have $-a_1 + u_1 < 0$ and $-a_2 + u_2 > 0$, so u_1 increases and the point will move due to the decrease of a_1 and increase of a_2 until it touches $\mathcal{SN}$ at L. At L the point jumps up to the right branch of Σ. The projection of the limit cycle on the slow plane (a_1, a_2) is a closed curve that touches the projection of $\mathcal{SN}$ at two points (a_{1J}, a_{2J}) and (a_{2J}, a_{1J}) symmetric to the line $a_1 = a_2$ (Figures 8C and 9B).

4.2.2. Release, escape, and winner-take-all in the competition model with smooth sigmoid gain function. Analogous to our analysis in section 4.1 we consider in the following the projection of the curve of jumping points $\mathcal{SN}$ on the slow plane (a_1, a_2). As mentioned above, if the up-to-down jump takes place at (a_{1J}, a_{2J}), then the down-to-up jump is at (a_{2J}, a_{1J}); so the projection is a closed curve symmetric to the diagonal $a_1 = a_2$.

We consider in (a_1, a_2) the curve Γ_0 of equations

$$A_1 = -\frac{1}{g}[F(u_{1J}) + \beta u_{2J}], \quad A_2 = -\frac{1}{g}[F(u_{2J}) + \beta u_{1J}]$$

with $F'(u_{1J})F'(u_{2J}) = \beta^2$. Then from (4.8) we have

$$a_{1J} = A_1 + \frac{I}{g}, \quad a_{2J} = A_2 + \frac{I}{g}.$$

For a given I, the projection of $\mathcal{SN}$ on the slow plane (say, $\Gamma = \Gamma_I$) is exactly the translation of Γ_0 with quantity $(I/g, I/g)$ along the first bisector. Obviously as I decreases, the curve Γ moves down on the upper-right–lower-left direction (Figure 9B). That is similar to the movement of the blue-red square for the case of the Heaviside step function in section 4.1 (Figure 5).

Let us consider now an oscillatory solution that exists for some $I \in (I^*_{hb}, I^{**}_{hb})$. If population 1 is dominant ($u_1 > u_2$), then $-a_1 + u_1 > 0$ and $-a_2 + u_2 < 0$ imply $u_1 - u_2 > a_1 - a_2$. On the other hand, since the trajectory is on Σ, we also have $a_1 = [I - F(u_1) - \beta u_2]/g$, $a_2 = [I - F(u_2) - \beta u_1]/g$, and thus $a_1 - a_2 = [\beta(u_1 - u_2) - (F(u_1) - F(u_2))]/g$. At the jumping point a_1 reaches its maximum and a_2 its minimum, so $0 < a_{1J} - a_{2J} < u_{1J} - u_{2J}$ can be written as

$$0 < W(u_{1J}) = \frac{1}{g}\left(\beta - \frac{F(u_{1J}) - F(u_{2J})}{u_{1J} - u_{2J}}\right) < 1. \tag{4.9}$$

The point (u_{1J}, u_{2J}), $u_{1J} > u_{2J}$, satisfies $F'(u_{1J})F'(u_{2J}) = \beta^2$ and describes the part of $\mathcal{SN}$ that corresponds to the upper knees. Let $u_{m\beta}, u_{M\beta} \in (0, 1)$ be the values defined by $u_{m\beta} < u^*_{hb} < u_0 < u^{**}_{hb} < u_{M\beta}$, $F'(u_{m\beta}) = F'(u_{M\beta}) = \beta^2/F'(u_0) = \beta^2 S'(\theta)$ (see the shape of F' in Figure 2C). Then the upper branch of $\mathcal{SN}$ results by gluing together three arcs on which (1) u_{1J} increases between u^{**}_{hb} and $u_{M\beta}$ (so u_{2J} decreases from u^{**}_{hb} to u_0); (2) u_{1J} decreases

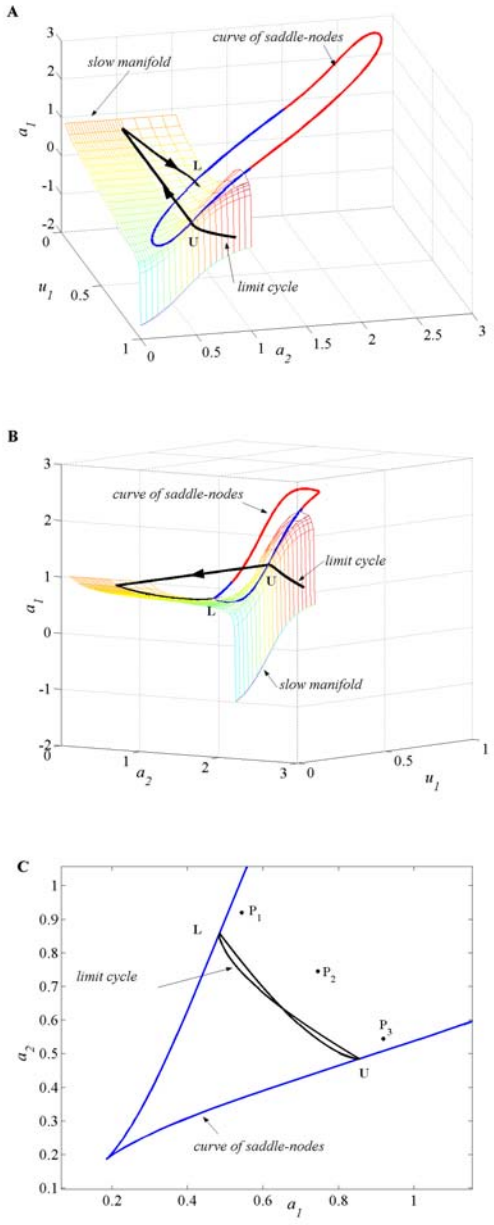


Figure 8. (A–B) *Limit cycle solution and the slow manifold* Σ *for system* (2.1) *with parameters* $I = 1.5$, $\beta = 1.1$, $g = 0.5$, $r = 10$, $\theta = 0.2$, *and* $\tau = 5000$. (C) *The projection of the limit cycle on the slow plane* (a_1, a_2); P_1, P_2, *and* P_3 *are the projections of the equilibrium points.*

from $u_{M\beta}$ to u_0 (and u_{2J} decreases from u_0 to $u_{m\beta}$); (3) u_{1J} continues to decrease from u_0 to u^*_{hb} (however, u_{2J} increases now between $u_{m\beta}$ and u^*_{hb}).

We can plot the expression $W(u_{1J})$ from (4.9) for $u_{1J} \in [u^{**}_{hb}, u_{M\beta}]$ and $u_{1J} \in [u^*_{hb}, u_{M\beta}]$ with the corresponding $u_{2J} = u_2(u_{1J})$ and check if the graph is below the horizontal line $W = 1$ (Figure 9A). If for all combinations (u_{1J}, u_{2J}) we have $W < 1$, then the winner-take-

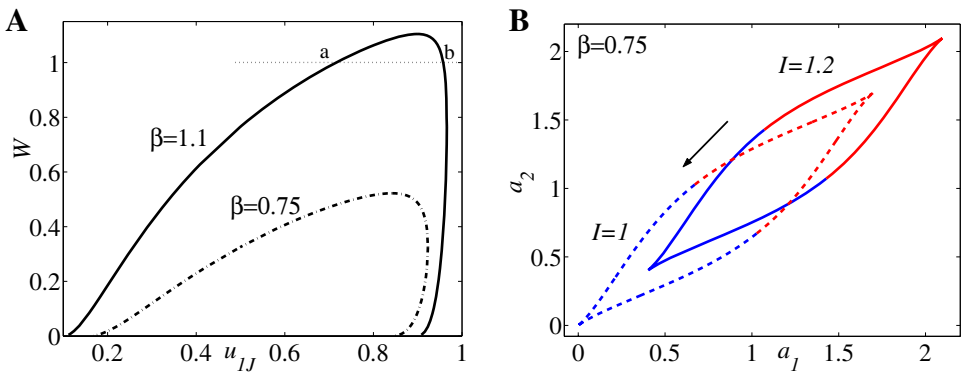


Figure 9. (A) *The graph of $W = W(u_{1J})$ for upper knees on $\mathcal{SN}$ computed for $g = 0.5$, $r = 10$, $\theta = 0.2$, and S as in (2.2). At $\beta = 0.75$ (dashed-dotted curve) the maximum value of W is less than unity: Winner-take-all regime does not exist; at $\beta = 1.1$ (thick curve) the maximum value of W is larger than unity: the values of I where winner-take-all occurs correspond to u_{1J} at points* **a** *and* **b**. (B) *Projection Γ of the curve of saddle-nodes on the slow plane (a_1, a_2) for $\beta = 0.75$, $g = 0.5$, $r = 10$, $\theta = 0.2$, S as in (2.2), and two values of input $I = 1.2$ and $I = 1$; Γ moves down on the first bisector direction as I decreases. The blue side is associated to the right branch of W in panel A, from $W = 0$ to W_{max} (an escape mechanism), and the red side is associated to the left branch of W from W_{max} to 0 (a release mechanism).*

all regime does not exist in the interval (I^*_{hb}, I^{**}_{hb}). Otherwise, for (u_{1J}, u_{2J}) where $W = 1$ we can compute the values of I where winner-take-all occurs (see below).

Remark 4.2. *Let us observe that $W(u_{1J}) = [\beta - F'(\xi)]/g$ for some intermediate ξ between u_{2J} and u_{1J}. At $u_{1J} = u_{2J} = u^{**}_{hb}$ or $u_{1J} = u_{2J} = u^*_{hb}$ (that represents exactly the equilibrium point $\mathbf{e}_I$ at $I = I^{**}_{hb}$ and I^*_{hb}, respectively) we have $F'(\xi) = \beta$; therefore, we can extend by continuity the function $W|_{[}u^{**}_{hb}, u_{M\beta}]$ at $u_{1J} = u^{**}_{hb}$ and the function $W|_{[}u^*_{hb}, u_{M\beta}]$ at $u_{1J} = u^*_{hb}$ by taking $W = 0$ at these edges.*

For gain function S as in (2.2) and different values of β we plotted the curve W; the curve W always has a maximum value W_{max} for some $u_{1J}^{Wmax} \in (u_0, u_{M\beta})$ with $u_2 = u_2(u_1) \in (u_{m\beta}, u_0)$. Here $\partial u_2/\partial u_1 = -[F'(u_2)F''(u_1)]/[F'(u_1)F''(u_2)] > 0$. W measures the relative distance between the values of slow variables against that between fast variables. Based on this observation we color $\mathcal{SN}$, and obviously Γ, in *blue* for (u_{1J}, u_{2J}) starting at $(u^{**}_{hb}, u^{**}_{hb})$ and varying until it reaches $(u_{1J}^{Wmax}, u_{2J}^{Wmax})$ and in *red* for (u_{1J}, u_{2J}) between $(u_{1J}^{Wmax}, u_{2J}^{Wmax})$ and (u^*_{hb}, u^*_{hb}) (Figures 7C, 9B). The first corresponds to the right branch of W from $W = 0$ to W_{max}, and the latter corresponds to the left branch of W from W_{max} to 0.

Well inside the interval that defines the blue part of Γ (that is, not too close to the value u_{1J}^{Wmax} that gives the maximum W) we have either $u_0 < u_{2J} < u^{**}_{hb} < u_{1J}$ or $u^*_{hb} < u_{2J} < u_0 < u^{**}_{hb} < u_{1J}$; so $F'(u_{2J}) < \beta < F'(u_{1J})$. However, we recall that $F'(u_{2J}) = 1/S'(I - \beta u_{1J} - g a_{2J})$ and $F'(u_{1J}) = 1/S'(I - \beta u_{2J} - g a_{1J})$. That means that the gain function S has at the jump a bigger slope at $I - \beta u_1 - g a_2$, the net input to the suppressed population, than at $I - \beta u_2 - g a_1$, the net input to the dominant population; in other words, the gain to the suppressed population falls in the range of steeper S. According to the definitions introduced at the beginning of section 4, this case corresponds to an *escape* type of dynamics.

On the red part of Γ there is an opposite behavior: at least away from the edge u_{1J}^{Wmax} we have either $u_{2J} < u^*_{hb} < u_0 < u_{1J} < u^{**}_{hb}$ or $u_{2J} < u^*_{hb} < u_{1J} < u_0 < u^{**}_{hb}$. That is,

$F'(u_{2J}) = 1/S'(I - \beta u_{1J} - g a_{2J}) > \beta > F'(u_{1J}) = 1/S'(I - \beta u_{2J} - g a_{1J})$. At the jump the gain $S'(I - \beta u_2 - g a_1)$ to the dominant population falls in the range of steeper S than that to the suppressed population. We called this a *release* mechanism.

In fact, the values u_{2J}^{Wmax} and u_{1J}^{Wmax} with $u_{2J}^{Wmax} < u_0 < u_{1J}^{Wmax}$ (and not u_{hb}^{*} and u_{hb}^{**}) determine what we generically call the "steeper" part of S. For the particular case of gain function symmetric to its inflection point (θ, u_0), i.e., for S such that $S(\theta + x) + S(\theta - x) = 1$ it can be shown that indeed $u_{2J}^{Wmax} = u_{hb}^{*}$ and $u_{1J}^{Wmax} = u_{hb}^{**}$ (for example, S as in (2.2); see Remark 4.3). However, in other cases that is not true anymore; then it is more difficult to interpret the terms escape and release for u_{1J} close to u_{1J}^{Wmax}, the passage point from one dynamical regime to another.

Occurrence of the winner-take-all regime. We can determine now the minimum value of β, say, β_{wta}, where the winner-take-all regime occurs in system (2.1). Moreover, for a given $\beta > \beta_{wta}$ we find the corresponding values of I that delineate this regime.

We note that if $I \in (I_{hb}^{*}, I_{pf}^{*}] \cup [I_{pf}^{**}, I_{hb}^{**})$, then system (2.1) has a unique equilibrium point and it belongs to the middle branch of Σ ($F'(u_I) < \beta$).

For $I > I_{pf}^{*}$ or $I < I_{pf}^{**}$ *sufficiently close* to the pitchfork bifurcation point, system (2.1) has three equilibria and all are situated again on the middle branch of Σ, inside $\mathcal{SN}$, the curve of lower and upper knees (Figure 7A and 7C). That is because $F'(u_{1p}) < \beta$ and $F'(u_{2p}) < \beta$, so $F'(u_{1p})F'(u_{2p}) < \beta^2$. A trajectory that starts on either the upper or the lower branch of Σ cannot approach an equilibrium point since it will first reach $\mathcal{SN}$, and so the system oscillates. However, for intermediate values of I between I_{pf}^{*} and I_{pf}^{**}, two of the equilibrium points may move on the lateral branches of Σ and become stable ($F'(u_{1p})F'(u_{2p}) > \beta^2$; see Figure 7D and 7F). That is the case where the winner-take-all regime occurs. We should point out that the equilibrium $\mathbf{e}_I = (u_I, u_I, u_I, u_I)$ always remains unstable for $I \in (I_{pf}^{*}, I_{pf}^{**}) \subset (I_{hb}^{*}, I_{hb}^{**})$. The boundary between oscillatory and winner-take-all dynamics is obtained when the equilibrium points (u_{1p}, u_{2p}) belong to $\mathcal{SN}$, that is, when on $\mathcal{SN}$ both $a_1 = u_1$ and $a_2 = u_2$ are true. We find in the following the values of I where winner-take-all appears.

The values of I, say, I_w, that delineate the winner-take-all regime are defined by

$$
\begin{aligned}
&F'(u_{1J})F'(u_{2J}) = \beta^2,\\
&\frac{1}{g}\left(\beta - \frac{F(u_{1J}) - F(u_{2J})}{u_{1J} - u_{2J}}\right) = 1,\\
&I_w = F(u_{1J}) + g u_{1J} + \beta u_{2J} \quad [\, = F(u_{2J}) + g u_{2J} + \beta u_{1J}\,].
\end{aligned}
\tag{4.10}
$$

The left-hand side of the second equation in (4.10) is in fact $W(u_{1J})$. Since the curve W always has a maximum value W_{max} for some $u_{1J}^{Wmax} \in (u_0, u_{M\beta})$, the line $W = 1$ can be either above the maximum or below it. If $W_{max} < 1$ (Figure 9A, $\beta = 0.75$), then there is no winner-take-all regime. If $W_{max} > 1$ (Figure 9A, $\beta = 1.1$), then there exist two values for u_{1J} where the curve W intersects the horizontal line $W = 1$.

The critical (minimum) value β_{wta} where the winner-take-all regime appears in system (2.1) results from the case of $W_{max} = 1$; i.e., it satisfies the conditions $W(u_{1J}) = 1$ and $W'(u_{1J}) = 0$. Thus the minimum value β_{wta} that introduces winner-take-all dynamics in system (2.1) is defined by

$$F'(u_{1J})F'(u_{2J}) = \beta^2, \qquad u_{1J} \neq u_{2J},$$
$$(F(u_{1J}) - F(u_{2J}))[F'(u_{1J})F''(u_{2J}) + F'(u_{2J})F''(u_{1J})]$$
$$= (u_{1J} - u_{2J})[F'(u_{1J})^2F''(u_{2J}) + F'(u_{2J})^2F''(u_{1J})],$$
$$\beta_{wta} = g + \frac{F(u_{1J}) - F(u_{2J})}{u_{1J} - u_{2J}}. \tag{4.11}$$

Remark 4.3. *We note that due to its particular form* (2.2)*, S is symmetric about the point (θ, u_0) with $u_0 = 0.5$. The graph of F' is symmetric to the vertical line $u = 0.5$, i.e., $F'(1-u) = F'(u)$ for all $u \in (0,1)$, and so $F''(1-u) = -F''(u)$. In this particular case we have $F'(u^*_{hb}) = F'(u^{**}_{hb}) = \beta$, so $u^*_{hb} = 1 - u^{**}_{hb}$ and $F''(u^*_{hb}) = -F''(u^{**}_{hb})$. According to* (4.11) *the maximum value W_{max} is obtained exactly at $u_{1J} = u^{**}_{hb}$, $u_{2J} = u^*_{hb}$.*

For gain function S as in (2.2) and different values of β we plotted the curve W and determined the interval $[I^*_w, I^{**}_w]$ where winner-take-all occurs. For example, choosing $r = 10$, $\theta = 0.2$, $g = 0.5$, and $\beta = 1.1$, we have $W_{max} = 1.1046$ and $I^*_w = 0.697$ (computed at $u_{1J} = 0.7158$, $u_{2J} = 0.0424$) and $I^{**}_w = 1.303$ (computed at $u_{1J} = 0.9576$, $u_{2J} = 0.2842$). The minimum value of β for the winner-take-all regime is $\beta_{wta} = 1.0387$. As expected from our analysis in section 3.3, the value of β that guarantees existence of multiple equilibria in system (2.2), i.e., $\beta_{pf} = g + 1/S'(\theta) = 0.9$, is smaller than β_{wta}.

For the same choice of parameters as above ($\beta = 1.1$, $g = 0.5$, and S as in (2.2) with $r = 10$, $\theta = 0.2$), we plot the projection of the limit cycle and the curve of saddle-nodes on the slow plane (a_1, a_2) for different values of parameter I. Figure 10A gives the bifurcation diagram of activity u_1 versus input strength I for $\tau = 100$. In the rest of the panels we choose $\tau = 5000$ to mimic the singular limit cycle solution with jumping points exactly on the curve of saddle-nodes (for smaller τ, e.g., $\tau = 100$, the jumping points do not belong to the curve of saddle-nodes but fall close to it). In this case we have

$$I^*_{hb} = 0.1434,\ I^*_{pf} = 0.4064,\ I^*_w = 0.697,\ I^{**}_w = 1.303,\ I^{**}_{pf} = 1.5936,\ I^{**}_{hb} = 1.8566.$$

At larger values of I ($I = 1.7$ and $I = 1.5$) oscillation is due to an escape mechanism (Figure 10B–C); at an intermediate value $I = 1$ the system is in the winner-take-all regime (Figure 10D); at smaller values of I ($I = 0.5$ and $I = 0.3$) oscillation is due to a release mechanism (Figure 10E–F). Besides the limit cycle other important trajectories are the equilibrium points. There are three unstable equilibria for $I = 1.5$ and $I = 0.5$ but only one equilibrium point at $I = 1.7$ and $I = 0.3$. At $I = 1$ two out of three equilibria are stable.

Remark 4.4. *Equations* (4.11) *and* (4.10)*, used to determine the critical β where the winner-take-all regime exists in system* (2.1) *and then, for $\beta > \beta_{wta}$, to estimate I^*_w and I^{**}_w, prove to be reliable. The estimations obtained by this method are in excellent agreement with the results found in system* (2.1)*'s numerical simulations for both symmetric and asymmetric gain functions. The latter case is discussed in section* 5*.*

5. Neuronal competition models that favor the escape (or release) dynamical regime.

As seen in section 4.1, the dynamical scheme of T versus I is symmetric to $I = \theta + \frac{\beta+g}{2}$ for S, the Heaviside step function. When the winner-take-all regime exists, its corresponding I-input interval is equally split around the value $\theta + \frac{\beta+g}{2}$, that is, $\theta + g \leq I \leq \theta + \beta$ as in Theorem 4.1. Moreover, an equal input range is found for both release and escape mechanisms:

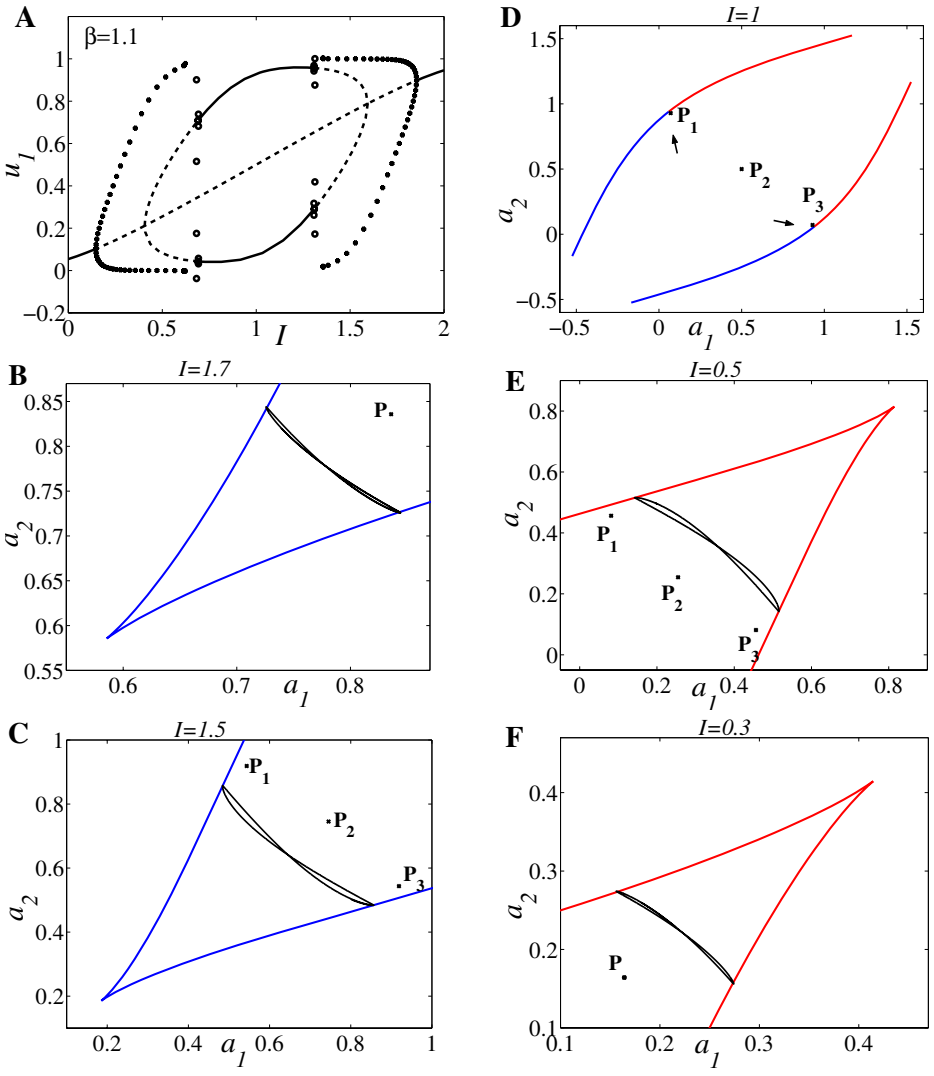


Figure 10. (A) *Bifurcation diagram of activity* u_1 *versus input* I *for* $\beta = 1.1$, $g = 0.5$, $r = 10$, $\theta = 0.2$, $\tau = 100$. (B–F) *Projection of the limit cycle and the curve of saddle-nodes on the slow plane* (a_1, a_2) *for different values of parameter* I *chosen according to the bifurcation diagram in panel A; however, in order to mimic the singular limit cycle solution we choose here* $\tau = 5000$. *Symbol* $*$ *indicates the location of equilibria. At large values of* I *oscillation is due to an escape mechanism:* (B) $I = 1.7$, (C) $I = 1.5$; (D) *at intermediate value* $I = 1$ *winner-take-all dynamics is observed; then at low values of* I *oscillation is due to a release mechanism:* (E) $I = 0.5$, (F) $I = 0.3$. *Single (panels B, F) or multiple (panels C, E) unstable equilibria can coexist with the limit cycle.*

$\theta + \frac{g}{2} < I < \theta + g$ and $\theta + \beta < I < \theta + \beta + \frac{g}{2}$ as in Theorem 4.1, or $\theta + \frac{g}{2} < I < \theta + \frac{\beta+g}{2}$ and $\theta + \frac{\beta+g}{2} < I < \theta + \beta + \frac{g}{2}$ as in Theorem 4.2. Therefore, it seems reasonable to ask to what extent the symmetry of the I-T dynamical scheme about a specific value I^* relates to the geometry of S and, more generally, to the form of the equations in (2.1). We address this question in the following and find a heuristic method to reduce one of the two ranges of

escape and release mechanisms while still maintaining the other one.

First let us note that we obtain a result similar to that in section 4.1 for any smooth sigmoid S as long as it is symmetric about its *threshold*. The threshold (say, th) is defined as the value where the gain function reaches its middle point: for S taking values between 0 and 1, it is $S(th) = 0.5$. The terminology comes from the fact that if a net input to one population is below the threshold ($x < th$), then it determines a weak effective response ($S(x) < 1/2$); at equilibrium that would correspond to an inactive state. On the other hand, a net input above the threshold ($x > th$) determines a strong effective response ($S(x) > 1/2$) which, at equilibrium, corresponds to an active state. The *symmetry condition* of S with respect to the threshold is described mathematically by the equality $S(th + x) + S(th - x) = 1$ for any real x or, equivalently, $S(x) + S(2\,th - x) = 1$. The gain function defined by (2.2) is such an example: in this case the threshold is exactly the inflection point θ ($S''(\theta) = 0$; $S(\theta) = 0.5$).

Theorem 5.1. *If the gain function S satisfies $S(\theta + x) + S(\theta - x) = 1$, then system (2.1) with input I^* is diffeomorphic equivalent to system (2.1) with input $I = 2\theta + \beta + g - I^*$.*

Proof. Due to the symmetry of S, equation $\dot{\tilde{u}}_1 = -\tilde{u}_1 + S(I^* - \beta\tilde{u}_2 - g\tilde{a}_1)$ with the change $u_1 = 1 - \tilde{u}_1$, $u_2 = 1 - \tilde{u}_2$, $a_1 = 1 - \tilde{a}_1$, $a_2 = 1 - \tilde{a}_2$ becomes $\dot{u}_1 = -u_1 + 1 - S(I^* - \beta(1 - u_2) - g(1 - a_1)) = -u_1 + S((2\theta + \beta + g - I^*) - \beta u_2 - g a_1)$; on the other hand, equation $\tau\dot{\tilde{a}}_1 = -\tilde{a}_1 + \tilde{u}_1$ becomes $\tau\dot{a}_1 = -a_1 + u_1$. Therefore, system (2.1) with input I^* is diffeomorphic equivalent to system (2.1) with input $(2\theta + \beta + g - I^*)$. ∎

Remark 5.1. *Theorem 5.1 implies that system (2.1) has the same type of solutions for any two values of input strength I_1 and I_1^* such that $\frac{1}{2}(I_1 + I_1^*) = \theta + \frac{\beta+g}{2}$. Moreover, if at I_1 and I_1^* an oscillatory solution of period T_1 and T_1^* exists, then, due to the diffeomorphism, $T_1 = T_1^*$. Obviously, if $T_1 < T_2$ for $I_1 < I_2 < \theta + \frac{\beta+g}{2}$, then $T_1^* < T_2^*$ for the corresponding values $I_1^* > I_2^* > \theta + \frac{\beta+g}{2}$. Therefore, for symmetric S to its inflection point (threshold in this case), the intervals of I for regions II and IV (see Figure 3F or H) have the same length and are symmetric to the line $I = \theta + \frac{\beta+g}{2}$.*

In order to explore the effect the asymmetry of S has on the bifurcation diagram, we consider S to be

$$S(x) = \begin{cases} 2u_0 / \left(1 + e^{-\frac{r}{2u_0}(x-\theta)}\right), & x \le \theta, \\ 1 - 2(1-u_0) / \left(1 + e^{-\frac{r}{2(1-u_0)}(\theta - x)}\right), & x > \theta, \end{cases} \tag{5.1}$$

with $u_0 \in (0, 1)$.

Therefore the inflection point θ and the threshold th satisfy $S(\theta) = u_0$ and $th > \theta$ if $u_0 < 1/2$, respectively, $th < \theta$ if $u_0 > 1/2$.

The graphs of S as in (5.1) and their corresponding F' (where $F = S^{-1}$) are drawn in Figure 11A–B with parameter values $r = 10$, $\theta = 0.2$, and $u_0 = 0.1, 0.5, 0.9$. Then the I-T bifurcation diagram for the competition model (2.1) is constructed at $\beta = 0.75$ (Figure 11C). Numerical results show that the symmetry of this bifurcation diagram is indeed a direct consequence of the symmetry of S to its threshold. Such is the case $u_0 = 0.5$. For other choices of u_0 one of the two regions that correspond to the increasing and decreasing I-T branch is favored (it is wider)—the former when $u_0 > 1/2$ and the latter when $u_0 < 1/2$.

Recall from section 4 the definition (4.8) for the curve $\mathcal{SN}$ of jumping points (with its projection Γ on the plane of slow variables (a_1, a_2)) and the definition (4.9) for W that char-

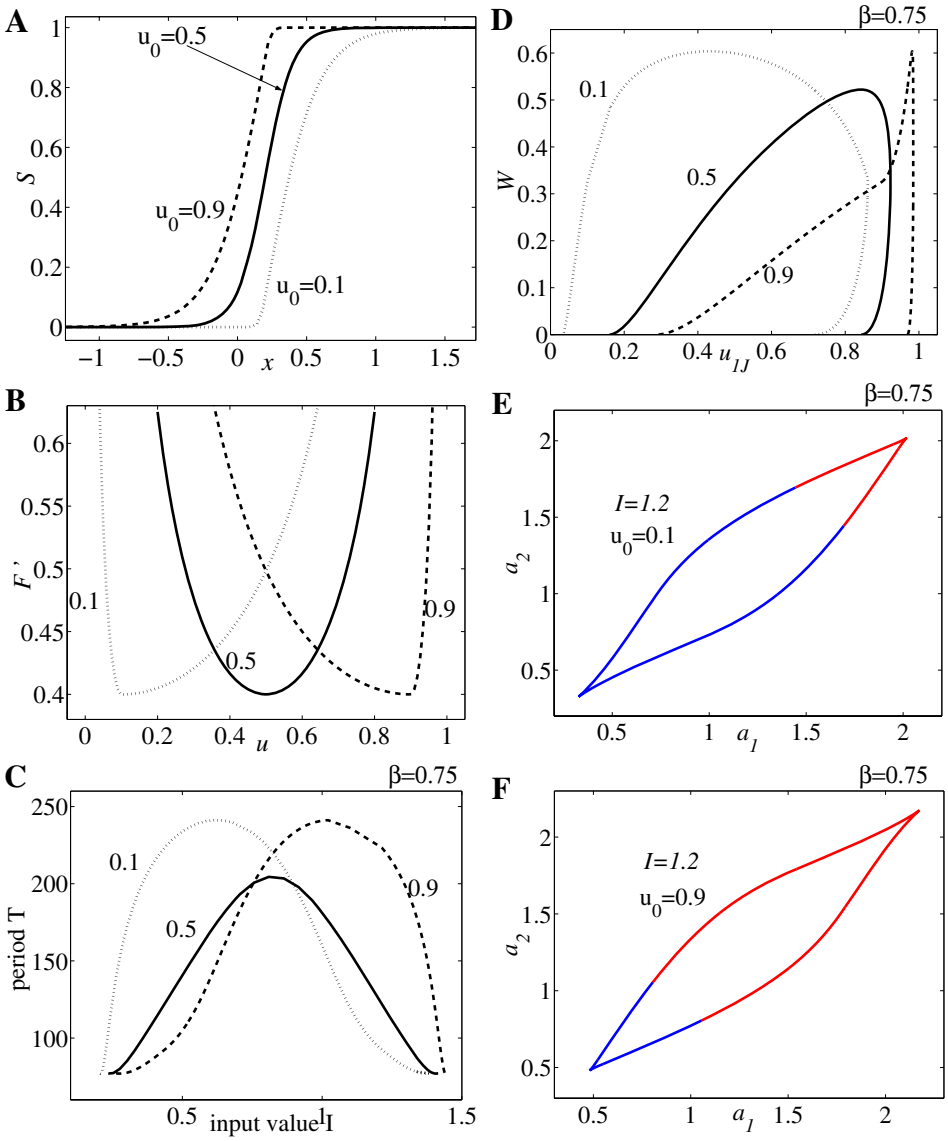


Figure 11. *Consider parameter values $r = 10$, $\theta = 0.2$, $g = 0.5$, $\tau = 100$ and gain functions S as in (5.1) with $u_0 = 0.1, 0.5, 0.9$.* (A) *Graphs of S.* (B) *Their corresponding F' where $F = S^{-1}$.* (C) *Bifurcation diagram for period T versus input I in system (2.1) as $\beta = 0.75$.* (D) *The graph of $W = W(u_{1J})$ for upper knees on $\mathcal{SN}$ computed as $\beta = 0.75$.* (E–F) *Projection Γ of the curve of saddle-nodes on the slow plane (a_1, a_2) for $\beta = 0.75$, $I = 1.2$, and $u_0 = 0.1, 0.9$. The blue side is associated to the escape mechanism and the red side to the release mechanism. The asymmetry of S to the threshold favors* (E) *escape if $u_0 < 1/2$ and* (F) *release if $u_0 > 1/2$.*

acterizes escape, release, and winner-take-all regimes. The shape of W changes dramatically with u_0 (Figure 11D, $\beta = 0.75$) but not at all with β. However, for any fixed u_0 the curve W always has a unique maximum point; this maximum moves down as β decreases.

The asymmetry of S leads to an asymmetry of the curve Γ: for $u_0 < 1/2$ we have $u_0 < u_{1J}^{Wmax} < u_{hb}^{**}$, $u_{2J}^{Wmax} < u_{hb}^{*}$, and the curves $\mathcal{SN}$ and Γ have a longer blue part than

red part; the escape mechanism is favored (Figure 11E, $u_0 = 0.1$). On the contrary, for $u_0 > 1/2$ the release mechanism is favored since the red part of Γ is longer: $u_{1J}^{Wmax} > u_{hb}^{**}$ and $u_{hb}^{*} < u_{2J}^{Wmax} < u_0$ (Figure 11F, $u_0 = 0.9$). We can understand this specific behavior by looking at the network populations' dynamics about the threshold: in case of $u_0 < 1/2$ the maximum gain to each population is reached for some net input *below* the threshold ($S'(\theta) = \max S'$ with $S(\theta) = u_0$, so $\theta < th$); then about the threshold th the gain S' has a decreasing trend. Consequently, there is a wider range for inputs where the inactive population has a significant gain compared to the active population, and so the escape mechanism is favored (the suppressed population regains control on its own, thus becoming active). On the other hand, if $u_0 > 1/2$, the maximum gain to the network's populations is reached for some net input *above* the threshold ($\theta > th$) and the gain has an increasing trend in the vicinity of th. Therefore, the release mechanism is indeed much more easily obtained.

Remark 5.2. *The function S defined by (5.1) does not entirely satisfy the hypotheses introduced in section 2. In this case $F'''(u_0)$ does not exist anymore at $u_0 \neq 1/2$ even if we still have F as $\mathcal{C}^2(0,1)$ and $\mathcal{C}^\infty((0,1) \setminus \{u_0\})$. Nevertheless this property does not affect our analytical results (e.g., the expansion we used in section 3 to prove the existence of stable oscillatory solutions still makes sense since it is done locally about some point u^* different than u_0). Consequently, all the results found for system (2.1) in previous sections remain valid.*

5.1. A competition model that favors escape. In [31] we investigated four distinct neuronal competition models: a model by Wilson [37], one by Laing and Chow (the LC-model [17]), and two other variations of the LC-model that we called depression-LC and adaptation-LC. The latter is exactly the system (2.1) with symmetric sigmoid function. Besides the models' commonalities we also noticed some differences: in some cases the bifurcation diagrams show a preference of the system to the escape mechanism (the region of I that corresponds to the decreasing I-T branch is wider; see Figures 3 and 4 in [31]). Moreover, for sufficiently low inhibition in Wilson's and the depression-LC models the increasing (release-related) branch can disappear completely. Based on the results obtained in the present paper, we can explain those numerical observations.

Let us take, for example, Wilson's model for binocular rivalry [37, 31]. Since the time-scale for inhibition is much shorter than the time-scale for the (excitatory) firing rate, we can assume that the inhibitory population tracks the excitatory population almost instantaneously. Thus Wilson's model becomes equivalent to a system of the form

$$
\begin{aligned}
\dot{u}_1 &= -u_1 + \frac{\gamma(I - \beta u_2)_+^2}{(\theta + a_1)^2 + (I - \beta u_2)_+^2}, \\
\dot{u}_2 &= -u_2 + \frac{\gamma(I - \beta u_1)_+^2}{(\theta + a_2)^2 + (I - \beta u_1)_+^2}, \\
\tau \dot{a}_1 &= -a_1 + g u_1, \\
\tau \dot{a}_2 &= -a_2 + g u_2,
\end{aligned}
\tag{5.2}
$$

where γ is a positive constant and $[x]_+$ is defined as $[x]_+ = 0$ if $x < 0$ and $[x]_+ = x$ if $x \geq 0$. As in (2.1), parameters β and g represent here the strength of the inhibition and adaptation; I is the external input strength. We see that in the differential equations for u_j the nonlinearity

is introduced through a function

$$\tilde{S}(x;\Theta) = \frac{\gamma[x]_+^2}{\Theta^2 + [x]_+^2},$$

that is, $\dot{u}_1 = -u_1 + \tilde{S}(I - \beta u_2;\, \theta + a_1)$ and $\dot{u}_2 = -u_2 + \tilde{S}(I - \beta u_1;\, \theta + a_2)$. We note that $\tilde{S}$ is asymptotic to γ as $x \to \infty$ and satisfies $\tilde{S} = 0$ for $x \le 0$ and $\tilde{S}(\Theta;\Theta) = \gamma/2$, which means Θ is the threshold. To compare system (5.2)'s dynamics with that of (2.1), we will assume without loss of generality that $\gamma = 1$.

By the change of variables $w_1 = \beta u_1 - I$, $w_2 = \beta u_2 - I$, $A_1 = \beta a_1/g - I$, $A_2 = \beta a_2/g - I$, the system (5.2) is diffeomorphic equivalent to

$$\begin{aligned} \dot{w}_1 &= -w_1 + \beta\tilde{S}(-w_2;\theta_1) - I, \\ \dot{w}_2 &= -w_2 + \beta\tilde{S}(-w_1;\theta_2) - I, \\ \tau\dot{A}_1 &= -A_1 + w_1, \\ \tau\dot{A}_2 &= -A_2 + w_2 \end{aligned} \tag{5.3}$$

with $\theta_1(t) = \theta + \frac{g}{\beta}(A_1(t) + I)$ and $\theta_2(t) = \theta + \frac{g}{\beta}(A_2(t) + I)$.

On the other hand, let us consider the system (2.1) with S as in (2.2). By a similar change of variables $w_1 = \beta u_1 - I$, $w_2 = \beta u_2 - I$, $A_1 = \beta a_1 - I$, $A_2 = \beta a_2 - I$ this system is diffeomorphic equivalent to

$$\begin{aligned} \dot{w}_1 &= -w_1 + \beta S(-w_2;\theta_1) - I, \\ \dot{w}_2 &= -w_2 + \beta S(-w_1;\theta_2) - I, \\ \tau\dot{A}_1 &= -A_1 + w_1, \\ \tau\dot{A}_2 &= -A_2 + w_2 \end{aligned} \tag{5.4}$$

with $\theta_1(t)$ and $\theta_2(t)$ as above. The threshold of S (rewritten as $S(x;\theta) = 1/(1 + e^{-r(x-\theta)})$) is θ.

As we can see, up to the specific expression of the gain function, Wilson's and the adaptation-LC models are equivalent. The reason Wilson's model shows preference to the escape mechanism instead of release (while for the adaptation-LC model the interval ranges for escape and release dynamics have equal length) resides in the asymmetric shape of $\tilde{S}$ with respect to its threshold. The threshold is Θ but the maximum gain is obtained at $\frac{\Theta}{\sqrt{3}} < \Theta$ ($\tilde{S}''_{xx}(x;\Theta) = 0$ at $x = \frac{\Theta}{\sqrt{3}}$). The resulting behavior is similar to that of S as in (5.1) with $u_0 < 1/2$ ($S''(\theta) = 0$, $S(\theta) = u_0$, and $\theta < th$). The role of θ is played by $\frac{\Theta}{\sqrt{3}}$ and the corresponding value for u_0 is $\tilde{S}(\frac{\Theta}{\sqrt{3}};\Theta) = 1/4$.

Remark 5.3. *In fact, the difference between $\tilde{S}$ and asymmetric S from (5.1) is more subtle. Take again $\gamma = 1$; the restriction of $\tilde{S}$ on $(0,\infty)$ is invertible with inverse $\tilde{F}$ defined on $(0,1)$ by $\tilde{F}(u;\Theta) = \Theta\sqrt{\frac{u}{1-u}}$. In systems (5.3) and (5.4) the graph of $\tilde{F}'_u$, $\tilde{F}'_u(u;\theta_j) = \frac{\theta_j}{2(1-u)\sqrt{u(1-u)}}$ has a well-like shape similar to that of the graph of $F'_u(u;\theta_j) = \frac{4u_0^2}{ru(2u_0-u)}$ if $u \in (0,u_0]$ and*

$\frac{4(1-u_0)^2}{r(1-u)(1-2u_0+u)}$ *if* $u \in (u_0, 1)$. *However,* $\tilde{F}'_u$ *depends on* θ_j *(that implicitly means dependence on the slow variable) while* F'_u *does not. That explains why, for* S *as in* (5.1)*, when they exist, the Hopf bifurcation points always come in pairs (and oscillations start at both points with the same frequency—see* (3.5) *and* (3.6) *in section* 3*); so release and escape oscillatory regimes (even if not equally balanced) always coexist for* (2.1)*. On the contrary, in Wilson's model* (5.2) *the dependence of* $\tilde{F}'_u$ *on the slow variable allows us to find a (reduced) parameter regime where only the escape mechanism is possible* [31].

5.2. Competition models with nonlinear slow negative feedback. The local analysis pursued in section 3 shows that the uniform equilibrium point $\mathbf{e}_I$ can lose its stability through either a Hopf bifurcation (at exactly two values I^*_{hb} and I^{**}_{hb}) or a pitchfork bifurcation (at exactly two values I^*_{pf} and I^{**}_{pf}). This result comes from the intersection of the graph of F' (which has a well-like shape) with the straight horizontal lines $y = \frac{\beta}{1+\frac{1}{\tau}}$ and $y = \beta - g$, respectively. This type of intersection restricts the possibilities to either $I^*_{hb} < I^*_{pf} < I^{**}_{pf} < I^{**}_{hb}$ or $I^*_{pf} < I^*_{hb} < I^{**}_{hb} < I^{**}_{pf}$. The first case corresponds to a feedback/adaptation-dominated neuronal competition model and ensures the existence of stable oscillations in (2.1). Moreover, these appear with the same frequency $\omega = \frac{1}{\tau}\sqrt{\frac{g(\tau+1)}{\beta} - 1}$ at both values I^*_{hb} and I^{**}_{hb}, and they are due to two different mechanisms: escape (for larger values of I) and release (for lower values of I). Thus we conclude that in system (2.1) escape and release oscillatory regimes *always coexist*. The choice of asymmetric gain function with respect to its threshold helps to reduce one or another regime but cannot eliminate it completely.

Nevertheless there is a way to modify (2.1) such that the new obtained system shows preference to either the escape or release mechanim; moreover, as for Wilson's model, in some parameter regime we can find only escape-based oscillations (or, vice versa, only release-based oscillations). In this sense we consider the neuronal competition model with nonlinear slow negative feedback

$$\begin{aligned} \dot{u}_1 &= -u_1 + S(I - \beta u_2 - g a_1), \\ \dot{u}_2 &= -u_2 + S(I - \beta u_1 - g a_2), \\ \tau \dot{a}_1 &= -a_1 + a_\infty(u_1), \\ \tau \dot{a}_2 &= -a_2 + a_\infty(u_2) \end{aligned} \tag{5.5}$$

with S as in (2.2) and

$$a_\infty(x; \theta_a) = 1/\left(1 + e^{-r_a(x-\theta_a)}\right). \tag{5.6}$$

We obtain the following result, which is similar to Theorem 5.1.

Theorem 5.2. *Consider the nonlinear term* $a_\infty(x; \theta_a)$ *defined by* (5.6)*. If the gain function* S *satisfies* $S(\theta + x) + S(\theta - x) = 1$*, then system* (5.5) *with input* I^* *and slow equation nonlinearity* $a_\infty(x; \theta_a)$ *is diffeomorphic equivalent to system* (5.5) *with input* $I = 2\theta + \beta + g - I^*$ *and slow nonlinearity* $a_\infty(x; 1 - \theta_a)$.

Proof. With the change of variables $u_1 = 1 - \tilde{u}_1$, $u_2 = 1 - \tilde{u}_2$, $a_1 = 1 - \tilde{a}_1$, $a_2 = 1 - \tilde{a}_2$, system (5.5) with input I^* and nonlinear term in the slow equation $a_\infty(x; \theta_a)$ takes the form (5.5) with input $(2\theta + \beta + g - I^*)$ and $a_\infty(x; 1 - \theta_a)$. ∎

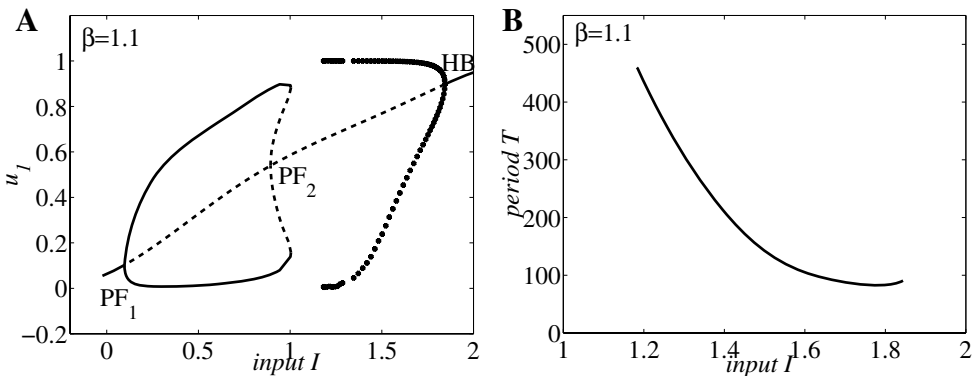


Figure 12. *Bifurcation diagrams for neuronal competition model (5.5): (A) population activity u_1 versus input strength I; (B) period T of the network oscillations versus I. Parameter values are $\beta = 1.1$, $g = 0.5$, $\tau = 100$, $\theta = 0.2$, $r = 10$, $r_a = 10$, and $\theta_a = 0.7$. Functions S and a_∞ are defined by (2.2) and (5.6).*

If $\theta_a = 1/2$, we have $a_\infty(x;\theta_a) = a_\infty(x;1-\theta_a)$, and thus the system (5.5) has the same type of solutions for any two values of input strength I_1 and I_1^* such that $\frac{1}{2}(I_1 + I_1^*) = \theta + \frac{\beta+g}{2}$. As for the case of linear adaptation (where the choice was $a_\infty(u) = u$), the intervals of I where the period of oscillations increases and decreases have the same length and are symmetric to the line $I = \theta + \frac{\beta+g}{2}$.

The symmetry of the I-T bifurcation diagram can be destroyed by choosing $\theta_a \neq 1/2$. As numerical simulations of system (5.5) show, the choice of $\theta_a > 1/2$ leads to a preference of the system for the escape mechanism (decreasing T versus I). On the contrary, for $\theta_a < 1/2$ the system shows preference to the release mechanism (increasing T versus I).

For parameters $\beta = 1.1$, $g = 0.5$, $\tau = 100$, and $\theta = 0.2$, $r = 10$ for S and $r_a = 10$ and $\theta_a = 0.7$ for a_∞, we plot in Figure 12 the bifurcation diagram of population activity u_1 versus the stimulus strength I and then the period T versus I. In Figure 12A we observe that the Hopf bifurcation point that existed for low value of I in case of linear adaptation disappears now. Instead we find a supercritical pitchfork bifurcation where stable nonuniform equilibrium points are born (they correspond to the winner-take-all case). The increasing branch of T versus I graph disappears while the decreasing branch is still present (Figure 12B). We find only four dynamical regimes (as opposed to the five described in Figure 3F and G): fusion (equal activity levels) for large and low input, winner-take-all, and oscillations with decreasing T as function of I for intermediate values of stimulus strength.

By choosing $\theta_a = 0.3$ the bifurcation diagrams in Figure 12 are virtually mirrored along the stimulus axis; in this case only the increasing I-T branch exists.

We explain these numerical results through an analytical approach. Similarly to the local analysis in section 3, we note that system (5.5) has a unique uniform equilibrium $\mathbf{e}_I = (u_I, u_I, a_\infty(u_I), a_\infty(u_I))$ for any real I. The value $u_I \in (0,1)$ is defined by equation $I = F(u_I) + \beta u_I + g a_\infty(u_I)$ and decreases with a decrease in I; moreover, $\lim_{I\to\infty} u_I = 1$ and $\lim_{I\to-\infty} u_I = 0$. The characteristic equation of the linearization matrix about $\mathbf{e}_I$ is a product of two factors: $\lambda^2 + \lambda\left(1 + \frac{1}{\tau} + \frac{\beta}{F'(u_I)}\right) + \frac{1}{\tau}\left(1 + \frac{ga'_\infty(u_I)+\beta}{F'(u_I)}\right) = 0$ and $\lambda^2 + \lambda\left(1 + \frac{1}{\tau} - \frac{\beta}{F'(u_I)}\right) + \frac{1}{\tau}\left(1 + \frac{ga'_\infty(u_I)-\beta}{F'(u_I)}\right) = 0$. Thus two of the eigenvalues always have negative real part, while the

other two show the type of stability for $\mathbf{e}_I$. Decreasing I, the stability of $\mathbf{e}_I$ is lost either through a pair of purely imaginary eigenvalues at $F'(u_I) = \beta/(1+\frac{1}{\tau})$, $F'(u_I) + ga'_\infty(u_I) > \beta$ or through a zero eigenvalue at $F'(u_I) + ga'_\infty(u_I) = \beta$ and $F'(u_I) > \beta/(1+\frac{1}{\tau})$.

Assume now that θ_a is greater than $1/2$ and close to 1. For large I the corresponding fixed point $\mathbf{e}_I$ has u_I close to 1 and in the vicinity of θ_a. Since this is the steeper (almost linear) part of a_∞, we can approximate $a_\infty(u_I) \approx ku_I$. The condition $F'(u_I) = \beta/(1+\frac{1}{\tau})$, $F'(u_I) > \beta - ga'_\infty(u_I) = \beta - gk$, can be attained (at least for adaptation-dominated systems) and stable oscillations occur. On the other hand, for small I, the fixed point $\mathbf{e}_I$ has u_I close to 0 and further away from θ_a. The function a_∞ is almost constant there, so $a'_\infty(u_I) \approx 0$. Then $F'(u_I) + ga'_\infty(u_I) \approx \beta > \beta/(1+\frac{1}{\tau})$ and the stability of $\mathbf{e}_I$ is lost through a zero eigenvalue (a pitchfork bifurcation).

Remark 5.4. *We note that the choice of $\theta_a > 1/2$ sufficiently close to 1 helps the slow variable for the suppressed population, say, a_1, to change faster than the slow variable for the dominant population, say, a_2. Since $a_\infty(u_1)$ is approximately constant to zero and $a_\infty(u_2)$ falls onto the linear part of the sigmoid, a_1 decays almost exponentially according to $\tau\dot{a}_1 \approx -a_1$, while a_2 grows more slowly according to $\tau\dot{a}_2 \approx -a_2 + ku_2$. The input-output function of the suppressed population changes faster, favoring escape. An opposite effect is obtained for $\theta_a < 1/2$.*

6. Discussion. We investigated a class of competition models that describe rhythmic (alternating) phenomena that arise in a range of neural contexts including perception of ambiguous sensory stimuli (such as binocular rivalry) and motor coordination (as in CPGs). These models rely on mutual inhibition between populations of neurons and a slow process in the form of spike frequency adaptation and/or synaptic depression, and they have the following commonalities [31].

A decrease in the strength of the input to the noise-free system leads to five possible types of dynamics: (i) at high values of input strength both competing populations are active at equal levels; (ii) by decreasing the input strength the system enters an oscillatory regime with the oscillation period increasing as input decreases; (iii) then for lower input the system is in a winner-take-all regime where only one population is dominant while the other is suppressed forever; (iv) continuing to decrease the input strength, the system oscillates again; in this region the oscillation period decreases as the input decreases; (v) at low values of input, oscillations disappear and both competing populations are inactive at an equal level. In addition, for weak inhibition the winner-take-all regime does not occur; however, the period of oscillations still depends on the input strength in a nonmonotonic fashion.

In a computational study of a model similar to (2.1), the authors of [23] also report transitions between simultaneous activity (single equilibrium), oscillations, and winner-take-all. The bifurcation diagram (Figure 4) in the parameter plane (I, β) resembles Figure 9 in [23]. However, Moldakarimov et al. were interested mostly in how the system's dynamics changes with inhibition strength (an internal parameter of the system) and not with stimulus strength. By analyzing the influence of stimulus on the oscillations' frequency/period, we provide a refined characterization of possible behaviors in this class of competition models.

The five dynamical regimes mentioned above were illustrated in Figure 3 for system (2.1), our choice as a particular example. Despite the fact that it has some limitations such as a lack of recurrent excitation, a symmetric gain function, and a nonsaturating $a_\infty(u)$, system (2.1)

has the advantage of being simple enough to allow for a thorough analytical investigation while still displaying the dynamical characteristics found in a larger class of models. Thus, for (2.1) we proved the existence of oscillations and we showed that they are antiphase as expected for a network of two competing populations. Moreover, by considering the period of oscillations T as a function of input strength I, we proved that the I-T graph is nonmonotonic. We associated the increasing I-T branch with a release mechanism and the decreasing branch with an escape mechanism. Release occurs for lower values of I. It means that during oscillation the dominant population loses control due to accumulating slow negative feedback and it becomes unable to suppress its competitor; consequently the latter becomes active and thereby takes the role of suppressor. On the contrary, escape occurs for higher values of I and it means that the suppressed population regains control on its own, starts to inhibit its competitor, and forces it into the down state. Moreover, using singular perturbation techniques, we characterized in the limiting case the conditions for occurrence of winner-take-all dynamics at intermediate values of stimulus strength.

Our presumption is that the potential for alternation of percepts depends on neuronal competition. If competition were significantly reduced or eliminated (say, effectively making β very small in the model), alternations would not occur in the presence of a stimulus. That is, we suppose that an isolated population would not oscillate. To satisfy this constraint we have disallowed recurrent excitation in (2.1): this precludes oscillations in an isolated population for any input value I. Perhaps for this goal the complete elimination of recurrent excitation is an extreme way to satisfy the constraint. Alternatively, we could consider systems with fast equations of the form $\dot{u}_j = -u_j + S(I + \alpha u_j - \beta u_k - g a_j)$ and allow for some recurrent excitation but not strong enough to let an isolated population oscillate (e.g., take $\alpha < (1 + \frac{1}{\tau})/S'(\theta)$). This modification, however, does not affect our conclusion on the nonmonotonicity of the period of oscillations versus input strength curve (not shown): both "release" and "escape" branches still appear.

Other modifications of (2.1) that favor either release or escape as responsible for oscillations were discussed in section 5. One extension of (2.1) allows the gain function to be asymmetric with respect to the threshold. This maintains one of the two oscillatory regimes while reducing the other one. A different rendition of (2.1) invokes a nonlinear slow negative feedback (a sigmoidal-shaped $a_\infty(u)$) and completely eliminates one of the two regimes. The existence of the saturating branches for a sigmoidal $a_\infty(u)$ introduces an asymmetry in the system; thus under some specific conditions either the suppressed population recovers from slow negative feedback faster than the dominant population accumulates its own negative feedback (favoring escape, see Figure 12), or the reverse occurs. Interestingly, adding noise to these specially designed models, we automatically recover the nonmonotonicity of the period versus input curve [31].

Oscillations in mutually inhibitory neuronal networks based on fast-slow dynamics [27] as well as the terms "release" and "escape" [36, 32] were previously discussed for neuronal networks in the presence of local autocatalysis. The autocatalysis was either an intrinsic process (like voltage-gated persistent inward currents) or a synaptic process (like intrapopulation recurrent excitation). Other models assumed networks of excitatory cells interacting through a global inhibitory feedback that typically produce the winner-take-all dynamics; the inhibition was dynamic with a slow time-scale and induced more complicated oscillatory patterns

with one cell being active for a while and then spontaneously turning off and allowing another one to take over [8]. Contrary to these examples, there is no autocatalysis in the neuronal competition models we investigate here. Instead the alternation is a combined result of two processes: mutual inhibition that acts effectively as a fast positive feedback (disinhibition) and a slow negative feedback (adaptation, but that could alternatively be synaptic depression).

Appendix A. Normal form for Hopf bifurcation. To construct the normal form for the Hopf bifurcation, in section 3 we use the expansion of $S^{(k)}(F(u_I))$, $k = 1, 2, 3, \ldots$, with respect to ε. That is obtained as follows: take, for example,

$$f_1(I) \stackrel{def}{=} S'(F(u_I)) = f_1(I^* + \varepsilon^2\alpha) = f_1(I^*) + f_1'(I^*)\varepsilon^2\alpha + \mathcal{O}(\varepsilon^4).$$

Here $f_1(I^*) = S'(F(u^*)) = 1/F'(u^*) = (1 + \frac{1}{\tau})/\beta$ and $f_1'(I) = S''(F(u_I)) \cdot F'(u_I) \cdot du_I/dI$.

Based on (3.2), $f_1'(I) = S''(F(u_I)) \cdot F'(u_I)/(\beta + g + F'(u_I))$. On the other hand, since $u \equiv S(F(u))$, we have $1 \equiv S'(F(u)) \cdot F'(u)$, and further $0 \equiv S''(F(u)) \cdot F'(u)^2 + S'(F(u)) \cdot F''(u)$, i.e., $S''(F(u)) = -F''(u)/F'(u)^3$. Therefore,

$$f_1(I) = \frac{1 + 1/\tau}{\beta} - \frac{F''(u^*)}{F'(u^*)^2(\beta + g + F'(u^*))}\varepsilon^2\alpha + \mathcal{O}(\varepsilon^4)\,.$$

Similarly, we compute

$$f_2(I) \stackrel{def}{=} S''(F(u_I)) = f_2(I^* + \varepsilon^2\alpha) = f_2(I^*) + \mathcal{O}(\varepsilon^2) = -\frac{F''(u^*)}{F'(u^*)^3} + \mathcal{O}(\varepsilon^2)$$

and $f_3(I) \stackrel{def}{=} S'''(F(u_I)) = f_3(I^* + \varepsilon^2\alpha) = f_3(I^*) + \mathcal{O}(\varepsilon^2) = S'''(F(u^*)) + \mathcal{O}(\varepsilon^2)$. From $S''(F(u)) = -F''(u)/F'(u)^3$ we obtain $S'''(F(u)) = [3F''(u)^2 - F'(u) \cdot F'''(u)]/F'(u^*)^5$, so

$$f_3(I) = \frac{3F''(u^*)^2 - F'(u^*) \cdot F'''(u^*)}{F'(u^*)^5} + \mathcal{O}(\varepsilon^2).$$

Normal form. Let us now present the main steps in the algorithm for the construction of the normal form starting with

$$L_0V_0 = \varepsilon[\mathcal{B}(V_0, V_0) - L_0V_1] + \varepsilon^2[\mathcal{C}(V_0, V_0, V_0) + 2\mathcal{B}(V_0, V_1) + \Lambda V_0 - L_0V_2] + \mathcal{O}(\varepsilon^3)\,. \tag{A.1}$$

In the limit $\varepsilon \to 0$, the vector V_0 is a solution of the linear system $L_0V_0 = 0$ with two eigenvalues $\lambda_{1,2}$ of negative real part and two purely imaginary eigenvalues $\lambda_{3,4} = \pm i\omega$. Thus V_0 belongs to the center manifold; i.e., for an eigenvector $\xi \in \boldsymbol{C}^4$ of $\mathcal{A}_0$ that satisfies $\mathcal{A}_0\xi = i\omega\xi$, say,

$$\xi = (-\tau\omega + i,\, \tau\omega - i,\, i,\, -i)^{\mathrm{T}}, \tag{A.2}$$

the solution V_0 takes the form

$$V_0(t) = w(t)\xi e^{i\omega t} + \bar{w}(t)\bar{\xi}e^{-i\omega t}.$$

However, since $L_0V_0 = \mathcal{O}(\varepsilon)$, $w(t)$ is ε-dependent, and it can be written in slow time $s = \varepsilon^2 t$ as $w = w(s)$. In the singular perturbation expansion, $w(s) = w(s)_{|\varepsilon=0} + \frac{dw}{ds}_{|\varepsilon=0}\varepsilon^2 t + \mathcal{O}(\varepsilon^4)$.

With notation $Z = w(s)_{|\varepsilon=0}$, $Z' = \frac{dw}{ds}_{|\varepsilon=0}$, and so on (here $'$ stands for the derivative with respect to the slow time), we have $w = Z + \varepsilon^2 t Z' + \mathcal{O}(\varepsilon^4)$ and

$$V_0 = \left(Z\xi e^{i\omega t} + \bar{Z}\bar{\xi}e^{-i\omega t}\right) + \varepsilon^2 t\left(Z'\xi e^{i\omega t} + \bar{Z}'\bar{\xi}e^{-i\omega t}\right) + \mathcal{O}(\varepsilon^4).$$

Then we compute $L_0\left(te^{i\omega t}\xi\right) = e^{i\omega t}\xi$, $L_0\left(te^{-i\omega t}\bar{\xi}\right) = e^{-i\omega t}\bar{\xi}$ and $\mathcal{B}\left(\xi e^{i\omega t}, \xi e^{i\omega t}\right) = \frac{1}{2}Bb^2\boldsymbol{p}\,e^{2i\omega t}$, $\mathcal{B}\left(\bar{\xi}e^{-i\omega t}, \bar{\xi}e^{-i\omega t}\right) = \frac{1}{2}B\bar{b}^{\,2}\boldsymbol{p}\,e^{-2i\omega t}$, $\mathcal{B}\left(\xi e^{i\omega t}, \bar{\xi}e^{-i\omega t}\right) = \frac{1}{2}B|b|^2\boldsymbol{p}$, where $b = \beta\tau\omega + i(g-\beta)$ and $\boldsymbol{p} = (1,1,0,0)^{\mathrm{T}}$. Equation (A.1) becomes

$$L_0V_1 = \frac{1}{2}Z^2Bb^2\boldsymbol{p}\,e^{2i\omega t} + \frac{1}{2}\bar{Z}^2B\bar{b}^{\,2}\boldsymbol{p}\,e^{-2i\omega t} + Z\bar{Z}B|b|^2\boldsymbol{p} + \varepsilon\left[-Z'\xi e^{i\omega t} - \bar{Z}'\bar{\xi}e^{-i\omega t} + \mathcal{C}(V_0,V_0,V_0) + 2\mathcal{B}(V_0,V_1) + \Lambda V_0 - L_0V_2\right] + \mathcal{O}(\varepsilon^2),$$

so we look for solutions V_1 of the form

$$V_1 = w^2\xi_1 e^{2i\omega t} + 2w\bar{w}\xi_2 + \bar{w}^2\xi_3 e^{-2i\omega t} = Z^2\xi_1 e^{2i\omega t} + 2Z\bar{Z}\xi_2 + \bar{Z}^2\xi_3 e^{-2i\omega t} + \mathcal{O}(\varepsilon^2).$$

From the singular perturbation expression we determine $\xi_1 = \frac{1}{2}Bb^2(2i\omega I - \mathcal{A}_0)^{-1}\boldsymbol{p}$, $\xi_2 = -\frac{1}{2}B|b|^2\mathcal{A}_0^{-1}\boldsymbol{p}$, and $\xi_3 = \bar{\xi}_1$, that is,

$$\xi_1 = \frac{Bb^2\psi}{2|\psi|^2}(1 + 2i\tau\omega, 1 + 2i\tau\omega, 1, 1)^{\mathrm{T}}, \quad \xi_2 = \frac{B|b|^2}{2\left[1 + \left(\frac{g}{\beta}+1\right)\left(1+\frac{1}{\tau}\right)\right]}(1,1,1,1)^{\mathrm{T}},$$

where $\psi = 1 - 4\omega^2\tau + \left(\frac{g}{\beta}+1\right)\left(1+\frac{1}{\tau}\right) - 4i\omega(\tau+1)$.

Then we compute $\mathcal{C}(V_0,V_0,V_0)$, $\mathcal{B}(V_0,V_1)$ and $\Lambda V_0 = -\alpha A(Zb\,e^{i\omega t} + \bar{Z}\bar{b}\,e^{-i\omega t})\boldsymbol{q} + \mathcal{O}(\varepsilon^2)$ with $\boldsymbol{q} = (1,-1,0,0)^{\mathrm{T}}$ and obtain

$$L_0V_2 = -(Z'\xi + \alpha AZb\boldsymbol{q})\,e^{i\omega t} + Z^2\bar{Z}b^2\bar{b}\boldsymbol{q}\,e^{i\omega t}\left(-\frac{D}{2} + \frac{(\beta+g)B^2}{1+\left(\frac{g}{\beta}+1\right)\left(1+\frac{1}{\tau}\right)} + \frac{B^2(\beta+g+2i\beta\omega\tau)\psi}{2|\psi|^2}\right) + e^{3i\omega t}(\ldots) + cc + \mathcal{O}(\varepsilon).$$

In order for the solution V_2 to exist, the right-hand side of the above equation should be orthogonal on the eigenvectors of the adjoint operator $L_0^* = -\frac{d}{dt} - \mathcal{A}_0^{\mathrm{T}}$ on the space of periodic solutions $V(t) = V(t + \frac{2\pi}{\omega})$ with inner product

$$\langle V, W\rangle = \frac{\omega}{2\pi}\int_0^{\frac{2\pi}{\omega}}\sum_{i=1}^{4} v_i(t)\bar{w}_i(t)\,dt, \quad V = (v_i)^{\mathrm{T}}_{i=\overline{1,4}},\ W = (w_i)^{\mathrm{T}}_{i=\overline{1,4}}.$$

That means the right-hand side should be orthogonal on $\left\{\eta e^{i\omega t}, \bar{\eta}e^{-i\omega t}\right\}$ with η solution of $\mathcal{A}_0^{\mathrm{T}}\eta = -i\omega\,\eta$ and $\xi\cdot\bar{\eta} = 1$; that is,

$$\eta = -\frac{1}{4\omega\tau}(1,\,-1,\,-1 - i\omega\tau,\,1 + i\omega\tau)^{\mathrm{T}}.$$

We obtain the normal form $Z' = \alpha A\varphi Z - \mathcal{L}Z^2\bar{Z}$ with $\mathcal{L}$ as in Theorem 3.2.

By rescaling $I - I^* = \varepsilon^2\alpha$ and $z(t) = \varepsilon Z(\varepsilon^2 t) = \varepsilon Z(s)$, we have $\dot{z} = dz/dt = \varepsilon \frac{dZ}{ds}\frac{ds}{dt} = \varepsilon^3 Z'$; the above differential equation becomes exactly (3.11).

First Lyapunov coefficient. We would like to determine sufficient conditions for the Hopf bifurcation to be supercritical, i.e., for $\mathrm{Re}(\mathcal{L}) > 0$.

We now use inequality (3.3) to show that the first term in the sum that defines $\mathcal{L}$ has positive real part:

$$\begin{aligned}\mathrm{Re}\left(\varphi\psi\left(\frac{\beta+g}{2} + i\beta\omega\tau\right)\right) &= (2\beta g)\tau + \frac{1}{4}(7g^2 + 5\beta g - 6\beta^2) + \frac{1}{4\tau}(7g^2 - 6\beta g + 3\beta^2)\\ &> 2\beta^2 - 2\beta g + \frac{1}{4}(7g^2 + 5\beta g - 6\beta^2) + \frac{1}{4\tau}(7g^2 - 6\beta g + 3\beta^2)\\ &= \frac{1}{4}(7g^2 - 3\beta g + 2\beta^2) + \frac{1}{4\tau}(7g^2 - 6\beta g + 3\beta^2) > 0\,. \end{aligned} \tag{A.3}$$

For the second term, (3.3) implies

$$l_1 = \frac{2(\beta+g)}{1 + \left(\frac{g}{\beta}+1\right)\left(1+\frac{1}{\tau}\right)} = \frac{2}{\frac{1}{\beta+g} + \frac{1}{F'(u^*)}} > \left(1 + \frac{1}{\tau+1}\right)F'(u^*) > F'(u^*).$$

Therefore, the real part of the second term satisfies

$$\beta\tau^2\omega^2|\varphi|^2B^2\left(l_1 - \frac{D}{B^2}\right) > \beta\tau^2\omega^2|\varphi|^2B^2F'(u^*)\left(\frac{F'(u^*)F'''(u^*)}{F''(u^*)^2} - 2\right). \tag{A.4}$$

Appendix B. Normal form for pitchfork bifurcation. The construction of this normal form follows similar steps to those in section 3.1 and Appendix A. Here $F'(u^\circ) = \beta - g$ with $u^\circ \in \{u^*_{pf}, u^{**}_{pf}\}$ and $I^\circ \in \{I^*_{pf}, I^{**}_{pf}\}$.

The operators $\mathcal{B}$, $\mathcal{C}$, Λ, and L_0 are defined in the same way as in section 3.1 with coefficients A, B, D as in (3.10). However, the derivatives of F at the bifurcation point u° take different values, so $S'(F(u_I)) = 1/(\beta - g) + \alpha A\varepsilon^2 + \mathcal{O}(\varepsilon^4)$, $S''(F(u_I)) = B + \mathcal{O}(\varepsilon^2)$, and $S'''(F(u_I)) = D + \mathcal{O}(\varepsilon^2)$ with A, B, and D evaluated at either u^*_{pf} or u^{**}_{pf}. The matrix of the linearized system is now

$$\mathcal{A}_0 = \begin{pmatrix} -1 & -\frac{\beta}{\beta-g} & -\frac{g}{\beta-g} & 0\\ -\frac{\beta}{\beta-g} & -1 & 0 & -\frac{g}{\beta-g}\\ \frac{1}{\tau} & 0 & -\frac{1}{\tau} & 0\\ 0 & \frac{1}{\tau} & 0 & -\frac{1}{\tau}\end{pmatrix}$$

and has a zero eigenvalue. Therefore, we find an eigenvector ξ ($\mathcal{A}_0\xi = 0$) and an eigenvector η of the adjoint matrix ($\mathcal{A}_0^{\mathrm{T}}\eta = 0$) such that $\xi \cdot \eta = 1$. They are

$$\xi = (1, -1, 1, -1)^{\mathrm{T}}, \quad \eta = \frac{1}{2(\beta - g(\tau+1))}(\beta - g, -\beta + g, -\tau g, \tau g)^{\mathrm{T}}.$$

With the perturbation $I - I^\circ = \varepsilon^2\alpha$, $V(t) = \varepsilon V_0(t) + \varepsilon^2 V_1(t) + \varepsilon^3 V_2(t) + \cdots$, we obtain that V_0 belongs to the eigenspace, that is, $V_0 = w(t)\xi$. The expansion with respect to the slow time ($s = \varepsilon^2 t$, $w = w(s)$, $Z = w(s)_{|\varepsilon=0}$, etc.) implies $V_0 = (Z + \varepsilon^2 tZ' + \mathcal{O}(\varepsilon^4))\xi$ and

$$L_0V_1 = Z^2\mathcal{B}(\xi,\xi) + \varepsilon[-Z'\xi + Z^3\mathcal{C}(\xi,\xi,\xi) + 2\mathcal{B}(V_0, V_1) + Z\Lambda\xi - L_0V_2] + \mathcal{O}(\varepsilon^2).$$

The vector V_1 is chosen to be of the form $V_1 = w^2\xi_1 = Z^2\xi_1 + \mathcal{O}(\varepsilon^2)$ with ξ_1 orthogonal on η. It results in $\xi_1 = -\mathcal{A}_0^{-1}\mathcal{B}(\xi,\xi)$, i.e., $\xi_1 = \frac{B^2(\beta-g)^3}{4\beta}(1,1,1,1)^{\mathrm{T}}$. The normal form is

$$Z' = (\Lambda\xi\cdot\eta)Z + [\mathcal{C}(\xi,\xi,\xi)\cdot\eta + 2\mathcal{B}(\xi,\xi_1)\cdot\eta]Z^3.$$

From $\Lambda\xi = \alpha A(\beta-g)\boldsymbol{q}$, $\mathcal{C}(\xi,\xi,\xi) = \frac{D}{6}(\beta-g)^3\boldsymbol{q}$, $\mathcal{B}(\xi,\xi_1) = -\frac{B^2}{8\beta}(\beta-g)^4(\beta+g)\boldsymbol{q}$ (where $\boldsymbol{q} = (1,-1,0,0)^{\mathrm{T}}$), and by rescaling $I - I^\circ = \varepsilon^2\alpha$ and $z(t) = \varepsilon Z(\varepsilon^2 t) = \varepsilon Z(s)$, we obtain exactly (3.15).

Appendix C. Additional material. To illustrate system (2.1)'s dynamics under the *escape* and *release* mechanisms we run numerical simulations in XPPAUT [7, 9] with S as in (2.2), $r = 10$, $\theta = 0.2$, and parameters $\beta = 1.1$, $g = 0.5$, $\tau = 100$ as in Figure 3F. Then in the fast plane (u_1, u_2) we obtain the trajectory of the point on the rivalry limit cycle: the point is drawn as a black thick dot; the u_1-nullcline is colored in red, and the u_2-nullcline is colored in blue.

70584_01.gif [3.7MB] illustrates the escape mechanism for $I = 1.5$.

70584_02.gif [3.8MB] illustrates the release mechanism for $I = 0.5$.

The small black square corresponds to the slow plane (a_1, a_2), where we included in green the projection of the limit cycle trajectory. This picture shows how the slow negative feedback accumulates for the dominant population and then how it recovers for the suppressed population (e.g., if u_1 is ON and u_2 is OFF, then a_1 increases and a_2 decreases). Then the cycle repeats.

Acknowledgment. We thank Sukbin Lim for helpful discussions.

REFERENCES

[1] T. J. Andrews and D. Purves, *Similarities in normal and binocular rivalrous viewing*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 9905–9908.
[2] R. Blake, *A neural theory of binocular rivalry*, Psychol. Rev., 96 (1989), pp. 145–167.
[3] J. W. Brascamp, R. van Ee, A. J. Noest, R. H. Jacobs, and A. V. van den Berg, *The time course of binocular rivalry reveals a fundamental role of noise*, J. Vis., 6 (2006), pp. 1244–1256.
[4] R. Curtu and B. Ermentrout, *Oscillations in a refractory neural net*, J. Math. Biol., 43 (2001), pp. 81–100.
[5] R. Curtu and B. Ermentrout, *Pattern formation in a network of excitatory and inhibitory cells with adaptation*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 191–231.
[6] R. L. Calabrese, *Half–center oscillators underlying rhythmic movements*, in The Handbook of Brain Theory and Neural Networks, M. A. Arbib, ed., MIT Press, Cambridge, MA, 1995, pp. 444–447.
[7] B. Ermentrout, http://www.math.pitt.edu/~bard/xpp/xpp.html (2006).
[8] B. Ermentrout, *Complex dynamics in winner–take–all neural nets with slow inhibition*, Neural Networks, 5 (1992), pp. 415–431.
[9] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.
[10] A. W. Freeman, *Multistage model for binocular rivalry*, J. Neurophysiol., 94 (2005), pp. 4412–4420.
[11] P. A. Getting, *Comparative analysis of invertebrate central pattern generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. H. Cohen, S. Rossignol, and S. Grillnet, eds., Wiley, New York, 1988, pp. 101–127.
[12] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, Birkhäuser, Basel, 2004.

[13] S. Grossberg, *Pattern formation by the global limits of a nonlinear competitive interaction in n dimensions*, J. Math. Biol., 4 (1977), pp. 237–256.
[14] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Field*, Springer-Verlag, New York, 1983.
[15] K. Gurney, T. J. Prescott, and P. Redgrave, *A computational model action selection in the basal ganglia* (I): *A new functional anatomy*, Biol. Cybernet., 84 (2001), pp. 401–410.
[16] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd ed., Springer-Verlag, New York, 1998.
[17] C. R. Laing and C. C. Chow, *A spiking neuron model for binocular rivalry*, J. Comput. Neurosci., 12 (2002), pp. 39–53.
[18] S. H. Lee, R. Blake, and D. J. Heeger, *Traveling waves of activity in primary visual cortex during binocular rivalry*, Nat. Neurosci., 8 (2005), pp. 22–23.
[19] W. J. M. Levelt, *On Binocular Rivalry*, Psychological Studies, Minor Series 2, Mouton, The Hague, The Netherlands, 1968.
[20] N. K. Logothetis, D. A. Leopold, and D. L. Sheinberg, *What is rivaling during binocular rivalry?*, Nature, 380 (1996), pp. 621–624.
[21] Z. H. Mao and S. G. Massaquoi, *Dynamics of winner–take–all competition in recurrent neural networks with lateral inhibition*, IEEE Trans. Neural Networks, 18 (2007), pp. 55–69.
[22] J. Maynard Smith, *Models in Ecology*, Cambridge University Press, Cambridge, UK, 1974, pp. 59–68.
[23] S. Moldakarimov, J. E. Rollenhagen, C. R. Olson, and C. C. Chow, *Competitive dynamics in cortical responses to visual stimuli*, J. Neurophysiol., 94 (2005), pp. 3388–3396.
[24] R. Moreno–Bote, J. Rinzel, and N. Rubin, *Noise–induced alternations in an attractor network model of perceptual bi–stability*, J. Neurophysiol., 98 (2007), pp. 1125–1139.
[25] T. J. Mueller and R. Blake, *A fresh look at the temporal dynamics of binocular rivalry*, Biol. Cybernet., 61 (1989), pp. 223–232.
[26] A. Polonsky, R. Blake, J. Braun, and D. J. Heeger, *Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry*, Nat. Neurosci., 3 (2000), pp. 1153–1159.
[27] J. Rubin and D. Terman, *Geometric singular perturbation analysis of neuronal dynamics*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., Elsevier, Amsterdam, 2002, pp. 93–146.
[28] N. Rubin, *Binocular rivalry and perceptual multi–stability*, Trends in Neurosci., 26 (2003), pp. 289–291.
[29] N. Rubin and J. M. Hupe, *Dynamics of perceptual bi–stability: Plaids and binocular rivalry compared*, in Binocular Rivalry and Bistable Perception, D. Alais and R. Blake, eds., MIT Press, Cambridge, MA, 2004.
[30] E. Salinas, *Background synaptic activity as a switch between dynamical states in a network*, Neural Comput., 15 (2003), pp. 1439–1475.
[31] A. Shpiro, R. Curtu, J. Rinzel, and N. Rubin, *Dynamical characteristics common to neuronal competition models*, J. Neurophysiol., 97 (2007), pp. 462–473.
[32] F. K. Skinner, N. Kopell, and E. Marder, *Mechanisms for oscillation and frequency control in reciprocally inhibitory model neural networks*, J. Comput. Neurosci., 1 (1994), pp. 69–87.
[33] A. L. Taylor, G. W. Cottrell, and W. B. Kristan, Jr., *Analysis of oscillations in a reciprocally inhibitory network with synaptic depression*, Neural Comput., 14 (2002), pp. 561–581.
[34] F. Tong, K. Nakayama, J. T. Vaughan, and N. Kanwisher, *Binocular rivalry and visual awareness in human extrastriate cortex*, Neuron, 21 (1998), pp. 753–759.
[35] A. van Ooyen, *Competition in the development of nerve connections: A review of models*, Network: Comput. in Neural Syst., 12 (2001), pp. 1–47.
[36] X.-J. Wang and J. Rinzel, *Alternating and synchronous rhythms in reciprocally inhibitory model neurons*, Neural Comput., 4 (1992), pp. 84–97.
[37] H. R. Wilson, *Computational evidence for a rivalry hierarchy in vision*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 14499–14503.
[38] H. R. Wilson, R. Blake, and S. H. Lee, *Dynamics of travelling waves in visual perception*, Nature, 412 (2001), pp. 907–910.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 651–711
© 2008 Society for Industrial and Applied Mathematics

Classification of Spatially Localized Oscillations in Periodically Forced Dissipative Systems*

J. Burke†, A. Yochelis‡, and E. Knobloch†

Abstract. Formation of spatially localized oscillations in parametrically driven systems is studied, focusing on the dominant 2:1 resonance tongue. Both damped and self-excited oscillatory media are considered. Near the primary subharmonic instability such systems are described by the forced complex Ginzburg–Landau equation. The technique of spatial dynamics is used to identify three basic types of coherent states described by this equation—small amplitude oscillons, large amplitude reciprocal oscillons resembling holes in an oscillating background, and fronts connecting two spatially homogeneous states oscillating out of phase. In many cases all three solution types are found in overlapping parameter regimes, and multiple solutions of each type may be simultaneously stable. The origin of this behavior can be traced to the formation of a heteroclinic cycle in space between the finite amplitude spatially homogeneous phase-locked oscillation and the zero state. The results provide an almost complete classification of the properties of spatially localized states within the one-dimensional forced complex Ginzburg–Landau equation as a function of the coefficients.

Key words. forced complex Ginzburg–Landau equation, 2:1 resonance, spatial dynamics, localized states, oscillons

AMS subject classifications. 35B32, 35B60, 35G20

DOI. 10.1137/070698191

1. Introduction.

The study of parametrically driven oscillatory systems has a long history dating back to Faraday's experiments in 1831 [16]. A single periodically driven nonlinear oscillator exhibits a number of intriguing phenomena when the natural oscillation frequency and the driving frequency are close to strong resonance [2, 23]. These include frequency locking, hysteresis, and various types of chaotic oscillations [2, 19, 23]. Many of these phenomena are present in spatially extended systems as well, including charge-density waves [6], autocatalytic surface reactions [14], cardiac activity [20], the Belousov–Zhabotinsky (BZ) chemical reaction [34], and optical parametric oscillators [54]. Such systems admit, in addition, different types of spatially structured oscillations such as different types of standing waves and spirals [10, 25, 32, 39, 41, 42, 43, 46, 53, 60]. These phenomena are usually easiest to see in the vicinity of the subharmonic resonance, so called because the system responds with half the forcing frequency. In the following we refer to this resonance as the 2:1 resonance, and the resonance tongue containing homogeneous oscillations phase-locked to half the forcing frequency as the 2:1 resonance tongue. As is well known, these phase-locked states typically

*Received by the editors July 23, 2007; accepted for publication (in revised form) by B. Sandstede March 13, 2008; published electronically July 16, 2008. This work was supported by NASA under grant NNCO4GA47G and NSF under grants DMS-0305968 and DMS-0605238.
http://www.siam.org/journals/siads/7-3/69819.html
†Department of Physics, University of California, Berkeley, CA 94720 (burkej8@socrates.berkeley.edu, knobloch@px1.berkeley.edu).
‡Department of Medicine (Cardiology), University of California, Los Angeles, CA 90095 (yochelis@ucla.edu).

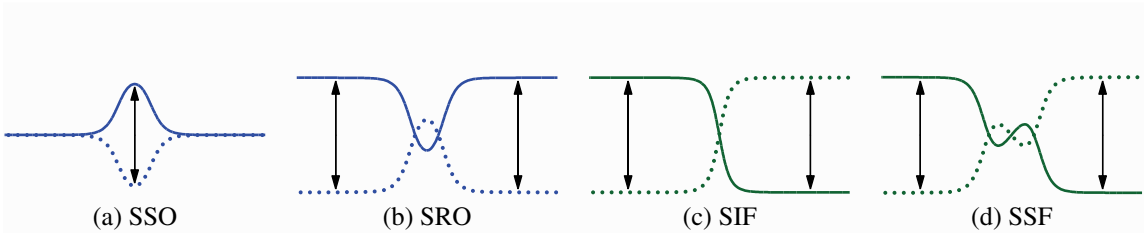


Figure 1. *Sketch of* (a) *a stable standard oscillon (SSO),* (b) *a stable reciprocal oscillon (SRO),* (c) *a stable Ising front (SIF), and* (d) *a stable structured front (SSF). The arrows indicate that after one period of forcing the profile switches from the solid profile to the dotted profile and hence returns to the original configuration only after two periods of the forcing. Although the front in* (d) *is also of Ising type, the terminology used emphasizes the difference in the front profiles. Despite their lack of reflection symmetry, the SIFs and SSFs do not drift. This is in contrast to the so-called Bloch fronts (not shown) which do drift (see text).*

coexist with the trivial state along one of the boundaries of the resonance tongue [17, 46].

Recent experiments on parametrically driven granular media have revealed, in addition, the presence of *spatially localized* oscillations that have been called *oscillons* [1, 35, 38, 52, 50]. These oscillations are embedded in a stationary background (Figure 1(a)) and also oscillate with half the forcing frequency [35, 52]. In the following we refer to states of this type as *standard* oscillons. Additional experiments have identified other types of spatially localized structures [4, 44]. These include hole-like states in a background oscillating state (Figure 1(b)) referred to here as *reciprocal oscillons* [57]. In such states both the hole and the background oscillate with half the forcing frequency, and hence both oscillate synchronously. In addition, experiments on the BZ reaction subjected to stroboscopic optical forcing and on vertically driven granular media reveal the presence of monotonic (or Ising) fronts connecting domains of spatially homogeneous oscillations, each phase-locked to half the forcing frequency but 180° out of phase. Such Ising fronts are often found in the same parameter regime as the reciprocal oscillons already mentioned, i.e., near the boundary of the 2:1 resonance tongue [40]. The Ising fronts can in turn undergo instabilities leading to traveling (Bloch) fronts [11, 18]. In two space dimensions, a transverse instability of Ising fronts may lead to the formation of standing wave labyrinths [60], while in the Bloch front case spiral turbulence was observed [39].

In this paper we show that much of this phenomenology can be captured by the forced complex Ginzburg–Landau (FCGL) equation [10, 15] valid near onset of the primary subharmonic instability. We use this equation to provide a comprehensive classification of the properties of the three types of spatially localized structures mentioned above—standard oscillons, reciprocal oscillons, and fronts—in different parameter regimes. Since the FCGL equation provides a spatial unfolding of the 2:1 strong resonance familiar from dynamical systems theory [19], it can be considered the "normal form" for this resonance in spatially extended systems. As such it arises naturally and inevitably in the applications mentioned above and will arise at small amplitude in all extended oscillatory systems driven at frequency close to twice the natural oscillation frequency. Thus the only difference between these applications lies in the values of the coefficients which are system-specific and have to be computed in terms of physical parameters via standard techniques. Examples can be found in [37, 49] for the case of optical parametric oscillators; see also [5]. In cases where the primitive field equations (or the system parameters) are not known, as frequently occurs in chemical [33, 60] and granular [4, 55]

systems, the FCGL equation is frequently used as a model equation to describe the observed behavior, particularly in two or more space dimensions.

The presence of spatially localized states in the FCGL equation should come as no surprise. As already mentioned, the parametric forcing generally leads to coexistence between a spatially homogeneous phase-locked oscillation and the trivial state along one of the boundaries of the resonance tongue. In these circumstances one can expect to find fronts connecting these two states, and by combining such fronts back-to-back one can construct spatially localized states. In the following we think of structures of this type as heteroclinic *cycles* in space and employ *spatial dynamics* [9] coupled with numerical branch following techniques to establish the presence of such cycles in appropriate parameter regimes [57]. In addition, we demonstrate, following [28], that such cycles can be responsible for the presence of multiple stable standard and reciprocal oscillons in overlapping parameter regions. We also demonstrate that different types of stable monotonic (Figure 1(c)) and structured (Figure 1(d)) fronts, i.e., different types of heteroclinic orbits, can be present in this region as well. Thus the presence of heteroclinic cycles is responsible, under appropriate conditions, for the profusion of stable homoclinic and heteroclinic structures in certain regions of parameter space. These include states we refer to as single-pulse states (discussed above), as well as a variety of so-called multipulse states that resemble bound states of the single-pulse states, and different types of fronts.

Our classification divides naturally into two parts corresponding, in the absence of parametric forcing, to damped and self-excited oscillations. In section 2 we introduce the FCGL equation describing the 2:1 resonance in one spatial dimension and review the properties of spatially homogeneous phase-locked states, i.e., the states that oscillate with half the frequency of the forcing. In section 3 we focus on the damped oscillatory regime and identify analytically bifurcations to small amplitude spatially localized states of homoclinic type, all of which are unstable. We next present the results of numerical continuation, which allows us to follow these states toward larger amplitude; these results indicate that these solutions terminate in a (spatial) heteroclinic bifurcation and show that in some parameter regimes the homoclinic states acquire stability in the vicinity of this bifurcation. We identify the resulting states with the observed standard oscillons. In addition, we find two further types of spatially localized states, corresponding to the reciprocal oscillons and front-like states. The former bifurcate from the spatially homogeneous phase-locked states at finite amplitude and are also homoclinic; in contrast, the front-like states correspond to heteroclinic connections between two phase-locked states 180° out of phase. Both types may terminate in the same heteroclinic bifurcation as the standard oscillons. We show that the reciprocal oscillons may acquire stability near this global bifurcation, while the various structured fronts may do so as well. In section 4 we present parallel results for the case of self-excited oscillations and discuss the impact of the bifurcation to Bloch fronts on the stability of both Ising and structured fronts. The paper concludes in section 5 with a discussion of the results and their significance for ongoing experiments. Details of the analytical computations are relegated to several appendices.

2. The forced complex Ginzburg–Landau equation. We consider a continuous system in one spatial dimension near a bifurcation to spatially homogeneous oscillations with natural frequency ω in the presence of spatially homogeneous forcing with frequency Ω. As is well known, interesting dynamical behavior is associated with strong resonances of the form $\Omega : \omega = n : 1$, $1 \leq n \leq 4$ [2, 23]. Of these the resonance tongue associated with the subharmonic

resonance $\Omega \approx 2\omega$ is the broadest and therefore ideally suited for experimental study [14, 18, 32, 39, 53]. Inside the resonance tongue the system responds to the forcing with oscillations at frequency $\Omega/2$, corresponding to *phase-locked* states [2, 23]. Outside of the tongue the frequency difference $|\omega - \Omega/2|$ is too large and the response frequency is no longer locked to the forcing frequency.

In the present paper we are interested in including large-scale spatial modulation in this theory. Thus we suppose that a dynamical observable w takes the form

$$w = w_0 + Ae^{i\Omega t/2} + c.c. + \cdots, \tag{2.1}$$

where w_0 represents the equilibrium state, $A(x,t)$ is a complex amplitude, and the ellipses denote higher order terms. The oscillation amplitude $A(x,t)$ obeys the FCGL equation [10, 15]

$$A_t = (\mu + i\nu)A - (1 + i\beta)|A|^2A + (1 + i\alpha)A_{xx} + \gamma\bar{A}, \tag{2.2}$$

where μ represents the distance from onset of the oscillatory instability, ν is the detuning from the unforced frequency, and α, β, and γ represent dispersion, nonlinear frequency correction, and the forcing amplitude, respectively. Here $\bar{A}$ is the complex conjugate of A. Equation (2.2) can be derived by standard asymptotic methods from the relevant governing equations provided the amplitude A remains small, i.e., provided the system is close to onset of spontaneous spatially homogeneous oscillations and the forcing amplitude is suitably small. In the absence of forcing, these oscillations grow when $\mu > 0$ but decay when $\mu < 0$.

With the exception of the trivial solution $A = 0$, stationary solutions of (2.2) are all "phase-locked" in the sense that they correspond to observables w in the original system that oscillate at exactly half the driving frequency, $\Omega/2$. We use the term *uniform* state to refer to a phase-locked state that is independent of x. States of this type take the form $A = R\exp(i\phi)$, where

$$R^2 = \left(R^{\pm}\right)^2 \equiv \frac{\beta(\nu - \nu_\beta) \pm \sqrt{\rho_\beta^2\gamma^2 - (\nu - \beta\mu)^2}}{\rho_\beta^2}, \tag{2.3}$$

and $\phi = \phi^{\pm}$ solves

$$\cos 2\phi^{\pm} = \frac{\left(R^{\pm}\right)^2 - \mu}{\gamma}, \quad \sin 2\phi^{\pm} = \frac{\nu - \beta\left(R^{\pm}\right)^2}{\gamma}. \tag{2.4}$$

The uniform phase-locked states exist provided $R^2 > 0$. Here $\rho_\beta \equiv \sqrt{1+\beta^2}$ and

$$\nu_\beta \equiv -\frac{\mu}{\beta}. \tag{2.5}$$

In the following we denote the two branches of solutions in $0 \le \phi^{\pm} < \pi$ by $A_u^{\pm}$, so $\operatorname{Im} A_u^{\pm} \ge 0$; the solutions in $\pi \le \phi^{\pm} < 2\pi$ which lie exactly out of phase with $A_u^{\pm}$ correspond to $-A_u^{\pm}$.

When $\nu > \nu_\beta$, the A_u^- states appear via a subcritical bifurcation from $A = 0$ when the forcing amplitude γ reaches $\gamma = \gamma_0$,

$$\gamma_0 \equiv \sqrt{\mu^2 + \nu^2}, \tag{2.6}$$

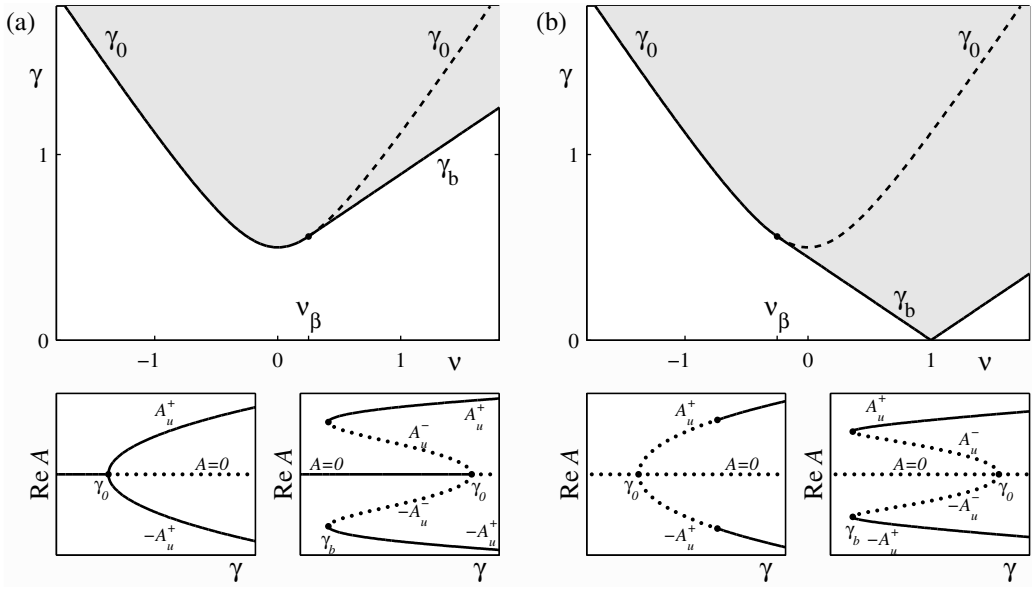


Figure 2. *The boundary of the* 2:1 *resonance tongue in the* (ν,γ) *plane when* $\alpha = 1$, $\beta = 2$, *and* (a) $\mu = -0.5$, (b) $\mu = 0.5$. *The line* $\gamma = \gamma_0$ *is plotted as solid (dashed) when the bifurcation to uniform phase-locked states is supercritical (subcritical). The shaded region indicates the existence of spatially uniform solutions* A_u^+. *The lower panels show typical bifurcation diagrams when the bifurcation at* γ_0 *is supercritical* ($\nu < \nu_\beta$) *and subcritical* ($\nu > \nu_\beta$). *In the bifurcation diagrams solid (dotted) lines represent solutions which are stable (unstable) with respect to uniform perturbations. In* (b) *the* A_u^+ *state may be stabilized either at* γ_b *or at a Hopf bifurcation.*

and are unstable. These states annihilate with the A_u^+ in a saddle-node bifurcation at $\gamma = \gamma_b$,

$$\gamma_b \equiv \frac{|\nu - \beta\mu|}{\rho_\beta}. \tag{2.7}$$

In contrast, when $\nu < \nu_\beta$, the A_u^+ states bifurcate supercritically from $A = 0$ at γ_0 and the saddle-node bifurcation is absent. The line $\gamma = \gamma_b$ is tangent to the curve $\gamma = \gamma_0$ at $\nu = \nu_\beta$, as shown in Figure 2.

The large number of parameters in (2.2) is responsible for a wide range of behavior. In section 3 we consider the case of damped oscillations, $\mu < 0$, followed by the self-excited case, $\mu > 0$, in section 4. Without loss of generality we restrict both discussions to the case $\beta > 0$ but allow α to be positive or negative. The resulting half-space splits into five regions (Figure 3), independently of μ, each characterized by distinct behavior in the (ν,γ) plane; we are free to choose $\gamma > 0$ but must allow ν to be positive or negative. Bifurcation diagrams showing branches of stationary solutions $A(x)$ as a function of the parameter γ correspond to vertical (constant ν) slices through this plane. In general, the types of solutions present in such diagrams are determined by the value of ν relative to a few critical values, such as ν_β.

3. Damped oscillatory regime. In this section we examine solutions to (2.2) in the damped case, $\mu < 0$. The resonance region containing the states $A_u^\pm$ is shown in Figure 2(a) and is characteristic of the uniform solutions found in all five regions of the (α,β) parameter

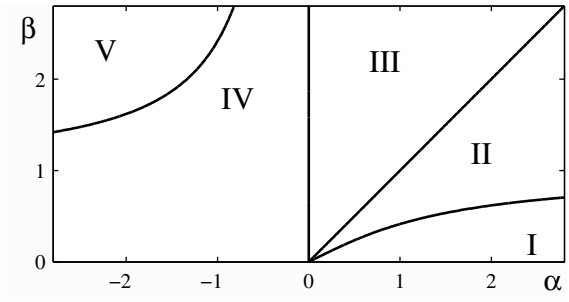


Figure 3. *The (α, β) parameter plane splits into five regions based on the behavior of the solutions. The line $\alpha = 0$ divides* III *from* IV*, while the line $\alpha = \beta$ divides* II *from* III*. The curves dividing* I *from* II *and* IV *from* V *are given by $z(\alpha, \beta) = 0$, cf. (3.18); $z < 0$ in* I *and* V *but $z > 0$ in* II–IV*. These results are independent of the sign and value of μ.*

plane introduced in Figure 3, provided that $\mu < 0$. In the following we refer to these as Regions I^-–V^-, where the superscript refers to the sign of μ. The $A = 0$ state is stable to uniform perturbations in $\gamma < \gamma_0$ and unstable in $\gamma > \gamma_0$. The A_u^+ state is present in the shaded region and is stable with respect to spatially uniform perturbations; the A_u^- state is present in $\gamma_b < \gamma < \gamma_0$ and is unstable. Thus the parametric forcing is responsible for the creation of a region of *bistability* between $A = 0$ and A_u^+, defined by $\nu > \nu_\beta$, $\gamma_b < \gamma < \gamma_0$ (Figure 2(a)). This region plays an important role in what follows.

There are two types of nonuniform solutions of (2.2)—spatially periodic and spatially localized. To study these, we rewrite (2.2) in terms of the real and imaginary parts of the amplitude A, $U \equiv \mathrm{Re}\, A$ and $V \equiv \mathrm{Im}\, A$:

$$\begin{bmatrix} U_t \\ V_t \end{bmatrix} = (\mathcal{L} + \mathcal{N}) \begin{bmatrix} U \\ V \end{bmatrix}. \tag{3.1}$$

Here $\mathcal{L}$ is the linear operator

$$\mathcal{L} = \begin{bmatrix} \mu + \gamma & -\nu \\ \nu & \mu - \gamma \end{bmatrix} + \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{xx}, \tag{3.2}$$

while $\mathcal{N}$ includes the nonlinear terms:

$$\mathcal{N} = -\left(U^2 + V^2\right) \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}. \tag{3.3}$$

The steady states of this system satisfy a fourth order ordinary differential equation (ODE) in x that can be studied using a combination of bifurcation theory and numerical branch following techniques. It is important to recognize that (3.1) is *reversible* in the sense that it is invariant under the spatial reflection $x \to -x$. In the following we think of x as a time-like variable and classify the states of interest as either homoclinic or heteroclinic orbits depending on the behavior of $A(x)$ as $x \to \pm\infty$. In addition, we distinguish between "small" amplitude states, which bifurcate from the trivial state $A = 0$, and "large" amplitude states that are related to the finite amplitude uniform phase-locked states A_u^+. Small amplitude orbits

homoclinic to $A = 0$ correspond to standard oscillons, while large amplitude orbits homoclinic to A_u^+ correspond to reciprocal oscillons. Heteroclinic orbits between $\pm A_u^+$ correspond to the observed fronts. In either case a localized solution must leave a homogeneous state as x increases from $x = -\infty$ and must approach a homogeneous state (same or different) as $x \to \infty$. Thus we examine both the stability of the uniform states in time as well as their stability in x, i.e., their temporal *and* their spatial eigenvalues.

3.1. Small amplitude states for $\alpha > 0$. In this subsection we analyze, analytically and numerically, small amplitude states which bifurcate from the $A = 0$ state. We first consider the case $\alpha > 0$ spanning the Regions I^-–III^- of the (α, β) plane shown in Figure 3. The behavior in $\alpha < 0$, spanning Regions IV^- and V^-, is described in a subsequent section.

3.1.1. Temporal stability of $A = 0$. The (complex) growth rate s of an infinitesimal perturbation of the trivial state $A = 0$ with wavenumber k,

$$\begin{bmatrix} U \\ V \end{bmatrix} = \epsilon \begin{bmatrix} u \\ v \end{bmatrix} e^{ikx+st} + c.c., \tag{3.4}$$

where $\epsilon \ll 1$ and u, v are constants, is given by

$$s^2 + 2s(k^2 - \mu) + k^4\rho_\alpha^2 - 2k^2\alpha(\nu - \nu_\alpha) + \gamma_0^2 - \gamma^2 = 0. \tag{3.5}$$

Here $\rho_\alpha \equiv \sqrt{1 + \alpha^2}$ and

$$\nu_\alpha \equiv -\frac{\mu}{\alpha}. \tag{3.6}$$

The $A = 0$ state is stable at any value of the parameters such that $\operatorname{Re} s < 0$ for all k. Analysis of the dispersion relation reveals three bifurcations of interest. The first occurs at γ_0 and is a pitchfork bifurcation to the uniform phase-locked states $A_u^\pm$. This bifurcation is associated with the change in stability of the $k = 0$ mode. The growth rate of this mode near γ_0, for $|k| \ll 1$, is given by

$$s(k, \gamma) \approx -\frac{\alpha}{\mu}(\nu - \nu_\alpha)k^2 - \frac{\gamma_0}{\mu}(\gamma - \gamma_0) \tag{3.7}$$

and indicates that the $A = 0$ state is stable for all $\gamma < \gamma_0$ provided $\nu < \nu_\alpha$; at γ_0 this state loses stability with respect to uniform disturbances and is unstable to a range of wavenumbers in $\gamma > \gamma_0$, as shown in Figure 4(a). In contrast, when $\nu > \nu_\alpha$, the $A = 0$ state is already unstable to spatially *nonuniform* perturbations when the $k = 0$ bifurcation takes place (Figure 4(b)).

The second bifurcation associated with the dispersion relation (3.5) is a finite wavenumber (Turing) bifurcation, present when $\nu > \nu_\alpha$ (Figure 4(b)). Here the $A = 0$ state is stable for small γ and first loses stability at $\gamma = \gamma_a$,

$$\gamma_a \equiv \frac{|\nu - \mu\alpha|}{\rho_\alpha}, \tag{3.8}$$

to perturbations with wavenumber k_a,

$$k_a \equiv \frac{\sqrt{\alpha(\nu - \nu_\alpha)}}{\rho_\alpha}. \tag{3.9}$$

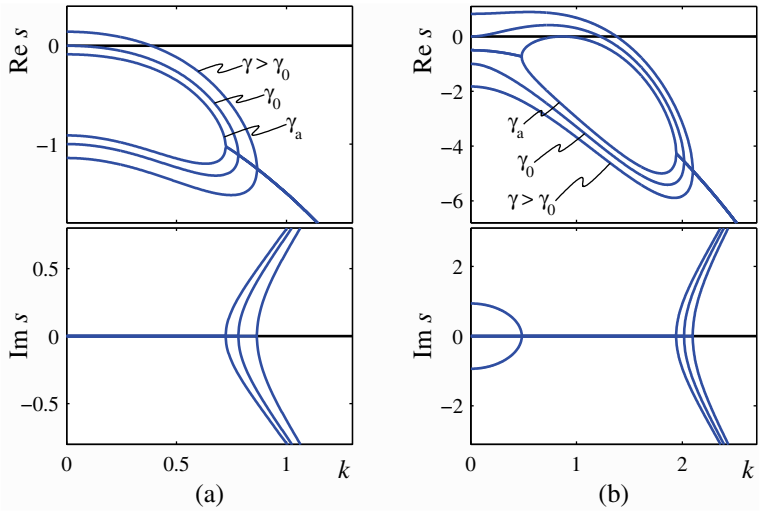


Figure 4. *Dispersion relation for the $A = 0$ state showing the growth rate* $\mathrm{Re}\, s$ *and oscillation frequency* $\mathrm{Im}\, s$ *of perturbations as a function of the wavenumber k for several values of γ. When $\alpha > 0$,* (a) *corresponds to $\nu < \nu_\alpha$ and* (b) *to $\nu > \nu_\alpha$, and vice versa when $\alpha < 0$.*

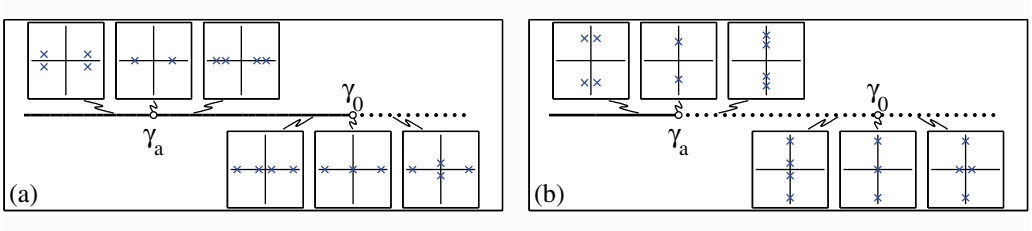


Figure 5. *Temporal stability along the $A = 0$ branch as a function of γ: solid (dotted) lines represent stable (unstable) solutions. The insets show the spatial eigenvalues λ of $A = 0$ in the complex λ plane. When $\alpha > 0$,* (a) *corresponds to $\nu < \nu_\alpha$ and* (b) *to $\nu > \nu_\alpha$, and vice versa when $\alpha < 0$.*

The resulting bifurcation produces a branch of spatially periodic states with wavenumber k_a. When $\gamma > \gamma_a$, the $A = 0$ state is unstable to a range of wavenumbers (which includes $k = 0$ once $\gamma > \gamma_0$).

The above results are summarized in Figure 5: the state $A = 0$ is stable in time (solid line) for $\gamma < \gamma_0$ when $\nu < \nu_\alpha$ and for $\gamma < \gamma_a$ when $\nu > \nu_\alpha$. The dots indicate that $A = 0$ is unstable.

The third bifurcation, mentioned briefly in the introduction, is a Hopf bifurcation corresponding to $\mathrm{Re}\, s = 0$, $\mathrm{Im}\, s \neq 0$ at $k = 0$. This bifurcation occurs at $\mu = 0$ provided $\gamma < \nu$ and produces a branch of spatially uniform oscillations. This mode and its interaction with the Turing mode have been studied in [58, 59] but are not considered in the present paper, which focuses on values of μ away from $\mu = 0$.

The line $\gamma = \gamma_a$ defined in (3.8) is tangent to the curve $\gamma = \gamma_0$ at ν_α. Figures 6 and 7 show this line in relation to γ_0 and γ_b for Regions II^- and III^- in the $\alpha > 0$ quadrant of Figure 3. A similar plot for Region I^- is omitted since the behavior resembles that found in Region II^-. In

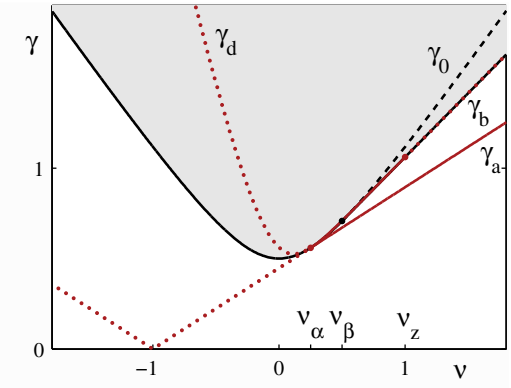


Figure 6. *The (ν,γ) plane for Region* II^-. *The curves γ_a and γ_d correspond to bifurcations in $\nu > \nu_\alpha$ and $\nu_\alpha < \nu < \nu_z$ (solid lines), and to Belyakov–Devaney points elsewhere (dotted lines). Thus γ_d forms the lower boundary of the region (shaded) of stable spatially uniform solutions A_u^+ in the interval $\nu_\alpha < \nu < \nu_z$ (barely visible). The picture for Region* I^- *is qualitatively similar except for the absence of the tangency at ν_z since $\nu_z < \nu_\beta$. Thus in* I^- *γ_d forms the lower boundary of the shaded region everywhere in $\nu > \nu_\alpha$. Other notation is the same as in Figure 2. Parameters: $\mu = -0.5$, $\alpha = 2$, $\beta = 1$.*

these figures γ_a is shown as a solid line in $\nu > \nu_\alpha$ where it corresponds to a Turing bifurcation. For later reference it is also plotted (as a dotted line) in $\nu < \nu_\alpha$. Thus the $A = 0$ state is stable up to γ_a when this is plotted as a solid line and up to γ_0 elsewhere. The value of ν_β and the location of γ_b are determined by nonlinearity and have no bearing on the linear stability of $A = 0$ discussed above, but the value of ν_α relative to ν_β will play an important role in what follows. We therefore distinguish between the case $\nu_\alpha < \nu_\beta$ (i.e., $\alpha > \beta$, spanning Regions I^- and II^-) and the case $\nu_\alpha > \nu_\beta$ (i.e., $\alpha < \beta$, Region III^-). Although γ_a must lie below γ_0, it is clear from these figures that it may lie either above or below γ_b. For future reference we include in these figures a third critical value, ν_z, as well as a curve labeled γ_d; both are related to finite amplitude effects discussed below.

For spatially nonuniform steady states $A(x)$, whether spatially extended or localized, the linearized stability problem constitutes an eigenvalue problem for the growth rate s. This problem has in general an infinite number of solutions, although only the eigenvalues with the largest growth rates are of interest. In the following we omit the details of the required computations (these are similar to those in [7]) but indicate whether a particular solution is stable or unstable. In the case of instability we indicate the number and shape of the eigenfunctions whose eigenvalues s have positive real part. For spatially localized states the two commonest instabilities are amplitude and phase instabilities. The former is characterized by an eigenfunction with the same parity (even or odd) as the steady state $A(x)$, while the latter has the opposite parity. In both instances the corresponding eigenvalues are real.

3.1.2. Spatial eigenvalues of $A = 0$. As mentioned above, a necessary condition for the presence of an exponentially localized state that approaches $A = 0$ as $x \to \pm\infty$ is that this state have at least one spatial eigenvalue with positive real part and one with negative real part. In this section we identify the regions in the (ν,γ) plane where this is the case. This amounts to examining the stability in x of the fixed point $A = 0$. The spatial eigenvalues λ

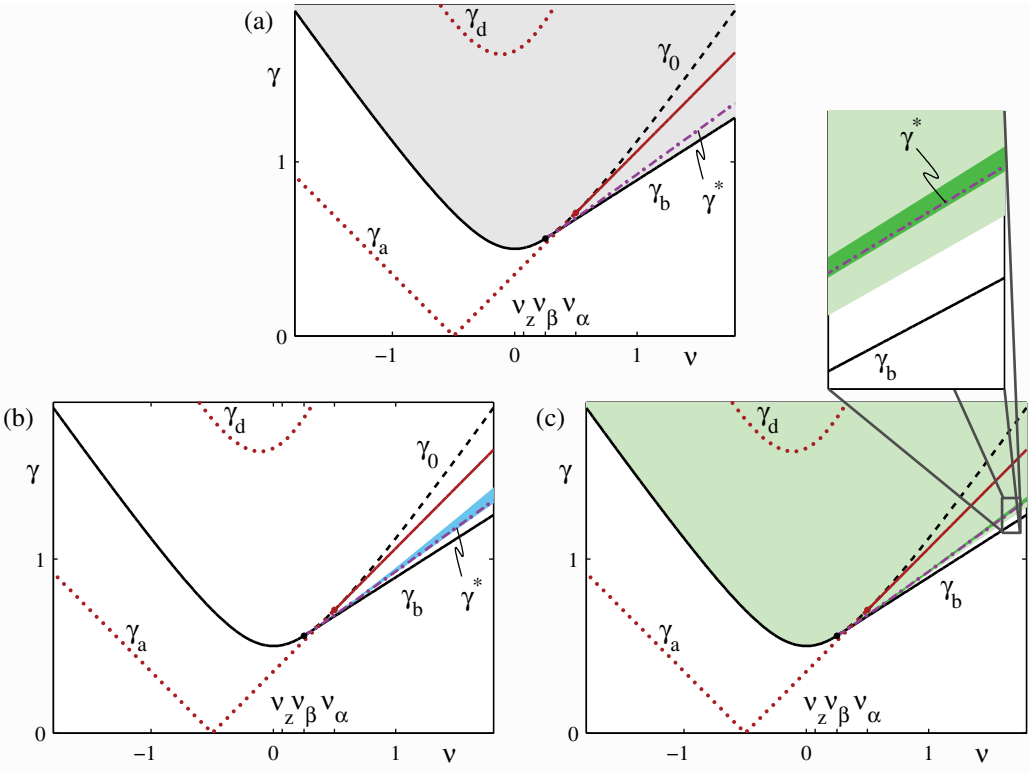


Figure 7. *The* (ν,γ) *plane for Region* III$^-$. *In* (a) *shading indicates the presence of stable uniform states* A_u^+, *while in* (b) *it indicates the existence of stable reciprocal oscillons (SROs). In* (c) *the light shading indicates existence of stable Ising fronts (SIFs), while the darker shading indicates the (very narrow) region of existence of stable structured fronts (SSFs), shown more clearly in the inset. The heteroclinic cycle forms along the dot-dashed line* γ^*. *The remaining notation is the same as in Figure* 6. *Parameters:* $\mu = -0.5$, $\alpha = 1$, $\beta = 2$.

of this fixed point are determined by linearizing (3.1) around the $A = 0$ state:

$$\begin{bmatrix} U \\ V \end{bmatrix} = \epsilon \begin{bmatrix} u \\ v \end{bmatrix} e^{\lambda x}. \tag{3.10}$$

The four eigenvalues satisfy a quadratic equation for λ^2,

$$\rho_\alpha^2 \lambda^4 + 2\alpha(\nu - \nu_\alpha)\lambda^2 + (\gamma_0^2 - \gamma^2) = 0. \tag{3.11}$$

In particular, if λ is an eigenvalue, then $\bar{\lambda}$ and $-\lambda$ are as well. Consequently there are four possibilities: (i) the spatial eigenvalues are real, (ii) there is a quartet of complex eigenvalues, (iii) the spatial eigenvalues are imaginary, and (iv) two eigenvalues are real and two are imaginary. The transition from (i) to (ii) is via a Belyakov–Devaney point [9], while the transition from (ii) to (iii) corresponds to a (spatial) Hopf bifurcation with 1:1 resonance [26].

Localized states correspond to intersections of the stable and unstable manifolds of $A = 0$. In cases (i) and (ii) this state has a two-dimensional stable and a two-dimensional unstable manifold. These manifolds are transverse to the two-dimensional fixed point subspace of the

symmetry $x \to -x$ and $A \to A$ and hence intersect in a structurally stable way [56]; i.e., we expect localized states in cases (i) and (ii) only. In contrast, in case (iv) the stable and unstable manifolds are only one-dimensional and homoclinic orbits are exceptional [30, 36].

The motion of the eigenvalues in the complex λ plane as γ varies depends critically on the value of ν relative to $\nu_\alpha > 0$. We begin by observing that for small γ the eigenvalues form a complex quartet as in case (ii). When $\nu < \nu_\alpha$ (Figure 5(a)) these eigenvalues collide on the real axis in a Belyakov–Devaney point when $\gamma = \gamma_a$. Because this collision occurs away from the imaginary λ axis, it does not correspond to a local bifurcation of the $A = 0$ state. Above γ_a the eigenvalues split but remain on the real axis, as in case (i). As γ continues to increase, two of the eigenvalues move toward the origin and collide at $\lambda = 0$ when $\gamma = \gamma_0$. This collision does correspond to a bifurcation, and in fact it is at γ_0 that the uniform phase-locked states $A_u^\pm$ bifurcate from A_0. For $\gamma > \gamma_0$ the zero eigenvalues split along the imaginary axis, resulting in eigenvalue structure (iv). In contrast, when $\nu > \nu_\alpha$ (Figure 5(b)), the complex eigenvalues that collide at γ_a do so on the imaginary axis at $\lambda = \pm i k_a$. This is the Hopf bifurcation with 1:1 resonance (in space) or equivalently the Turing bifurcation (in time) mentioned above. In the language of spatial dynamics the spatially extended states produced at this bifurcation are referred to as *spatially periodic* states with wavenumber k_a. Above γ_a the eigenvalues split but remain imaginary, as in case (iii). Two of these collide at the origin when $\gamma = \gamma_0$ and split along the real axis, resulting, as before, in eigenvalue structure of type (iv). Thus when $\nu < \nu_\alpha$, spatially localized states may exist everywhere in $\gamma < \gamma_0$, while for $\nu > \nu_\alpha$ they are expected in $\gamma < \gamma_a$ only. These regimes correspond precisely to those where the $A = 0$ state is stable in time, as indicated in Figure 5. This observation is significant since localized states in general inherit any instabilities of the background asymptotic state (but see [3, 48]).

3.1.3. Localized states bifurcating from $\gamma = \gamma_0$. The uniform phase-locked states A_u^- bifurcate from $A = 0$ at γ_0, and this bifurcation is subcritical if $\nu > \nu_\beta$. In Appendix A we show that if in addition $\nu < \nu_\alpha$ (Figure 5(a)), then at γ_0 there is in addition a subcritical bifurcation to *localized* states. Near γ_0 these take the form

$$A_{L,0}^\pm(x) = \pm(\eta_0 + i)\sqrt{\frac{\gamma_0 - \gamma}{b_0/2}}\,\operatorname{sech}\left(\sqrt{\frac{\gamma_0 - \gamma}{-a_0}}\,x\right), \tag{3.12}$$

where $\eta_0 \equiv (\gamma_0 - \mu)/\nu$, $a_0 \equiv \alpha(\nu - \nu_\alpha)/\gamma_0 < 0$, and $b_0 \equiv 2\beta(\nu - \nu_\beta)(\gamma_0 - \mu)/\nu^2 > 0$. These localized states are biasymptotic to $A = 0$ as $x \to \pm\infty$ and are present in Region III$^-$ of the (α, β) plane only, provided $\nu_\beta < \nu < \nu_\alpha$, $\gamma \le \gamma_0$ (Figure 7).

These analytical solutions can be followed away from γ_0 using numerical continuation [13]. Figure 8 shows the resulting bifurcation diagram typical of $\nu_\beta < \nu < \nu_\alpha$ in Region III$^-$. In this and all subsequent bifurcation diagrams we plot solutions in terms of their norms

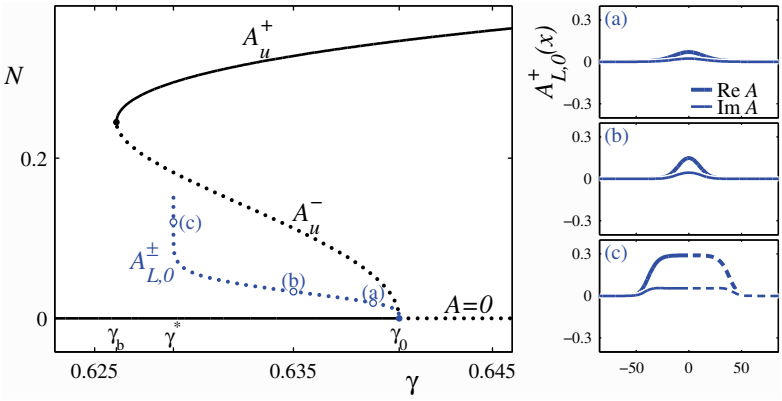


Figure 8. *Bifurcation diagram for $\nu_\beta < \nu < \nu_\alpha$ in Region* III$^-$*, showing the branches of localized states $A_{L,0}^\pm$ found by continuing the solution (3.12) away from γ_0, as well as the uniform states $A_u^\pm$. Solid (dotted) lines correspond to stable (unstable) solutions. Right panels show profiles of $A_{L,0}^+$ at locations* (a), (b), *and* (c), *respectively; panel* (c) *shows a broad homoclinic orbit near $\gamma^* \approx 0.6289$ decomposed into a pair of nearly heteroclinic connections, one connecting $A = 0$ to A_u^+ (solid line) and the other connecting A_u^+ to $A = 0$ (dashed line). Parameters: $\mu = -0.5$, $\alpha = 1$, $\beta = 2$, $\nu = 0.4$ with* (a) $\gamma = 0.639$, (b) $\gamma = 0.635$, (c) $\gamma = 0.6289$.

$$N = \sqrt{\frac{1}{L}\int_{-L/2}^{L/2} \{|A|^2 + |\partial_x A|^2\}\, \mathrm{d}x}, \tag{3.13}$$

where L is the large but finite spatial period used in the numerical computations. Typically $L = 200$. Figure 8 shows not only the uniform phase-locked states $A_u^\pm$ but also the branches of spatially localized solutions $A_{L,0}^\pm$ found by continuing the analytical result (3.12) away from γ_0. Solutions along these branches are always even in x, $A_{L,0}^\pm(-x) = A_{L,0}^\pm(x)$. In addition the two branches of solutions are related by the symmetry $A_{L,0}^+(x) = -A_{L,0}^-(x)$ and therefore always have identical norms in the bifurcation diagram. The right panels in Figure 8 show sample solutions along the $A_{L,0}^+$ branch. Near γ_0 the profile is broad (Figure 8(a)). As γ decreases away from γ_0, the solution first contracts in x forming a well-localized state (Figure 8(b)). However, as γ approaches the value γ^*, these localized states broaden again, with $A_{L,0}^+$ spending more and more "time" near the uniform value A_u^+. The resulting broad homoclinic orbit can be thought of as a pair of nearly heteroclinic orbits, the first connecting $A = 0$ to A_u^+ and the second connecting A_u^+ back to $A = 0$ (Figure 8(c)). We refer to this pair of heteroclinic connections as a heteroclinic *cycle*. Numerical branch following reveals that the branches of localized states $A_{L,0}^\pm$ approach γ^* monotonically from above; as a result the $A_{L,0}^\pm$ solutions remain *unstable* to an amplitude (even) mode throughout their range of existence. In $\gamma < \gamma^*$ the broad localized states of the type shown in Figure 8(c) evolve in time by decreasing their width and eventually collapsing to the (stable) uniform profile $A = 0$. In contrast, in $\gamma > \gamma^*$ states of this type grow in width, filling more and more of the domain with the (stable) uniform phase-locked state A_u^+.

The point γ^* at which a heteroclinic cycle is present is analogous to the so-called Maxwell point in the theory of variational systems [7, 45], where it corresponds to the equal energy case. It turns out that the multiplicity of steady states near γ^* is determined by the spatial

eigenvalues of the two states connected by the cycle and hence is independent of the type of system.

3.1.4. Localized states bifurcating from $\gamma = \gamma_a$. Recall that for $\nu > \nu_\alpha$ the four spatial eigenvalues of the trivial state form a complex quartet when $\gamma < \gamma_a$, collide at $\lambda = \pm ik_a$ when $\gamma = \gamma_a$, and then split but remain imaginary when $\gamma > \gamma_a$. For this case, weakly nonlinear analysis in the vicinity of γ_a (see Appendix B) reveals that there is a branch of small amplitude spatially periodic solutions given by

$$A_{P,a}(x) = (\eta_a + i)\sqrt{\frac{\gamma_a - \gamma}{b_a}}\cos(k_a x + \varphi), \tag{3.14}$$

where the phase φ is arbitrary, $b_a \equiv 6\eta_a(\beta - \alpha)$, and, since $\alpha > 0$,

$$\eta_a = \alpha + \rho_\alpha > 0 \tag{3.15}$$

(see (B.7) in Appendix B). Consequently this bifurcation is supercritical whenever $\beta < \alpha$; in this case no additional *localized* states bifurcate from γ_a. This is the case in Regions I^- and II^- of Figure 3. In contrast, when $\beta > \alpha$ (Region III^-), the bifurcation to spatially periodic states is subcritical, and localized spatial *oscillations* of the form

$$A_{L,a}(x) = (\eta_a + i)\sqrt{\frac{\gamma_a - \gamma}{b_a/2}}\,\mathrm{sech}\left(\sqrt{\frac{\gamma_a - \gamma}{a_a}}x\right)\cos(k_a x + \varphi) \tag{3.16}$$

bifurcate from $A = 0$ simultaneously with the spatially periodic states and in the same direction. Here $a_a \equiv 2\rho_\alpha^2 k_a^2/\gamma_a > 0$. Bifurcations of this type occur in Figure 7 in the region $\nu > \nu_\alpha$ with the localized oscillations present below the line $\gamma = \gamma_a$. Within the asymptotic analysis, the phase φ in (3.16) is again arbitrary, but terms beyond all orders select solutions with the phases $\varphi = 0, \pi/2, \pi, 3\pi/2$ [8, 31]. In the following we therefore distinguish between families of even ($A_{L,a}^{\mathrm{even}}$; $\varphi = 0, \pi$) and odd ($A_{L,a}^{\mathrm{odd}}$; $\varphi = \pi/2, 3\pi/2$) parity homoclinic solutions.

The analytical solutions in (3.14) and (3.16) can be followed away from γ_a using numerical continuation, as shown in Figure 9. The left panel of the figure shows the branch of spatially periodic states $A_{P,a}$ as well as the branches of both even and odd localized states $A_{L,a}$. The branch of uniform phase-locked states $A_u^\pm$ is shown for reference. The right panels show sample solutions along the two branches of localized states. The small amplitude localized states are broad near γ_a (Figure 9(a),(c)); as γ decreases away from γ_a, the envelope of these states contracts in x, forming well-localized packets (Figure 9(b),(d)). As γ decreases toward γ^*, the profiles of both even and odd states broaden. Near γ^* the even localized states $A_{L,a}^{\mathrm{even}}$ can be thought of as a pair of nearly heteroclinic orbits, the first of which connects $A = 0$ to A_u^+ with the second connecting A_u^+ back to $A = 0$ (Figure 9(c)). In contrast, $A_{L,a}^{\mathrm{odd}}$ near γ^* consists of three parts: a pair of nearly heteroclinic orbits connecting $A = 0$ to $\pm A_u^+$ on either side of a heteroclinic connection between A_u^+ and $-A_u^+$ (Figure 9(f)). The numerical results again suggest that at γ^* there exist heteroclinic cycles connecting $A = 0$ and $\pm A_u^+$. As in Figure 8, the branches of localized states approach γ^* monotonically from above, and the solutions are always unstable. The $A_{L,a}^{\mathrm{even}}$ branch is unstable to a single amplitude (even) mode, while the $A_{L,a}^{\mathrm{odd}}$ branch is unstable to two modes, an amplitude (odd) mode and a phase (even)

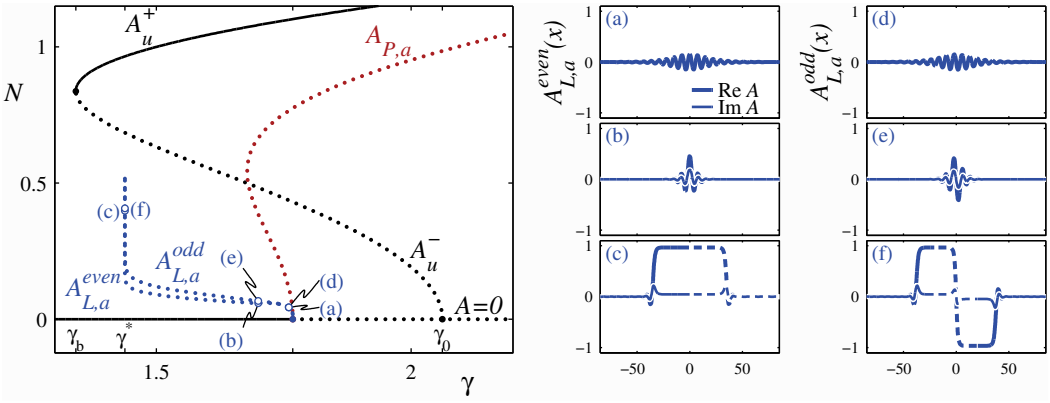


Figure 9. *Bifurcation diagram for $\nu > \nu_\alpha$ in Region* III$^-$*, showing branches of odd and even parity localized states $A_{L,a}$ found by continuation of the solutions (3.16) away from γ_a, as well as the spatially periodic states $A_{P,a}$ and the uniform states $A_u^\pm$. Solid (dotted) lines correspond to stable (unstable) solutions.* (a)–(f) *Profiles of $A_{L,a}$ at the locations indicated in the left panel. Panel* (c) *shows that the broad homoclinic orbits present near $\gamma^* \approx 1.437$ on the $A_{L,a}^{\rm even}$ branch can be thought of as pairs of nearly heteroclinic connections between $A = 0$ and A_u^+ (solid and dashed lines). Panel* (f) *shows that the corresponding odd parity states take the form of a pair of nearly heteroclinic connections between $A = 0$ and $\pm A_u^+$ (solid lines), with a nearly heteroclinic orbit from A_u^+ to $-A_u^+$ (dashed) in between. Parameters: $\mu = -0.5$, $\alpha = 1$, $\beta = 2$, $\nu = 2$, with* (a),(d) $\gamma = 1.76$, (b),(e) $\gamma = 1.70$, (c),(f) $\gamma = 1.437$.

mode. The $A_{P,a}$ states are likewise always unstable; although the nature of this instability varies throughout Region III$^-$, the unstable modes typically correspond to long wavelength disturbances.

The heteroclinic cycles identified in the bifurcation diagrams in Figures 8 and 9 occur along the line γ^* in Figure 7, corresponding to Region III$^-$; the line emerges from the tangency at ν_β that creates the saddle-node bifurcation and passes continuously from $\nu_\beta < \nu < \nu_\alpha$ (Figure 8) to $\nu > \nu_\alpha$ (Figure 9).

3.1.5. Small amplitude states for $\alpha < 0$. When $\alpha < 0$ as in Regions IV$^-$ and V$^-$, the critical value ν_α is negative. This change in sign requires a reinterpretation of the equations derived above. In this case the dispersion relation (3.5) predicts that the Turing bifurcation of the $A = 0$ state at γ_a is present when $\nu < \nu_\alpha$ instead of $\nu > \nu_\alpha$. Figures 10 and 11 show the (ν, γ) plane for Regions IV$^-$ and V$^-$, respectively, with γ_a plotted as a solid line in $\nu < \nu_\alpha$, where it corresponds to a bifurcation point. Because the analysis at γ_a performed in Appendix B now applies in $\nu < \nu_\alpha$, the expression for η_a in (B.7) reduces to

$$\eta_a = \alpha - \rho_\alpha < 0. \tag{3.17}$$

Moreover, when $\nu > \nu_\alpha$, the spatial eigenvalues at γ_0 are real, and the collision of eigenvalues at γ_a occurs on the real axis; localized states are therefore expected in $\gamma < \gamma_0$ (as in Figure 5(a)). In contrast, when $\nu < \nu_\alpha$, the eigenvalues at γ_0 are imaginary and the collision at γ_a occurs on the imaginary axis. In this case localized states are expected in $\gamma < \gamma_a$ (as in Figure 5(b)).

The change in sign of η_a also affects the bifurcations from γ_a present in $\alpha < 0$. As $\beta > \alpha$ everywhere in this quadrant the bifurcation to the spatially periodic states $A_{P,a}$ is supercritical

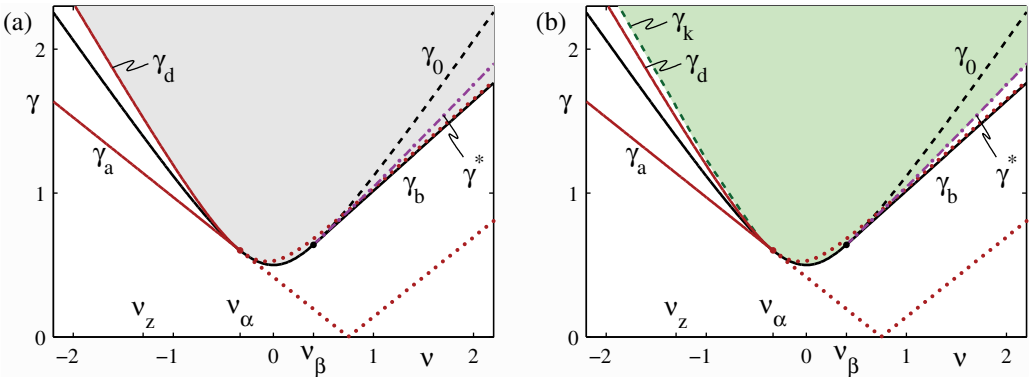


Figure 10. *The (ν,γ) plane in Region* IV^-. *In* (a) *shading indicates the presence of stable uniform states A_u^+, while in* (b) *it indicates the SIF region. The line γ_k in* (b) *corresponds to saddle-node bifurcations of large amplitude fronts referred to as kinks. Parameters: $\mu=-0.5$, $\alpha=-1.5$, $\beta=1.25$.*

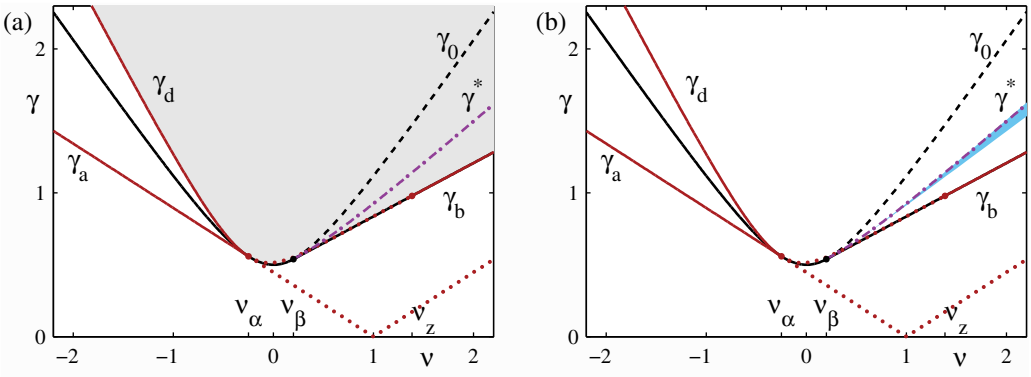


Figure 11. *The (ν,γ) plane in Region* V^-. *In* (a) *shading indicates the presence of stable uniform states A_u^+, while in* (b) *it indicates the SSO region. Parameters: $\mu=-0.5$, $\alpha=-2$, $\beta=2.5$.*

throughout Regions IV^- and V^-. Moreover, no bifurcations to localized spatial oscillations of the form (3.16) can take place.

Expression (3.12) remains valid in $\alpha<0$: in both Regions IV^- and V^-, $\nu_\beta>\nu_\alpha$ and hence a branch of small amplitude single-peaked localized states bifurcates subcritically at γ_0 whenever $\nu>\nu_\beta$; such states are therefore present in $\gamma<\gamma_0$ (see Figures 10 and 11). Thus the only small amplitude localized states in Regions IV^- and V^- are those associated with γ_0.

We can follow the branches of analytically known small amplitude states away from γ_0 using numerical continuation. In Regions IV^- and V^- this branch behaves like the corresponding branch in Figure 8 and approaches γ^* monotonically whenever ν is near ν_β. However, new behavior is found when $\nu\gg\nu_\beta$ (Figure 12): the branch of localized states $A_{L,0}^{\pm}$ now approaches γ^* in an oscillatory fashion, undergoing an infinite sequence of saddle-node bifurcations as it winds toward it; cf. [28]. At each saddle-node bifurcation the stability of the localized states changes: the localized states near γ_0 are unstable, but there is a region surrounding γ^*, between the first and second saddle-node bifurcations, within which there is a finite multiplicity

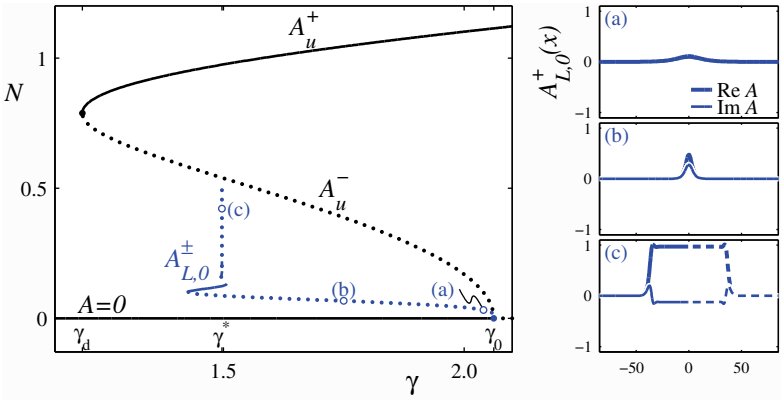


Figure 12. *Bifurcation diagram for $\nu \gg \nu_\beta$ in Region V^-, showing the branches of localized states $A_{L,0}^{\pm}$ found by continuing the solution (3.12) away from γ_0, as well as the uniform states $A_u^{\pm}$. Solid (dotted) lines correspond to stable (unstable) solutions. A closeup of the behavior near $\gamma^* \approx 1.496$ is shown in Figure 13. (a)–(c) Profiles of $A_{L,0}^+$ at the locations indicated in the left panel. Parameters: $\mu = -0.5$, $\alpha = -2$, $\beta = 2.5$, $\nu = 2$; (a) $\gamma = 2.04$, (b) $\gamma = 1.75$, (c) $\gamma = 1.496$.*

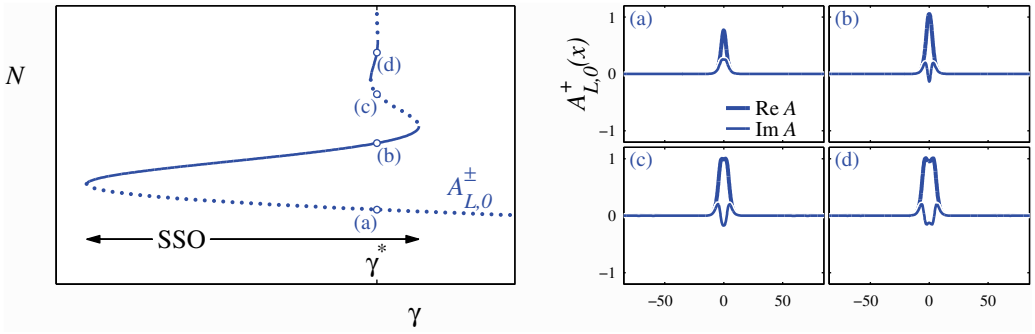


Figure 13. *Detail of Figure 12 with sample profiles of $A_{L,0}^+$ near $\gamma = \gamma^*$ (see text). In the left panel solid (dotted) lines correspond to stable (unstable) solutions.*

of *stable* localized states (Figure 13). We label this region SSO since it contains states that correspond in the physical system (2.1) to stable standard oscillons, i.e., stable localized states that are biasymptotic to $A = 0$ as $x \to \pm\infty$.

The location of the SSO states in the (ν, γ) plane in Region V^- is shown in Figure 11(b) and takes the form of a wedge straddling γ^*, increasing in width as ν increases. The emergence of γ^* from the tangency at ν_β is difficult to make out in Figure 11(b); a closeup for a different set of parameters elsewhere in Region V^- is shown in Figure 14. Near ν_β, in the domain labeled (a) in Figure 14, the branch of localized states $A_{L,0}^{\pm}$ approaches γ^* monotonically from γ_0. The SSO region appears only when ν exceeds ν_β sufficiently, in the domain labeled (b) in Figure 14. Numerically we find that the SSO region first appears near the crossing of γ^* and the curve γ_d (to be defined in the next section). Moreover, there is good theoretical reason to believe that the SSO region in fact emerges from this codimension-two point, although it is initially exponentially thin and hence numerically invisible. There is a similar SSO

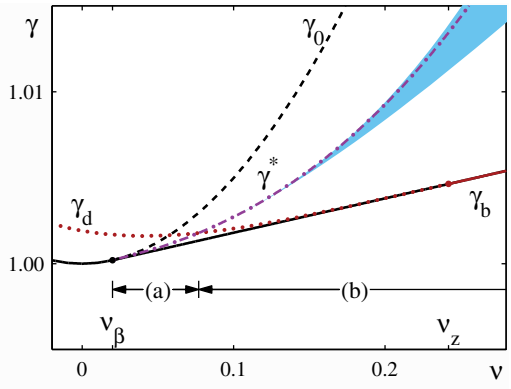


Figure 14. *Closeup of the SSO region in the (ν,γ) plane in Region* V^-. *In region* (a) $\gamma^* < \gamma_d$ *and the approach of small amplitude localized states to* γ^* *is monotonic; in* (b) $\gamma^* > \gamma_d$ *and the approach is via a series of saddle-node bifurcations, the first two of which define the boundaries of the SSO region (shaded). Parameters:* $\mu = -1$, $\alpha = -5$, $\beta = 50$.

region present in Region IV^-, but it is orders of magnitude thinner than the one shown for Region V^-.

3.2. Large amplitude states in $z > 0$. In the previous subsection we found it convenient to consider separately the small amplitude behavior in $\alpha > 0$ and $\alpha < 0$. In this section we turn to a study of *large amplitude* states, discussing separately the cases $z > 0$ and $z < 0$, where

$$z(\alpha,\beta) \equiv -\alpha(1-\beta^2) + 2\beta \tag{3.18}$$

(see below). The former spans Regions II^-–IV^- in the (α,β) plane of Figure 3, while the latter inequality characterizes Regions I^- and V^-.

3.2.1. Stability of $A = A_u^{\pm}$. The stability of the uniform phase-locked states in time is determined by linearizing (3.1) about these states and looking for solutions of the form e^{ikx+st}. The resulting dispersion relation is

$$\begin{aligned} &s^2 - 2s\left(\mu - k^2 - 2|A_u^{\pm}|^2\right) + (1+\alpha^2)k^4 + 2\left[2(1+\alpha\beta)|A_u^{\pm}|^2 - (\mu+\alpha\nu)\right]k^2 \\ &+ 4\left[(\mu+\beta\nu)|A_u^{\pm}|^2 + \gamma^2 - \gamma_0^2\right] = 0. \end{aligned} \tag{3.19}$$

Consider first the stability near γ_0, where the uniform phase-locked states bifurcate from $A = 0$. The amplitude of these states near this point is given by

$$|A_u^{\pm}(\gamma)|^2 = -\frac{\gamma_0(\gamma-\gamma_0)}{\beta(\nu-\nu_\beta)} + \dots. \tag{3.20}$$

This expression is valid both in the subcritical case ($\gamma < \gamma_0$, $\nu > \nu_\beta$), in which A_u^- bifurcates from γ_0, and in the supercritical case ($\gamma > \gamma_0$, $\nu < \nu_\beta$), in which A_u^+ bifurcates from γ_0. For small wavenumbers the growth rate near this bifurcation is given by

$$s(k,\gamma) \approx -\frac{\alpha}{\mu}(\nu-\nu_\alpha)k^2 + \frac{2\gamma_0}{\mu}(\gamma-\gamma_0). \tag{3.21}$$

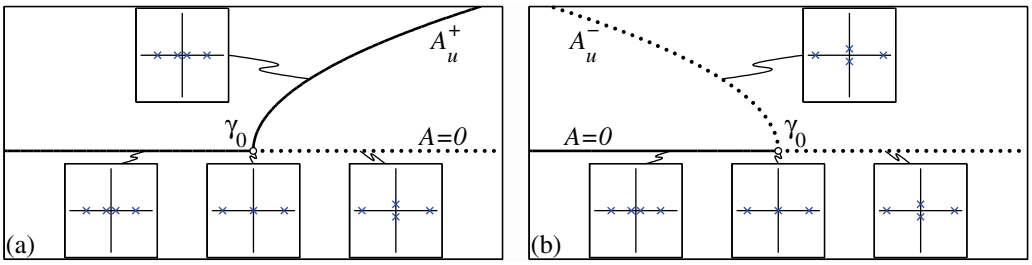


Figure 15. *Temporal stability of $A = 0$ and the uniform phase-locked states $A_u^\pm$ near γ_0 when $\alpha > 0$, $\nu < \nu_\alpha$, and the bifurcation at γ_0 is* (a) *supercritical or* (b) *subcritical: Solid (dotted) lines represent stable (unstable) solutions. The insets show the spatial eigenvalues λ of these states in the complex λ plane.*

Thus, when the bifurcation is subcritical, the A_u^- state which emerges from γ_0 is unstable to the $k = 0$ mode, as shown in Figure 15(b). When the bifurcation is supercritical, the A_u^+ state that emerges from γ_0 inherits the stability of the $A = 0$ state from $\gamma < \gamma_0$: when $\alpha > 0$, it is stable when $\nu < \nu_\alpha$ and unstable when $\nu > \nu_\alpha$. An example of the former is shown in Figure 15(a).

We next analyze the dispersion relation near the saddle-node bifurcation of uniform phase-locked states at $\gamma = \gamma_b$, present in $\nu > \nu_\beta$. At this point the amplitude $|A_u(\gamma_b)|$ of these states is given by

$$|A_u(\gamma_b)|^2 = \frac{\beta(\nu - \nu_\beta)}{\rho_\beta^2}, \tag{3.22}$$

while nearby it is

$$|A_u^\pm|^2 = |A_u(\gamma_b)|^2 \pm \frac{\sqrt{2\gamma_b(\gamma - \gamma_b)}}{\rho_\beta} + \dots. \tag{3.23}$$

The dispersion relation near the saddle-node, valid for small k, reduces to

$$s(k,\gamma) \approx -\frac{z(\nu - \nu_z)k^2 \pm 2\rho_\beta^3 |A_u(\gamma_b)|^2 \sqrt{2\gamma_b(\gamma - \gamma_b)}}{\rho_\beta^2(2|A_u(\gamma_b)|^2 - \mu)}, \tag{3.24}$$

where

$$\nu_z \equiv -\frac{(1-\beta^2) + 2\alpha\beta}{z}\mu \tag{3.25}$$

and z is as defined in (3.18). The "+" sign in this expression refers to the upper branch A_u^+, and the "−" sign to the lower branch A_u^-. Once again the A_u^- branch is always unstable with respect to the $k = 0$ mode. Since $\mu < 0$ and $z > 0$, the A_u^+ branch is stable near the saddle-node when $\nu > \nu_z$ but is unstable when $\nu < \nu_z$. The resulting stability assignments in the neighborhood of the saddle-node of uniform phase-locked states are shown in Figure 16.

It follows that the behavior of the system at a given value of ν depends on the location of ν_z relative to ν_β. In Region II$^-$ (Figure 6), ν_z is greater than ν_β; consequently, for $\nu_\beta < \nu < \nu_z$

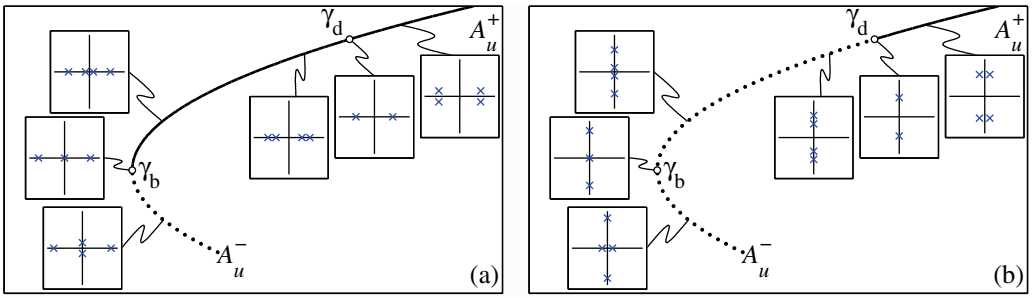


Figure 16. *Temporal stability of the uniform phase-locked states: Solid (dotted) lines represent stable (unstable) solutions. The insets show the spatial eigenvalues of the uniform states in a neighborhood of γ_b and γ_d. When $z > 0$,* (a) *corresponds to $\nu > \nu_z$ and* (b) *to $\nu < \nu_z$, and vice versa for $z < 0$.*

the behavior near the saddle-node is described by Figure 16(b), while for $\nu > \nu_z$ it is described by Figure 16(a). In contrast, in Regions III$^-$ (Figure 7) and IV$^-$ (Figure 10), ν_z is less than ν_β, and the behavior shown in Figure 16(a) applies whenever a saddle-node is present ($\nu > \nu_\beta$).

Since the A_u^+ state is necessarily stable at large γ, the presence of instability near γ_b implies the presence of a further bifurcation which stabilizes the A_u^+ state as γ increases. This bifurcation is a Turing bifurcation and occurs at $\gamma = \gamma_d$ (Figure 16(b)), where

$$\gamma_d^2 = \gamma_b^2 + \frac{\left\{ z(\mu^2 + \nu\nu_z) + \mathrm{sgn}[\alpha - \beta]\mu\rho_\alpha\rho_\beta^2\gamma_0 \right\}^2}{4\mu^2\rho_\beta^2(\alpha - \beta)^2}; \tag{3.26}$$

the associated wavenumber is given by

$$k_d = \frac{\sqrt{\mu + \alpha\nu - 2|A_u^+(\gamma_d)|^2(1 + \alpha\beta)}}{\rho_\alpha}. \tag{3.27}$$

These expressions apply when the A_u^+ state emerges from a supercritical bifurcation at γ_0 as well: an initially unstable state ($\nu > \nu_\alpha$) stabilizes as γ increases at γ_d.

In Region II$^-$ (Figure 6), the Turing bifurcation is present in $\nu_\alpha < \nu < \nu_z$. The curve $\gamma = \gamma_d$ is tangent at ν_α to the line $\gamma = \gamma_a$ (and hence also to the curve $\gamma = \gamma_0$), corresponding to the simultaneous creation of both Turing instabilities (of $A = 0$ and $A = A_u^\pm$) at γ_0. At ν_z the curve $\gamma = \gamma_d$ is tangent to the line $\gamma = \gamma_b$, corresponding to the annihilation of the Turing bifurcation as it merges with the saddle-node of uniform phase-locked states. Between these tangencies γ_d corresponds to bifurcations and hence forms the lower boundary of the stable uniform phase-locked states within the resonance tongue. In Figure 6 we indicate this portion of γ_d with a solid line (barely visible) to distinguish it from portions where it does not correspond to bifurcations (shown dotted). In some places in Figure 6 the $\gamma = \gamma_d$ curve is difficult to distinguish from the other curves present; this is especially true at large values of ν where $\gamma_d - \gamma_b \ll 1$.

In Region III$^-$ (Figure 7), the Turing bifurcation is absent. The curve $\gamma = \gamma_d$ always lies above γ_0, but no tangencies are present. Finally, in Region IV$^-$ (Figure 10), the Turing bifurcation reappears along with the tangency of $\gamma = \gamma_d$ to $\gamma = \gamma_a$ (and $\gamma = \gamma_0$) at ν_α. In this

case the bifurcations are present in $\nu < \nu_\alpha$ (solid line), and at the tangency the two Turing bifurcations merge and annihilate. At large ν the curve $\gamma = \gamma_d$ (dotted) approaches $\gamma = \gamma_b$ and becomes difficult to distinguish.

3.2.2. Spatial eigenvalues of $A = A_u^\pm$. In this subsection we study the spatial eigenvalues of the uniform phase-locked states to identify the parameter regimes with possible exponentially localized states biasymptotic to $A = A_u^\pm$. These eigenvalues are determined by linearizing (2.2) about $A_u^\pm$ and satisfy

$$(1+\alpha^2)\lambda^4 + 2\lambda^2\left[\mu + \alpha\nu - 2(1+\alpha\beta)|A_u^\pm|^2\right] + 4\left[(\mu+\beta\nu)|A_u^\pm|^2 + \gamma^2 - \gamma_0^2\right] = 0. \tag{3.28}$$

As was the case above when considering the stability in time, we first examine the spatial eigenvalues near γ_0 where the uniform phase-locked states bifurcate from $A = 0$. At γ_0 there are two zero eigenvalues and two order one eigenvalues given by

$$\lambda = \pm\Lambda_0 \equiv \pm\frac{\sqrt{-2\alpha(\nu-\nu_\alpha)}}{\rho_\alpha}. \tag{3.29}$$

When $\alpha > 0$, these eigenvalues are real if $\nu < \nu_\alpha$ and imaginary if $\nu > \nu_\alpha$, in agreement with the eigenvalue analysis along the $A = 0$ branch described in section 3.1.2. Using (3.20) for the amplitude of the uniform phase-locked states near $\gamma = \gamma_0$, the small spatial eigenvalues along this branch are given approximately by

$$\lambda^2 = \frac{4\gamma_0(\gamma-\gamma_0)}{\rho_\alpha^2\Lambda_0^2} \tag{3.30}$$

and hence are of the same sign as Λ_0^2 when $\gamma > \gamma_0$ and of the opposite sign when $\gamma < \gamma_0$. Thus a supercritical bifurcation to uniform phase-locked states at γ_0 implies that the zero eigenvalues split "toward" the large eigenvalues $\pm\Lambda_0$ and all four eigenvalues are either real or imaginary (the eigenvalue structure (i) or (iii) in section 3.1.2). In contrast, when this bifurcation is subcritical, the zero eigenvalues split in the direction "opposite" to the large eigenvalues, and so two are real and two are imaginary (the eigenvalue structure (iv) in section 3.1.2). These spatial eigenvalues are shown in the insets in Figure 15 for the case $\nu < \nu_\alpha$, where Λ_0 is real; a similar figure holds in the case $\nu > \nu_\alpha$, where Λ_0 is imaginary.

For the case shown in Figure 15(a), the spatial eigenvalues of the trivial state are such that no localized states biasymptotic to $A = 0$ can exist in $\gamma > \gamma_0$. But in this same range of γ the eigenvalues of the uniform phase-locked state A_u^+ can support localized states, and indeed, as shown in the next section, localized states in the form of *fronts* bifurcate supercritically from γ_0 whenever the eigenvalue structure in Figure 15(a) applies.

We next compute the eigenvalues of the uniform phase-locked states near the saddle-node bifurcation at $\gamma = \gamma_b$. At the saddle-node, (3.28) has a pair of zero eigenvalues and a pair of order one eigenvalues given by

$$\lambda = \pm\Lambda_b \equiv \pm\frac{\sqrt{2z(\nu-\nu_z)}}{\rho_\alpha\rho_\beta}. \tag{3.31}$$

Since $z > 0$ these eigenvalues are real when $\nu > \nu_z$ (Figure 16(a)) and imaginary when $\nu < \nu_z$ (Figure 16(b)). Using (3.23) for the amplitude along the $A_u^{\pm}$ branches near the saddle-node, the zero eigenvalues become

$$\lambda^2 = \pm \frac{4|A_u(\gamma_b)|^2}{\Lambda_b^2} \frac{\rho_\beta}{\rho_\alpha^2} \sqrt{2\gamma_b(\gamma - \gamma_b)}. \tag{3.32}$$

Here the "+" sign refers to the upper branch A_u^+, and the "−" sign to the lower branch A_u^-. As shown in Figure 16, along the A_u^+ branch the zero eigenvalues split "toward" the large eigenvalues and hence all four are either real or imaginary, while along the A_u^- branch the zero eigenvalues split in the "opposite" direction, and two eigenvalues are always real and two are always imaginary. We conclude that when the bifurcation to uniform phase-locked states is subcritical, the eigenvalues of the A_u^- states are always of type (iv). Thus we do not expect localized states involving A_u^-. On the other hand, a branch of localized states involving A_u^+ may emerge from the saddle-node at $\gamma = \gamma_b$ when the eigenvalue structure is of the type shown in Figure 16(a) but not in the case shown in Figure 16(b).

The eigenvalue analysis predicts one final bifurcation that can occur along the A_u^+ branch of uniform phase-locked states: the four eigenvalues, which are either all real or all imaginary, collide pairwise at $\gamma = \gamma_d$. The spatial eigenvalues at this collision are $\lambda = \pm ik_d$. In the case in which the bifurcation to uniform phase-locked states at γ_0 is subcritical and the A_u^+ branch emerges from the saddle-node at γ_b, the collision occurs on the imaginary axis whenever the eigenvalues $\pm\Lambda_b$ lie on the imaginary axis, as in Figure 16(b). If the eigenvalues $\pm\Lambda_b$ are real, the collision occurs instead on the real axis, as in Figure 16(a). In contrast, when the A_u^+ branch emerges directly from a supercritical bifurcation at γ_0, the eigenvalue collision occurs on the imaginary axis when the eigenvalues $\pm\Lambda_0$ are imaginary, as in Figure 5(b), and on the real axis when they are real, as in Figure 5(a). When the collision at γ_d occurs on the real axis, it corresponds to a global bifurcation, the Belyakov–Devaney point. When the collision at γ_d occurs on the imaginary axis, it corresponds to a local bifurcation point; this is precisely the Turing bifurcation at γ_d identified in section 3.2.1. In addition to identifying the appearance of spatially extended Turing patterns, the eigenvalue analysis suggests that spatially localized states that approach A_u^+ may also emerge from this bifurcation point into $\gamma > \gamma_d$. Thus localized states may emerge *either* from γ_b *or* from γ_d but not both.

A comparison of the eigenvalue structure in Figures 5(a) and 16(a) shows that bifurcations that occur at γ_0 and γ_b are similar in nature; at both bifurcation points there are two zero eigenvalues and two real eigenvalues, and localized states may exist nearby in $\gamma < \gamma_0$ or $\gamma > \gamma_b$, where the eigenvalue structure is of type (i). A comparison of Figures 5(b) and 16(b) shows that the bifurcations that can occur at γ_a and γ_d are also similar. Both are reversible Hopf bifurcations with 1:1 resonance [26], and localized states may exist in $\gamma < \gamma_a$ or $\gamma > \gamma_d$, where the eigenvalue structure is of type (ii). However, the small amplitude localized states that bifurcate from γ_a and γ_0 are biasymptotic to $A = 0$, while the large amplitude states that bifurcate from γ_b and γ_d are biasymptotic to A_u^+. As such, the former correspond to standard oscillons, while the latter represent new states we refer to as reciprocal oscillons [57].

3.2.3. Fronts bifurcating from $\gamma = \gamma_0$. In section 3.1.3 we examined the "small amplitude" spatially localized states that bifurcate from $A = 0$ at γ_0 whenever the bifurcation

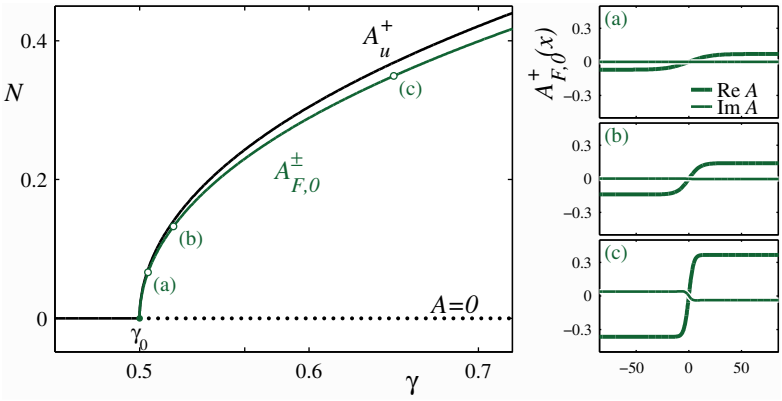


Figure 17. *Bifurcation diagram for $\nu < \nu_\alpha$ in Region* II^-*, showing the branches of front-like states $A_{F,0}^{\pm}$ found by continuing the solution (3.33) away from γ_0, together with the uniform states A_u^+. The norm of $A_{F,0}^{\pm}$ has been rescaled to distinguish it from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions.* (a)–(c) *Sample front profiles. Parameters:* $\mu = -0.5$, $\alpha = 2$, $\beta = 1$, $\nu = 0$; (a) $\gamma = 0.505$, (b) $\gamma = 0.52$, (c) $\gamma = 0.65$.

to the uniform phase-locked states is subcritical ($\nu > \nu_\beta$). These states are biasymptotic to $A = 0$ as $x \to \pm\infty$. In this section we examine the so-called *large amplitude* states that bifurcate from $A = 0$ at γ_0 whenever the bifurcation to the uniform phase-locked states is supercritical ($\nu < \nu_\beta$). Appendix A shows that such states are present in $\gamma > \gamma_0$ in Region II^- when $\nu < \nu_\alpha$ (Figure 6), in Region III^- when $\nu < \nu_\beta$ (Figure 7), and in Region IV^- when $\nu_\alpha < \nu < \nu_\beta$ (Figure 10) and take the form

$$A_{F,0}^{\pm}(x) = \pm(\eta_0 + i)\sqrt{\frac{\gamma - \gamma_0}{-b_0}}\tanh\left(\sqrt{\frac{\gamma - \gamma_0}{-2a_0}}x\right), \tag{3.33}$$

where $a_0 < 0$, $b_0 < 0$. The "+" branch refers to fronts which asymptote to the uniform phase-locked state A_u^+ as $x \to \infty$ and to $-A_u^+$ as $x \to -\infty$, while the reverse is true for the "−" branch. Thus both solutions have odd parity under spatial reflection and are *fronts*.

The analytical solutions (3.33) can be followed away from γ_0 using numerical continuation. Since heteroclinic cycles never form at these values of ν, the fronts extend to arbitrarily large γ (Figure 17). Near γ_0 the fronts are broad (Figure 17(a)) but become more localized as γ increases (Figure 17(b),(c)). In the following we refer to these fronts as Ising fronts. Such fronts are typically (although not necessarily) monotonic in x and are linearly stable to all perturbations, even and odd.

3.2.4. Localized states bifurcating from $\gamma = \gamma_b$. The weakly nonlinear analysis performed in Appendix C shows that when $\nu > \nu_z$ a branch of spatially localized states exists near the saddle-node bifurcation at $\gamma = \gamma_b$. These states are biasymptotic to A_u^+ and are given by

$$A_{L,b}(x) = A_u^+ - 3\Upsilon_1(\xi_b + i)\sqrt{\gamma - \gamma_b}\,\mathrm{sech}^2\left\{\left(\frac{\Upsilon_1}{2a_b/b_b}\right)^{1/2}(\gamma - \gamma_b)^{1/4}\,x\right\}; \tag{3.34}$$

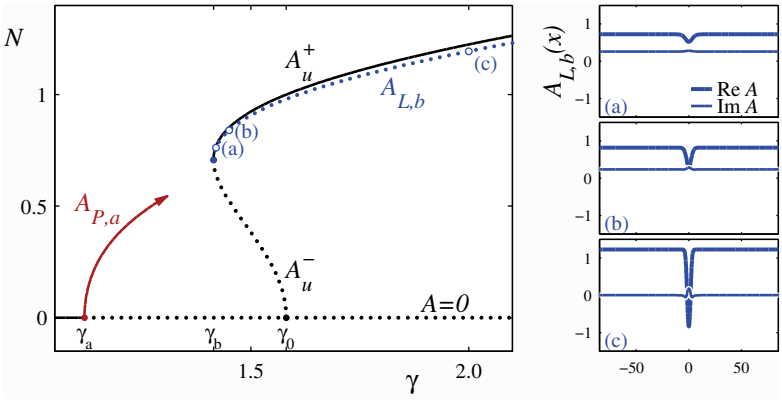


Figure 18. *Bifurcation diagram for $\nu > \nu_z$ in Region* II^-*, showing the branch of localized states $A_{L,b}$ found by numerical continuation of the solution* (3.34) *away from γ_b, as well as the spatially periodic states $A_{P,a}$ and the uniform states $A_u^{\pm}$. The norm of $A_{L,b}$ has been rescaled to distinguish it from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions.* (a)–(c) *Solution profiles at the locations indicated in the left panel. Parameters: $\mu = -0.5$, $\alpha = 2$, $\beta = 1$, $\nu = 1.5$, with* (a) $\gamma = 1.42$, (b) $\gamma = 1.45$, (c) $\gamma = 2$.

the constants in this expression are defined in Appendix C. There is a similar branch of localized states biasymptotic to $-A_u^+$ given by $-A_{L,b}(x)$. In Regions III^- and IV^- $\nu_z < \nu_\beta$, and hence the condition $\nu > \nu_z$ automatically holds whenever a saddle-node is present. A branch of localized states of this type therefore emerges from γ_b when $\nu > \nu_z$ in Region II^- (Figure 6) but is present whenever there is a saddle-node in Region III^- (Figure 7) and Region IV^- (Figure 10).

The analytical solution (3.34) can be followed away from γ_b using numerical continuation. The behavior away from the saddle-node depends critically on whether or not a heteroclinic cycle forms as γ increases away from γ_b. In the absence of such a cycle the branch of localized states simply extends to arbitrarily large γ, a situation that occurs when $\nu > \nu_z$ in Region II^-, as shown in Figure 18. The solution profile near γ_b is broad and forms a shallow dip in an otherwise uniform A_u^+ background (Figure 18(a)). Away from the saddle-node the solution contracts in x, forming a well-defined hole in the A_u^+ background (Figure 18(c)), but remains unstable throughout owing to a single unstable amplitude mode of even parity. The supercritical branch of stable spatially periodic states $A_{P,a}$ is shown for clarity only near γ_a.

When a heteroclinic cycle is present, the branch of localized states emerging from the saddle-node can approach γ^* either monotonically or in an oscillatory fashion involving a sequence of saddle-node bifurcations. In Region IV^- only the former is observed (Figure 19), while in Region III^- both possibilities may occur (Figure 20). The heteroclinic cycle that forms with increasing γ is *identical* to that responsible for the termination of the small amplitude states, as can be seen by comparing, for example, the profiles in Figure 20(c) and Figure 9(c),(f); this is a manifestation of the reciprocity already noted in [12, 27].

When the approach to γ^* is monotonic, the localized states are always unstable to a single amplitude (even) mode. On the other hand, when the approach is oscillatory (observed only in Region III^-), the stability changes at successive saddle-node bifurcations, as illustrated in Figure 21. As a result multiple stable *reciprocal oscillons* are present in a region straddling

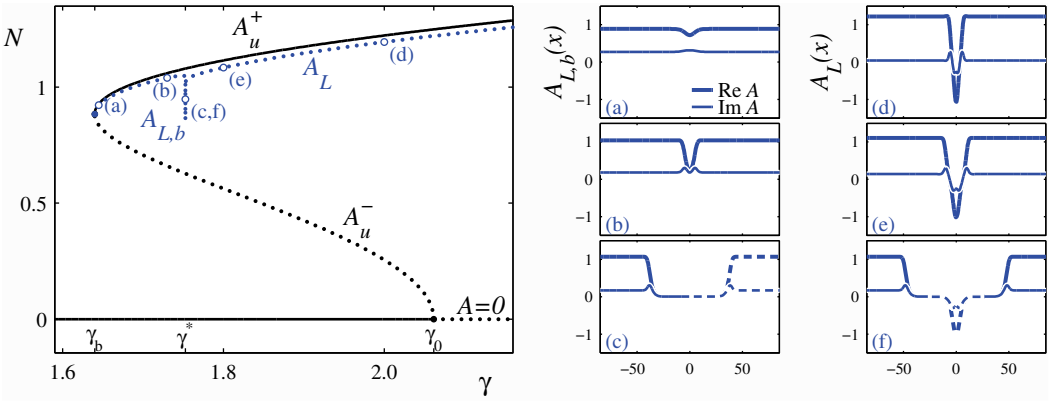


Figure 19. *Bifurcation diagram for $\nu > \nu_\beta$ in Region* IV$^-$*, showing the branch of localized states $A_{L,b}$ found by numerical continuation of the solution (3.34) away from γ_b, as well as a branch A_L found using the homotopy method and the uniform states $A_u^\pm$. The norms of the localized states have been rescaled to distinguish them from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions.* (a)–(c) *Solution profiles along $A_{L,b}$ at the locations indicated in the left panel.* (d)–(f) *Solution profiles along A_L at the locations indicated in the left panel. Parameters: $\mu = -0.5$, $\alpha = -1.5$, $\beta = 1.25$, $\nu = 2$, with* (a) $\gamma = 1.645$, (b) $\gamma = 1.730$, (c),(f) $\gamma = 1.752$, (d) $\gamma = 2$, (e) $\gamma = 1.8$.

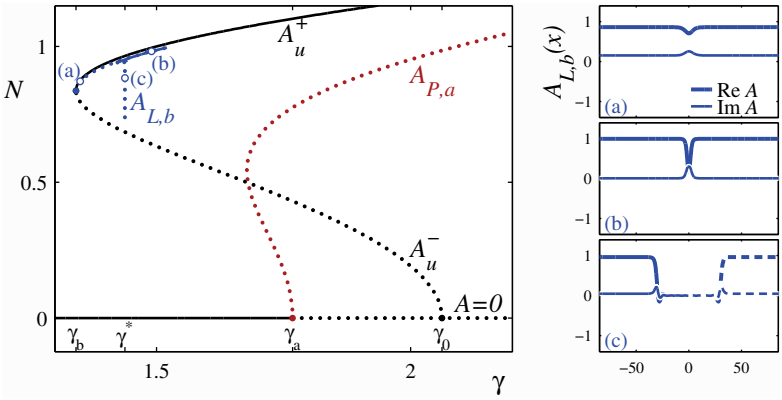


Figure 20. *Bifurcation diagram for $\nu \gg \nu_\beta$ in Region* III$^-$*, showing the branch of localized states $A_{L,b}$ found by numerical continuation of the solution (3.34) away from γ_b, as well as the spatially periodic states $A_{P,a}$ and the uniform states $A_u^\pm$. The norm of $A_{L,b}$ has been rescaled to distinguish it from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions. A closeup of the behavior near $\gamma^* \approx 1.437$ is shown in Figure 21.* (a)–(c) *Solution profiles at the locations indicated in the left panel. Parameters: $\mu = -0.5$, $\alpha = 1$, $\beta = 2$, $\nu = 2$, with* (a) $\gamma = 1.35$, (b) $\gamma = 1.49$, (c) $\gamma = 1.437$.

γ^*; in Figure 21 we label this region SRO. The extent of the SRO region in the (ν, γ) plane in Region III$^-$ is shown in Figure 7(b). The SRO region takes the form of a wedge straddling the line γ^* and increases in width as ν increases. As with the SSO region described in the previous section, the SRO region is not created at ν_β together with the line γ^* but appears only when ν sufficiently exceeds ν_β. A closeup of this behavior is shown in Figure 22 for a different set of parameters elsewhere in Region III$^-$. In this figure it is easier to see that the SRO region first appears only after the line γ^* crosses the line $\gamma = \gamma_a$.

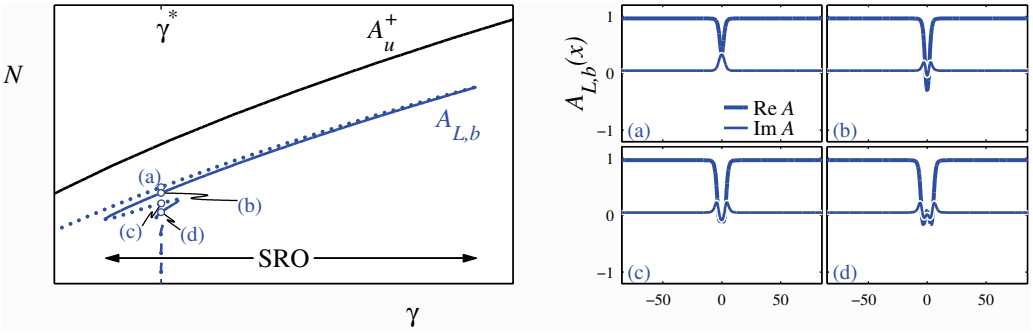


Figure 21. *Detail of Figure 20, including sample profiles near $\gamma = \gamma^*$ (see text). In the left panel solid (dotted) lines correspond to stable (unstable) solutions.*

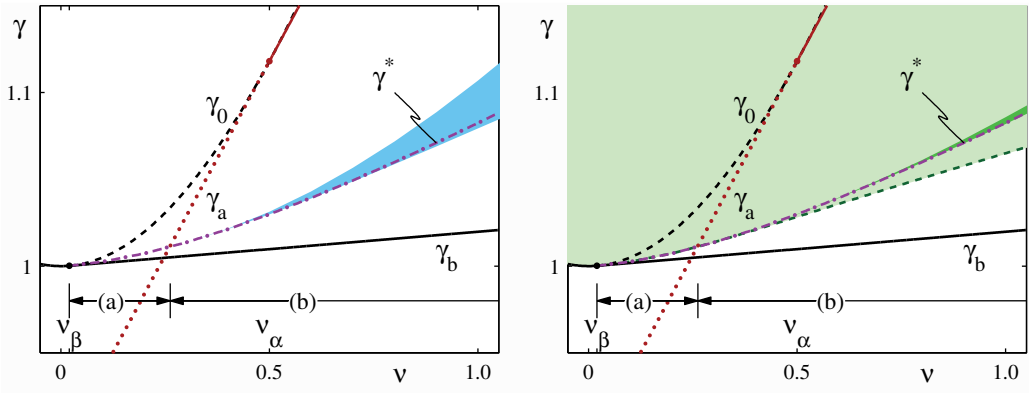


Figure 22. *Detail of the crossing of γ^* and γ_a in the (ν, γ) plane of Region III$^-$. In the domain labeled* (a)*, $\gamma^* > \gamma_a$ and the approach of large amplitude localized states and fronts to γ^* is monotonic; in* (b)*, $\gamma^* < \gamma_a$ and the approach is via a series of saddle-node bifurcations. Left panel: The SRO region (shaded) defined by the first two saddle-node bifurcations on the branch of localized states. Right panel: The SIF region (lightly shaded) extends down to the first saddle-node bifurcation on the front branch, while the SSF region (dark) extends from the second saddle-node to the third. Parameters: $\mu = -1$, $\alpha = 2$, $\beta = 50$.*

3.2.5. Localized states bifurcating from $\gamma = \gamma_d$. A collision of spatial eigenvalues on the imaginary axis at γ_d corresponds to a bifurcation point analogous to γ_a. For this case weakly nonlinear analysis in the vicinity of γ_d (Appendix D) reveals that there is a branch of spatially periodic solutions given by

$$A_{P,d}(x) = A_u^+ + 2(\xi_d + i)\sqrt{\frac{\gamma - \gamma_d}{b_d}}\cos(k_d x + \varphi) \tag{3.35}$$

and a branch of localized spatial oscillations of the form

$$A_{L,d}(x) = A_u^+ + 2(\xi_d + i)\sqrt{\frac{\gamma - \gamma_d}{b_d/2}}\,\mathrm{sech}\left\{\sqrt{\frac{\gamma - \gamma_d}{a_d}}x\right\}\cos(k_d x + \varphi)\,, \tag{3.36}$$

both of which bifurcate simultaneously from γ_d. Two similar branches also bifurcate at γ_d from the $-A_u^+$ branch. The expressions for ξ_d, a_d, and b_d are unwieldy (see Appendix D)

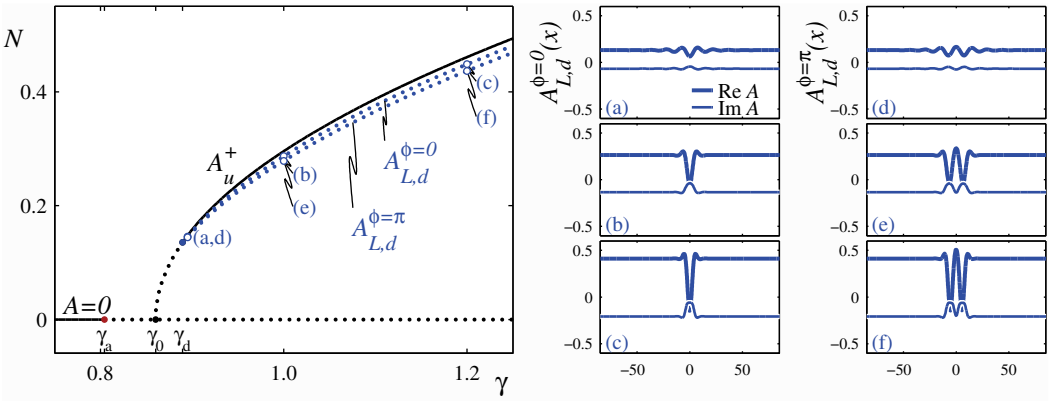


Figure 23. *Bifurcation diagram for $\nu < \nu_\alpha$ in Region* IV$^-$*, showing the two branches of localized states $A_{L,d}^{\varphi=0}$ and $A_{L,d}^{\varphi=\pi}$, as well as the branch of uniform phase-locked states A_u^+. For clarity the branches of spatially periodic states $A_{P,a}$ and $A_{P,d}$ which bifurcate from γ_a and γ_d, respectively, are omitted. The norms of the localized states have been rescaled to distinguish them from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions.* (a)–(c) *Sample localized profiles along $A_{L,d}^{\varphi=0}$.* (d)–(f) *Sample localized profiles along $A_{L,d}^{\varphi=\pi}$. Parameters: $\mu = -0.5$, $\alpha = -1.5$, $\beta = 1.25$, $\nu = -0.7$, with* (a),(d) $\gamma = 0.895$, (b),(e) $\gamma = 1$, (c),(f) $\gamma = 1.2$.

but can easily be evaluated for specific parameter choices. We find that $a_d, b_d > 0$ for all values of ν at which the collision at γ_d occurs on the imaginary axis. Thus, whenever γ_d is a bifurcation point, both $A_{P,d}$ and $A_{L,d}$ emerge into $\gamma > \gamma_d$. This is the case in Region II$^-$ when $\nu_\alpha < \nu < \nu_z$ and in Region IV$^-$ when $\nu < \nu_\alpha$; as already noted, γ_d never corresponds to a bifurcation in Region III$^-$.

The phase φ in (3.35) is arbitrary. Within the asymptotic analysis the phase φ in (3.36) is also arbitrary, but terms beyond all orders select solutions with phases $\varphi = 0, \pi$. We refer to these two solutions as $A_{L,d}^{\varphi=0}$ and $A_{L,d}^{\varphi=\pi}$, respectively. These states are distinct, unrelated by symmetry, and *even* in x. Odd states of the type present in (3.16) are excluded here since these localized states represent small perturbations of a *nontrivial* uniform background state.

The two branches of solutions described analytically by (3.36) can be followed away from the bifurcation point using numerical continuation. Since these solutions never apply when a heteroclinic cycle is present, the corresponding solution branch extends to arbitrarily large γ, as shown in Figure 23. The localized states are broad near γ_d (Figure 23(a),(d)) but contract in x as γ increases, forming well-localized packets. At large γ the $\varphi = 0$ solution corresponds to a single hole in the uniform A_u^+ background (Figure 23(c)), while the $\varphi = \pi$ solution corresponds to two adjacent holes (Figure 23(f)); we think of the latter as a bound state of two holes. Both branches are unstable throughout: the $\varphi = 0$ solutions are unstable to a single amplitude (even) mode, while the $\varphi = \pi$ solutions are unstable to both an amplitude (even) mode and a phase (odd) mode.

3.2.6. Additional large amplitude branches. All the branches of states presented thus far were found by numerically continuing an approximate analytical solution away from a local bifurcation. Some of these branches terminate in global bifurcations at γ^*. In this section we show that there are, in addition, branches of large amplitude localized states

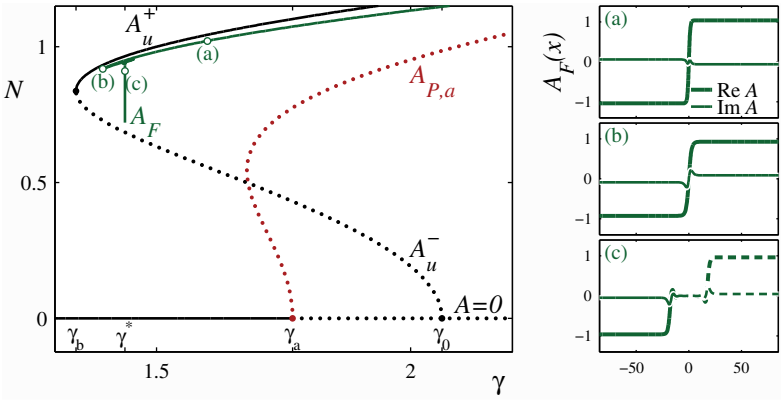


Figure 24. *Bifurcation diagram for $\nu \gg \nu_\beta$ in Region* III$^-$*, showing the branches of fronts A_F, as well as the spatially periodic states $A_{P,a}$ and the uniform states $A_u^\pm$. The norm of A_F has been rescaled to distinguish it from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions. A closeup of the behavior near $\gamma^* \approx 1.437$ is shown in Figure 25.* (a)–(c) *Sample localized profiles along A_F. The wide profile in* (c) *near γ^* can be thought of as a pair of nearly heteroclinic orbits between $A = 0$ and $\pm A_u^+$ (solid and dashed lines). Parameters: $\mu = -0.5$, $\alpha = 1$, $\beta = 2$, $\nu = 2$, with* (a) $\gamma = 1.6$, (b) $\gamma = 1.394$, (c) $\gamma = 1.437$.

and fronts for which we have no approximate analytical solution. We locate these branches using a homotopy method: starting from an analytically known local solution, we follow the corresponding solution branch to a large value of γ; we then change another parameter such as α or ν at fixed large γ to move into a regime where no analytical solution exists; finally, we follow this branch of large amplitude states back toward small γ. In the following we use this technique to locate new classes of both localized and front-like states, with a particular interest in locating stable states in the neighborhood of the line γ^*.

We find that the behavior of the large amplitude front-like states present at large γ depends on the presence or absence of γ^*. When such a point is present, as in Region III$^-$ for $\nu > \nu_\beta$, the branch of fronts may approach γ^* from above either monotonically or through a series of saddle-node bifurcations. Numerically we observe that the former occurs when $\gamma^* > \gamma_a$, a condition that holds only very close to ν_β. The latter occurs when $\gamma^* < \gamma_a$ and is the case for all $\nu \gg \nu_\beta$. A bifurcation diagram with this behavior is shown in Figure 24. In this case, the fronts at large γ are simple monotonic (Ising) transitions from A_u^+ to $-A_u^+$, but as one passes between adjacent saddle-nodes the fronts develop extra structure near their midpoint, and we refer to the resulting profiles as *structured* fronts. The Ising fronts are stable at large γ, but stability switches at each saddle-node. Thus stable Ising fronts (SIFs) exist for all γ above the first saddle-node, and stable structured fronts (SSFs) exist between the second and third saddle-nodes. These regions are labeled SIF and SSF in the closeup shown in Figure 25.

In the absence of γ^* our numerics suggest that the branch of Ising fronts always terminates, as γ decreases, in a saddle-node bifurcation at $\gamma = \gamma_k$ involving a second branch of front-like states that also comes in from large γ. These new states resemble kinks [24] and are always unstable with respect to an amplitude (odd) mode. An example of this behavior taken from Region II$^-$ is shown in Figure 26. These kink fronts differ fundamentally from the structured fronts created near γ^*, despite the fact that both are nonmonotonic transitions between $\pm A_u^+$.

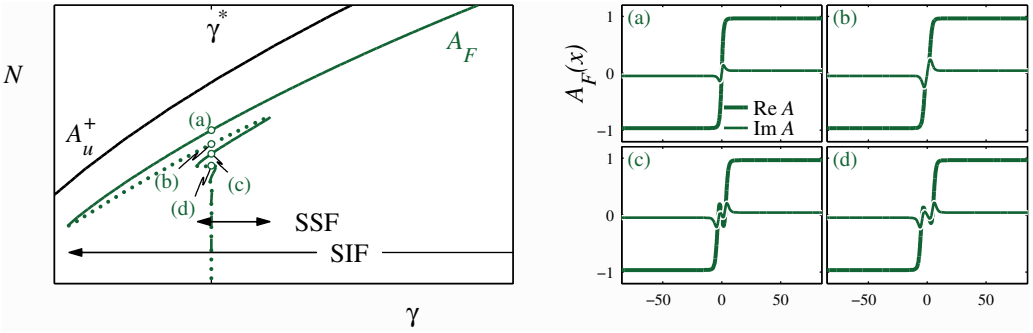


Figure 25. *Detail from Figure 24 of the A_F branch as it approaches γ^*. The significance of the regions SIF and SSF is explained in the text.*

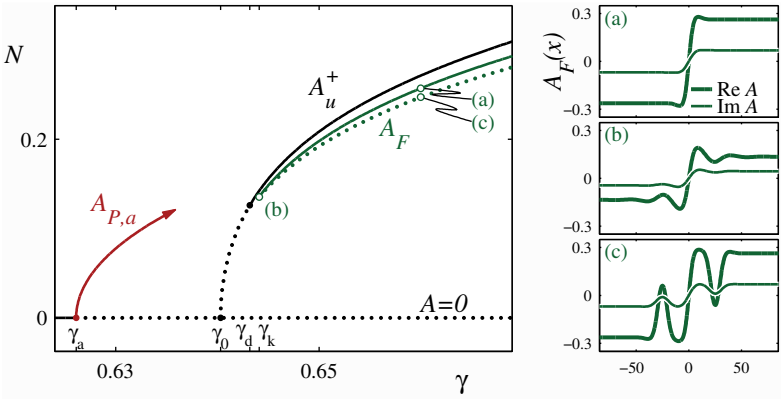


Figure 26. *Bifurcation diagram for $\nu_\alpha < \nu < \nu_\beta$ in Region* II^-*, showing the branch of front states A_F, as well as the spatially periodic states $A_{P,a}$ and the uniform states $A_u^\pm$. The norm of A_F has been rescaled to distinguish it from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions.* (a)–(c) *Sample localized profiles. Parameters: $\mu = -0.5$, $\alpha = 2$, $\beta = 1$, $\nu = 0.4$, with* (a),(c) $\gamma = 0.66$, (b) $\gamma = 0.644$.

The latter result when an Ising front develops extra structure near the midpoint at $A = 0$, while the former are the result of nucleation as the nontrivial states on either side of the front develop holes which are bound to the front.

In Region II^- (Figure 6) there is no γ^* line in the (ν, γ) plane and the only stable fronts are monotonic. In $\nu < \nu_\alpha$ the SIF region extends down to γ_0, where the fronts $A_{F,0}^\pm$ bifurcate directly from $A = 0$. In $\nu > \nu_\alpha$ the lower boundary of the SIF region consists of the line of saddle-node bifurcations where the monotonic fronts collide with the kinks. This saddle-node bifurcation always appears to lie (slightly) above γ_d, although in $\nu > \nu_z$ both this saddle-node and γ_d approach γ_b, and we can no longer determine the precise relation between the boundary of the SIF region and γ_d.

Figure 7(c) shows the region of existence of stable fronts in Region III^- and the associated line γ^*. The SIFs extend down to γ_0 in $\nu < \nu_\beta$. Above ν_β the behavior is difficult to make out in Figure 7(c) but easier to see in Figure 22, corresponding to a value of the parameters elsewhere in Region III^-. In the domain directly above ν_β, labeled (a) in this figure, the

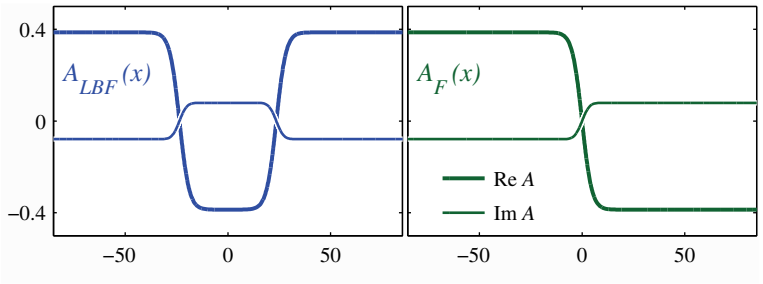


Figure 27. *Comparison of the profiles of an LBF and an (Ising) front that coexist at the same value of the parameters in Region* I^-. *Parameters:* $\mu = -0.5$, $\alpha = 2$, $\beta = 0.5$, $\nu = -0.2$, $\gamma = 0.712$.

approach to γ^* is monotonic and the SIF region extends down only as far as γ^*. In the domain labeled (b) the approach of the branch of fronts to γ^* is through a series of saddle-nodes and the SIF region extends beyond γ^* to the first of these, which lies above γ_b. In the latter case Region III^- also contains an SSF region. The regime of existence is difficult to make out in Figure 7(c), but in Figure 22 it is clear that the SSF region occupies a thin wedge around the line γ^* in the domain labeled (b). A comparison of Figures 7(b) and 7(c) (or the closeups in Figure 22) shows the intimate relation between the SRO and SSF regions: both are associated with nonmonotonic approach of a branch of large amplitude states to γ^* and are present at identical values of ν, although the widths of the regions (as determined by the locations of the saddle-nodes) are different.

Figure 10(b) does the same for Region IV^-. In $\nu_\alpha < \nu < \nu_\beta$ the SIF region extends down to γ_0, where the fronts $A^{\pm}_{F,0}$ bifurcate directly from $A = 0$. Below ν_α the boundary of this region consists of the line of saddle-nodes γ_k where the monotonic fronts collide with the kinks; again this saddle-node always lies above γ_d. Above ν_β the large amplitude fronts always approach γ^* monotonically from above, and the line γ^* forms the boundary of the SIF region.

The homotopy method can also be applied to the localized states. In Region II^-, the branch of large amplitude localized states $A_{L,d}$ that emerges from γ_d in $\nu_\alpha < \nu < \nu_z$ and the branch of $A_{L,b}$ that emerges from γ_b when $\nu > \nu_z$ each extends to large γ, where the profiles resemble deep holes in a uniform A^+_u background. When the localized states are continued numerically in ν below ν_α and then followed back to small γ we find that this hole deepens and broadens as the localized state fills with $-A^+_u$. Eventually the localized state becomes wide enough that it resembles a pair of bound fronts between $\pm A^+_u$. We refer to such solutions as localized bound fronts (LBFs), A_{LBF}. A profile on the A_{LBF} branch in Region I^- is shown in Figure 27 along with the Ising front at the same value of the parameters; the profiles are nearly identical over the entire width of the front. The separation between the two fronts increases as γ approaches γ_d from above. Although the numerical results are inconclusive, we conjecture that this branch terminates at γ_d in a global bifurcation analogous to the Belyakov–Devaney bifurcation. In this case we have two symmetrically related fixed points so that the "single-pulse" orbit which persists in both $\gamma < \gamma_d$ and $\gamma > \gamma_d$ is the front connecting $\pm A^+_u$. We interpret the LBFs as the simplest of the infinite multiplicity of multipulse states expected to exist in $\gamma > \gamma_d$.

LBFs are also present in Region III$^-$, though only above the Belyakov–Devaney point γ_d. As a result we find no additional localized states in the neighborhood of γ^* since it always lies far below γ_d.

In Region IV$^-$ the branch of large amplitude localized states $A_{L,b}$ that emerge from γ_b in $\nu > \nu_\beta$ approaches γ^* monotonically from below, while the branch $A_{L,d}$ that emerges from γ_d in $\nu < \nu_\alpha$ extends to large γ. The latter can be followed in ν not only into the domain $\nu_\alpha < \nu < \nu_\beta$, where no analytical results apply, but also into $\nu > \nu_\beta$ corresponding to the previously inaccessible range of parameter space above the line γ^*. When numerically continued back toward small γ we find two types of behavior. At values of ν such that $\gamma_d > \gamma^*$, the resulting solutions resemble the LBFs and are again found numerically only above γ_d. This is so for $\nu_\alpha < \nu < \nu_\beta$ as well as at some values of ν slightly above ν_β. When ν exceeds ν_β sufficiently that $\gamma^* > \gamma_d$, the branch of localized states approaches γ^* monotonically from above. It may be of interest to recall that in this same region the *small amplitude* localized states $A^{\pm}_{L,0}$ approach this same γ^* in a series of saddle-node bifurcations (Figure 12). These new large amplitude localized states, labeled A_L in Figure 19, exist in addition to those that bifurcate from the saddle-node at γ_b and approach γ^* from below, and are everywhere unstable to an amplitude mode. The A_L profile at γ^* consists of a heteroclinic cycle involving three parts (Figure 19(f)): a front from A^+_u to $A = 0$, a localized state biasymptotic to $A = 0$, and another front from $A = 0$ to A^+_u. This limiting profile is distinct from the limiting profile (Figure 19(c)) along the $A_{L,b}$ branch that also terminates at γ^*, although the segments of these profiles resembling fronts between A^+_u and $A = 0$ are identical.

3.3. Large amplitude states in $z < 0$. When $z < 0$, as in Regions I$^-$ and V$^-$, the expressions presented above remain valid, but we must reinterpret the results. Specifically, in (3.31) the order one eigenvalues at γ_b are now real when $\nu < \nu_z$ and in this case large amplitude localized states biasymptotic to A^+_u are expected to emerge from the saddle-node into $\gamma > \gamma_b$ (as in Figure 16(a)). The collision of eigenvalues at γ_d occurs on the imaginary axis and corresponds to a bifurcation point when $\nu > \nu_z$ (as in Figure 16(b)).

In Region I$^-$, ν_z, as defined in (3.25), is always less than ν_β, and the behavior near the saddle-node is necessarily of the type shown in Figure 16(b). In this region the Turing bifurcations of the $A = 0$ and A^+_u states are created in the tangency of γ_d and γ_a to γ_0 at ν_α, and γ_d remains a bifurcation point in all $\nu > \nu_\alpha$, where it forms the boundary of the stable region of uniform phase-locked states within the resonance tongue. The large amplitude localized states and fronts are similar to those found in Region II$^-$, including the LBFs in $\nu < \nu_\alpha$ found using the homotopy method. Like Region II$^-$, Region I$^-$ contains no heteroclinic cycles, and no SRO or SSF regions are present.

The behavior in Region V$^-$ resembles very closely that of Region IV$^-$. The one notable exception is that in Region IV$^-$ the Turing bifurcation at γ_d was present only in $\nu < \alpha$, while in Region V$^-$ it is also present in $\nu > \nu_z$ where it defines the boundary of the region of stable uniform phase-locked states. Nevertheless the large amplitude states observed in Region V$^-$ match those described previously for Region IV$^-$, including the necessarily monotonic approach of both localized states and fronts to γ^*.

3.4. Pinning near γ^*. Thus far we have identified numerically two types of behavior near γ^*: a branch of states can approach γ^* either monotonically or through a series of saddle-node

bifurcations. We have exhibited examples of both types of behavior in our discussion of the termination of small and large amplitude localized states, and of fronts. In this subsection we show that the type of behavior observed is determined by the spatial eigenvalues of the spatially uniform solutions connected by the heteroclinic cycles present at γ^* [28].

We consider first the case of large amplitude localized states, such as those that emerge from γ_b, shown in Figures 19 and 20. Near the saddle-node at γ_b these states represent small perturbations of the A_u^+ state, and these grow in amplitude as γ approaches γ^*. Near this point the inner region of the localized state approaches $A = 0$ and we can think of the localized state as a combination of two fronts, one from A_u^+ to $A = 0$ and a second from $A = 0$ back to A_u^+. If the eigenvalues of the $A = 0$ state are complex, the approach to $A = 0$ will be via a decaying oscillation in x; the departure from $A = 0$ will likewise be via a growing oscillation in x. These oscillatory tails can be seen in the profile shown in Figure 20(c). Such oscillations are responsible for the mutual "pinning" of the fronts that is, in turn, responsible for the series of saddle-node bifurcations that must occur as the fronts move apart and the heteroclinic cycle is approached. When the fronts are close to one another the interaction between them is strong, and the pinning interval is therefore broad; when the fronts are far apart they interact only weakly and the distance between successive saddle-node bifurcations shrinks. The widest of these pinning intervals is the SRO region labeled in Figure 21. When the eigenvalues of the $A = 0$ state are instead real, the approach to (and departure from) the $A = 0$ state is exponential, and no oscillatory tails are present (Figure 19(b)). In this case pinning is absent, and the branch of localized states approaches γ^* monotonically.

There are three regions with γ^* lines: Regions III$^-$, IV$^-$, and V$^-$. In Regions IV$^-$ (Figure 10) and V$^-$ (Figure 11), the line γ^* always lies in $\nu > \nu_\beta$ and always lies between γ_a and γ_0, in a region where the eigenvalues of $A = 0$ are real. In these two regions large amplitude localized states experience no pinning, and the branches of these states must therefore approach γ^* monotonically. In Region III$^-$ (Figure 7), the line γ^* falls above γ_a when it is first created at ν_β, but as ν increases γ^* crosses γ_a somewhere in the range $\nu_\beta < \nu < \nu_\alpha$ and lies below γ_a in $\nu > \nu_\alpha$. This crossing is shown in detail in Figure 22. In a very small range in ν below this crossing, labeled (a) in the figure, the eigenvalues of $A = 0$ are real and there is no pinning; above this crossing, in the range labeled (b), the eigenvalues are complex, resulting in pinning and an associated SRO interval. The width of this interval increases from zero as ν increases beyond the crossing point.

We can also use the same pinning argument to understand the creation of the SSF region shown in Figure 7(c) and in the right panel of Figure 22. The monotonic fronts that exist at large γ split as γ approaches γ^*, creating a broad inner region which fills with the $A = 0$ state (Figure 24(c)). When the spatial eigenvalues of this state are complex the fronts experience pinning and the approach of the branch of fronts to γ^* will be via a series of saddle-node bifurcations which define the SSF region. Because these same eigenvalues are responsible for the SRO region of the large amplitude localized states, we expect the SSF and SRO regions to be present in the *same* range of values of ν, as confirmed by Figure 22.

A similar argument applies in the case of small amplitude localized states biasymptotic to $A = 0$, but in this case it is the A_u^+ (or $-A_u^+$) state which fills the inner region near γ^*. As a result the behavior near γ^*, shown, for example, in Figures 8 and 12, is determined by the spatial eigenvalues of A_u^+. In Region III$^-$ the line γ^* always lies below γ_d, and the eigenvalues

of A_u^+ at γ^* are therefore real. It follows that in this region the small amplitude localized states cannot exhibit pinning and hence that the corresponding solution branch must approach γ^* monotonically, as in Figures 8 and 9. In Regions IV$^-$ and V$^-$ the γ^* line falls below γ_d when it is first created in the tangency at ν_β but crosses γ_d as ν increases. This crossing is shown in detail in Figure 14. It follows that when ν is sufficiently larger than ν_β, the approach to γ^* becomes oscillatory, resulting in the SSO region shown in Figure 11(b).

3.5. Another type of pinning. In the preceding subsections we have seen a number of examples of what might be called classical pinning regions. The fundamental object responsible for this behavior is a heteroclinic cycle between two distinct states, the $A = 0$ and $A = A_u^+$ uniform states [28]. The LBFs can likewise be understood in terms of pinning, but in this case the fundamental fronts are connections between $\pm A_u^+$ found in the SIF regions. As such the resulting homoclinic orbits could be called *homoclinic* cycles because the two states visited by the trajectory are related by symmetry. If the spatial eigenvalues of the state A_u^+ are complex, the two fronts will possess oscillatory tails that will interact and lead to preferred separations. This notion is consistent with the numerical observation that these LBF states are only found above γ_d, where the spatial eigenvalues are complex. Even when γ_d does not correspond to a bifurcation point, the spatial eigenvalues below γ_d are real and no pinning can take place.

There are two important observations that distinguish this case from the previous results. First, the heteroclinic orbits responsible for the classical pinning regions exist only along the line γ^*, whereas the orbits responsible for the LBF states exist throughout the entire SIF region. Thus, while the line γ^* organizes the regions SSO, SRO, and SSF, it is less clear what organizes the various LBF states. Second, due to symmetry, the Ising fronts that exist within the SIF region do not drift. As such the pinning force does not oppose any natural tendency of the bound fronts to drift but selects only the preferred separations.

It is also worth mentioning the states which result when the unstable kink fronts described above are evolved in time. The holes on either side of the front deepen into a stable state which resembles five bound Ising fronts (the original front plus two from the deepening of each hole). Numerical simulations show that other stable states consisting of three or seven bound Ising fronts also exist within much of the SIF regions. It is easy to imagine a large multiplicity of fronts and localized states created in this manner. Thus although the focus of this paper is on the stable states associated with the pinning of fronts near the γ^* line, it is clear that bound fronts play an important role in the full enumeration of stable solutions to (2.2).

3.6. Multipulse branches near $\gamma = \gamma^*$. We have already seen that the region around γ^* contains a large multiplicity of states associated with the pinning of fronts as some branches wind toward γ^*. This region also contains a large multiplicity of branches beyond those already discussed, referred to as multipulse branches [29]. Figure 9 shows two heteroclinic cycles emerging monotonically from γ^* toward small amplitude. The simplest of these is $A_{L,a}^{\text{even}}$, consisting of two bound heteroclinic connections from $A = 0$ to A_u^+ (Figure 9(c)). The second slightly more complicated cycle, $A_{L,a}^{\text{odd}}$, consists of three parts: a heteroclinic connection from $A = 0$ to A_u^+, followed by a connection from A_u^+ to $-A_u^+$, and finally a connection from $-A_u^+$ back to $A = 0$ (Figure 9(f)). The existence of the front between A_u^+ and $-A_u^+$ at γ^* is a result of pinning and is responsible for the series of saddle-nodes as the large amplitude Ising

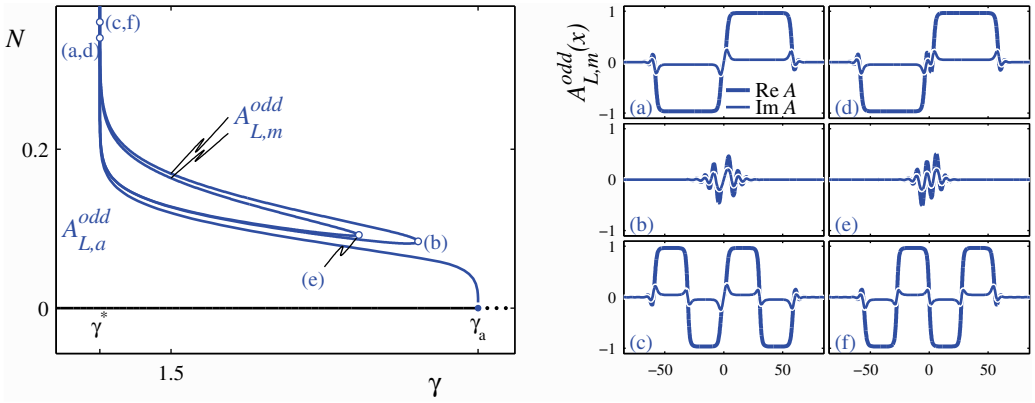


Figure 28. *Bifurcation diagram for $\nu \gg \nu_\beta$ in Region* III$^-$*, showing branches of odd multipulse states $A^{\rm odd}_{L,m}$ together with the branch $A^{\rm odd}_{L,a}$ of odd single-pulse states. For clarity the localized states are drawn as solid lines even though they are everywhere unstable.* (a)–(f) *Sample localized profiles. Parameters:* $\mu = -0.5$, $\alpha = 1$, $\beta = 2$, $\nu = 2$; (a),(c),(d),(f) $\gamma = 1.437$; (b) $\gamma = 1.715$; (e) $\gamma = 1.663$.

fronts approach γ^* (Figure 25). In fact, the connection between A^+_u and $-A^+_u$ in Figure 9(f) is identical to the front profile shown in Figure 25(a). There are actually (infinitely) many fronts connecting $\pm A^+_u$ at γ^*, each of which can be used to assemble a different heteroclinic cycle at this point. All cycles assembled in this way consist of two pulses and are odd. Figures 28(a),(d) show two examples, assembled using the fronts shown in panels (b) and (c) of Figure 25, respectively. Figure 28 also shows the result of continuing these cycles numerically away from γ^* and toward small amplitude. The cycle $A^{\rm odd}_{L,a}$ is the only odd branch that terminates in a local bifurcation at γ_a. Both of the other branches of odd states shown in the figure terminate at saddle-node bifurcations in $\gamma^* < \gamma < \gamma_a$. The other branches that collide at these saddle-nodes also trace back to γ^*; these are also odd but consist of four pulses—not two. These four-pulse states can each be decomposed into a heteroclinic cycle made of five fronts. The profile shown, for example, in Figure 28(c) consists of a front from $A = 0$ to A^+_u, followed by three fronts between A^+_u and $-A^+_u$, and finally a front from $-A^+_u$ back to $A = 0$. Of the three intermediate fronts, two correspond to the profile shown in Figure 25(a) and the third to the profile shown in Figure 25(b). We refer to these odd parity localized multipulse states in general as $A^{\rm odd}_{L,m}$.

Pinning also causes the branch of large amplitude localized states $A_{L,b}$ shown in Figure 20 to approach γ^* in a series of saddle-node bifurcations, resulting in an infinite series of orbits homoclinic to A^+_u at γ^*. These profiles are shown in Figure 21. These homoclinic orbits can be used to assemble additional heteroclinic cycles at γ^* which have even parity, corresponding to even multipulse localized states $A^{\rm even}_{L,m}$. The profiles of two such cycles are shown in Figures 29(a),(d) and are assembled using the localized states shown in Figures 21(a),(b), respectively. Figure 29 shows the result of numerically continuing these even heteroclinic cycles away from γ^*; the branch $A^{\rm even}_{L,a}$ is shown for reference. As before, we find that the branches of two-pulse states terminate in $\gamma^* < \gamma < \gamma_a$ in saddle-node bifurcations involving other branches which, when followed back to γ^*, correspond to even heteroclinic cycles that are yet more complex. For example, the profile shown in Figure 29(c) consists of five parts: a homoclinic orbit to

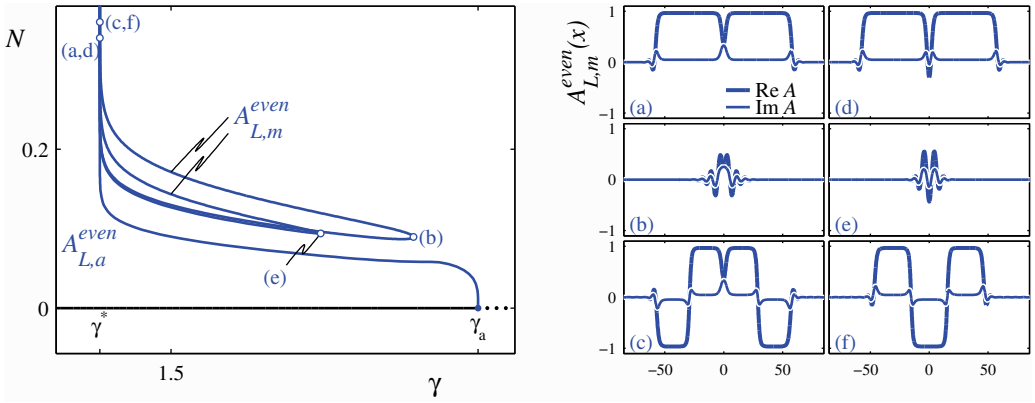


Figure 29. *Bifurcation diagram for $\nu \gg \nu_\beta$ in Region* III$^-$*, showing branches of even multipulse states $A^{\rm even}_{L,m}$ together with the branch $A^{\rm even}_{L,a}$ of even single-pulse states. For clarity the localized states are drawn as solid lines even though they are everywhere unstable.* (a)–(f) *Sample localized profiles. Parameters: $\mu = -0.5$, $\alpha = 1$, $\beta = 2$, $\nu = 2$;* (a),(c),(d),(f) *$\gamma = 1.437$;* (b) *$\gamma = 1.711$;* (e) *$\gamma = 1.630$.*

A_u^+ sandwiched between two heteroclinic fronts connecting $\pm A_u^+$ and two more heteroclinic connections from A_u^- to $A = 0$.

The description of multipulse states is more complicated when $\nu < \nu_\alpha$ and the bifurcation from $A = 0$ to localized states occurs at γ_0 instead of γ_a, and depends on the location of γ^* relative to γ_a. When $\gamma^* < \gamma_a$, the multipulse states in the neighborhood of γ^* can be assembled as described above, but when these states are numerically continued into $\gamma > \gamma^*$, they do not collide pairwise in saddle-node bifurcations. Instead the multipulse branches all terminate at the Belyakov–Devaney point γ_a, which lies between γ^* and γ_0. As the branches approach γ_a, the distance between the individual pulses that make up each profile is large and they interact only weakly. When instead $\gamma^* > \gamma_a$, the large amplitude localized states and fronts approach γ^* monotonically. In this case no small amplitude multipulse states are present.

The results in Figures 28 and 29 give a glimpse of the large number of complex multipulse localized states biasymptotic to $A = 0$ that are present near γ^* when the large amplitude states experience pinning. Likewise, at parameter values for which the small amplitude localized states experience pinning ($\gamma_d < \gamma^*$) it is possible to assemble these pieces into complex heteroclinic cycles that asymptote to $\pm A_u^+$. These can be followed away from γ^* to large amplitude. An example is shown in Figure 19. Evidently the A_L profile shown in Figure 19(f) can now be interpreted as a *heteroclinic* cycle to A_u^+ involving three parts: two fronts on either side of a localized state. Although many complex heteroclinic cycles can be assembled in this manner, we find numerically that it is difficult to continue all but the simplest ones away from γ^*. Should any multipulse branches of this type exist in $\gamma < \gamma^*$, we expect that they terminate pairwise above γ_d when this is a Turing bifurcation and at γ_d when it is a Belyakov–Devaney point.

4. Self-excited oscillatory regime. In this section we describe solutions to (3.1) in the self-excited oscillatory case, $\mu > 0$. In this case we refer to the five regions of the (α, β) plane shown in Figure 3 as Regions I$^+$–V$^+$. As in the damped case, the bifurcation to uniform

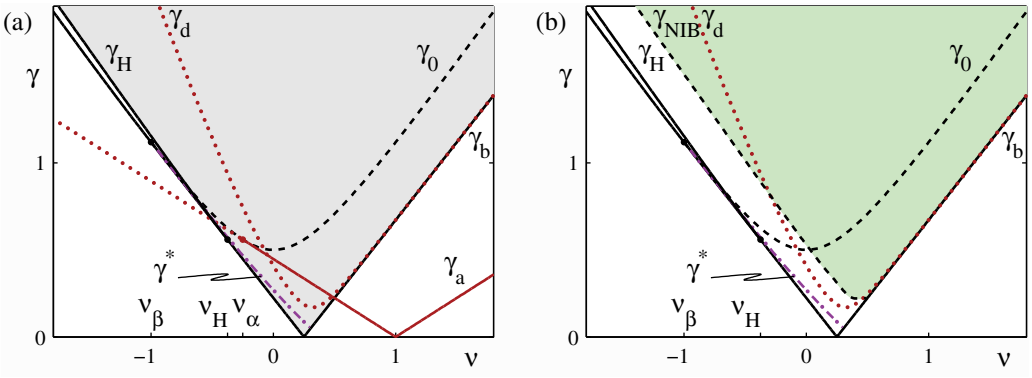


Figure 30. *The (ν,γ) plane in Region I^+. In* (a) *shading indicates the presence of stable uniform states A_u^+. In* (b) *the curve $\gamma=\gamma_{NIB}$ marks the location of the nonequilibrium Ising–Bloch bifurcation which forms the left boundary of the SIF region (shaded). Parameters: $\mu=0.5$, $\alpha=2$, $\beta=0.5$. For these parameters, $\nu_z=2.75$ and falls outside the range shown.*

phase-locked states at γ_0 is supercritical when $\nu<\nu_\beta$ and subcritical when $\nu>\nu_\beta$. But in the self-excited case $\nu_\beta<0$, and the boundary of the resonance tongue in the (ν,γ) plane differs qualitatively from that in the damped case, extending down to $\gamma_b=0$ at $\nu=\beta\mu$ (Figure 2(b)). Each region is described below. For each region we describe the stability of the uniform solutions in both t and x. We next consider the effect of the sign change in μ on the analytical predictions of localized states and fronts, and then present numerical results which show the extent in the (ν,γ) plane of stable states of either type.

4.1. Region I^+. The resonance tongue in this region is shown in Figure 30(a). When $\mu>0$ the $A=0$ state is always unstable to a range of wavenumbers owing to the presence of a $k=0$ Hopf bifurcation; the A_u^- state is also everywhere unstable. However, we expect that when $|\mu|$ is small the effect of changing the sign of μ should be confined to small amplitude states. Indeed, we find that the A_u^+ state remains stable at sufficiently large γ, although it may lose stability with decreasing γ. This loss of stability defines a (new) boundary of the resonance tongue and may arise in one of two ways. When $\nu<\nu_H$,

$$\nu_H\equiv-\frac{(1-\beta^2)}{2\beta}\mu, \tag{4.1}$$

the uniform phase-locked states A_u^+ lose stability at a $k=0$ Hopf bifurcation that occurs at $\gamma=\gamma_H$, where

$$\gamma_H^2=\gamma_b^2+\rho_b^2\left[\frac{\mu}{2}-|A_u^+(\gamma_b)|^2\right]^2. \tag{4.2}$$

The curve $\gamma=\gamma_H$ is shown in Figure 30(a) and forms the lower stability boundary for the uniform phase-locked states when $\nu<\nu_H$. At $\nu=\nu_H$ this curve is tangent to γ_b, and the Hopf bifurcation turns into a Takens–Bogdanov bifurcation. Note that $\nu_H>\nu_\beta$ for all values of β, independent of α. It follows that the above description of the Hopf bifurcation in Region I^+

also applies in Regions II^+–V^+. The dispersion relation (3.19) shows that the Hopf bifurcation occurs when

$$\mu - 2|A_u^+(\gamma_H)|^2 = 0 \tag{4.3}$$

and hence only in $\mu > 0$.

The second possibility is that, as γ decreases, the A_u^+ state loses stability with respect to a Turing bifurcation at $\gamma = \gamma_d$. Equation (3.26) for the location of this bifurcation remains valid in $\mu > 0$ and applies provided the critical wavenumber k_d in (3.27) is real. In Region I^+ the Turing bifurcation is present in $\nu > \nu_z$, at which point the curve $\gamma = \gamma_d$ is tangent to the line $\gamma = \gamma_b$. Between ν_H and ν_z the uniform phase-locked states remain stable down to the boundary of the resonance tongue at the saddle-node at γ_b. Thus the A_u^+ phase-locked state is stable in a subregion of the resonance tongue defined by $\gamma > \gamma_H$ when $\nu < \nu_H$, $\gamma > \gamma_b$ when $\nu_H < \nu < \nu_z$, and $\gamma > \gamma_d$ when $\nu > \nu_z$.

The analysis of the spatial eigenvalues in the damped case also applies to the self-excited case. Since $\alpha > 0$ in Region I^+, the spatial eigenvalue structure shown in Figure 5(a) applies to the $A = 0$ state when $\nu < \nu_\alpha$, and that shown in Figure 5(b) applies when $\nu > \nu_\alpha$. The spatial eigenvalues of the uniform phase-locked states in $\nu < \nu_\beta$, where A_u^+ bifurcates supercritically from γ_0, are shown in Figure 15(a). Since $z < 0$ in Region I^+, the spatial eigenvalues in the neighborhood of the saddle-node are as shown in Figure 16(b) whenever $\nu_\beta < \nu < \nu_z$ and, as shown in Figure 16(a), whenever $\nu > \nu_z$. Of course, the stability assignments in these figures do not necessarily carry over to the case $\mu > 0$.

Figure 30 also shows the location of the line $\gamma = \gamma^*$ at which heteroclinic connections between $A = 0$ and $A = A_u^+$ are present. The location of this line in the (ν, γ) plane differs qualitatively from the $\mu < 0$ case. As before, the line emerges from the tangency at ν_β, but because of the location of ν_β and the orientation of the resonance tongue, it extends toward smaller γ as ν increases and then terminates in a codimension-two point at which $\gamma^* = \gamma_b$, slightly above $\nu = \beta\mu$, where $\gamma_b = 0$. However, the line always lies below γ_d, and hence the spatial eigenvalues of the uniform phase-locked state A_u^+ at $\gamma = \gamma^*$ are real (of type (i)). As a result we do not expect localized solutions biasymptotic to $A = 0$ to exhibit pinning as the branch approaches γ^* and the profile fills with the A_u^+ state. The spatial eigenvalues of the $A = 0$ state at γ^* are also of type (i) near the tangency at ν_β, but γ^* crosses $\gamma = \gamma_a$ somewhere in $\nu_\beta < \nu < \nu_\alpha$, implying that in part of the domain in ν the spatial eigenvalues of the trivial state at γ^* are of type (ii). Thus the pinning which was responsible for the SRO and SSF regions in $\mu < 0$ may create similar structures in $\mu > 0$ as well.

Several of the analytical solutions described in section 3 also apply in Region I^+. We begin with the states bifurcating from the $A = 0$ state even though we expect these to inherit the instability of this state. The solution (3.14) describing the spatially periodic extended states $A_{P,a}$ bifurcating from γ_a is valid in Region I^+ when $\nu > \nu_\alpha$, but the sign of η_a and hence the nature of the bifurcation depends on the value of ν relative to $\alpha\mu$. This critical value corresponds to the point at which $\gamma_a(\nu) = 0$ and is easily located graphically in Figure 30. Below this value, the general expression for η_a in (B.7) reduces to $\eta_a = \alpha + \rho_\alpha > 0$, and the bifurcation to spatially periodic states is subcritical. Above $\nu = \alpha\mu$ we have instead $\eta_a = \alpha - \rho_\alpha < 0$, and the bifurcation is supercritical. In the former case there is in addition a subcritical bifurcation from γ_a to spatially localized states $A_{L,a}$ as described by (3.16); in the latter no such

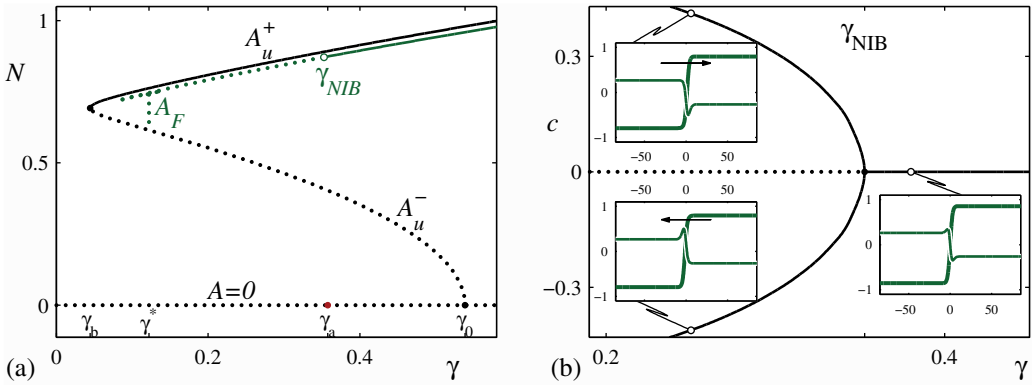


Figure 31. (a) *Bifurcation diagram in Region* I^+ *at a value of* ν *such that* $\gamma^* < \gamma_a$*, showing the branches of fronts* A_F *as well as the uniform states* $A_u^{\pm}$*. The spatially periodic branch* $A_{P,a}$ *that bifurcates from* $A = 0$ *at* γ_a *is not shown. The norm of* A_F *has been rescaled to distinguish it from the uniform states.* (b) *Bifurcation diagram showing the NIB bifurcation. The order parameter in this diagram is the front speed c; this speed vanishes for Ising fronts but is nonzero for Bloch fronts. The insets show profiles of the fronts, with the arrows indicating the drift direction for the Bloch fronts. Solid (dotted) lines correspond to stable (unstable) solutions. Parameters:* $\mu = 0.5$, $\alpha = 2$, $\beta = 0.5$, $\nu = 0.2$.

localized states are present. When ν lies below ν_α but above ν_β, (3.12) describes the branch of small amplitude spatially localized states $A_{L,0}^{\pm}$ which emerge subcritically from γ_0.

The branches of localized states, $A_{L,0}$ and $A_{L,a}$, can be followed away from their respective bifurcation points using numerical continuation. For values of ν for which heteroclinic cycles are present, $A_{L,0}$ and $A_{L,a}$ always approach γ^* monotonically in a manner very similar to that shown in Figures 8 and 9. This is consistent with the expectation based on the spatial eigenvalues of the A_u^+ state. However, the stability of these localized states in Region I^+ is different from that in the $\mu < 0$ case. As before, there is an unstable amplitude mode with a localized eigenfunction, but now there is in addition an infinite spectrum of unstable modes with extended eigenfunctions, corresponding to instabilities inherited from the background $A = 0$ state. There is also a range in ν, above the termination of the γ^* line but below $\nu = \alpha\mu$, in which the $A_{L,a}$ branch emerges subcritically from γ_a but no heteroclinic cycle is present. In this case the branch of localized states interacts instead with a secondary branch of spatially periodic states that bifurcate from the subcritical $A_{P,a}$ states. The details of this interaction and the (unstable) stationary states that result are beyond the scope of this paper but resemble related behavior already studied in the context of the Swift–Hohenberg equation [8].

An examination of the large amplitude localized states reveals several new examples of interesting behavior. Large amplitude Ising fronts are found in Region I^+ using the homotopy method described above. The starting point for this method is the branch of fronts $A_{F,0}$ in (3.33) which bifurcate supercritically from γ_0 when $\nu < \nu_\beta$ and extend to arbitrarily large γ. At other values of ν, the behavior of the branch of fronts as γ decreases depends on whether or not a heteroclinic cycle forms. When it does, the branch of fronts approaches γ^* either monotonically (near ν_β, where $\gamma^* > \gamma_a$ and the spatial eigenvalues of $A = 0$ are real) or in a series of saddle-nodes (when $\gamma^* < \gamma_a$ and these eigenvalues are complex). An example of the

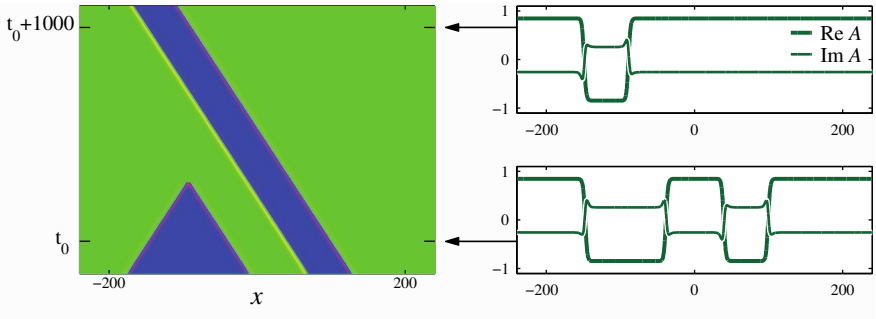


Figure 32. *Space-time plot showing the evolution of four Bloch fronts, shown in terms of* $\mathrm{Im}\,A$. *The frames on the right show the profiles of* $\mathrm{Re}\,A$ *(thick lines) and* $\mathrm{Im}\,A$ *(thin lines) at two particular times. The asymmetry of each Bloch front, which determines the drift direction, is most apparent in the latter. Parameters are as in Figure 31, with* $\gamma = 0.4$. *The solution was generated on the periodic domain* $x \in [-240, 240]$ *using a spectral method with* 1024 *modes.*

latter is shown in Figure 31(a). The front profiles that make up this branch look similar to those shown in Figure 24. At large γ these correspond to monotonic Ising fronts between A_u^+ and $-A_u^+$ but near γ^* the profiles become structured and eventually approach a heteroclinic cycle involving the state $A = 0$ as the branch approaches γ^* in a series of saddle-nodes. But in Region I^+, all the fronts that exist near γ^* are unstable due to a nonequilibrium Ising–Bloch (NIB) bifurcation [11] at $\gamma_{NIB} > \gamma^*$, where the Ising fronts lose stability with respect to an (even) phase mode. Below γ_{NIB} stability is transferred to a pair of counterpropagating Bloch fronts, distinguished from the stationary Ising fronts by their lack of odd parity. As shown in Figure 31(b) the speed c of the resulting Bloch fronts increases as the square root of the distance from γ_{NIB} [11].

The motion of the Bloch fronts is shown in Figure 32. When two such fronts traveling in opposite directions collide, they annihilate; fronts traveling in the same direction move at identical speeds and therefore never collide. We surmise that the pinning mechanism discussed in section 3.5 for Ising fronts also applies to Bloch fronts, although a detailed description of this mechanism will be given elsewhere. We have, however, found examples of Bloch fronts whose drift velocity is identically zero (i.e., pinned Bloch fronts) and in the neighborhood of such states identified stationary localized states similar to the LBF states described above (Figure 27) but resembling bound Bloch fronts.

At sufficiently large values of ν where no heteroclinic cycle forms, the branch of Ising fronts annihilates instead with a branch of kinks in a saddle-node bifurcation, as shown previously in Figure 26. The region of existence of SIFs in Region I^+ is shown in Figure 30(b). Above $\nu \approx 0.5$ the Ising fronts are stable all the way down to the saddle-node, where they join with the kinks. Below this value of ν the fronts always lose stability with respect to an NIB bifurcation prior to the saddle-node bifurcation. The line $\gamma = \gamma_{NIB}$ is shown in Figure 30(b) and forms the boundary of the SIF region. The figure shows that the NIB bifurcation prevents the formation of a wedge of *stable* structured front (SSF) states around γ^*.

A similar story applies with regard to the large amplitude localized states in Region I^+. When $\nu_\beta < \nu < \nu_z$ and the eigenvalue structure near the saddle-node is of the type shown in

Figure 16(a), (3.34) describes the branch of localized states biasymptotic to A_u^+ which bifurcate from the saddle-node at γ_b. Above ν_z the spatial eigenvalues of the uniform phase-locked states are as shown in Figure 16(b), and in this case γ_d is a bifurcation point producing the states described by (3.35) and (3.36). We find numerically that in this region the coefficients a_d and b_d are again strictly positive and hence that the bifurcation at γ_d is always subcritical, with the spatially periodic and localized states emerging into $\gamma > \gamma_d$. In either case the global behavior of the $A_{L,b}$ and $A_{L,d}$ branches depends on whether a heteroclinic cycle forms. When no such cycle forms, these branches extend to arbitrarily large γ and are always unstable. In contrast, when it does form, the branch of localized states approaches the associated γ^* either monotonically or in a series of saddle-nodes as in the example shown in Figure 20. But unlike the earlier examples, in Region I^+ the branch of localized states does not change stability at each saddle-node: as expected the broad localized states resembling heteroclinic cycles between $A = A_u^+$ and $A = 0$ inherit the instability of the $A = 0$ state, but even the narrow profiles, such as the analogue of the profile shown in Figure 21(b), are unstable. In time these unstable localized states evolve into deeper localized states that split into two (Bloch) fronts and drift away from each other. Thus, despite the multiplicity of localized states present in this region, no wedge of *stable* reciprocal oscillons (SROs) around γ^* results.

4.2. Region II^+. The behavior in Region II^+ mimics that of Region I^+, including the presence and general orientation of the line γ^*. The one notable difference is that the Turing bifurcation of the uniform phase-locked states A_u^+ at γ_d is absent in this region. Thus the large amplitude localized states always bifurcate from the saddle-node at γ_b. But the important conclusions regarding the presence of the NIB bifurcation, the extent of the SIF region, and the lack of SSF and SRO regions still apply.

4.3. Region III^+. The resonance tongue in Region III^+ is shown in Figure 33(a). The existence and stability of the uniform phase-locked states are similar to those in Region I^+. At large γ these are stable. Above ν_H the stable region extends all the way to the boundary of the resonance tongue at γ_b, while below ν_H the stable region extends only down to γ_H, where the uniform state loses stability with respect to a Hopf bifurcation. In Region III^+ the Turing bifurcation at γ_d, present when ν lies between ν_α and ν_z, falls below the Hopf bifurcation and is therefore never part of the boundary of the stable uniform states.

The γ^* line in Region III^+ is drastically different when compared to Regions I^+ and II^+. In this case it is created in a codimension-two point at which $\gamma^* = \gamma_b$ and extends to arbitrarily large γ as ν increases. Because γ^* always lies below γ_a, the spatial eigenvalues of $A = 0$ at γ^* are complex. As a result branches of large amplitude localized states approach γ^* in a series of saddle-nodes. Unlike Regions I^+ and II^+, in Region III^+ *some* of the resulting localized states are stable despite the instability of the $A = 0$ state [47]. The shaded region in Figure 33(b) shows the location of this truncated SRO region. At large values of ν the branch of large amplitude localized states $A_{L,b}$ that emerges from the saddle-node at γ_b resembles that in Figure 20, including stability assignments: the branch is initially unstable with respect to a single (even) amplitude mode, and gains and loses stability repeatedly at successive saddle-node bifurcations as the branch winds toward γ^*. However, when the localized state becomes sufficiently wide, it loses stability in a Hopf bifurcation to an oscillatory mode with a spatially localized eigenfunction. Provided this instability occurs below the second saddle-node, the

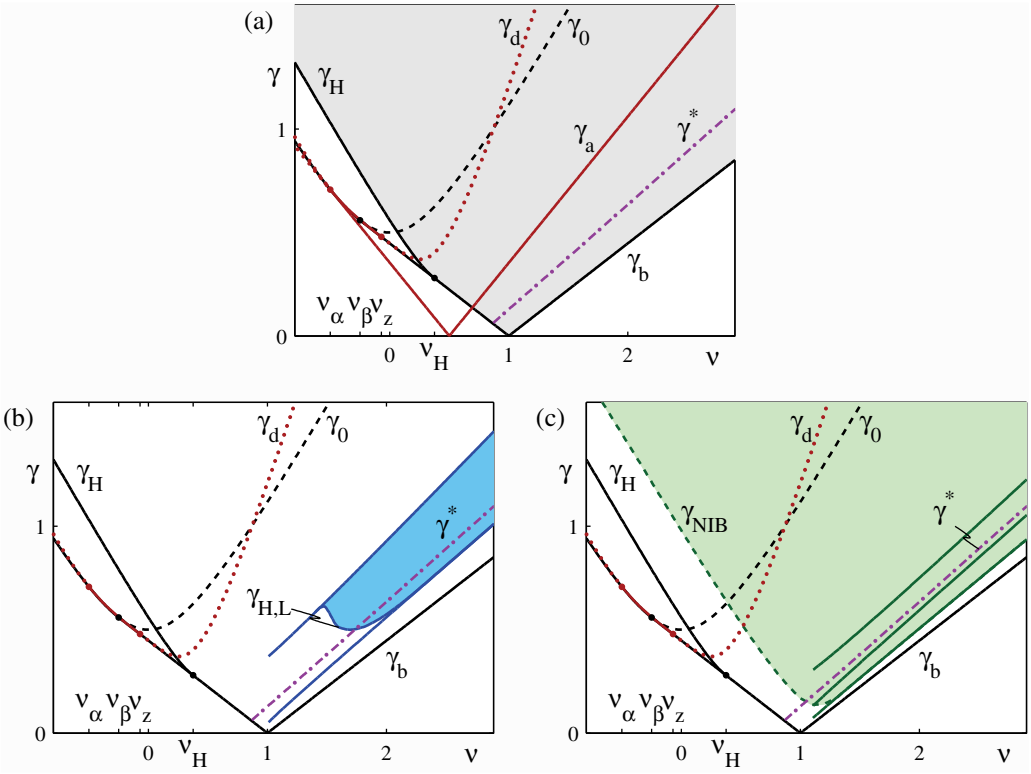


Figure 33. *The (ν,γ) plane in Region III^+. In* (a) *shading indicates the presence of stable uniform states A_u^+. In* (b) *shading indicates the SRO region, while in* (c) *it indicates the SIF region. The three unlabeled lines in* (c) *mark the first three saddle-nodes as the fronts approach γ^*. An SSF region exists near the γ^* line, between the second and third saddle-node, at larger values of ν. Parameters: $\mu = 0.5$, $\alpha = 1$, $\beta = 2$.*

SRO region spans the entire range between the first and second saddle-nodes. At large ν the Hopf bifurcation does in fact occur far "down" the $A_{L,b}$ branch, but as ν decreases this instability occurs farther "up" the branch. Figure 34 shows a bifurcation diagram at a value of ν sufficiently small that the Hopf bifurcation lies between the first two saddle-nodes. In this case the truncated SRO region extends only from the first saddle-node down to the localized Hopf bifurcation, labeled $\gamma_{H,L}$. Figure 33(b) shows that the line of localized Hopf bifurcations $\gamma = \gamma_{H,L}$ forms the boundary of the SRO region in the (ν,γ) plane. At sufficiently small ν the SRO region is completely absent because the Hopf bifurcation occurs before the first saddle-node. For reference, the location of the first two saddle-nodes is still indicated in the figure even when they do not correspond to the boundary of the SRO region.

The behavior of the fronts in Region III^+ is shown in Figure 33(c). In $\nu < \nu_\alpha$ the fronts $A_{F,0}$ emerge supercritically from γ_0 and extend to arbitrarily large γ. Using the homotopy method, we find that at large positive values of ν, where heteroclinic cycles are present, the complex eigenvalues of the $A = 0$ state result in front pinning and hence in the creation of structured fronts. At intermediate values of ν the branch of fronts terminates, as γ decreases, slightly above γ_b; here the branch becomes difficult to track numerically, and the nature of this bifurcation remains unclear. The fronts are stable at large γ and either lose stability in

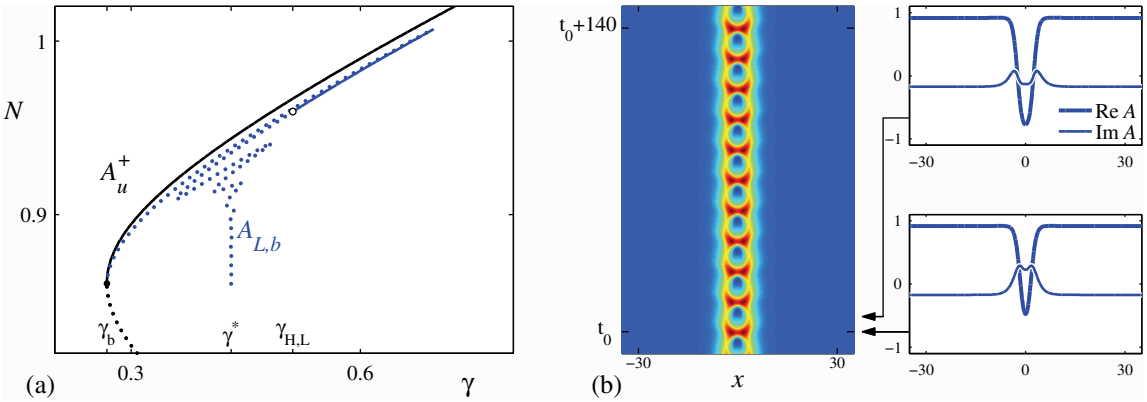


Figure 34. (a) *Bifurcation diagram in Region* III^+ *showing the branch of localized states* $A_{L,b}$ *as well as the uniform states* $A_u^{\pm}$. *The branch of localized states undergoes a localized Hopf bifurcation at* $\gamma_{H,L} \approx 0.51$. *The norm of the branch of localized states has been rescaled to distinguish it from the uniform states. Solid (dotted) lines correspond to stable (unstable) solutions.* (b) *Space-time plot showing a (stable) spatially localized temporal oscillation in* $\gamma < \gamma_{H,L}$, *shown in terms of* $\mathrm{Im}\, A$. *The oscillation period is* $T \approx 14$. *The frames on the right show the profile at two particular times separated by half a period. Parameters:* $\mu = 0.5$, $\alpha = 1$, $\beta = 2$, $\nu = 1.6$. *In* (b), $\gamma = 0.4$ *and the solution was generated on the periodic domain* $x \in [-70, 70]$ *using a spectral method with* 512 *modes; only half this domain is shown.*

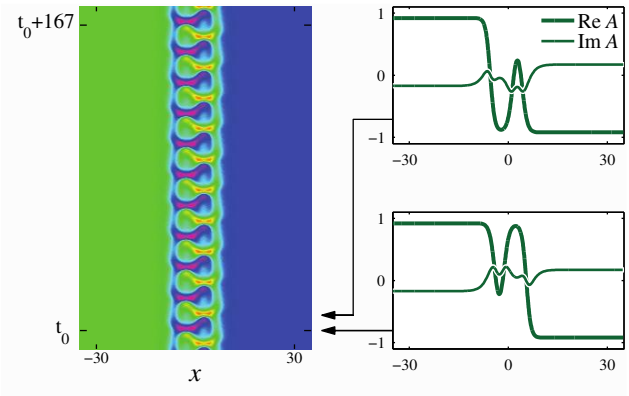


Figure 35. *Space-time plot showing a (stable) oscillating front, shown in terms of* $\mathrm{Im}\, A$. *The oscillation period is* $T \approx 16.7$; *the frames on the right show the profile at two particular times separated by half a period. Parameters are identical to those in Figure* 34(b).

an NIB bifurcation as γ decreases or, when ν is sufficiently large, remain stable down to the first saddle-node associated with the pinning region. As with the large amplitude localized states, the SSF region is truncated at small values of ν by a Hopf bifurcation. This bifurcation moves farther down the branch as ν increases and, at large values of ν (larger than the domain shown in Figure 33(c)), an SSF region does appear.

Although in this paper we are primarily interested in time-independent solutions of (2.2), it is noteworthy that some of the unstable localized states near γ^* are found to evolve in time into structures that remain spatially localized but are time-periodic. These oscillations are

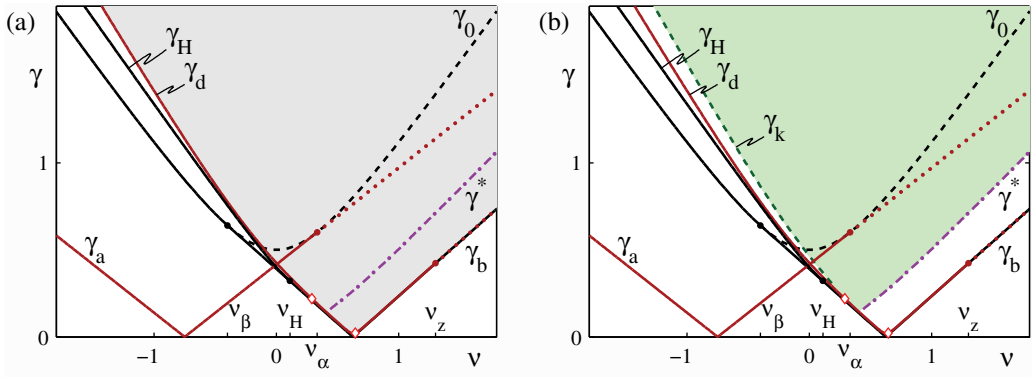


Figure 36. *The (ν,γ) plane in Region* IV$^+$. *In* (a) *shading indicates the presence of stable uniform states A_u^+, while in* (b) *it indicates the SIF region. The bifurcation along γ_d is supercritical between locations indicated by $\diamond$. In* (b) *the line γ_k corresponds to saddle-node bifurcations of the kink-like states. Parameters: $\mu = 0.5$, $\alpha = -1.5$, $\beta = 1.25$.*

confined to the inhomogeneous part of the profile, while the uniform part remains stationary (Figure 34(b)). Such states belong to the branch of localized periodic solutions (not shown) that emerges from the localized Hopf bifurcation at $\gamma_{H,L}$. In addition, the Hopf instability of the fronts in the neighborhood of γ^* leads to localized oscillating fronts (Figure 35). In either case the frequency of the oscillations is not in general rationally related to the driving frequency, and hence states of this type correspond to multifrequency oscillations of the observable w in the original system.

4.4. Region IV$^+$. The resonance tongue in Region IV$^+$, shown in Figure 36(a), resembles that in Region III$^+$, as does the location of the line $\gamma = \gamma^*$ in $\nu > \beta\mu$. However, since $\alpha < 0$ in Region IV$^+$ (and in Region V$^+$), the critical value ν_α is now positive, and the analytical result (3.14) describing the bifurcation at γ_a to spatially periodic states is valid in $\nu < \nu_\alpha$. In this range, the line $\gamma = \gamma_a$ is shown as a solid line to indicate that it corresponds to bifurcation points. For $\nu < \alpha\mu$, η_a reduces to $\alpha - \sqrt{1+\alpha^2} < 0$ and the bifurcation to spatially periodic states at γ_a is supercritical, while for $\nu > \alpha\mu$, η_a reduces to $\alpha + \sqrt{1+\alpha^2} > 0$ and the bifurcation is subcritical. In the former case no localized states are present near γ_a, while in the latter case there is in addition a subcritical bifurcation to spatially localized states $A_{L,a}$ as described by (3.16). Likewise, in $\nu > \nu_\alpha$ there is a subcritical bifurcation to spatially localized states $A_{L,0}$, as described by (3.12), which emerge from γ_0.

The uniform phase-locked states A_u^+ exist in $\gamma > \gamma_0$, $\nu < \nu_\beta$ and in $\gamma > \gamma_b$, $\nu > \nu_\beta$. The locations of the Hopf (γ_H, present in $\nu < \nu_H$) and Turing (γ_d, present in $\nu < \nu_z$) bifurcations of the uniform phase-locked states, as defined in (4.2) and (3.26), respectively, are also shown in the figure. An inspection reveals that γ_d always lies above γ_H. As a result the boundary of the stable uniform phase-locked states in $\nu < \nu_z$ is γ_d, and the Hopf bifurcation does not play an important role. Above ν_z the uniform states remain stable all the way to the saddle-node at γ_b.

Likewise, the spatial eigenvalues of the uniform phase-locked states are of the type shown in Figure 16(b) whenever $\nu < \nu_z$, in the range where γ_d is a bifurcation point. The spatially

periodic states $A_{P,d}$ described by (3.35) bifurcate subcritically (into $\gamma > \gamma_d$) over most of the domain. There is, however, a range of ν values for which this bifurcation is supercritical ($\gamma < \gamma_d$), as indicated in Figure 36(a). In the former case the coefficients a_d, b_d defined in Appendix D are both positive, and there is in addition a branch of spatially localized states $A_{L,d}$, given by (3.36), that emerges subcritically from γ_d. In the latter case $a_d > 0$ while $b_d < 0$, and no localized states bifurcate from γ_d. When $\nu > \nu_z$, the spatial eigenvalues in the neighborhood of γ_d and γ_b are described instead by Figure 16(a). In this case the localized states $A_{L,b}$ given by (3.34) emerge from the saddle-node.

The γ^* line for this region is also shown in the figure and emerges from a codimension-two point defined by $\gamma^* = \gamma_d$ that lies slightly below $\nu = \beta\mu$ (at which $\gamma_b = 0$). In contrast, in Region III$^+$ the analogous codimension-two point occurs at $\gamma^* = \gamma_b$. It is clear from the figure that the codimension-two point lies between the two $\diamond$'s where no localized states are known analytically, although a supercritical bifurcation to spatially periodic solutions $A_{P,d}$ is present. Because the γ^* line lies above γ_d in a region with complex spatial eigenvalues of the A_u^+ state, we expect branches of small amplitude localized states near γ^* to experience pinning. While this is confirmed numerically, these SSO-like states inherit the instability of the $A = 0$ background and are always unstable. Over the domain shown in the figure (up to about $\nu = 2$) the γ^* line lies below γ_a, in a region with complex spatial eigenvalues of the $A = 0$ state. Thus branches of large amplitude fronts and localized states near γ^* are expected to exhibit pinning. At larger values of ν the γ^* line crosses above γ_a, the spatial eigenvalues become real, and pinning cannot occur.

We consider first the behavior of the large amplitude localized states, found numerically by continuing one of the analytically known solutions. The localized states $A_{L,d}$ that emerge from γ_d below the leftmost $\diamond$ in the figure extend to arbitrarily large γ since, at these values of ν, no heteroclinic cycles form. The localized states $A_{L,d}$ that emerge from γ_d above the rightmost $\diamond$ in the figure, and the localized states $A_{L,b}$ that emerge from γ_b when $\nu > \nu_z$, always approach a heteroclinic cycle and do so as expected either in a series of saddle-node bifurcations or monotonically. These solutions are unstable near their creation at γ_d or γ_b to a single amplitude mode; as γ increases, an additional localized Hopf mode destabilizes the solutions below the γ^* line, thereby eliminating a potential SRO region.

As for the fronts, these exist at large γ and can be found using the homotopy method described above. Several types of behavior are observed as γ decreases. Below $\nu \approx 0.2$ the Ising fronts terminate in a saddle-node bifurcation involving kink fronts. Above $\nu \approx 0.5$ a heteroclinic cycle is present and the Ising fronts approach γ^* as expected either in a series of saddle-node bifurcations or monotonically. Between these two limits the behavior remains unclear. The Ising fronts are stable at large γ. At large negative values of ν these fronts remain stable down to the saddle-node bifurcation, where this branch merges with (unstable) kinks. At less negative values the Ising fronts appear to lose stability above this saddle-node to an NIB bifurcation. At large positive values of ν the Ising fronts also always lose stability with decreasing γ, this time to a localized Hopf mode. The Hopf instability occurs above the first saddle-node when the approach to γ^* is through a series of saddle-nodes and above γ^* when the approach is monotonic. As a result the boundary of the SIF region near γ^* consists of the line of localized Hopf bifurcations, and the structured fronts that may be present near γ^* are all unstable.

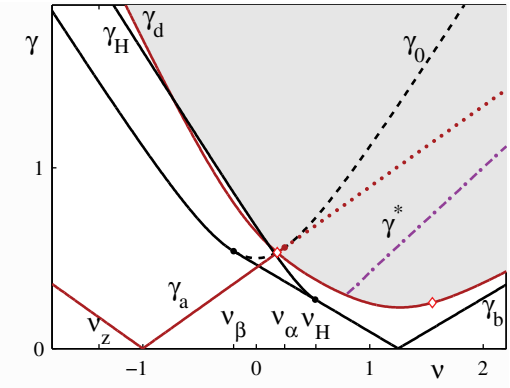


Figure 37. *The* (ν, γ) *plane in Region* V^+. *Parameters:* $\mu = 0.5$, $\alpha = -2$, $\beta = 2.5$.

In summary, Region IV^+ exhibits pinning of both small and large amplitude states near the γ^* line, leading to a multiplicity of spatially localized but *unstable* states in its vicinity, in regions resembling the SSO, SRO, and SSF regions already identified for $\mu < 0$.

4.5. Region V^+. The resonance tongue in Region V^+ is shown in Figure 37 and resembles Region IV^+. The main difference relates to γ_d. First, the tangency at ν_z is now absent, and hence γ_d is always a bifurcation point. It follows that the localized states $A_{L,b}$ emerging from the saddle-node at γ_b are never present in this region. Instead the branches of localized states $A_{L,d}$ emerge subcritically from γ_d, but only outside of the region marked by the two $\diamond$'s in the figure. Between these two points the bifurcation to spatially periodic states $A_{P,d}$ is supercritical, and no localized states are predicted by local analysis. The second major difference compared to Region IV^+ is that now the line γ_H of Hopf bifurcations of the uniform phase-locked states intersects the line γ_d (twice). Hence, there is a range in ν where the boundary of the stable uniform solutions is made up of γ_H instead of γ_d. The codimension-two points at which $\gamma_d = \gamma_H$ have not been studied, but an analogous interaction between a Turing bifurcation and a Hopf bifurcation of the $A = 0$ state has been considered [58, 59] and is known to create interesting dynamics.

The behavior of the fronts and localized states observed by continuing these states numerically also resembles that described for Region IV^+. Both large and small amplitude states approach the γ^* line in a series of saddle-nodes creating SSO-like, SRO-like, and SSF-like regions, but the resulting states are always unstable because the respective branch undergoes a localized Hopf or an NIB bifurcation before reaching γ^*.

5. Discussion and conclusions. The present study was largely motivated by the discovery in several distinct systems of subharmonic spatially localized temporal oscillations called oscillons [4, 40, 44, 52]. The experiments indicate that oscillons of this type are present inside the 2:1 resonance tongue, and it is natural in these circumstances to examine the vicinity of the boundary of this region, where homogeneous phase-locked oscillations coexist with the trivial state. This coexistence region is produced in response to the forcing and is found even in systems exhibiting supercritical dynamics in the absence of forcing. In the absence of primitive equations describing chemical or granular media, we have chosen to examine the

possible existence of spatially localized structures within the framework of the forced complex Ginzburg–Landau (FCGL) equation. This equation can be viewed as the normal form for periodically forced oscillations near the 2:1 resonance in extended dissipative systems. As such the FCGL equation describes the large-scale dynamics of all extended systems driven sufficiently close to the onset of a primary oscillatory instability by a sufficiently weak periodic force sufficiently close to the 2:1 temporal resonance. Within this equation the oscillon problem reduces to the study of time-independent spatially localized solutions for the amplitude of these oscillations; i.e., the background oscillation at half the forcing frequency is factored out. If this is restored and the corresponding states are recovered from (2.1), we see that the localized structures we have called standard oscillons do indeed resemble the type of localized temporal oscillation observed in the experiments, while the reciprocal oscillons resemble holes in a background that also oscillates with half the forcing frequency. Likewise, the Ising fronts [11] become phase kinks, connecting two nontrivial states oscillating exactly out of phase.

5.1. Summary of the results. Our study employs the technique of spatial dynamics. We have seen that this approach coupled with numerical branch following provides a powerful technique for analyzing problems of this type. The spatial dynamics approach allows us to view spatially periodic states as periodic orbits in space and identifies spatially localized states with homoclinic (oscillons) or heteroclinic (fronts) orbits. More importantly, this approach leads one to focus on the spatial eigenvalues of the spatially homogeneous states, and these in turn allow one to identify the parameter regimes where localized structures are likely and to interpret the large multiplicity of coexisting states as a consequence of a mechanism we refer to as pinning. The same mechanism is, in addition, responsible for stabilizing some of these states against time-dependent perturbations. As shown in the appendices, we have used bifurcations in the spatial eigenvalues to construct analytically both spatially periodic and spatially localized states, which can in turn be used to initialize numerical continuation. These techniques led us to identify parameter regimes in which the various localized structures are all organized around a special point in parameter space corresponding to the formation of a heteroclinic cycle in space. We have called this point γ^*. Among the results from our analysis are the following:

1. Identification of two distinct types of "small amplitude" oscillons, referred to collectively as standard oscillons. Both are biasymptotic to the rest state $A = 0$ as $x \to \pm\infty$. The first of these bifurcates from $A = 0$ at the phase-locking threshold γ_0 and resembles a single bump; near the heteroclinic cycle at γ^* these may be stable (SSO region in Regions IV^- and V^-). The second consists of localized spatial oscillations of even or odd parity that emerge from the Turing bifurcation at γ_a and are never stable.
2. Identification of two distinct types of "large amplitude oscillons," biasymptotic to A_u^+ or $-A_u^+$ as $x \to \pm\infty$, referred to collectively as reciprocal oscillons. The first of these emerges from the boundary of the fundamental 2:1 resonance tongue at γ_b and resembles a hole in a uniform background; near γ^* these may be stable (SRO region in Regions III^- and III^+). The second type is an even parity localized spatial oscillation on a uniform background that emerges from the Turing bifurcation at γ_d and is never stable.
3. Identification of two classes of large amplitude front-like states called Ising fronts and

Table 1

Summary of the behavior in the various regions of the (α, β) plane. Each region is defined uniquely by columns 2–5. In column 6 "yes" ("no") indicates the presence (absence) of a γ^* line. In column 7 "yes" indicates that an SSO region forms around a part of the γ^* line, "no" indicates that an SSO region is absent (either because there is no pinning at γ^*, or because additional instabilities eliminate stable standard oscillons), while "." indicates the absence of an SSO region due to the absence of the γ^* line. The same notation is used in column 8 to indicate the existence of SRO and SSF regions around γ^*. The final column emphasizes that an SIF region is always present, regardless of the presence of γ^*.

Region		sgn(μ)	sgn(α)	sgn($\alpha-\beta$)	sgn(z)	γ^* line	SSO region	SRO/SSF regions	SIF region
I^-		$-$	$+$	$+$	$-$	no	.	.	yes
II^-	(Fig. 6)	$-$	$+$	$+$	$+$	no	.	.	yes
III^-	(Fig. 7)	$-$	$+$	$-$	$+$	yes	no	yes	yes
IV^-	(Fig. 10)	$-$	$-$	$-$	$+$	yes	yes	no	yes
V^-	(Fig. 11)	$-$	$-$	$-$	$-$	yes	yes	no	yes
I^+	(Fig. 30)	$+$	$+$	$+$	$-$	yes	no	no	yes
II^+		$+$	$+$	$+$	$+$	yes	no	no	yes
III^+	(Fig. 33)	$+$	$+$	$-$	$+$	yes	no	yes	yes
IV^+	(Fig. 36)	$+$	$-$	$-$	$+$	yes	no	no	yes
V^+	(Fig. 37)	$+$	$-$	$-$	$-$	yes	no	no	yes

structured fronts, respectively; the latter possess internal structure in the front region (cf. [61]). These states are stable in regions called SIF (present in all five regions of the (α, β) plane) and SSF (present only in Regions III^- and III^+), respectively. The SRO and SSF regions overlap, while the SSO region is present for distinct parameter values.

4. Identification of a heteroclinic cycle at γ^* between the trivial and the phase-locked states A_u^+ in the subcritical region of the fundamental 2:1 resonance tongue, and elucidation of its role as an organizing center not only for the standard oscillons, but also for reciprocal oscillons and structured fronts. In addition, a variety of more complex states referred to as multipulse states also emanates from the vicinity of γ^*.
5. Detailed discussion of the differences between the damped ($\mu < 0$) and self-excited ($\mu > 0$) cases, and in particular between the stability properties of the localized structures in these two cases; identification of stable large amplitude fronts even in the absence of a heteroclinic cycle, including pinned and moving Bloch fronts.

The existence and stability results for localized states are summarized in the broadest terms in Table 1. Regions I–V are defined in terms of analytically accessible properties of the spatially uniform phase-locked states only. As a result, the existence and stability regions of some of the other states may extend across the region boundaries.

5.2. Connection to experiments. The standard oscillons observed, for example, in vertically vibrated granular media [52] resemble the single bump standard oscillons identified in the FCGL equation (2.2). In the first explanation of this phenomenon, a second field was required to stabilize states of this type [51]. The present work shows that this type of standard oscillon can be stable in damped driven systems even in the absence of a second field (SSO region in Regions IV^- and V^-).

In contrast, experiments on the optically forced BZ reaction [40, 44] have revealed the

presence of spot-like structures embedded in a background state that oscillates 180° out of phase with the spot. These states resemble the states we have called reciprocal oscillons and, in particular, the states $A_{L,b}$ and A_L shown in Figures 18(c) and 19(d),(e), respectively. However, our results indicate that these states are unstable unless the parameters are chosen to lie in Region III$^+$, close to γ^*, with $\nu \gg \nu_\beta$ (Figure 33(b)). Thus it may be more likely that the observed spots may in fact correspond to the bound fronts A_{LBF} shown in Figure 27, should states of this type prove to be stable.

Existing experiments on the BZ system [40] also reveal that the diameter of the observed reciprocal oscillons is comparable to the width of the Ising fronts connecting out-of-phase spatially homogeneous oscillations. The experiments reveal, moreover, that the reciprocal oscillons exist in a narrow parameter range near the onset of labyrinthine patterns, which in turn form as finite amplitude patterns near the boundary of the 2:1 resonance tongue [4, 60]. These results are in qualitative agreement with simulations of a modified version of (2.2) that includes interfacial tension [4]. In a companion paper [57] we present results of numerical integration of (2.2) in time in two spatial dimensions. These results suggest that the reciprocal oscillons and front-like structures described here in one spatial dimension possess analogues in two spatial dimensions and show that stable spot-like states and stable front-like phase kinks may indeed have similar widths and that both are found within the pinning region identified in one spatial dimension. We use these results to conjecture that the observed labyrinthine patterns are in fact the result of a transverse instability [4, 21, 22, 58] of the front-like states associated with the reciprocal oscillons in the pinning region. This aspect of the problem will be pursued in a future publication.

5.3. Future directions. From a broader theoretical perspective, the properties of the FCGL equation summarized above resemble those discussed at greater length for the Swift–Hohenberg equation (see [7] and references therein). This equation is also a bistable reversible fourth order system in space. However, there are important differences between the two sets of equations, the most significant being the fact that the Swift–Hohenberg equation is variational. Although the FCGL equation can be reduced to the Swift–Hohenberg equation in specific parameter regimes [4], in other regimes the stability properties of the two systems will in general differ. In particular, as we have seen in this paper, the FCGL equation can exhibit persistent time dependence, unlike the Swift–Hohenberg equation.

Although the Swift–Hohenberg equation is known to exhibit bistability between uniform states of the type studied in this paper, recent work on this equation has focused on the consequences of bistability between the trivial state and a spatially *periodic* state. As a result, the fronts that bound a spatially localized state at either side can lock to the underlying spatial structure, and the heteroclinic bifurcation opens out into a pinning *region*. This pinning region is defined by pairs of branches of localized states that snake back and forth across the Maxwell point, giving rise to a characteristic "snakes-and-ladders" structure [7, 8]. In contrast, when the bistability is between two spatially homogeneous states, pinning arises only from the oscillating *tails*. In this case the bounding fronts interact ever more weakly as the distance between them increases, and the pinning region necessarily collapses to the Maxwell point. Thus no true snaking occurs. Our study of the FCGL equation indicates that the variational structure, while helpful in the physical interpretation of the pinning region [8], is not a prerequisite for its presence in more general forced dissipative systems.

Indeed, our results indicate that similar behavior will be present in other systems exhibiting bistability between two spatially uniform states, provided that the system is reversible under spatial reflection and of fourth or higher order in space. The reversibility property renders the homoclinic connections corresponding to the standard and reciprocal oscillons structurally stable (codimension-zero) while making the creation of the heteroclinic cycle between $A = 0$ and $A = A_u^+$ of codimension one and hence is responsible for making the behavior associated with the pinning or snaking region of codimension one. In contrast, in nonreversible systems such as those describing traveling pulses in the comoving frame [62], homoclinic connections to homogeneous equilibria such as $A = 0$ and $A = A_u^+$ are of codimension one, while the formation of a pair of heteroclinic connections is of codimension two. Despite this difference, the behavior near each global connection, once formed, resembles that described here for the FCGL equation [62].

The detailed survey of spatially localized oscillations presented in this paper provides the necessary theoretical background for future experimental and theoretical explorations of spatially extended parametrically driven dissipative systems and their association with structured fronts, Bloch fronts, and labyrinthine patterns.

Appendix A. Weakly nonlinear analysis near $\gamma = \gamma_0$. At $\gamma = \gamma_0$ the $A = 0$ state has two zero spatial eigenvalues and two nonzero spatial eigenvalues. In the following we assume that $\mu < 0$ and note that when $\alpha > 0$ and $\nu > \nu_\alpha$ the two nonzero eigenvalues are imaginary. In this case we do not expect localized states nearby (Figure 5(b)). On the other hand, if $\alpha > 0$, $\nu < \nu_\alpha$, the two nonzero eigenvalues are real and for $\gamma < \gamma_0$ the zero eigenvalues split along the real axis (Figure 5(a)). Thus localized states may exist in $\gamma < \gamma_0$. To find these states we write $\gamma = \gamma_0 + \epsilon^2\delta$, where $\epsilon \ll 1$ and δ is an order one quantity. Writing $A \equiv U + iV$, the appropriate expansion of the fields U and V is

$$\begin{bmatrix} U \\ V \end{bmatrix} = \epsilon \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \epsilon^3 \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} \cdots , \tag{A.1}$$

where all quantities depend on x via the slow spatial variable $X \equiv \epsilon x$ only. With these scalings the linear operator in (3.1) becomes $\mathcal{L} = \mathcal{L}_0 + \epsilon^2\mathcal{L}_2$, where

$$\mathcal{L}_0 = \begin{bmatrix} \mu + \gamma_0 & -\nu \\ \nu & \mu - \gamma_0 \end{bmatrix}, \quad \mathcal{L}_2 = \begin{bmatrix} \delta & 0 \\ 0 & -\delta \end{bmatrix} + \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{XX}, \tag{A.2}$$

while the nonlinear operator becomes $\mathcal{N}(U, V) = \epsilon^2\mathcal{N}(u_1, v_1) + \ldots$.

At order ϵ stationary solutions to (3.1) satisfy

$$\mathcal{L}_0 \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{A.3}$$

implying that

$$\begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} \eta_0 \\ 1 \end{bmatrix} B(X), \tag{A.4}$$

where $\eta_0 \equiv (\gamma_0 - \mu)/\nu$ and $B(X)$ is a real-valued function of X. At order ϵ^3 we obtain

$$\mathcal{L}_0 \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} = -\{\mathcal{L}_2 + \mathcal{N}(u_1, v_1)\} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix}. \tag{A.5}$$

Since $\mathcal{L}_0$ is singular, (A.5) requires a solvability condition. The required condition is obtained by taking the scalar product of the equation with the adjoint null eigenvector of $\mathcal{L}_0$, viz.,

$$\Xi_0 = \begin{bmatrix} -\eta_0 & 1 \end{bmatrix}. \tag{A.6}$$

The result can be written in the form

$$a_0 B_{XX} = \delta B + b_0 B^3, \tag{A.7}$$

where $a_0 = \alpha(\nu - \nu_\alpha)/\gamma_0$, $b_0 = 2\beta(\nu - \nu_\beta)(\gamma_0 - \mu)/\nu^2$. This equation admits spatially homogeneous solutions $B = \sqrt{-\delta/b_0}$ or, equivalently,

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \eta_0 \\ 1 \end{bmatrix} \sqrt{\frac{\gamma_0 - \gamma}{b_0}} + \ldots, \tag{A.8}$$

where the omitted terms are of higher order in $|\gamma_0 - \gamma|$. This solution bifurcates supercritically when $b_0 < 0$ (i.e., $\nu < \nu_\beta$) and corresponds then to the A_u^+ states discussed in section 2. When $b_0 > 0$ (i.e., $\nu > \nu_\beta$) the bifurcation is subcritical and the homogeneous solutions then correspond to the A_u^- states.

Equation (A.7) also admits several different space-dependent solutions. Based on the eigenvalue analysis, we expect localized states only when $\nu < \nu_\alpha$, corresponding to $a_0 < 0$. When, in addition, $b_0 < 0$ (i.e., $\nu < \nu_\beta$), there is a supercritical bifurcation to front-like states of the form

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \eta_0 \\ 1 \end{bmatrix} \sqrt{\frac{\gamma - \gamma_0}{-b_0}} \tanh\left(\sqrt{\frac{\gamma - \gamma_0}{-2a_0}} x\right) + \ldots. \tag{A.9}$$

In contrast, when $b_0 > 0$ (i.e., $\nu > \nu_\beta$), there is a subcritical bifurcation to localized states of the form

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \eta_0 \\ 1 \end{bmatrix} \sqrt{\frac{\gamma_0 - \gamma}{b_0/2}} \operatorname{sech}\left(\sqrt{\frac{\gamma_0 - \gamma}{-a_0}} x\right) + \ldots. \tag{A.10}$$

Solutions of the latter type exist only if $\alpha < \beta$, i.e., in Region III (since $\alpha > 0$, by assumption). Evidently at fixed values of the parameters α, β, μ, ν at most one of these solution branches can be present.

Equation (A.7) has localized solutions analogous to (A.9) and (A.10) even when $a_0 > 0$ ($\nu > \nu_\alpha$) where localized states are not expected. This expectation is confirmed by the numerics: the approximate solutions constructed by the above method when $a_0 > 0$ cannot be extended away from $\gamma = \gamma_0$ and do not represent valid solutions of the original FCGL equation (2.2).

A similar discussion applies when $\alpha < 0$ but is omitted.

Appendix B. Weakly nonlinear analysis near $\gamma = \gamma_a$. At $\gamma = \gamma_a$ the four spatial eigenvalues of the $A = 0$ state collide at $\lambda = \pm ik_a$, each of double multiplicity, where

$$k_a \equiv \frac{\sqrt{\alpha(\nu - \nu_\alpha)}}{\rho_\alpha}. \tag{B.1}$$

As in the previous appendix we assume that $\mu < 0$, $\alpha > 0$. In this case, k_a is real when $\nu > \nu_\alpha$ and the eigenvalue collision occurs on the imaginary axis: below γ_a the eigenvalues form a complex quartet and above γ_a they are imaginary (Figure 5(b)). In the theory of reversible systems this situation is referred to as the reversible Hopf bifurcation [26], although here it takes place in space instead of time. The theory predicts the presence of localized states in $\gamma < \gamma_a$, consisting of oscillations with wavenumber k_a modulated on a large scale owing to the small real part of λ. In contrast, when $\nu < \nu_\alpha$, the wavenumber k_a is imaginary and the eigenvalue collision occurs on the real axis; thus no bifurcation occurs (Figure 5(a)).

To find the localized states expected when $\nu > \nu_\alpha$, we set $\gamma = \gamma_a + \epsilon^2\delta$, where $\epsilon \ll 1$ and δ is an order one quantity, and allow all quantities to depend on both the short scale x and the long scale $X = \epsilon x$. The appropriate expansion of the fields is

$$\begin{bmatrix} U \\ V \end{bmatrix} = \epsilon \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \epsilon^2 \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} + \epsilon^3 \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} \cdots . \tag{B.2}$$

With this scaling the linear operator in (3.1) becomes $\mathcal{L} = \mathcal{L}_0 + \epsilon\mathcal{L}_1 + \epsilon^2\mathcal{L}_2$, where

$$\mathcal{L}_0 = \begin{bmatrix} \mu + \gamma_a & -\nu \\ \nu & \mu - \gamma_a \end{bmatrix} + \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{xx}, \tag{B.3}$$

$$\mathcal{L}_1 = 2 \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_x \partial_X, \quad \mathcal{L}_2 = \begin{bmatrix} \delta & 0 \\ 0 & -\delta \end{bmatrix} + \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{XX}, \tag{B.4}$$

while the nonlinear operator becomes $\mathcal{N}(U, V) = \epsilon^2\mathcal{N}(u_1, v_1) + \ldots$.

At order ϵ stationary solutions to (3.1) satisfy

$$\mathcal{L}_0 \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{B.5}$$

and hence

$$\begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} \eta_a \\ 1 \end{bmatrix} B(X)\, e^{ik_a x} + c.c., \tag{B.6}$$

where $B(X)$ is a complex-valued function of X and

$$\eta_a \equiv \alpha + \mathrm{sgn}(\nu - \alpha\mu)\rho_\alpha. \tag{B.7}$$

Since $\alpha > 0$ and $\nu > \nu_\alpha$, the argument of sgn(.) is always positive, and

$$\eta_a = \alpha + \rho_\alpha > 0. \tag{B.8}$$

Expression (B.1) for k_a is also obtained at this order.

The order ϵ^2 terms in (3.1) are

$$\mathcal{L}_0 \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} = -\mathcal{L}_1 \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} \tag{B.9}$$

or

$$\mathcal{L}_0 \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} = -2i\rho_\alpha k_a \begin{bmatrix} 1 \\ \eta_a \end{bmatrix} B_X e^{ik_a x} + c.c. \tag{B.10}$$

The solvability condition for this equation is obtained by taking the scalar product with the adjoint null eigenvector,

$$\mathbf{\Xi}_a = \begin{bmatrix} -\eta_a & 1 \end{bmatrix} e^{-ik_a x} + c.c., \tag{B.11}$$

and integrating over the real line. Since this condition is automatically satisfied, we solve (B.10) and obtain

$$\begin{bmatrix} u_2 \\ v_2 \end{bmatrix} = \begin{bmatrix} \eta_a \\ 1 \end{bmatrix} C(X) e^{ik_a x} + 2ik_a \frac{\rho_\alpha^2}{\gamma_a} \begin{bmatrix} 0 \\ 1 \end{bmatrix} B_X e^{ik_a x} + c.c., \tag{B.12}$$

where $C(X)$ is an arbitrary complex-valued function that may be set to zero.

Proceeding to order ϵ^3, we obtain

$$\mathcal{L}_0 \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} = -\left\{\mathcal{L}_2 + \mathcal{N}(u_1, v_1)\right\} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} - \mathcal{L}_1 \begin{bmatrix} u_2 \\ v_2 \end{bmatrix}. \tag{B.13}$$

The solvability condition for this equation takes the form

$$a_a B_{XX} = -\delta B - b_a B|B|^2, \tag{B.14}$$

where $a_a = 2\rho_\alpha^2 k_a^2/\gamma_a$, $b_a = 6\eta_a(\beta - \alpha)$. The spatially periodic solution $B = \sqrt{-\delta/b_a}\, e^{i\varphi}$, or, equivalently,

$$\begin{bmatrix} U \\ V \end{bmatrix} = 2 \begin{bmatrix} \eta_a \\ 1 \end{bmatrix} \sqrt{\frac{\gamma_a - \gamma}{b_a}} \cos\left(k_a x + \varphi\right), \tag{B.15}$$

is always present when $\nu > \nu_\alpha$ and bifurcates subcritically when $b_a > 0$ (i.e., $\beta > \alpha$) or supercritically when $b_a < 0$ (i.e., $\beta < \alpha$). The phase φ is arbitrary and is a result of spatial translation invariance of (2.2).

In addition to the periodic states, there are two other types of solutions that are possible, depending on the sign of the coefficient b_a. When $b_a < 0$ (i.e., $\beta < \alpha$), there is a supercritical bifurcation to front-like states of the form

$$\begin{bmatrix} U \\ V \end{bmatrix} = 2 \begin{bmatrix} \eta_a \\ 1 \end{bmatrix} \sqrt{\frac{\gamma - \gamma_a}{-b_a}} \tanh\left(\sqrt{\frac{\gamma - \gamma_a}{2a_a}} x\right) \cos\left(k_a x + \varphi\right). \tag{B.16}$$

These fronts connect two out-of-phase periodic states of the type described by (B.15). However, despite their intrinsic interest, solutions of this type cannot be computed by our techniques, and their role in the dynamics of (2.2) remains unclear. Solutions of this type are therefore omitted from all bifurcation diagrams.

In contrast, when $b_a > 0$ (i.e., $\beta > \alpha$), there is a subcritical bifurcation to localized states of the form

$$\begin{bmatrix} U \\ V \end{bmatrix} = 2 \begin{bmatrix} \eta_a \\ 1 \end{bmatrix} \sqrt{\frac{\gamma_a - \gamma}{b_a/2}} \operatorname{sech}\left(\sqrt{\frac{\gamma_a - \gamma}{a_a}} x\right) \cos\left(k_a x + \varphi\right), \tag{B.17}$$

and no front-like states are present. The spatial phase φ remains arbitrary at the level of (B.14), but this is no longer the case when terms beyond all orders are included. These terms select the phases $\varphi = 0, \pi/2, \pi, 3\pi/2$; these correspond to odd ($\phi = \pi/2, 3\pi/2$) and even ($\phi = 0, \pi$) solutions of (B.17); cf. [8].

Appendix C. Weakly nonlinear analysis near $\gamma = \gamma_b$. In this appendix we assume that $\mu < 0$, $z > 0$. In this case a saddle-node bifurcation involving the uniform phase-locked states A_u^+ and A_u^- occurs at $\gamma = \gamma_b$ whenever $\nu > \nu_\beta$. At this point the uniform state has two zero spatial eigenvalues and two nonzero spatial eigenvalues, and the nonzero eigenvalues are real provided $\nu > \nu_z$ (Figure 16(a)). Along the A_u^+ branch the zero eigenvalues split along the real axis, and localized states may exist in the form of orbits homoclinic to A_u^+. To find these states we write $\gamma = \gamma_b + \epsilon^2\delta$, where $\epsilon \ll 1$ and $\delta > 0$ is an order one quantity. With $A = U + iV$ the localized states of interest can be written in the form

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix}^+ + \begin{bmatrix} u \\ v \end{bmatrix}, \tag{C.1}$$

where the first term is the uniform phase-locked state A_u^+ and the second corresponds to the space-dependent terms that decay to zero as $x \to \pm\infty$.

Based on the scaling defined above the uniform phase-locked states A_u^+ can be approximated by the series

$$\begin{bmatrix} U \\ V \end{bmatrix}^+ = \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} + \epsilon \begin{bmatrix} U_1 \\ V_1 \end{bmatrix} + \epsilon^2 \begin{bmatrix} U_2 \\ V_2 \end{bmatrix} + \dots, \tag{C.2}$$

where

$$\begin{bmatrix} U_0 \\ V_0 \end{bmatrix} = \begin{bmatrix} \eta_b \\ 1 \end{bmatrix} \Upsilon_0, \qquad \begin{bmatrix} U_1 \\ V_1 \end{bmatrix} = \sqrt{\delta} \begin{bmatrix} \xi_b \\ 1 \end{bmatrix} \Upsilon_1, \tag{C.3}$$

$$\eta_b = \beta + \operatorname{sgn}(\nu - \beta\mu)\rho_\beta, \quad \xi_b = \frac{\eta_b \nu + (1 - \beta\eta_b)|A_u(\gamma_b)|^2}{\nu - (\beta + \eta_b)|A_u(\gamma_b)|^2}, \tag{C.4}$$

$$\Upsilon_0 = \frac{|A_u(\gamma_b)|}{\sqrt{1 + \eta_b^2}}, \quad \Upsilon_1 = \operatorname{sgn}[\xi_b\eta_b + 1]\sqrt{\frac{\eta_b}{(\xi_b\eta_b + 1)(\xi_b - \eta_b)}}. \tag{C.5}$$

Since $\mu < 0$ and $\nu > \nu_\beta$, the argument of sgn(.) in (C.4) is always positive, and η_b reduces to

$$\eta_b = \beta + \rho_\beta. \tag{C.6}$$

The sign of the coefficient Υ_1 in (C.5) is chosen to ensure that the expansion in (C.2) corresponds to the A_u^+ state; the opposite sign yields the A_u^- state.

The second term in (C.1) can be expanded as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \epsilon \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \epsilon^2 \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} + \dots, \tag{C.7}$$

where all quantities depend on x via the slow spatial scale $X \equiv \epsilon^{1/2} x$. The linear operator in (3.1) takes the form $\mathcal{L} = \mathcal{L}_0 + \epsilon \mathcal{L}_1 + \epsilon^2 \mathcal{L}_2$, where

$$\mathcal{L}_0 = \begin{bmatrix} \mu + \gamma_b & -\nu \\ \nu & \mu - \gamma_b \end{bmatrix}, \quad \mathcal{L}_1 = \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{XX}, \quad \mathcal{L}_2 = \begin{bmatrix} \delta & 0 \\ 0 & -\delta \end{bmatrix}, \tag{C.8}$$

while the nonlinear terms take the form $\mathcal{N} = \mathcal{N}_0 + \epsilon \mathcal{N}_1 + \epsilon^2 \mathcal{N}_2 + \dots$, where

$$\mathcal{N}_0 = -\begin{bmatrix} U_0 & V_0 \end{bmatrix} \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}, \quad \mathcal{N}_1 = -2 \begin{bmatrix} U_0 & V_0 \end{bmatrix} \begin{bmatrix} U_1 + u_1 \\ V_1 + v_1 \end{bmatrix} \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}, \tag{C.9}$$

$$\mathcal{N}_2 = -\left\{ \begin{bmatrix} U_1 + u_1 & V_1 + v_1 \end{bmatrix} \begin{bmatrix} U_1 + u_1 \\ V_1 + v_1 \end{bmatrix} + 2 \begin{bmatrix} U_0 & V_0 \end{bmatrix} \begin{bmatrix} U_2 + u_2 \\ V_2 + v_2 \end{bmatrix} \right\} \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}. \tag{C.10}$$

At order ϵ^0 stationary solutions to (3.1) satisfy

$$\{\mathcal{L}_0 + \mathcal{N}_0\} \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{C.11}$$

an equality that holds by virtue of the definition of U_0 and V_0. At order ϵ we obtain

$$\{\mathcal{L}_0 + \mathcal{N}_0\} \begin{bmatrix} U_1 + u_1 \\ V_1 + v_1 \end{bmatrix} = -\{\mathcal{L}_1 + \mathcal{N}_1\} \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}. \tag{C.12}$$

The X-independent terms in this equation cancel by virtue of the definition of U_1 and V_1, leaving

$$\left\{ \mathcal{L}_0 + \mathcal{N}_0 - 2 \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} U_0^2 & U_0 V_0 \\ U_0 V_0 & V_0^2 \end{bmatrix} \right\} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{C.13}$$

Thus

$$\begin{bmatrix} u_1 \\ v_1 \end{bmatrix} = \begin{bmatrix} \xi_b \\ 1 \end{bmatrix} B(X), \tag{C.14}$$

where $B(X)$ is an unknown real-valued function of X.

Proceeding to order ϵ^2, we obtain

$$\{\mathcal{L}_0+\mathcal{N}_0\}\begin{bmatrix}U_2+u_2\\V_2+v_2\end{bmatrix}=-\{\mathcal{L}_1+\mathcal{N}_1\}\begin{bmatrix}U_1+u_1\\V_1+v_1\end{bmatrix}-\{\mathcal{L}_2+\mathcal{N}_2\}\begin{bmatrix}U_0\\V_0\end{bmatrix}. \tag{C.15}$$

Again the X-independent terms cancel. The solvability condition for this equation is obtained by taking the scalar product with

$$\Xi_b=\begin{bmatrix}-\eta_b & 1\end{bmatrix} \tag{C.16}$$

to eliminate the u_2, v_2 terms, leaving

$$a_b B_{XX}=b_b\left(2V_1B+B^2\right), \tag{C.17}$$

where

$$a_b=1+\alpha\xi_b+\alpha\eta_b-\eta_b\xi_b,\quad b_b=-\frac{\Upsilon_0(1+\eta_b^2)}{\Upsilon_1^2}. \tag{C.18}$$

The latter quantity is always negative. Equation (C.17) admits spatially homogeneous solutions $B=-2V_1$, or

$$\begin{bmatrix}U\\V\end{bmatrix}=\begin{bmatrix}U_0\\V_0\end{bmatrix}-\epsilon\begin{bmatrix}U_1\\V_1\end{bmatrix}+\dots, \tag{C.19}$$

corresponding to the other branch of uniform phase-locked states, A_u^-. In addition, (C.17) admits a branch of X-dependent localized states

$$B(X)=-3\Upsilon_1\sqrt{\delta}\,\mathrm{sech}^2\left\{\left(\frac{\Upsilon_1\sqrt{\delta}}{2a_b/b_b}\right)^{1/2}X\right\} \tag{C.20}$$

corresponding to

$$\begin{bmatrix}U\\V\end{bmatrix}=\begin{bmatrix}U\\V\end{bmatrix}^{+}-3\Upsilon_1\sqrt{\gamma-\gamma_b}\begin{bmatrix}\xi_b\\1\end{bmatrix}\mathrm{sech}^2\left\{(\gamma-\gamma_b)^{1/4}\left(\frac{\Upsilon_1}{2a_b/b_b}\right)^{1/2}x\right\}. \tag{C.21}$$

The two spatial eigenvalues of the $B=0$ state in (C.17) correspond to those described previously in (3.32), providing a relation between the eigenvalue analysis and asymptotic analysis:

$$\frac{\Upsilon_1 b_b}{2a_b}=\frac{|A_u(\gamma_b)|^2}{\Lambda_b^2}\frac{\rho_\beta}{\rho_\alpha^2}\sqrt{2\gamma_b}. \tag{C.22}$$

It follows that the solution (C.21) exists whenever the order one spatial eigenvalues at the saddle-node bifurcation are real, which is consistent with the assumptions made at the beginning of this appendix. In addition, (C.22) yields a useful relation between the sign of Λ_b^2 and

the signs of Υ_1 and a_b.

Appendix D. Weakly nonlinear analysis near $\gamma = \gamma_d$. In this appendix we assume that $\mu < 0$ and $z > 0$. In this case, when the bifurcation at γ_0 is supercritical, the four spatial eigenvalues of the A_u^+ branch are either all real or all imaginary between γ_0 and γ_d, and likewise between γ_b and γ_d when the bifurcation at γ_0 is subcritical. At γ_d these eigenvalues collide pairwise on either the real or the imaginary axis and form a complex quartet for $\gamma > \gamma_d$. When this collision occurs on the imaginary axis, γ_d corresponds to a bifurcation from which localized states may emerge in the form of orbits homoclinic to A_u^+ in $\gamma > \gamma_d$. To find these states we write $\gamma = \gamma_d + \epsilon^2\delta$, where $\epsilon \ll 1$ and δ is an order one quantity. With $A = U + iV$ the localized states of interest can be written

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix}^+ + \begin{bmatrix} u \\ v \end{bmatrix}, \tag{D.1}$$

where the first term corresponds to the uniform phase-locked states and the second to the x-dependent terms, which decay to zero as $x \to \pm\infty$.

Based on the scaling defined above, the uniform phase-locked states A_u^+ can be approximated by the series

$$\begin{bmatrix} U \\ V \end{bmatrix}^+ = \begin{bmatrix} U_0 \\ V_0 \end{bmatrix} + \epsilon^2 \begin{bmatrix} U_2 \\ V_2 \end{bmatrix} + \dots, \tag{D.2}$$

while the x-dependent terms can be expanded as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \epsilon \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \epsilon^2 \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} + \epsilon^3 \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} + \dots. \tag{D.3}$$

All quantities in (D.3) depend on both the short spatial scale x and the long spatial scale $X \equiv \epsilon x$. The linear operator in (3.1) takes the form $\mathcal{L} = \mathcal{L}_0 + \epsilon\mathcal{L}_1 + \epsilon^2\mathcal{L}_2$, where

$$\mathcal{L}_0 = \begin{bmatrix} \mu + \gamma_d & -\nu \\ \nu & \mu - \gamma_d \end{bmatrix} + \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{xx}, \tag{D.4}$$

$$\mathcal{L}_1 = 2 \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{Xx}, \quad \mathcal{L}_2 = \begin{bmatrix} \delta & 0 \\ 0 & -\delta \end{bmatrix} + \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{XX}, \tag{D.5}$$

while the nonlinear terms take the form $\mathcal{N} = \mathcal{N}_0 + \epsilon\mathcal{N}_1 + \epsilon^2\mathcal{N}_2 + \epsilon^3\mathcal{N}_3 + \dots$, where

$$\mathcal{N}_0 = -(U_0^2 + V_0^2) \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}, \tag{D.6}$$

$$\mathcal{N}_1 = -2(u_1U_0 + v_1V_0) \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}, \tag{D.7}$$

$$\mathcal{N}_2 = -\left\{ \begin{bmatrix} u_1 & v_1 \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + 2 \begin{bmatrix} U_0 & V_0 \end{bmatrix} \begin{bmatrix} U_2 + u_2 \\ V_2 + v_2 \end{bmatrix} \right\} \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}, \tag{D.8}$$

$$\mathcal{N}_3 = -2\left\{ \begin{bmatrix} U_0 & V_0 \end{bmatrix} \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} + \begin{bmatrix} u_1 & v_1 \end{bmatrix} \begin{bmatrix} U_2 + u_2 \\ V_2 + v_2 \end{bmatrix} \right\} \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix}. \tag{D.9}$$

At order ϵ^0 stationary solutions to (3.1) satisfy

$$\{\mathcal{L}_0+\mathcal{N}_0\}\begin{bmatrix}U_0\\V_0\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}, \tag{D.10}$$

implying that

$$\begin{bmatrix}U_0\\V_0\end{bmatrix}=\begin{bmatrix}\eta_d\\1\end{bmatrix}\Upsilon_0, \tag{D.11}$$

where

$$\eta_d=\frac{\nu-\beta|A_u^+(\gamma_d)|^2}{\mu+\gamma_d-|A_u^+(\gamma_d)|^2},\quad \Upsilon_0=\frac{|A_u^+(\gamma_d)|}{\sqrt{1+\eta_d^2}}. \tag{D.12}$$

At order ϵ we obtain

$$\{\mathcal{L}_0+\mathcal{N}_0\}\begin{bmatrix}u_1\\v_1\end{bmatrix}=-\mathcal{N}_1\begin{bmatrix}U_0\\V_0\end{bmatrix} \tag{D.13}$$

or

$$\mathcal{M}\begin{bmatrix}u_1\\v_1\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}, \tag{D.14}$$

where

$$\mathcal{M}\equiv\mathcal{L}_0+\mathcal{N}_0-2\begin{bmatrix}1&-\beta\\\beta&1\end{bmatrix}\begin{bmatrix}U_0^2&U_0V_0\\U_0V_0&V_0^2\end{bmatrix}. \tag{D.15}$$

The critical wavenumber k_d given in (3.27) is determined by the solvability condition for this equation. With this condition satisfied, the solution is

$$\begin{bmatrix}u_1\\v_1\end{bmatrix}=\begin{bmatrix}\xi_d\\1\end{bmatrix}\left\{Be^{ik_dx}+\bar{B}e^{-ik_dx}\right\}, \tag{D.16}$$

where

$$\xi_d=\frac{\nu-\alpha k_d^2-\beta|A_u^+(\gamma_d)|^2+2(U_0V_0-\beta V_0^2)}{\mu+\gamma_d-k_d^2-|A_u^+(\gamma_d)|^2-2(U_0^2-\beta U_0V_0)} \tag{D.17}$$

and $B(X)$ is a complex-valued function of X.

Proceeding to order ϵ^2, we obtain

$$\{\mathcal{L}_0+\mathcal{N}_0\}\begin{bmatrix}U_2+u_2\\V_2+v_2\end{bmatrix}=-\{\mathcal{L}_1+\mathcal{N}_1\}\begin{bmatrix}u_1\\v_1\end{bmatrix}-\{\mathcal{L}_2+\mathcal{N}_2\}\begin{bmatrix}U_0\\V_0\end{bmatrix}. \tag{D.18}$$

The X-independent terms can be written in the form

$$\mathcal{M}\begin{bmatrix}U_2\\V_2\end{bmatrix}=-\begin{bmatrix}\delta&0\\0&-\delta\end{bmatrix}\begin{bmatrix}U_0\\V_0\end{bmatrix}. \tag{D.19}$$

It follows that the solution for U_2, V_2 takes the form

$$\begin{bmatrix} U_2 \\ V_2 \end{bmatrix} = \delta \begin{bmatrix} \tau_d \\ 1 \end{bmatrix} \Upsilon_2, \tag{D.20}$$

where the constants τ_d and Υ_2 are determined by solving (D.19):

$$\begin{bmatrix} \tau_d \\ 1 \end{bmatrix} \Upsilon_2 = -\mathcal{M}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}. \tag{D.21}$$

Here $\mathcal{M}^{-1}$ refers to the inverse of $\mathcal{M}$ after setting the ∂_{xx} terms in this operator to zero. We do not give the explicit form for these constants, but they can be easily evaluated numerically for any set of parameter values.

The remaining X-dependent terms in (D.18) are

$$\mathcal{M} \begin{bmatrix} u_2 \\ v_2 \end{bmatrix} = -2 \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \partial_{xX} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} + \left\{ 2 \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} u_1^2 & u_1 v_1 \\ u_1 v_1 & v_1^2 \end{bmatrix} + (u_1^2 + v_1^2) \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \right\} \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}. \tag{D.22}$$

With the Ansatz

$$\begin{bmatrix} u_2 \\ v_2 \end{bmatrix} = s_0 \begin{bmatrix} \chi_0 \\ 1 \end{bmatrix} |B_1|^2 + s_1 \begin{bmatrix} \chi_1 \\ 1 \end{bmatrix} \left\{ i B_X e^{ik_d x} - i \bar{B}_X e^{-ik_d x} \right\} + s_2 \begin{bmatrix} \chi_2 \\ 1 \end{bmatrix} \left\{ B_1^2 e^{2ik_d x} + \bar{B}_1^2 e^{-2ik_d x} \right\} \tag{D.23}$$

the coefficients s_i and χ_i are determined by solving (D.22) at each order in $e^{nik_d x}$ $(n = 0, 1, 2)$. For the $n = 0$ terms, the x-derivatives in the $\mathcal{M}$ operator vanish. If we call the resulting operator $\mathcal{M}_0$, the $n = 0$ part of the solution is given by

$$s_0 \begin{bmatrix} \chi_0 \\ 1 \end{bmatrix} = 2\mathcal{M}_0^{-1} \left\{ 2 \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \xi_d^2 & \xi_d \\ \xi_d & 1 \end{bmatrix} + (\xi_d^2 + 1) \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \right\} \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}. \tag{D.24}$$

Similarly, for the $n = 2$ terms we define $\mathcal{M}_2$ by replacing ∂_{xx} in $\mathcal{M}$ by $-4k_d^2$. Then

$$s_2 \begin{bmatrix} \chi_2 \\ 1 \end{bmatrix} = \mathcal{M}_2^{-1} \left\{ 2 \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \xi_d^2 & \xi_d \\ \xi_d & 1 \end{bmatrix} + (\xi_d^2 + 1) \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \right\} \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}. \tag{D.25}$$

Since the operator $\mathcal{M}_1$, defined by replacing ∂_{xx} in $\mathcal{M}$ by $-k_d^2$, is singular, the two components of the $n = 1$ equation,

$$s_1 \mathcal{M}_1 \begin{bmatrix} \chi_1 \\ 1 \end{bmatrix} + 2k_d \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} \xi_d \\ 1 \end{bmatrix} = 0, \tag{D.26}$$

are linearly related, leading to the relation

$$s_1 = -2k_d \frac{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} \xi_d \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 & 0 \end{bmatrix} \mathcal{M}_1 \begin{bmatrix} \chi_1 \\ 1 \end{bmatrix}}, \tag{D.27}$$

where χ_1 is arbitrary (subject to $\chi_1 \neq \xi_d$). Without loss of generality, we choose $\chi_1 = 0$.

Finally, at order ϵ^3 we obtain

$$\{\mathcal{L}_0 + \mathcal{N}_0\} \begin{bmatrix} u_3 \\ v_3 \end{bmatrix} = -\{\mathcal{L}_1 + \mathcal{N}_1\} \begin{bmatrix} U_2 + u_2 \\ V_2 + v_2 \end{bmatrix} - \{\mathcal{L}_2 + \mathcal{N}_2\} \begin{bmatrix} u_1 \\ v_1 \end{bmatrix} - \mathcal{N}_3 \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}. \tag{D.28}$$

The solvability condition at this order is obtained by multiplying this equation by the adjoint null eigenvector of $\mathcal{M}_1$ and integrating over x. This vector is given by

$$\begin{bmatrix} \Xi_d & 1 \end{bmatrix} e^{-ik_d x}, \tag{D.29}$$

where

$$\Xi_d = -\frac{\nu - \alpha k_d^2 - \beta |A_u^+(\gamma_d)|^2 - 2(\beta U_0^2 + U_0 V_0)}{\mu + \gamma_d - k_d^2 - |A_u^+(\gamma_d)|^2 - 2(U_0^2 - \beta U_0 V_0)}. \tag{D.30}$$

The resulting equation can be written

$$a_d B_{XX} = \delta B - b_d B |B|^2, \tag{D.31}$$

where

$$a_d = -f_1/f_2, \qquad b_d = -f_3/f_2, \tag{D.32}$$

and

$$f_1 = \begin{bmatrix} \Xi_d & 1 \end{bmatrix} \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} \xi_d \\ 1 \end{bmatrix} - 2k_d s_1 \begin{bmatrix} \Xi_d & 1 \end{bmatrix} \begin{bmatrix} 1 & -\alpha \\ \alpha & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{D.33}$$

$$\begin{aligned} f_2 = \Upsilon_0 \Upsilon_2 \begin{bmatrix} \Xi_d & 1 \end{bmatrix} \Bigg\{ & -2 \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \tau_d \eta_d & \tau_d \\ \eta_d & 1 \end{bmatrix} - 2 \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \eta_d \tau_d & \eta_d \\ \tau_d & 1 \end{bmatrix} \\ & - 2(\eta_d \tau_d + 1) \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \Bigg\} \begin{bmatrix} \xi_d \\ 1 \end{bmatrix} + \begin{bmatrix} \Xi_d & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \xi_d \\ 1 \end{bmatrix}, \end{aligned} \tag{D.34}$$

$$\begin{aligned} f_3 = -2\Upsilon_0 \begin{bmatrix} \Xi_d & 1 \end{bmatrix} \Bigg\{ & \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \left\{ s_0 \begin{bmatrix} \chi_0 \\ 1 \end{bmatrix} + s_2 \begin{bmatrix} \chi_2 \\ 1 \end{bmatrix} \right\} \begin{bmatrix} \xi_d & 1 \end{bmatrix} \\ & + \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \xi_d \\ 1 \end{bmatrix} \left\{ s_0 \begin{bmatrix} \chi_0 & 1 \end{bmatrix} + s_2 \begin{bmatrix} \chi_2 & 1 \end{bmatrix} \right\} \\ & + \left\{ s_0(1 + \xi_d \chi_0) + s_2(1 + \xi_d \chi_2) \right\} \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \Bigg\} \begin{bmatrix} \eta_d \\ 1 \end{bmatrix} \\ & - 3(1 + \xi_d^2) \begin{bmatrix} \Xi_d & 1 \end{bmatrix} \begin{bmatrix} 1 & -\beta \\ \beta & -1 \end{bmatrix} \begin{bmatrix} \xi_d \\ 1 \end{bmatrix}. \end{aligned} \tag{D.35}$$

The solutions to (D.31) depend on the signs of the coefficients a_d and b_d. Regardless of the sign of a_d, there is always a solution of the form $B = \sqrt{\delta/b_d} e^{i\varphi}$, corresponding to spatially periodic states:

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix}^+ + 2 \begin{bmatrix} \xi_d \\ 1 \end{bmatrix} \sqrt{\frac{\gamma - \gamma_d}{b_d}} \cos(k_d x + \varphi). \tag{D.36}$$

When $b_d > 0$ these periodic states bifurcate subcritically, toward $\gamma > \gamma_d$; when $b_d < 0$ the bifurcation is supercritical, toward $\gamma < \gamma_d$. Equation (D.31) also has localized solutions of the form

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix}^{+} + 2 \begin{bmatrix} \xi_d \\ 1 \end{bmatrix} \sqrt{\frac{\gamma - \gamma_d}{b_d/2}} \operatorname{sech}\left\{ \sqrt{\frac{\gamma - \gamma_d}{a_d}} x \right\} \cos\left(k_d x + \varphi\right). \tag{D.37}$$

When both a_d and b_d are positive, these solutions bifurcate subcritically, toward $\gamma > \gamma_d$; when these coefficients are both negative, (D.31) suggests that the corresponding solutions bifurcate supercritically, toward $\gamma < \gamma_d$. However, based on the spatial eigenvalue analysis, we expect localized solutions only for $\gamma > \gamma_d$, and indeed no localized solutions to (2.2) in $\gamma < \gamma_d$ have been found. Thus the localized solutions of (D.31) correspond to solutions of (2.2) only when a_d and b_d are both positive.

Acknowledgments. A. Yochelis thanks Vered Rom-Kedar and the Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, where some of this work was carried out.

REFERENCES

[1] H. Arbell and J. Fineberg, *Temporally harmonic oscillons in Newtonian fluids*, Phys. Rev. Lett., 85 (2000), pp. 756–759.
[2] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.
[3] O. Batiste, E. Knobloch, A. Alonso, and I. Mercader, *Spatially localized binary-fluid convection*, J. Fluid Mech., 560 (2006), pp. 149–158.
[4] D. Blair, I. S. Aranson, G. W. Crabtree, V. Vinokur, L. S. Tsimring, and C. Josserand, *Patterns in thin vibrated granular layers: Interfaces, hexagons, and superoscillons*, Phys. Rev. E, 61 (2000), pp. 5600–5610.
[5] V. A. Brazhnyi, V. V. Konotop, S. Coulibaly, and M. Taki, *Field patterns in periodically modulated optical parametric amplifiers and oscillators*, Chaos, 17 (2007), 037111.
[6] S. E. Brown, G. Mozurkewich, and G. Grüner, *Subharmonic Shapiro steps and devil's-staircase behavior in driven charge-density-wave systems*, Phys. Rev. Lett., 52 (1984), pp. 2277–2280.
[7] J. Burke and E. Knobloch, *Localized states in the generalized Swift-Hohenberg equation*, Phys. Rev. E (3), 73 (2006), 056211.
[8] J. Burke and E. Knobloch, *Snakes and ladders: Localized states in the Swift-Hohenberg equation*, Phys. Lett. A, 360 (2007), pp. 681–688.
[9] A. R. Champneys, *Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics*, Phys. D, 112 (1998), pp. 158–186.
[10] P. Coullet and K. Emilsson, *Strong resonances of spatially distributed oscillators: A laboratory to study patterns and defects*, Phys. D, 61 (1992), pp. 119–131.
[11] P. Coullet, J. Lega, B. Houchmanzadeh, and J. Lajzerowicz, *Breaking chirality in nonequilibrium systems*, Phys. Rev. Lett., 65 (1990), pp. 1352–1355.
[12] P. Coullet, C. Riera, and C. Tresser, *A new approach to data storage using localized structures*, Chaos, 14 (2004), pp. 193–198.
[13] E. Doedel, R. Paffenroth, A. Champneys, T. Fairgrieve, Y. Kuznetsov, B. Sandstede, and X. Wang, *AUTO2000: Continuation and Bifurcation Software for Ordinary Differential Equations (with HOMCONT)*, http://indy.cs.concordia.ca/auto/.
[14] M. Eiswirth and G. Ertl, *Forced oscillations of a self-oscillating surface reaction*, Phys. Rev. Lett., 60 (1988), pp. 1526–1529.
[15] C. Elphick, G. Iooss, and E. Tirapegui, *Normal form reduction for time-periodically driven differential equations*, Phys. Lett. A, 120 (1987), pp. 459–463.

[16] M. Faraday, *On a peculiar class of acoustical figures*, Philos. Trans. R. Soc. London Ser. A, 52 (1831), pp. 299–340.
[17] S. Fauve, *Pattern forming instabilities*, in Hydrodynamics and Nonlinear Instabilities, C. Godrèche and P. Manneville, eds., Cambridge University Press, Cambridge, UK, 1998, pp. 387–492.
[18] T. Frisch, S. Rica, P. Coullet, and J. M. Gilli, *Spiral waves in liquid crystal*, Phys. Rev. Lett., 72 (1994), pp. 1471–1474.
[19] J.-M. Gambaudo, *Perturbation of a Hopf bifurcation by an external time-periodic forcing*, J. Differential Equations, 57 (1985), pp. 172–199.
[20] L. Glass, *Synchronization and rhythmic processes in physiology*, Nature, 410 (2001), pp. 277–284.
[21] D. Gomila, P. Colet, G. L. Oppo, and M. San Miguel, *Stable droplets and growth laws close to the modulational instability of a domain wall*, Phys. Rev. Lett., 87 (2001), 194101.
[22] D. Gomila, P. Colet, M. San Miguel, and G. L. Oppo, *Domain wall dynamics: Growth laws, localized structures and stable droplets*, Eur. Phys. J. Special Topics, 146 (2007), pp. 71–86.
[23] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[24] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. Ahmer Wadee, C. J. Budd, and G. J. Lord, *Cellular buckling in long structures*, Nonlinear Dynam., 21 (2000), pp. 3–29.
[25] R. Imbihl and G. Ertl, *Oscillatory kinetics in heterogeneous catalysis*, Chem. Rev., 95 (1995), pp. 697–733.
[26] G. Iooss and M. C. Pérouème, *Perturbed homoclinic solutions in reversible* 1:1 *resonance vector fields*, J. Differential Equations, 102 (1993), pp. 62–88.
[27] E. Knobloch, *Spatially localized structures in dissipative systems: Open problems*, Nonlinearity, 21 (2008), pp. T45–T60.
[28] J. Knobloch and T. Wagenknecht, *Homoclinic snaking near a heteroclinic cycle in reversible systems*, Phys. D, 206 (2005), pp. 82–93.
[29] J. Knobloch and T. Wagenknecht, *Snaking of multiple homoclinic orbits in reversible systems*, preprint, University of Manchester, Manchester, UK, 2007.
[30] K. Kolossovski, A. R. Champneys, A. V. Buryak, and R. A. Sammut, *Multi-pulse embedded solitons as bound states of quasi-solitons*, Phys. D, 171 (2002), pp. 153–177.
[31] G. Kozyreff and S. J. Chapman, *Asymptotics of large bound states of localized structures*, Phys. Rev. Lett., 97 (2006), 044502.
[32] A. L. Lin, M. Bertram, K. Martinez, H. L. Swinney, A. Ardelea, and G. F. Carey, *Resonant phase patterns in a reaction-diffusion system*, Phys. Rev. Lett., 84 (2000), pp. 4240–4243.
[33] A. L. Lin, A. Hagberg, A. Ardelea, M. Bertram, H. L. Swinney, and E. Meron, *Four-phase patterns in forced oscillatory systems*, Phys. Rev. E, 62 (2000), pp. 3790–3798.
[34] A. L. Lin, A. Hagberg, E. Meron, and H. L. Swinney, *Resonance tongues and patterns in periodically forced reaction-diffusion systems*, Phys. Rev. E (3), 69 (2004), 066217.
[35] O. Lioubashevski, Y. Hamiel, A. Agnon, Z. Reches, and J. Fineberg, *Oscillons and propagating solitary waves in a vertically vibrated colloidal suspension*, Phys. Rev. Lett., 83 (1999), pp. 3190–3193.
[36] E. Lombardi, *Oscillatory Integrals and Phenomena Beyond All Algebraic Orders, with Applications to Homoclinic Orbits in Reversible Systems*, Lecture Notes in Math. 1741, Springer-Verlag, New York, 2000.
[37] S. Longhi, *Hydrodynamic equation model for degenerate optical parametric oscillators*, J. Modern Opt., 43 (1996), pp. 1089–1094.
[38] S. Longhi, *Spatial solitary waves in nondegenerate optical parametric oscillators near an inverted bifurcation*, Opt. Commun., 149 (1998), pp. 335–340.
[39] B. Marts, A. Hagberg, E. Meron, and A. L. Lin, *Bloch-front turbulence in a periodically forced Belousov-Zhabotinsky reaction*, Phys. Rev. Lett., 93 (2004), 108305.
[40] B. Marts, K. Martinez, and A. L. Lin, *Front dynamics in an oscillatory bistable Belousov-Zhabotinsky chemical reaction*, Phys. Rev. E, 70 (2004), 056223.
[41] F. Melo, P. Umbanhowar, and H. L. Swinney, *Transition to parametric wave patterns in a vertically oscillated granular layer*, Phys. Rev. Lett., 72 (1994), pp. 172–175.
[42] F. Melo, P. B. Umbanhowar, and H. L. Swinney, *Hexagons, kinks, and disorder in oscillated granular layers*, Phys. Rev. Lett., 75 (1995), pp. 3838–3841.

[43] H.-K. Park and M. Bär, *Spiral destabilization by resonant forcing*, Europhys. Lett., 65 (2004), pp. 837–843.
[44] V. Petrov, Q. Ouyang, and H. L. Swinney, *Resonant pattern formation in a chemical system*, Nature, 388 (1997), pp. 655–657.
[45] Y. Pomeau, *Front motion, metastability and subcritical bifurcations in hydrodynamics*, Phys. D, 23 (1986), pp. 3–11.
[46] H. Riecke, J. D. Crawford, and E. Knobloch, *Time-modulated oscillatory convection*, Phys. Rev. Lett., 61 (1988), pp. 1942–1945.
[47] B. Sandstede and A. Scheel, *Gluing unstable fronts and backs together can produce stable pulses*, Nonlinearity, 13 (2000), pp. 1465–1482.
[48] A. Spina, J. Toomre, and E. Knobloch, *Confined states in large-aspect-ratio thermosolutal convection*, Phys. Rev. E (3), 57 (1998), pp. 524–545.
[49] K. Staliunas, *Transverse pattern formation in optical parametric oscillators*, J. Modern Opt., 42 (1995), pp. 1261–1269.
[50] L. Stenflo and M. Y. Yu, *Origin of oscillons*, Nature, 384 (1996), p. 224.
[51] L. S. Tsimring and I. S. Aranson, *Localized and cellular patterns in a vibrated granular layer*, Phys. Rev. Lett., 79 (1997), pp. 213–216.
[52] P. B. Umbanhowar, F. Melo, and H. L. Swinney, *Localized excitations in a vertically vibrated granular layer*, Nature, 382 (1996), pp. 793–796.
[53] V. K. Vanag, A. M. Zhabotinsky, and I. R. Epstein, *Oscillatory clusters in the periodically illuminated, spatially extended Belousov-Zhabotinsky reaction*, Phys. Rev. Lett., 86 (2001), pp. 552–555.
[54] M. Vaupel, A. Maître, and C. Fabre, *Observation of pattern formation in optical parametric oscillators*, Phys. Rev. Lett., 83 (1999), pp. 5278–5281.
[55] D. M. Winterbottom, S. M. Cox, and P. C. Matthews, *Pattern formation in a model of a vibrated granular layer*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 63–78.
[56] P. D. Woods and A. R. Champneys, *Heteroclinic tangles and homoclinic snaking in the unfolding of a degenerate reversible Hamiltonian-Hopf bifurcation*, Phys. D, 129 (1999), pp. 147–170.
[57] A. Yochelis, J. Burke, and E. Knobloch, *Reciprocal oscillons and nonmonotonic fronts in forced nonequilibrium systems*, Phys. Rev. Lett., 97 (2006), 254501.
[58] A. Yochelis, C. Elphick, A. Hagberg, and E. Meron, *Two-phase resonant patterns in forced oscillatory systems: Boundaries, mechanisms and forms*, Phys. D, 199 (2004), pp. 201–222.
[59] A. Yochelis, C. Elphick, A. Hagberg, and E. Meron, *Frequency locking in extended systems: The impact of a Turing mode*, Europhys. Lett., 69 (2005), pp. 170–176.
[60] A. Yochelis, A. Hagberg, E. Meron, A. L. Lin, and H. L. Swinney, *Development of standing-wave labyrinthine patterns*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 236–247.
[61] M. A. Zaks, A. Podolny, A. A. Nepomnyashchy, and A. A. Golovin, *Periodic stationary patterns governed by a convective Cahn–Hilliard equation*, SIAM J. Appl. Math., 66 (2006), pp. 700–720.
[62] M. G. Zimmermann, S. O. Firle, M. A. Natiello, M. Hildebrand, M. Eiswirth, M. Bär, A. K. Bangia, and I. G. Kevrekidis, *Pulse bifurcation and transition to spatiotemporal chaos in an excitable reaction-diffusion model*, Phys. D, 110 (1997), pp. 92–104.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 712–754
© 2008 Society for Industrial and Applied Mathematics

Tangency Bifurcations of Global Poincaré Maps*

Clare M. Lee†, Pieter J. Collins‡, Bernd Krauskopf†, and Hinke M. Osinga†

Abstract. One tool to analyze the qualitative behavior of a periodic orbit of a vector field in $\mathbb{R}^n$ is to consider the Poincaré return map to an $(n-1)$-dimensional section. The image under the Poincaré map of a point on this section that lies near the periodic orbit is obtained by following the flow of the vector field until the next (local) intersection. It is well known that the Poincaré map defined on a section transverse to a periodic orbit is a diffeomorphism locally near the periodic orbit. However, in practice one often considers the Poincaré map not only locally but also on a much larger global and typically unbounded section. Generically, there are then points where the flow is tangent to the section, and these give rise to discontinuities of the Poincaré map. In fact, the orbits of some points may not even return to the section, in which case the Poincaré map is not defined at all. In this paper we study tangency bifurcations of invariant manifolds of Poincaré maps on global sections of vector fields in $\mathbb{R}^2$ and $\mathbb{R}^3$. At such a bifurcation the manifold becomes tangent to the section, which results in a qualitative change of the Poincaré map while the underlying flow itself does not undergo a bifurcation. Using tools from singularity theory, we present normal forms of the codimension-one tangency bifurcations in the neighborhood of the respective tangency point. The study of these bifurcations is motivated by and illustrated with the examples of the (unforced) Van der Pol oscillator and a system modeling a semiconductor laser with optical injection. Finally, we present a framework for the generalization of our normal-form results to arbitrary dimension and codimension.

Key words. Poincaré map, quadratic and cubic tangency, flowbox, normal forms, singularity theory

AMS subject classifications. 37C10, 37G25, 58K50

DOI. 10.1137/07069972X

1. Introduction.

In 1892 Henri Poincaré introduced the idea of a return map of a vector field—today generally referred to as a *Poincaré map*—while he was studying a restriction of the three-body problem [35, 36]. His aim was to find the motion of three bodies (one having negligible mass compared with the other two—for example, the Sun, Earth, and Moon) given only their initial positions, velocities, and masses. His work was fundamental for both the local and global analysis of nonlinear dynamical systems. In particular, he studied periodic orbits and their stability and introduced the first return map to a given local section as a new tool. In this setting the section in phase space is chosen transverse to the periodic orbit, and one considers the map that is defined locally on the section by following the flow until it

*Received by the editors August 9, 2007; accepted for publication (in revised form) by J. Guckenheimer March 19, 2008; published electronically July 16, 2008.
http://www.siam.org/journals/siads/7-3/69972.html

†Centre for Applied Nonlinear Mathematics, Department of Engineering Mathematics, University of Bristol, Queen's Building, Bristol BS8 1TR, UK (Clare.Lee@bristol.ac.uk, b.krauskopf@bristol.ac.uk, h.m.osinga@bristol.ac.uk). The last author was supported by an Advanced Research Fellowship grant from the Engineering and Physical Sciences Research Council (EPSRC).

‡Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands (pieter.collins@cwi.nl). This author was supported by a Vidi grant of the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

returns back to the section. The periodic orbit of the system corresponds to a fixed point of this Poincaré map. Since the Poincaré map is a diffeomorphism (a smooth map with a smooth inverse) in a neighborhood of this fixed point, the existence and stability analysis of periodic orbits in the full phase space reduces to the study of fixed points of local diffeomorphisms. This fact is used in the bifurcation analysis of periodic orbits in standard textbooks such as [17, 27, 39].

Today, the Poincaré map is a much-used tool in theoretical studies and in the evaluation of experiments alike. Generally speaking, one studies attractors and other invariant sets that one obtains by considering only the intersection points of the flow with a prespecified section. One way of obtaining such a representation is to plot measured quantities (for example, position, velocity, voltage, or current) whenever a designated quantity has a particular value, for example, when it crosses its average. Another common choice is to consider a section in the space of the first derivative of the flow and to plot the position of successive maxima/minima of the system; such maps can be seen, for example, in [1, 37]. In this way, one can illustrate classic transitions, such as the period-doubling route to chaos or the break-up of an invariant torus.

The above discussion already shows that in practice one typically chooses a suitable and typically unbounded section—we speak of a global section in this context. A common choice of global section is a (hyper)plane defined by one of the variables having a fixed value. The key point is that the first return map to the global section is not only defined locally with respect to an invariant set, which is why we refer to it as a *global Poincaré map*.

A special case is that of a stroboscopic map of a periodically forced system with angular frequency ω; well-known examples (see, for example, [17, 39]) include the forced damped pendulum and the forced Van der Pol and the forced Duffing oscillators. A periodically forced system can be written as an autonomous vector field by considering time t as another variable, which is, hence, periodic with period $2\pi/\omega$. The Poincaré section is then taken at $t = 2\pi/\omega$ so that the Poincaré map records the variables at regular time intervals. In a mechanical experiment, this can be achieved by stroboscopic illumination with the forcing frequency; hence we have the name stroboscopic map; see, for example, [6, 21]. Since the section is effectively taken in time and not in space, each orbit intersects the global section defined by $t = 2\pi/\omega$ and does so transversely. This means that the stroboscopic map is a well-defined global diffeomorphism on the entire section.

By contrast, for a vector field that is not periodically forced one cannot find a global section on which the Poincaré map is a global diffeomorphism. Given a general autonomous vector field and any global section Σ, there exists a nonempty set of codimension one—which we call the *tangency locus*—where the vector field is tangent to Σ; that is, the flow is not transverse to the section. Furthermore, there may be orbits that do not return to the section in forward or backward time. These two obstructions were already known to Birkhoff [3], who considered the problem of finding a Poincaré map in the context of Hamiltonian systems. His goal was to find a so-called complete section that is intersected by all trajectories so that the Poincaré map gives information about the entire dynamics. Birkhoff's result is that the Poincaré map is well defined, smooth, and complete if the section is such that the tangency locus is invariant under the first return; one also speaks of a Birkhoff section [11]. In this case it is sufficient to consider the Poincaré map on a compact region that is bounded by the tangency locus. For a

two-degree-of-freedom Hamiltonian system one obtains an area preserving map of the plane by means of restricting to a fixed-energy surface. However, for arbitrary Hamiltonian systems the condition that the tangency locus is invariant under the first return is not necessarily satisfied. To deal with this more general Hamiltonian situation, Dullin and Wittek [11] generalized Birkhoff's result by constructing what they call a W-section, which guarantees a weaker form of completeness in the sense that orbits either return to the W-section in finite time or have a limit in the W-section as time goes to infinity. Analyzing the properties of the associated Poincaré map requires the study of geometric properties of the flow in phase space in relation to the energy surface [4].

We consider here the properties of the Poincaré map on a global section of a general autonomous vector field, by which we mean that it does not have any special properties such as a Hamiltonian structure or preserved symmetries. Typically, there are regions of the section where the Poincaré first return map can be defined locally as a diffeomorphism by considering the kth return map to the global section for a suitable fixed k. Such regions are bounded by the tangency locus. Namely, the kth return map is discontinuous across the tangency locus. This fact was used in [22] to explain the emergence of an increasing number of discontinuities of the one-dimensional map approximation associated with a chaotic attractor, such as that of the Rössler system. However, when one allows the number of intersections k with the section to vary across the tangency locus, then the Poincaré map can be extended as a continuous map to an adjacent region. How the number k must be changed to ensure continuity of the Poincaré map can be determined from the condition that the integration time be continuous; see also [11]. This idea was used in [10] to compute one-dimensional invariant manifolds of the global Poincaré map across the tangency locus. By continuing orbit segments (with the associated integration time) as two-point boundary value problems it is possible to compute one-dimensional invariant manifolds of the global Poincaré map across the tangency locus without the need for manually changing the number k of returns to the section; see [13] for details and examples.

The specific topic addressed in this paper is the characterization of bifurcations of a global Poincaré map that do not correspond to bifurcations of the underlying flow. Such topological changes of the Poincaré map can be brought about either by changing a system parameter so that an invariant object changes in such a way that its intersection with the section is affected, or equivalently by changing the position of the section in the flow. Indeed, one needs to take some care to avoid drawing wrong conclusions from topological changes of phase portraits in a given section. A concrete example is the appearance of extra branches of intersections with a fixed section of a one-parameter family of periodic orbits. This may simply be due to the periodic orbit changing shape in the full phase space, which does not correspond to a bifurcation of the flow. The emergence of extra branches in a given section typically happens, for example, when the orbit approaches a saddle-focus equilibrium [16]. Another example is the intersection of an invariant torus with a section, which can take different forms, as is discussed in detail in section 4.

The invariant set in the Poincaré section may be even more complicated in applications. An example is the study by Peikert and Sadlo [32, 33] of one-dimensional invariant manifolds in a two-dimensional section through a vortex ring associated with a river power plant. The authors refer to "seemingly ring-shaped lobes [as] an artifact of the slice plane which does

not follow well the curved center line of the structure" [33, sect. 5.2]. Indeed, such "lobes" arise due to the way the section intersects a corresponding two-dimensional manifold, and this depends on the exact position of the section. In particular, the study of changes of the invariant set with the position of the section is important for the interpretation of experimental measurements, such as two-dimensional sections through a vortex structure by means of a laser sheet [14].

We consider here the bifurcations of a smooth invariant manifold of dimension ℓ as it interacts with a global section of a generic vector field with an n-dimensional phase space. We call these bifurcations *tangency bifurcations*, because they are generated by orbits that are tangent to the section at the bifurcation point. We analyze the tangency bifurcations locally, which means that there are no equilibria in the section. Therefore, we may consider a flowbox in a suitable neighborhood of the bifurcation point, which is simply a domain in phase space that is bounded by orbit segments and transverse codimension-one in- and out-sets. We use the flowbox theorem [31] to "straighten out" the flow by mapping it to the standard flowbox with parallel flow. Subsequently, we apply coordinate changes to map the section to a standard form as motivated by singularity theory; the key is to show that the coordinate changes can be chosen to preserve the flow lines in the standard flowbox. As a result, we obtain normal forms for the unfoldings of tangency bifurcations in terms of parameterized families of curved sections in the standard flowbox. Specifically, we treat all tangency bifurcations of codimension one for $n \leq 3$ and then discuss how the notion of a tangency bifurcation can be generalized for arbitrary n and ℓ.

Singularity theory has been used to classify special classes of dynamical systems, including impacting systems [9], piecewise-smooth systems [41], noninvertible maps (endomorphisms) of the plane [12, 18, 26], flows on manifolds with boundary [34, 38, 40], and diffeomorphisms with curves [5]. As is the case for any specific class of systems, singularity theory suggests normal forms, but it needs to be adapted to the situation at hand. In the case of a tangency bifurcation of a global Poincaré map, as studied here, this means the required coordinate transformation needs to map the section and the manifold to suitable normal forms, while also mapping the flow to the standard flow in the flowbox. In the absence of an invariant manifold (or when it is comprised of a single trajectory) we can make use of work by Sotomayor and Teixeira, who study generic flows on a manifold with a codimension-one boundary, namely, two-dimensional flows in [40] and three-dimensional flows in [38]. Sotomayor and Teixeira work in a completely general setting and investigate what conditions on the vector field and the section are needed for structural stability of the interaction between the flow and the boundary. We remark that by considering different ways of "completing" the flow on the other side of the boundary or Poincaré section, the results of Sotomayor and Teixeira and those presented here can also be applied to piecewise-smooth systems (where one defines a second flow as one wishes) [41] and to reversible systems (where the other flow is induced by an involution) [42]. Because in our setup the global section plays the role of their boundary, we can use the genericity results in [38, 40] for the context of tangency bifurcation of a global Poincaré map.

There are a number of ways in which our work differs from that of Sotomayor and Teixeira. First, they do not consider the flow on the "outside" of the boundary, while we are interested in the flow on both sides of the section. More importantly, we consider bifurcations of a "local piece" of an invariant one- or two-dimensional manifold in a local flowbox around a

bifurcation point. This piece of manifold could be part of a periodic orbit, invariant torus, or a stable or unstable manifold of a saddle object. While the tangency bifurcation remains locally the same, its global manifestation depends on the type of invariant manifold one is considering. For example, Sotomayor and Teixeira consider as separate the two cases that a periodic orbit or a one-dimensional separatrix has a tangency with the boundary, which we interpret as manifestations of the same interaction with the global section. Finally, two-dimensional invariant manifolds were not considered in [38]. By contrast, we present unfoldings of all tangency bifurcations of a two-dimensional invariant manifold with a two-dimensional Poincaré section. To this end, we first map the invariant manifold to a horizontal plane in the standard flowbox so that the section is deformed to a smooth surface. Our proofs then detail how to construct suitable coordinate transformations in the standard flowbox, where we distinguish between the direction of the flow and directions perpendicular to the flow. As a result, we obtain a normal form of the respective tangency bifurcation, which is unfolded by moving a smooth surface (that is, the section) up and down. We remark that, if one disregards the special direction given by the flow, tangencies between two smooth surfaces in $\mathbb{R}^3$ (or manifolds of codimension one, more generally) can be phrased in the language of divergent diagrams of smooth maps [29]. Here the manifold and the section are viewed as level sets of two smooth maps from $\mathbb{R}^3$ to $\mathbb{R}$, and singularity theory applied to this context yields normal forms as presented here. We stress again that the normal-form results obtained in [29] are not immediately applicable to tangency bifurcation of Poincaré maps because one must also show that the required coordinate changes can be chosen to respect the flow in the flowbox. Furthermore, the fact that the invariant manifold actually consists of trajectories is important: a tangency along a trajectory has different consequences for the Poincaré map than a tangency in the parameter direction.

In contrast to the mentioned theoretical studies, we not only present the unfoldings of the tangency bifurcations abstractly in a flowbox but also illustrate them with concrete examples of vector fields arising in applications. Specifically, we use the two-dimensional unforced Van der Pol oscillator to explain why a discontinuity arises in a kth return map and how a quadratic tangency bifurcation manifests itself for the classic case of a periodic orbit. Tangency bifurcations of two-dimensional invariant manifolds are illustrated with the example of a three-dimensional vector field model of a semiconductor laser with optical injection [44]. Specifically, we show how quadratic and cubic tangencies occur naturally in the intersection of a family of invariant tori with a planar section. Finally, we present a geometric model that explains how the cubic tangency as observed in a planar section for the injection laser can be transformed into the normal form in a flowbox.

This paper is organized as follows. In section 2 we discuss the motivation behind this paper and outline the problems that arise from globalizing a local Poincaré map; here we also provide formal definitions. In section 3 we consider the quadratic tangency bifurcation with a global section of a two-dimensional flow; this case is illustrated with the Van der Pol oscillator. Section 4 then deals with quadratic tangencies of invariant manifolds with a two-dimensional section of a three-dimensional flow; a model of a semiconductor laser with optical injection serves as a concrete example. In section 5 we consider the case of a cubic tangency bifurcation of a three-dimensional flow; what this bifurcation looks like in practice is again illustrated with the example of a semiconductor laser with optical injection. Section 6 discusses the

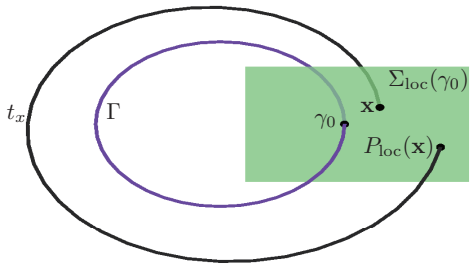


Figure 1. *A periodic orbit Γ intersecting a local section $\Sigma_{\rm loc}(\gamma_0)$ at the point γ_0. The local Poincaré map $P_{\rm loc}$ takes $\mathbf{x}$ to the next local intersection $P_{\rm loc}(\mathbf{x})$ with $\Sigma_{\rm loc}(\gamma_0)$.*

general case of a tangency bifurcation of an ℓ-dimensional manifold with a global section of an n-dimensional flow. Finally, in section 7 we draw some conclusions and discuss directions for future research.

2. Background and motivation.

Many readers will be familiar with the concept of a Poincaré map defined on a local section transverse to a periodic orbit. Consider a vector field

$$\dot{\mathbf{x}} = f(\mathbf{x}, \lambda), \quad \mathbf{x} \in X, \quad \lambda \in \mathbb{R}^m, \tag{2.1}$$

where X is the phase space, λ is a (vector-valued) parameter, and $f : X \to X$ is sufficiently smooth. For the purposes of this paper $X = \mathbb{R}^n$, where we mostly consider the cases $n = 2$ or $n = 3$. The flow associated with (2.1) is denoted by Φ, so that the orbit or trajectory through x is defined as

$$\mathcal{O}(x) = \{\Phi^t(x) \mid t \in \mathbb{R}\}. \tag{2.2}$$

We assume now that (2.1) has a periodic orbit Γ for some value of λ. Note that generically Γ is part of a (λ-dependent) family, but for the moment we consider the parameter λ as fixed. To obtain the standard definition of a local Poincaré map $P_{\rm loc}$, one chooses an $(n-1)$-dimensional submanifold Σ that intersects Γ transversely at an intersection point γ_0. The Poincaré map is then defined in some neighborhood $\Sigma_{\rm loc}(\gamma_0)$ of γ_0; see, for example, [27, 31]. A point $\mathbf{x} \in \Sigma_{\rm loc}(\gamma_0)$ is taken by the flow Φ to the next intersection of the forward trajectory $\{\Phi^t(\mathbf{x}) \mid \forall t \geq 0\}$ with $\Sigma_{\rm loc}(\gamma_0)$; see also Figure 1. That is,

$$\begin{aligned} P_{\rm loc} : \Sigma_{\rm loc}(\gamma_0) &\to \Sigma'_{\rm loc}(\gamma_0), \\ \mathbf{x} &\mapsto P_{\rm loc}(\mathbf{x}) := \Phi^{t_x}(\mathbf{x}), \end{aligned} \tag{2.3}$$

where $t_{\mathbf{x}} = \min\{t > 0 \mid \Phi^t(\mathbf{x}) \in \Sigma'_{\rm loc}(\gamma_0)\}$ is the time to the next local intersection and $\Sigma'_{\rm loc}(\gamma_0)$ is a neighborhood of $\Sigma_{\rm loc}(\gamma_0)$ such that $\Sigma_{\rm loc}(\gamma_0) \subseteq \Sigma'_{\rm loc}(\gamma_0)$. Note that γ_0 is a fixed point of $P_{\rm loc}$ and that $t_{\mathbf{x}}$ is close to the period of the periodic orbit Γ. Since the section Σ is chosen transverse to Γ at γ_0, it is always possible to choose $\Sigma_{\rm loc}(\gamma_0)$ such that $P_{\rm loc}$ is a well-defined diffeomorphism on $\Sigma_{\rm loc}(\gamma_0)$. Locally near γ_0 the dynamics of the vector field f are equivalent to the dynamics of $P_{\rm loc}$ on $\Sigma_{\rm loc}(\gamma_0)$ so that the local Poincaré map allows the study of basic bifurcations of the periodic orbit Γ; see, for example, [17, 27, 39].

In many applications, on the other hand, one is interested in more general invariant sets, including invariant tori and chaotic attractors. Therefore, one often considers some suitably chosen "large" and generally unbounded *global section* in phase space. For the purpose of this paper, we call Σ a global section if it is the image of a smooth embedding

$$F : \mathbb{R}^{n-1} \to \mathbb{R}^n, \tag{2.4}$$

where we assume that $F(\mathbb{R}^{n-1})$ divides the phase space $\mathbb{R}^n$ into two disjoint parts, that is to say, $|F(x)| \to \infty$ as $|x| \to \infty$. In other words, a global section $\Sigma = F(\mathbb{R}^{n-1})$ is a smooth manifold of dimension $n-1$ that is unbounded in all directions. Specifically, Σ is the unbounded image of the real line for $n = 2$ and of a two-dimensional plane for $n = 3$. It is common in applications to choose Σ simply as an $(n-1)$-dimensional hyperplane.

The *global Poincaré map* on the global section Σ is then defined as in (2.3), except that we now consider P on the entire global section Σ, that is,

$$\begin{aligned} P &: \Sigma \to \Sigma, \\ \mathbf{x} &\mapsto P(\mathbf{x}) := \Phi^{t_{\mathbf{x}}}(\mathbf{x}), \end{aligned} \tag{2.5}$$

where $t_{\mathbf{x}}$ is the time to first intersection with Σ; note that P is well defined at $\mathbf{x}$ if $0 \leq \inf\{t > 0 \mid \Phi^t(\mathbf{x}) \in \Sigma\} < \infty$.

As can already be inferred from Figure 1, a global section Σ typically has $k \geq 2$ intersections with a periodic orbit Γ. (Note that generically there is an even number of points in $\Gamma \cap \Sigma$.) In other words, the local Poincaré map P_{loc} defined on $\Sigma_{\text{loc}}(\gamma_0)$ coincides with the restriction to $\Sigma_{\text{loc}}(\gamma_0)$ of the kth iterate of the global Poincaré map as defined by (2.5). Indeed, P^k is a local diffeomorphism near any of the (transverse) intersection points of Γ with Σ.

It turns out that the global Poincaré map as defined by (2.5) is a diffeomorphism on the whole of Σ that describes the entire dynamics on $\mathbb{R}^n$ only under rather severe conditions (see, for example, [3, 30]), namely, only when

1. the flow Φ of (2.1) is transverse to Σ at every point $\mathbf{x} \in \Sigma$;
2. the forward orbit and the backward orbit through every point $\mathbf{x} \in \Sigma$ both have another intersection with Σ; and
3. every orbit of (2.1) intersects Σ.

In the literature Σ is often referred to as a global Poincaré section when these conditions are satisfied; see, for example, [45]. This is not to be confused with our notion of a global (i.e., nonlocal) section as defined in (2.4). Indeed, the above conditions are typically not satisfied for a global section in our sense.

Theorem 2.1. *For a generic vector field it is not possible to find a global Poincaré section Σ to which all points return and that is everywhere transverse to the flow, unless the phase space is a fibration over the circle.*

Theorem 2.1 is proved in [45] in two basic steps; see also [3, 7, 11, 30]. The first step is to show that if a global Poincaré section with the above properties exists, then the phase space X is topologically equivalent to the suspension manifold $[0,1] \times \Sigma/\!\sim$. Here, the quotient is taken with respect to the equivalence relation $\sim$, where $(0, x_1) \sim (1, x_2)$ for $x_1, x_2 \in \Sigma$ if $x_2 = P(x_1)$. The second step of the proof is to show that if Σ is a smooth manifold on which P is a diffeomorphism, then $[0,1] \times \Sigma/\!\sim$ can be written as a fibration over $\mathbb{S}^1$.

An immediate consequence of Theorem 2.1 is the following. Consider a generic vector field on $\mathbb{R}^n$. Then the vector field is equivalent to a periodically forced system if and only if one can find a global section $\Sigma = F(\mathbb{R}^{n-1})$ on which the Poincaré map is a diffeomorphism. In other words, diffeomorphisms on $\mathbb{R}^{n-1}$ correspond to periodically forced vector fields on $\mathbb{R}^n$ (written in autonomous form where time t is one of the axes). Indeed, periodically forced vector fields form an important subclass with well-known examples such as the forced damped pendulum, the forced Van der Pol equations, and the forced Duffing oscillator [17, 39]. The global Poincaré map in the sense discussed here is given as the stroboscopic map, that is, as the time-$2\pi/\omega$ map. The fact that the stroboscopic map is a global diffeomorphism is a particularly nice property of periodically forced vector fields. However, as was discussed above, this property is very special.

For a general (that is, not a periodically forced) vector field, there are points where the vector field is tangent to the global section Σ. This can already be deduced from the case of a section through a periodic orbit Γ. If one considers two consecutive (transverse) intersection points γ_0 and γ_1 of Γ with Σ, then the flow points in opposite directions (with respect to Σ) near γ_0 and γ_1, respectively. By continuity of the vector field, we must have at least one point on Σ where the vector field is tangent to the section. We define the *tangency locus* C as

$$C := \{\mathbf{x} \in \Sigma \mid f(\mathbf{x}) \cdot \vec{n}_\Sigma(\mathbf{x}) = 0\}, \tag{2.6}$$

where $\vec{n}_\Sigma(\mathbf{x})$ is the unit normal to Σ at the point $\mathbf{x}$. Note that if Σ is a hyperplane, the normal $\vec{n}_\Sigma(\mathbf{x})$ does not depend on $\mathbf{x}$. For the remainder of this paper we assume that $C \neq \emptyset$. It follows from the implicit function theorem [43] that C consists of smooth codimension-one submanifolds of Σ, provided that 0 is a regular point of $f(\mathbf{x}) \cdot \vec{n}_\Sigma(\mathbf{x})$. That is, for a one-dimensional section Σ ($n = 2$) the tangency locus C is generically a set of isolated points. For a two-dimensional section Σ ($n = 3$) it consists of smooth curves. Furthermore, in a two-dimensional section Σ there may be points of C where the flow has a cubic tangency (that is, a cusp singularity) with Σ; such points are generically isolated. These genericity statements follow from results by Sotomayor and Teixeira in [38, 40] on flows on two- and three-dimensional manifolds with boundary.

The importance of the tangency locus C lies in the realization that any kth-return map to Σ for any $k \geq 0$, that is, the Poincaré map P as defined by (2.5) or its kth iterate, is discontinuous across C; this was already noted by Birkhoff [3] in the context of Hamiltonian systems. The reason for this discontinuity is that the number of intersections with the section Σ changes due to the tangency; see also section 3.1. We remark that a kth-return map can be extended continuously across C, namely, by changing the number of global intersections and considering the $(k \pm 1)$st-return map in the adjoining region. The required number of global intersections is determined by the condition that the time $t_{\mathbf{x}}$ to the next (local) intersection depends continuously on the point $\mathbf{x}$ across C [10]; see also [11]. This approach of extending a Poincaré map across C by continuation of the entire orbit segment from Σ back to Σ, which includes the integration time $t_{\mathbf{x}}$, is especially useful when one wants to compute invariant manifolds of global Poincaré maps [13]. Note that a thus extended Poincaré map is only continuous across C but not smooth; see the discussion of Figure 3 in section 3.1. In other words, the existence of a tangency locus is indeed an obstacle to finding a Poincaré map that is a global diffeomorphism.

We consider here tangency bifurcations of the Poincaré map P on a global section Σ, which are characterized by the interaction of an invariant manifold M with the tangency locus $C \subset \Sigma$. The first step is to define an appropriate notion of topological equivalence.

Definition 2.2. *We are given two flows on two open neighborhoods U_1 and U_2 of $\mathbb{R}^n$ and $(n-1)$-dimensional smooth sections $\Sigma_1 \subset U_1$ and $\Sigma_2 \subset U_2$ with tangency loci C_1 and C_2, respectively. Suppose, further, that there are ℓ-dimensional invariant manifolds $M_1 \subset U_1$ and $M_2 \subset U_2$. We say that the flow on U_1 is* Σ-M-topologically equivalent *to that on U_2 if there exists a homeomorphism $h : U_1 \to U_2$ such that*

(E1) *h maps orbits in U_1 to orbits in U_2 and respects the direction of time;*
(E2) *h maps Σ_1 to Σ_2 and C_1 to C_2; and*
(E3) *h maps M_1 to M_2.*

Note that (E2) and (E3) ensure that $h|_{\Sigma_1}$ maps $M_1 \cap \Sigma_1$ to $M_2 \cap \Sigma_2$. For notational convenience, we refer to Σ-M-topological equivalence simply as topological equivalence in what follows. Similarly, we refer to the orbits of a flow on an open neighborhood U relative to $\Sigma, M \subset U$ simply as a phase portrait. Here we also assume that Σ divides U into two disjoint parts; this, by our definition of a global section, is satisfied for any sufficiently small U.

Following standard ideas of bifurcation theory [17, 27], we say that a phase portrait is structurally stable if any sufficiently close phase portrait is topologically equivalent. Here closeness between phase portraits is given by the C^1-topology of the underlying vector fields and the C^1-distance between the respective sections and invariant manifolds; compare with [38, 40]. Consequently, a bifurcation takes place when a phase portrait is not structurally stable, and it is of codimension $c \in \mathbb{N}$ if it is structurally stable in a c-parameter family (but not in a $(c-1)$-parameter family); one also speaks of an unfolding. Here two parameterized families of systems are topologically equivalent if the phase portraits of the two unfoldings are topologically equivalent according to Definition 2.2 and the underlying family of homeomorphisms can be chosen to depend continuously on the parameters.

Clearly bifurcations of the underlying flow in the usual sense (meaning that (E1) is violated) are also bifurcations of the Poincaré map P in the sense of Definition 2.2; these bifurcations are covered by standard bifurcation theory. However, there are new types of bifurcations in the sense of Definition 2.2, namely, those that correspond to violation of (E2) and/or (E3). Our interest here is in a class of such bifurcations—the tangency bifurcations—which are due to a qualitative change of $M \cap \Sigma$. In this paper we restrict to tangency bifurcations in $\mathbb{R}^n$ for $n \leq 3$ of codimension one, meaning that they occur structurally stably in one-parameter families. Tangency bifurcations do not involve equilibria in Σ so that we can consider the phase portrait in a flowbox near the interaction of the invariant manifold with the section. A flowbox does not contain any equilibria and is characterized by an in-set I and an out-set O transverse to the flow, with "sides" that consist of orbit segments. This is possible because a flow without equilibria is transient [34]; that is, each orbit leaves a compact connected manifold in finite positive and negative time. According to the flowbox theorem [31], the vector field in any given flowbox is conjugate (by means of a coordinate transformation that is as smooth as the vector field f) to parallel flow in the standard flowbox, which we define here for definiteness as

$$\left\{ \begin{array}{rcl} \dot{u} & = & 1, \\ \dot{v} & = & 0, \end{array} \right. \tag{2.7}$$

where $u \in [-1,1]$ and $v \in [-1,1]^{n-1}$. It follows that the standard in-set and the standard out-set are

$$I = \{u = -1\} \quad \text{and} \quad O = \{u = +1\}. \tag{2.8}$$

The basic idea behind transforming a given flowbox to the standard flowbox is that the flow lines are "straightened out." In the present setting, this means that the section Σ becomes curved. In other words, a normal form as considered here consists of a suitable family of curved sections that interact with a fixed invariant manifold, which is determined by a prespecified subset of the in-set I.

3. Two-dimensional flows.

We begin by explaining the main concepts with a concrete example in section 3.1, before presenting a general normal-form result in section 3.2.

3.1. The (unforced) Van der Pol oscillator.

As a concrete motivating example we consider the (unforced) Van der Pol oscillator [20, 39] of an RLC circuit, which is defined as the two-dimensional vector field

$$\begin{cases} \dot{x} &= y, \\ \dot{y} &= a(1-x^2)y - x, \end{cases} \tag{3.1}$$

where $a = 0.25$ is used throughout this paper. For this value of a, system (3.1) has an attracting periodic orbit Γ. As the global section we choose the horizontal line

$$\Sigma = \Sigma_s := \{(x,y) \in \mathbb{R}^2 \mid y = s\} \tag{3.2}$$

for some constant $s \in \mathbb{R}$. Figure 2 shows how the global section Σ (green) for $s = 1$ intersects the periodic orbit Γ (purple) transversely in two points, γ_0 and γ_1. Recall that a transverse intersection of Σ with Γ always leads to at least two intersection points, regardless of the choice of Σ. Near γ_0 the flow is upward through Σ, while near γ_1 it points down; at the points C the flow is tangent to Σ. Using (3.1) and (3.2), the points C at which the flow is tangent to Σ satisfy

$$\dot{y}|_{y=s} = 0 \quad \Leftrightarrow \quad \big(a(1-x^2)y - x\big)\big|_{y=s} = 0$$

so that

$$C = C(s) = \begin{cases} \{(0,0)\} & \text{if } s = 0, \\ \left\{\left(\frac{1 \pm \sqrt{1+4a^2s^2}}{-2as}, s\right)\right\} & \text{if } s \neq 0. \end{cases}$$

As is clear from this formula, changing the section a little (that is, a sufficiently small change of $s \neq 0$) does not lead to a qualitative change of the tangency set C. The two tangency points $(-2 \pm \sqrt{5}, 1)$ for $s = 1$ are labeled C in Figure 2. From the figure we can observe that the first intersection with Σ of the forward orbit $\mathcal{O}^+(\mathbf{x})$ of a point $\mathbf{x}$ below Σ (except for the equilibrium at the origin) always lies to the left of the tangency point $(-2 + \sqrt{5}, 1)$. In fact, in this case $\mathcal{O}^+(\mathbf{x})$ also always lies to the right of the other tangency point $(-2 - \sqrt{5}, 1)$. By contrast, the forward orbit of a point $\mathbf{x}$ above Σ intersects either to the left of $(-2 - \sqrt{5}, 1)$ or to the right of $(-2 + \sqrt{5}, 1)$, which depends on whether the orbit starts to the right or to the left of the backward orbit of the tangency point $(-2 - \sqrt{5}, 1)$. Note that this backward orbit divides the space into orbits that intersect Σ and orbits that miss.

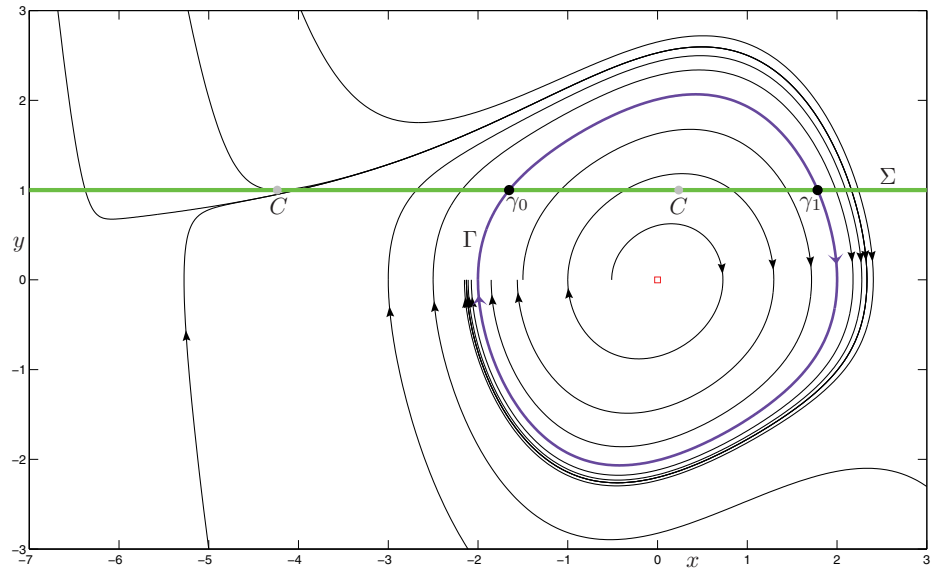


Figure 2. *Phase portrait of the (unforced) Van der Pol oscillator (3.1) for $a = 0.25$ with a global section Σ_1 (green line) at $y = 1$. The periodic orbit Γ (purple) intersects Σ_1 at the two points γ_0 and γ_1. The flow is tangent to Σ at the points denoted by C.*

Near γ_0 we can define a local Poincaré map P_{loc} on a local section $\Sigma_{\text{loc}}(\gamma_0) \subset \Sigma$, and similarly we can define P_{loc} near γ_1 on $\Sigma_{\text{loc}}(\gamma_1) \subset \Sigma$. Since there are two transverse intersections of Γ with Σ, we have $P_{\text{loc}} = P^2$ on both $\Sigma_{\text{loc}}(\gamma_0)$ and $\Sigma_{\text{loc}}(\gamma_1)$, where P is the global Poincaré map as defined by (2.5). Figure 3 illustrates that P^2 is indeed not a continuous map on the whole of Σ but is discontinuous at C. To see this we consider the tangency point $c := (-2 + \sqrt{5}, 1) \in C$. The image $P^2(c + \delta)$ of any point $c + \delta \in \Sigma_s$ for $\delta \geq 0$ lies closer to γ_1 than $c + \delta$. (The only exception is $P^2(\gamma_1) = \gamma_1$.) Note that $P^2(c)$ is the limit of $P^2(c + \delta)$ as $\delta \to 0$ and it lies closer to γ_1 than c. Now consider a point $c - \epsilon \in \Sigma$ for $\epsilon > 0$ small. As is shown in Figure 3, the image $P^2(c - \epsilon)$ lies closer to γ_0 than $c - \epsilon$. In the limit of $\epsilon \to 0$ the image $P^2(c-\epsilon)$ converges to $P^1(c)$ rather than to $P^2(c)$. Hence, P^2 is discontinuous across c.

It is straightforward to see that the discontinuity of P^2 is due to a discontinuity of the global Poincaré map P itself. Namely, we have that

$$\lim_{\epsilon \to 0} P(c - \epsilon) = c \neq P(c) = \lim_{\delta \to 0} P(c + \delta).$$

Note that this involves a discontinuity of the integration time, namely, the first-return time $t_{(c-\epsilon)}$ associated with $P(c - \epsilon)$ goes to 0 for $\epsilon \to 0$, while there is a nonzero integration time associated with the orbit segment connecting c with $P(c)$. As can be seen from Figure 3, the continuous extension of $P_{\text{loc}} = P^2$ from $\Sigma_{\text{loc}}(\gamma_1)$ across the point c to $\Sigma_{\text{loc}}(\gamma_0)$ is the map P^3. In particular, for this extension the integration time back to a local neighborhood of the section is continuous; see also [10, 13].

We now consider the interaction of the family of sections Σ_s with the periodic orbit Γ, that is, with an invariant manifold of the flow. As is shown in Figure 4, a codimension-one

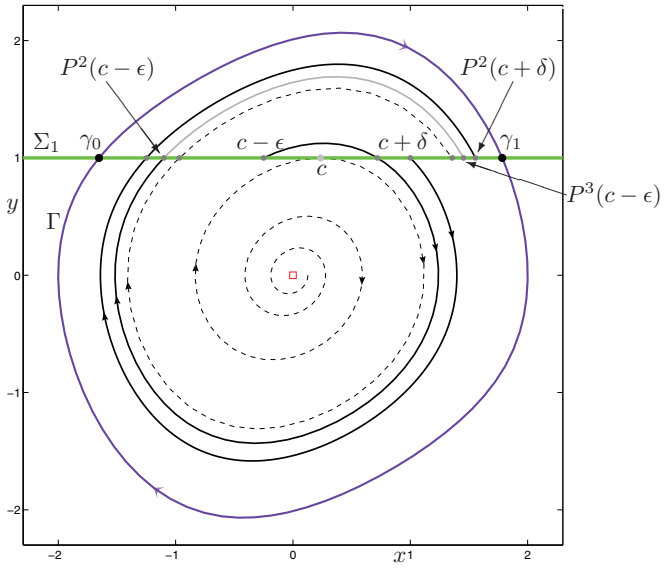


Figure 3. *Illustration of the discontinuity of P^2 for the Van der Pol oscillator (3.1) across a tangency point $c \in \Sigma_1$; compare with Figure 2.*

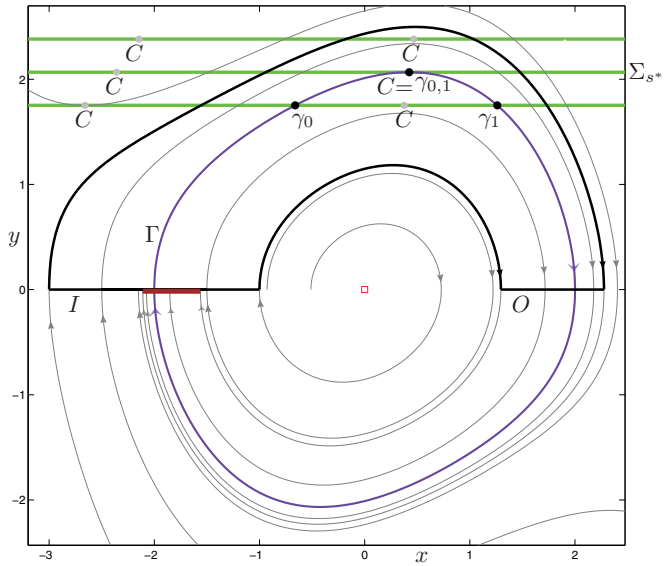


Figure 4. *A flowbox for the Van der Pol oscillator (3.1) near the quadratic tangency of the periodic orbit Γ with the section Σ_{s^*}; the three topologically different choices for Σ_s (green lines) are $s = 1.7518$, $s = 2.0670$, and $s = 2.3822$; compare with Figure 2.*

bifurcation of the Poincaré map occurs for the specific value s^* of s where Γ has a quadratic tangency with the section Σ_{s^*}; numerically we find that $s^* \approx 2.0670$. For $s < s^*$ there are

two transverse intersection points of Γ with Σ_s, as is the case in Figures 2 and 3. For $s > s^*$, on the other hand, Γ has no intersections at all with Σ_s so that the Poincaré map P does not give any information on the dynamics of Γ in this case.

To analyze the situation we consider a flowbox that includes part of the periodic orbit Γ near the tangency point with Σ; see Figure 4, where the flowbox is indicated by thick black lines. Note that the periodic orbit Γ inside the flowbox is simply a particular orbit that is determined by a single point of the in-set I. The fact that Γ is a periodic orbit cannot be deduced from the flowbox alone. Rather, one needs to take into account the global map R from O back to I, which is a contraction in this case since Γ is an attracting orbit. In Figure 4 the image $R(O)$ of the out-set is the thick brown segment on the in-set I.

The idea is now to consider only the flow inside the flowbox of Figure 4. For $s = 1.7518$ the section Σ_s intersects Γ in the two points γ_0 and γ_1, as for the case $s = 1$ of section 3.1. Orbits inside the flowbox that are sufficiently close to Γ intersect Σ_s twice, but orbits below the tangency point C in the flowbox do not intersect Σ_s at all. For $s = s^* \approx 2.0670$, orbits above Γ intersect Σ_s, but orbits below Γ (that is, below the point C in the flowbox) do not. This is due to the nontransverse quadratic tangency of Γ with Σ_{s^*}. Finally, for $s = 2.3822$, the periodic orbit Γ does not intersect the section Σ_s so that points sufficiently close to Γ do not intersect Σ_s either. The overall situation concerning the periodic orbit Γ follows when one takes the map R from O to I into account. Namely, under repeated re-entry into the flowbox all orbits eventually intersect Σ_s infinitely often for $s < s^*$, while all orbits eventually do not return to Σ_s for $s > s^*$. We stress that this division is determined locally inside the flowbox by the quadratic tangency between Γ and Σ_s at $s^* \approx 2.0670$.

3.2. Normal form of the quadratic tangency bifurcation. Generically, a tangency of an orbit of a vector field is quadratic [40], and we have the following result.

Proposition 3.1. *In any sufficiently small flowbox, the phase portrait near a quadratic tangency of an orbit of a flow in $\mathbb{R}^2$ with a global section is topologically equivalent to the phase portrait in the standard flowbox (2.7) for $n = 2$ given by the section*

$$\Sigma = \{(u, v) \mid v = 2\,u^2\}. \tag{3.3}$$

The proof of Proposition 3.1 can be found in [40] for the related situation of a flow on a two-dimensional manifold with boundary. Note that singularity theory guarantees that there is a smooth map that maps the section and the tangent orbit to the standard flowbox as stated. It requires an extra step to show that this map can be chosen in such a way that orbits map to orbits.

The local dynamics within the standard two-dimensional flowbox, that is, for $(u, v) \in [-1, 1]^2$, can be described by giving the transfer map ρ_I from the in-set I to the section Σ, the transfer map ρ_Σ from Σ to itself, and the transfer map ρ_O from Σ to the out-set O. Since not all orbits hit Σ, the map ρ_I is not everywhere defined; similarly, ρ_Σ is not defined for points that have no local returns to Σ. We have the following explicit formulae:

$$u = \rho_I(v) = \begin{cases} -\sqrt{v/2} & \text{for } 0 \le v \le 1, \\ \text{undefined} & \text{for } -1 \le v < 0, \end{cases} \tag{3.4}$$

$$\rho_\Sigma(u) = \begin{cases} -u & \text{for } -1 \le u \le 0, \\ \text{undefined} & \text{for } 0 < u \le 1, \end{cases}$$

$$v = \rho_O(u) = \begin{cases} 2u^2 & \text{for } 0 \le u \le 1, \\ \text{undefined} & \text{for } -1 \le u < 0. \end{cases}$$

Similarly, the associated transfer times are

$$\begin{aligned} \tau_I(v) &= 1 - \sqrt{v/2} \quad \text{for } 0 \le v \le 1, \\ \tau_\Sigma(u) &= 2u \quad \text{for } -1 \le u \le 0, \\ \tau_O(v) &= 1 - 2u^2 \quad \text{for } 0 \le u \le 1. \end{aligned} \tag{3.5}$$

For a general flow with quadratic tangency, (3.4) and (3.5) are normal forms that give the leading-order components of the respective transfer maps. The singularities of the above transfer maps expose the local dynamics within the flowbox. Notice that ρ_I has a quadratic singularity at $v = 0$ due to the grazing of the trajectory at the minimum of the parabolic section Σ, that is, at C.

We now consider the case that the flowbox contains a one-dimensional invariant manifold that has a quadratic tangent with a particular section, as was the case for the Van der Pol system in Figure 4. This *quadratic tangency bifurcation* is of codimension one, where we require the standard genericity condition that the dependence on the parameter is smooth and that the manifold crosses the section with positive speed. We have the following normal-form result.

Proposition 3.2. *In any sufficiently small flowbox, the unfolding of a quadratic tangency of a one-dimensional invariant manifold of a flow in $\mathbb{R}^2$ with a global section is topologically equivalent to the unfolding in the standard flowbox (2.7) for $n = 2$ given by the one-parameter family of sections*

$$\Sigma_s = \{(u, v + s) \mid v = 2\,u^2\}, \tag{3.6}$$

where the invariant manifold is the line

$$M = \{(u, v) \mid v = 0\}.$$

Proof. Suppose that the unfolding parameter of the quadratic tangency is η and the bifurcation takes place at $\eta = 0$. According to Proposition 3.1, any phase portrait of the unfolding in a flowbox near the quadratic tangency point is topologically equivalent to that given by $\widetilde{\Sigma} = \Sigma_0$ in the standard flowbox (2.7) for $n = 2$. Therefore, the invariant manifold (which is simply a single orbit) is mapped to a straight flowline $\widetilde{M} = \{v = s(\eta)\}$ in the standard flowbox, where $s(0) = 0$; furthermore, genericity of the dependence on η implies that $\frac{ds}{d\eta}(0) \neq 0$ so that $s(\eta)$ unfolds the bifurcation. The result follows by applying the coordinate change $(u, v) \mapsto (u, v - s(\eta))$, which maps $\widetilde{M}$ to M and $\widetilde{\Sigma}$ to $\Sigma_{-s(\eta)}$. The thus constructed η-family of coordinate changes is continuous by the genericity assumption on η. ∎

Figure 5 illustrates the quadratic tangency bifurcation of Proposition 3.2 by showing the standard flowbox (2.7) with the parabolic section given by (3.6). Figure 5(a) shows the generic case $M \cap \Sigma_s = \emptyset$, panel (b) is at the tangency where $M \cap \Sigma_s = C$ for $s = 0$, and panel (c) is the generic case $M \cap \Sigma_s = \{m_0, m_1\}$. Note that the in-set I (that is, the v-space) acts as the parameter space, because the relative position of M and Σ_s is uniquely determined by $M \cap I$ relative to the projection of C onto I; compare with Figure 4.

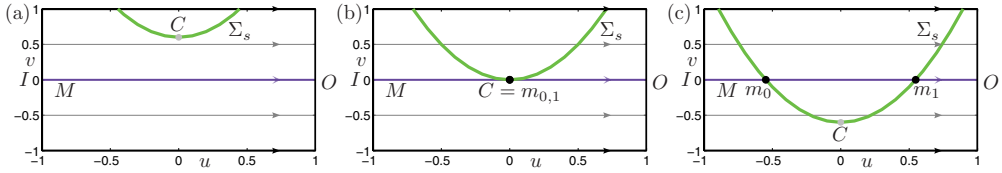


Figure 5. *Unfolding of a quadratic tangency of an invariant manifold M (purple) with a global section (green) given by the family of parabolic sections Σ_s in the standard flowbox for $s = 0.6$ (a), $s = 0$ (b), and $s = -0.6$ (c).*

3.3. Global manifestations of a quadratic tangency. The example of a quadratic tangency bifurcation of the Van der Pol system as discussed in section 3.1 is a specific global manifestation of this bifurcation. Namely, the invariant manifold M of Proposition 3.2 is actually a segment of the attracting periodic orbit Γ. This means that there is a map R from the out-set O back to the in-set I and that this map is a contraction. Therefore, the first return map P on Σ_s near a quadratic tangency of Γ is given by

$$P(u) = \begin{cases} \rho_\Sigma(u) & \text{for} \quad -1 \le u \le 0, \\ \rho_I \circ R \circ \rho_O(u) & \text{for} \quad 0 < u \le 1, \end{cases} \tag{3.7}$$

where u is the coordinate in the standard flowbox. It is now possible to determine the second return map as

$$P^2(u) = \begin{cases} \rho_I \circ R \circ \rho_O \circ \rho_\Sigma(u) & \text{for} \quad -1 \le u \le 0, \\ \rho_\Sigma \circ \rho_I \circ R \circ \rho_O(u) & \text{for} \quad 0 < u \le 1. \end{cases} \tag{3.8}$$

Note that P and P^2 have two branches, namely, one defined on the interval from I to C and one on the interval from C to O.

Figure 6 shows P and P^2 in a flowbox of the Van der Pol system (3.1) before, at, and after the tangency of the periodic orbit Γ that was illustrated in Figure 4. The two branches have been computed with numerical continuation, where the beginning point was varied along Σ_s toward C from the in-set and from the out-set, respectively. In all panels the endpoints of a branch that have an image under the map are shown as gray dots. The purple dots denote the intersection points $\gamma_{0,1}$ of $\Gamma \cap \Sigma_s$. Note that the slopes of all branches are less than 1 in absolute value, because the periodic orbit Γ is attracting.

We first discuss the consequence of the quadratic tangency for the first return map P, which is shown in the left column of Figure 6. Before the quadratic tangency the section Σ_s is above the periodic orbit Γ so that $\Gamma \cap \Sigma_s = \emptyset$. As a result, the points on Σ_s in the interval $[C, P^{-1}(C))$ do not return to the section under P; see panel (a1). At the quadratic tangency for $s = s^* \approx 2.0670$ the section Σ_s intersects Γ tangentially in a single point, which is a fixed point of P. Note from panel (b1) that P is continuous and both left and right differentiable; however, these derivatives differ at the point $\Sigma_s \cap \Gamma$ (purple dot). After the quadratic tangency, shown in panel (c1), the first return map P is well defined everywhere but is discontinuous at the tangency locus C. Furthermore, P has an attracting period-two orbit, namely, the two intersection points $\gamma_{0,1}$ (purple dots) of Γ with Σ_s. Note that during this entire transition the

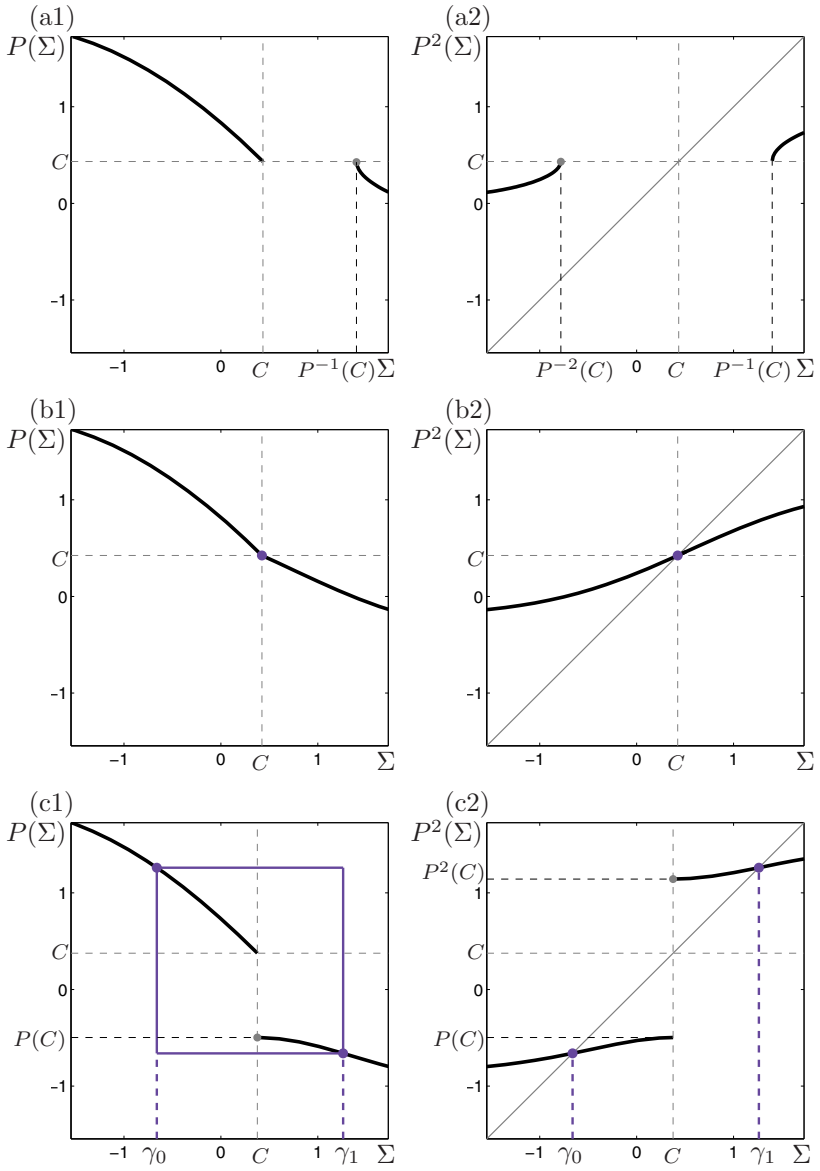


Figure 6. *The first return map P (left column) and second return map P^2 (right column) for the Van der Pol oscillator (3.1) on the sections Σ_s before, at, and after a quadratic tangency; compare with Figure 4. Rows* (a)–(c) *are for $s \approx 2.1359$, $s \approx 2.0670$, and $s \approx 1.7518$, respectively.*

left branch (from I to C) remains virtually unchanged, while the right branch (from C to O) changes topologically owing to the changing nature of $\rho_I \circ R \circ \rho_O$ with s.

The second return map P^2 is simply the second iterate of P; it is shown in the right column of Figure 6. Before the quadratic tangency, points in the interval $(P^{-2}(C), P^{-1}(C)]$ do not have an image under P^2; see panel (a2). At the tangency, the two branches come together in a fixed point (purple dot), at which P^2 is continuous and smooth. After the tangency bifurcation, the map P^2 is well defined everywhere but discontinuous at C. The intersection

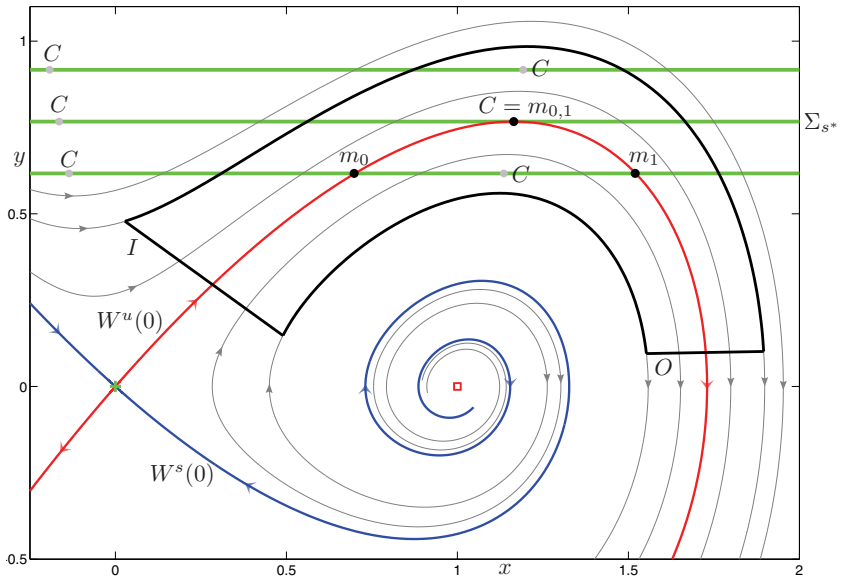


Figure 7. *Phase portrait of the vector field (3.9) for $\lambda = 0.25$ with the stable manifold $W^s(0)$ (blue) and the unstable manifold $W^u(0)$ (red) of the origin. Also shown are three different choices of horizontal global section Σ_s, before, at, and after a tangency of $W^u(0)$ with Σ_s, namely, for $s = 0.6169$, 0.7669, and 0.9169, respectively. The black flowbox shows that this bifurcation unfolds as described by the normal form.*

points $\gamma_{0,1}$ (purple dots) of Γ with Σ_s are fixed points under P^2. Indeed, the restrictions of P^2 to the neighborhoods $\Sigma_{\text{loc}}(\gamma_0)$ or $\Sigma_{\text{loc}}(\gamma_1)$ give the respective local Poincaré maps P_{loc}. Note that in these transitions both branches of P^2 are changing, because they both depend on the map R that describes the flow outside the flowbox.

As we have seen, the properties of P and P^2 depend on the map R from the out-set back to the in-set. In case of a repelling periodic orbit Γ, the map R^{-1} from the in-set back to the out-set is a contraction. Therefore, the respective maps P and P^2 can be obtained from Figure 6 by reversing the roles of in-set and out-set. Geometrically, this corresponds to a reflection in the antidiagonal of the respective panels.

An altogether different global situation arises when the invariant manifold M inside the flowbox is not a segment of a periodic orbit. A generic and dynamically relevant situation is that M is a segment of a global stable or unstable manifold of an equilibrium. As a concrete example we consider the vector field

$$\begin{cases} \dot{x} &= y, \\ \dot{y} &= \lambda y + x - x^2, \end{cases} \tag{3.9}$$

which was introduced in [19] as a system with a homoclinic bifurcation. For $\lambda = 0.25$ the phase portrait of (3.9) is as in Figure 7. The system has a saddle point at the origin with stable manifold $W^s(0)$ (blue) and unstable manifold $W^u(0)$ (red) as shown in Figure 7. We consider a family of horizontal sections $\Sigma_s = \{(x, y) \mid y = s\}$ (green), which has a tangency with $W^u(0)$ for $s^* \approx 0.7669$. When one restricts one's attention to a suitable flowbox (boldface curves),

then this tangency unfolds as described by the normal form in Proposition 3.2; compare with Figure 5. However, the invariant manifold M inside the flowbox is now a segment of $W^u(0)$ so that points of the out-section O move off to infinity. Hence, there exists no map R from O back to I and the intersection points m_0 and m_1 do not correspond to fixed points of the first return map P. In fact, the map P is defined only from I to C, where it is as shown in the left column of Figure 6.

Note that Teixeira [40] considered the related situations that a periodic orbit or a separatrix interacts with the boundary of a two-dimensional flow. However, he treated these two global situations as different cases. By contrast, we take the point of view that the local bifurcation mechanism in a suitable flowbox is actually the same, while the global dynamics is determined by the exact nature of the invariant manifold that is involved in the tangency with the section (or boundary) inside the flowbox.

4. Quadratic tangency bifurcation in three-dimensional flows. An invariant manifold of a three-dimensional flow can be either one- or two-dimensional. Therefore, there are more possibilities for interactions of an invariant manifold with a two-dimensional section. As before, we discuss interactions in the context of a family of two-dimensional sections Σ_s at a regular point of the critical locus C, that is, at a generic tangency point between a three-dimensional flow and the section. To this end, we consider the standard flowbox (2.7) for $n = 3$, where we take $(v_1, v_2) = v \in [-1, 1]^2$ as coordinates. We start with a straightforward generalization of Proposition 3.2.

Corollary 4.1. *In a sufficiently small flowbox, the unfolding of a quadratic tangency of a one-dimensional invariant manifold of a flow in $\mathbb{R}^3$ with a global section is topologically equivalent to the unfolding in the standard flowbox (2.7) for $n = 3$ given by the one-parameter family of sections*

$$\Sigma_s = \{(u, v_1, v_2 + s) \mid v_2 = 2\,u^2\}, \tag{4.1}$$

where the invariant manifold is the line

$$M = \{(u, v) \mid (v_1, v_2) = (0, 0)\}.$$

Note that this bifurcation is of codimension one, because it involves the interaction of a one-dimensional manifold with a one-dimensional fold curve C in $\mathbb{R}^3$. In the standard flowbox the section Σ_s is a parabolic cylinder with the straight line $C = C(s) = \{(u, v_1, v_2 + s) \mid u = 0 \text{ and } v_2 = 0\}$, along which the flow has a quadratic tangency. This represents the generic situation inside a flowbox in the absence of cusp points on C, that is, in a sufficiently small neighborhood of a generic quadratic tangency point; see also [38, Figure 5.1]. Note further that this unfolding reduces to the case for $n = 2$ of Proposition 3.2 by means of considering the two-dimensional slice for $v_1 = 0$. Therefore, the form of the transfer map ρ_I from the in-set I to the section Σ, the transfer map ρ_Σ from Σ to itself, and the transfer map ρ_O from Σ to the out-set O are as given in (3.4).

The more interesting possibility for $n = 3$ is the interaction of a two-dimensional invariant manifold M with a two-dimensional global section near a quadratic tangency of a single orbit. In this setting we do not consider the details of the dynamics on the part of M inside the flowbox but simply consider M as a smooth family of one-dimensional orbit segments. We

say that there is a *codimension-one quadratic tangency bifurcation* between the two surfaces M and Σ at the point $\mathbf{x}^* = M \cap \Sigma$ if $M \cap I$ has a quadratic tangency with the projection (along flowlines) of C onto I. As a genericity condition we require that the dependence on the parameter is smooth and that M crosses Σ with positive speed. Note that this definition is general and does not depend on the choice of flowbox. The two cases of the quadratic tangency bifurcation are distinguished by whether $M \cap I$ lies in the region to which Σ projects or not, which we refer to as the *saddle case* and the *minimax case*, respectively.

Proposition 4.2. *In any sufficiently small flowbox, the unfolding of a quadratic tangency of a two-dimensional invariant manifold of a flow in $\mathbb{R}^3$ with a global section is topologically equivalent to the unfolding in the standard flowbox (2.7) for $n = 3$ given by the one-parameter family of sections*

$$\Sigma_s = \{(u, v_1, v_2 + s) \mid v_2 = 2\,u^2 \pm 2\,v_1^2\}, \tag{4.2}$$

where the invariant manifold is the plane

$$M = \{(u, v_1, v_2) \mid v_2 = 0\}.$$

The plus sign in (4.2) gives the minimax case and the minus sign the saddle case of the bifurcation.

Proof. Suppose that the unfolding parameter of the quadratic tangency is η and the bifurcation takes place at $\eta = 0$. In light of Corollary 4.1 we may consider a manifold $\widetilde{M}$ in the standard flowbox with $\Sigma_{s(\eta)}$ as given by (4.1) such that the quadratic tangency for $\eta = 0$ takes place at the origin. Then $\widetilde{M} \cap I$ is (locally and for sufficiently small η) given by a function $\mu_\eta : v_1 \to v_2$ on the in-set I with a single minimum or maximum. Hence, the u-independent coordinate change $(u, v_1, v_2) \mapsto (u, v_1, v_2 - \mu_\eta(v_1) - s(\eta))$ maps $\widetilde{M}$ to the plane $\{v_2 = 0\}$. As a result, the image $\widetilde{\Sigma}_{s(\eta)}$ under this transformation is either a paraboloid or a saddle surface, where $s(\eta)$ is the v_2-value (vertical distance) of the maximum, the minimum, or the saddle point, respectively. Therefore, for each η there exists a v_2-dependent coordinate change of v_1 that leaves the origin invariant and a v_2-dependent coordinate change of u (as in the proof of Proposition 3.1) that together bring $\widetilde{\Sigma}_{s(\eta)}$ to the normal form given by (4.2). Again, a rescaling of time ensures that $\dot{u} = 1$. The thus constructed η-family of coordinate changes is continuous by the genericity assumption on η. ∎

From a singularity theory point of view, the two cases determined by the sign in (4.2) are the well-known transitions through a minimum or maximum (minimax for short), and through a saddle [2, 15, 46]. In particular, these are the only two generic cases, and the unfolding given in Proposition 4.2 applies in both cases. However, there is an additional ingredient that is important in the present context: one also needs to consider the direction of the flow on either side of the tangency locus C.

The minimax and the saddle cases of the quadratic tangency of Proposition 4.2 are illustrated in Figures 8 and 9, respectively. In these figures, the left-hand columns show the curve $M \cap I = \{v_2 = 0\}$ in the in-set I with the projection of Σ_s (green region) that is bounded by the projection $\widehat{C}$ of the fold curve C; any orbit in the green region intersects Σ twice in the flowbox. The middle columns show the section Σ_s (green) and the manifold M (pink) in the three-dimensional flowbox; the tangency locus C is the fold curve given by $u = 0$ in

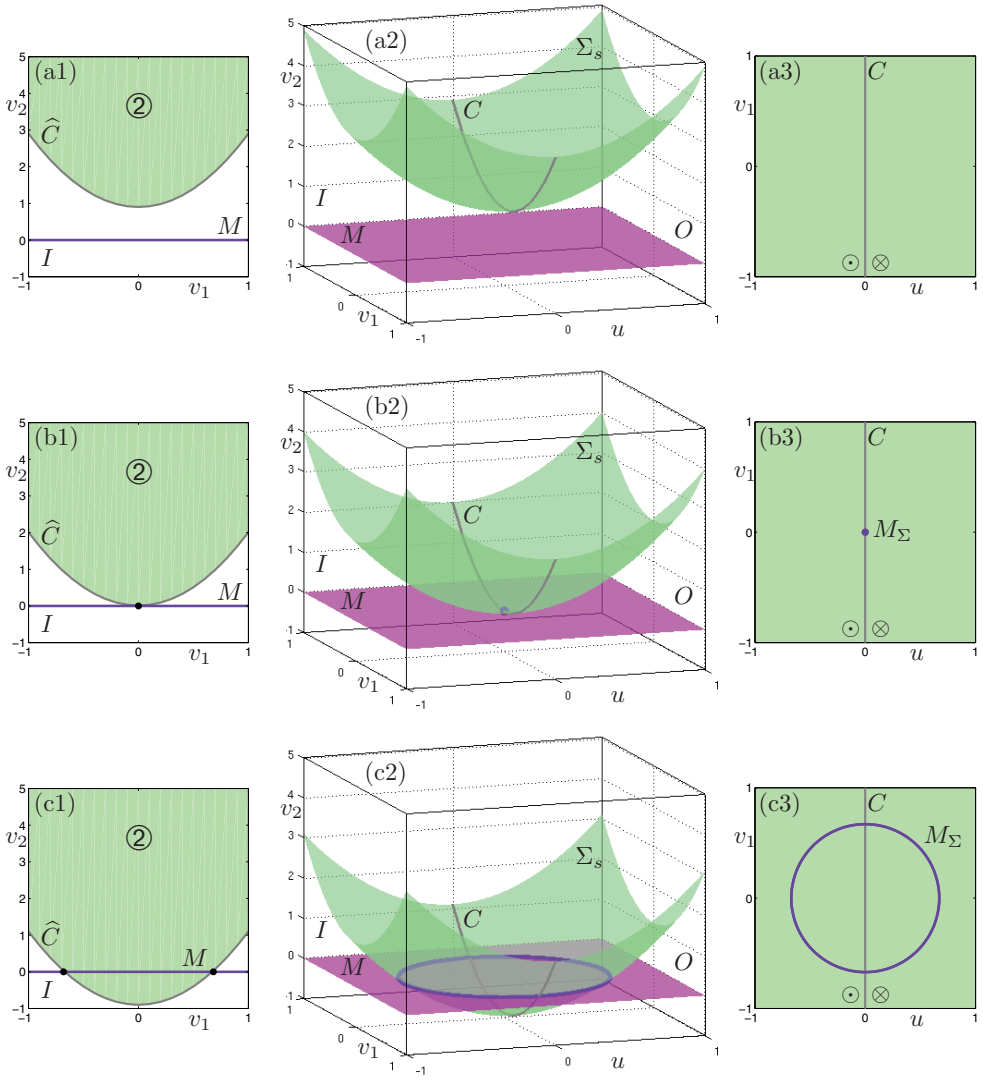


Figure 8. *The minimax case of the quadratic tangency bifurcation, as described by the plus sign in (4.2), before* (a)*, at* (b)*, and after* (c) *the bifurcation. The left column shows $M \cap I = \{v_2 = 0\}$ relative to the projection curve $\widehat{C}$ that bounds the projection of Σ_s (green region) onto the in-set I; the symbol ② indicates that there are two intersections with Σ in the green region. The middle column shows the section Σ_s (green) and the planar two-dimensional manifold M (pink), and the right column shows the intersection M_Σ in Σ_s. The tangency locus C (gray curve) separates the regions where the flow is upward (⊙) and downward (⊗). Rows* (a)–(c) *are for $s = 0.9$, $s = 0.0$, and $s = -0.9$, respectively.*

both cases. The right-hand columns show the corresponding intersections M_Σ (purple) in the global section Σ_s. The direction of the flow through Σ_s is indicated: the symbol ⊙ denotes upward flow and the symbol ⊗ downward flow (with respect to the normal to the section). Rows (a) through (c) are before, at, and after the respective bifurcations.

The minimax transition creates a topological circle M_Σ; see Figure 8(c). This circle is

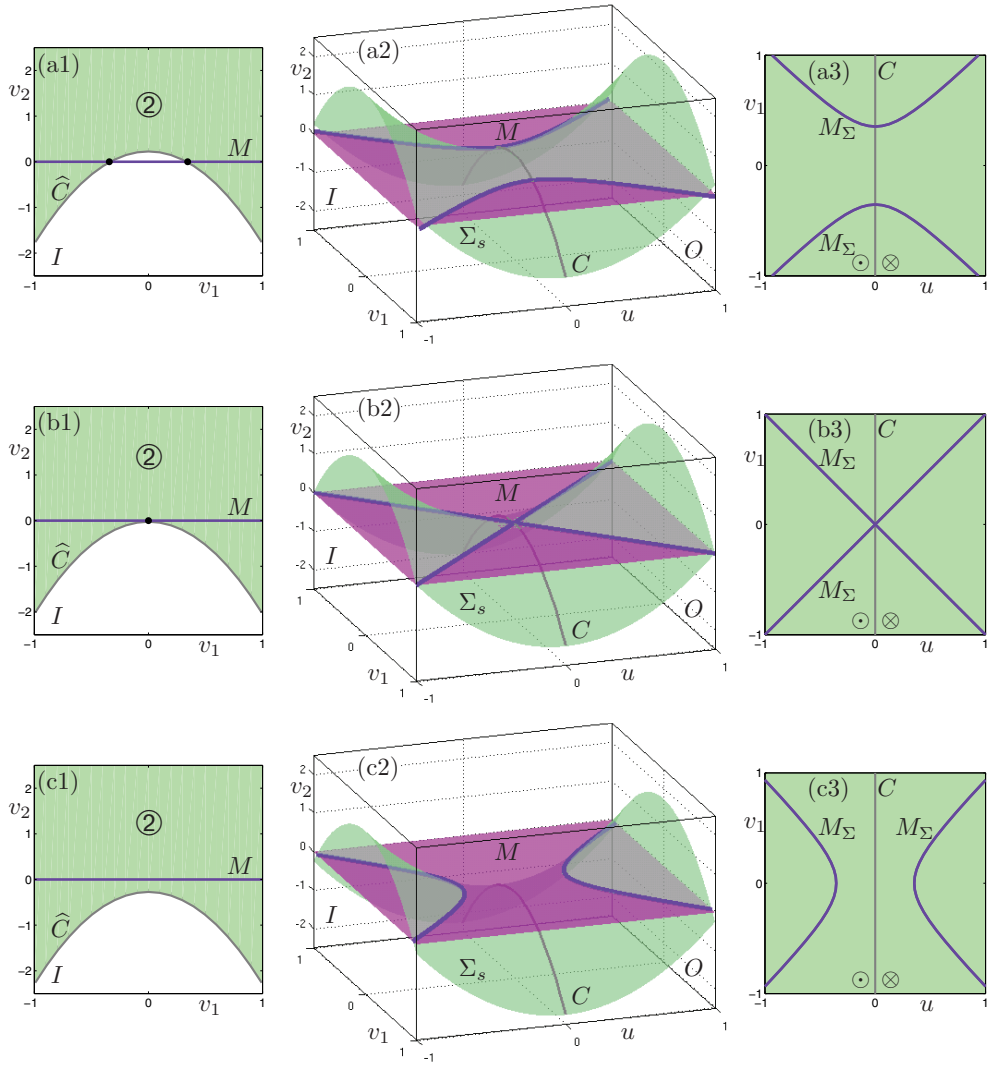


Figure 9. *The saddle case of the quadratic tangency bifurcation, as described by the minus sign in (4.2), before* (a)*, at* (b)*, and after* (c) *the bifurcation. The left column shows $M \cap I = \{v_2 = 0\}$ relative to the projection curve $\widehat{C}$ that bounds the projection of Σ_s (green region) onto the in-set I; the symbol ② indicates that there are two intersections with Σ in the green region. The middle column shows the section Σ_s (green) and the planar two-dimensional manifold M (pink), and the right column shows the intersection M_Σ in Σ_s. The tangency locus C (gray curve) separates the regions where the flow is upward ($\odot$) and downward ($\otimes$). Rows* (a)–(c) *are for $s = 0.25$, $s = 0.0$, and $s = -0.25$, respectively.*

divided by C into two parts: points on the left half of M_Σ are mapped under the flow to the right half of M_Σ; the points on $M_\Sigma \cap C$ intersect Σ only once inside the flowbox. By contrast, in the saddle transition we have that $M_\Sigma \neq \emptyset\ \forall s$; see Figure 9. The intersection of M with Σ_s in row (a) consists of two arcs that both cross C. Points on each arc to the left of C return to Σ_s on the same arc to the right of C, namely, at the same value of v_1. At the moment of the

saddle transition in row (b) the two arcs of M_Σ meet on C to form a cross. In row (c) M_Σ consists of two arcs in a different way; namely, there is an arc in the region of upward flow and one arc in the region of downward flow; points on the left arc are mapped diffeomorphically to the right arc.

4.1. Intersection of an invariant two-torus with a plane. As for the case $n = 2$ in section 3.2, the question arises as to how the minimax and the saddle transitions in a flowbox manifest themselves for a given three-dimensional vector field. This depends again on whether there is a map R from the two-dimensional out-set O back to the two-dimensional in-set I. A natural setting in which such a map may occur is that the segment M inside the flowbox is part of an invariant two-torus $\mathbb{T}$ of the underlying flow. In this case, the map R leaves M invariant, meaning that $M \cap O$ is mapped back to $M \cap I$. When the torus is attracting, R contracts the v_2-direction of the standard flowbox.

Importantly, bifurcations of the Poincaré map on a section can be brought about either by changes to the section or by changes to the torus itself. To show this, we consider a simple geometric example, given by the unit circle

$$\Gamma = \{(x, y, z) \in \mathbb{R}^3 \mid x = \cos\theta,\ y = 0,\ z = \sin\theta, \quad \text{where } 0 \le \theta < 2\pi\} \tag{4.3}$$

in the (x, z)-plane, surrounded by a tube of radius r that forms the torus

$$\begin{aligned}\mathbb{T} = \{(x, y, z) \in \mathbb{R}^3 \mid x = (1 + r\cos\phi)\cos\theta,\ y = r\sin\phi,\\ z = (1 + r\cos\phi)\sin\theta, \quad \text{where } 0 \le \theta, \phi < 2\pi\}.\end{aligned} \tag{4.4}$$

We may consider $\mathbb{T}$ and Γ as a geometric model of an invariant torus surrounding a periodic orbit of an unspecified vector field, as long as $\mathbb{T}$ does not have self-intersections, that is, for $0 < r < 1$. We now consider the family of global planar sections $\Sigma_s = \{(x, y, z) \in \mathbb{R}^3 \mid z = s\}$. While we are not specifying an underlying flow, we make the assumption that the tangency locus $C \subset \Sigma_s$ is given by the condition $x = 0$; note that this is consistent with the desired invariance of $\mathbb{T}$ and Γ. For specificity we further assume that the flow is upward ($\odot$) on Σ_s for $x < 0$ and downward ($\otimes$) for $x > 0$.

The situation for $r = 0.5$ and $s = 0.5$ is shown in Figure 10(a) in (x, y, z)-space, while panel (b) shows the invariant objects in the planar section Σ_s. Note that for the specific example of $\mathbb{T}$ as defined by (4.4) the family of intersection curves $\mathbb{T}_\Sigma$ comprises part of the family of Cassini ovals [28]. In particular, for these values of r and s the set $\mathbb{T}_\Sigma$ is a lemniscate, which means that Figure 10 shows a saddle transition of $\mathbb{T}_\Sigma$ locally near C; compare with Figure 9. This codimension-one situation can be unfolded in two different ways, namely, either by changing the section Σ, that is, by varying s, or by varying the radius r of the torus. In the former case, the section moves up and down through the torus $\mathbb{T}$ and the periodic orbit Γ. As is shown in the accompanying animation (69972_01.gif [8.13MB]), apart from the saddle transition at $s = 0.5$, one also encounters a quadratic tangency of Γ for $s = 1$ and a minimax transition of $\mathbb{T}$ for $s = 1$. On the other hand, when $\mathbb{T}_{0.5}$ remains fixed and the radius r of $\mathbb{T}$ is changed, then only the saddle transition is unfolded; see the accompanying animation (69972_02.gif [3.85MB]).

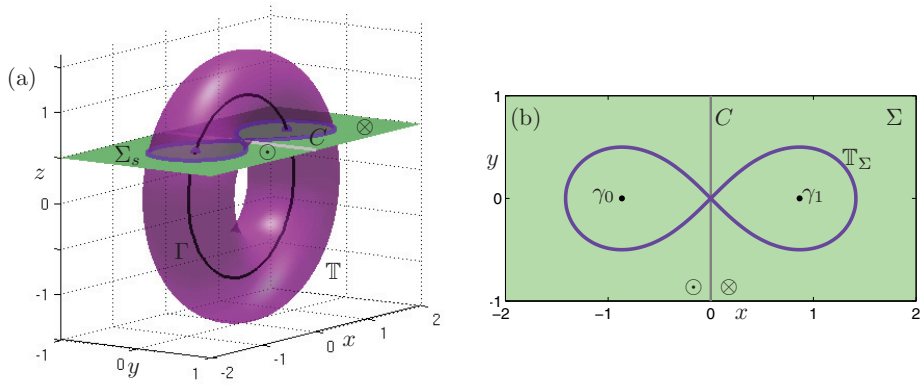


Figure 10. *Saddle transition in the geometric example (4.3)–(4.4) of an invariant torus $\mathbb{T}$ (pink) surrounding a periodic orbit Γ (black curve) with the planar section $\Sigma_{0.5}$ (green) when the radius of $\mathbb{T}$ is $r = 0.5$; the intersections $\mathbb{T}_\Sigma = \mathbb{T} \cap \Sigma_s$ are shown in purple and $\Gamma \cap \Sigma_s$ is indicated by black dots. Along C the flow is tangent to Σ, and the symbols $\odot$ and $\otimes$ indicate where the flow is upward and downward, respectively. The bifurcation unfolds when either the section or the radius r is changed, as is illustrated in the accompanying animations (69972_01.gif [8.13MB] and 69972_02.gif [3.85MB]).*

4.2. Saddle transitions in a semiconductor laser system. To demonstrate how quadratic tangencies arise in a practical example, we consider the model of a semiconductor laser with optical injection given by the equations

$$\begin{cases} \dot{E} = K + \left(\frac{1}{2}(1 + i\alpha)\, n - i\omega\right) E, \\ \dot{n} = -2\gamma n - (1 + 2Bn)\left(|E|^2 - 1\right), \end{cases} \tag{4.5}$$

where $E = E_x + iE_y$ is the complex electric field and n is the population inversion. The parameters α, B, and γ characterize the material properties of the laser, ω is the detuning, and K the injection field strength; see [44] for further details. In this paper $\alpha = 2$, $B = 0.015$, $\gamma = 0.035$, and $\omega = 0.43$ are used throughout, while K is varied.

As was done in [44], we choose the fixed global section

$$\Sigma = \{(E, n) \in \mathbb{C} \times \mathbb{R} \mid n = -0.1\}. \tag{4.6}$$

By numerical integration of (4.5) it can be found quite easily that, for a range of the injection strength K around 0.1139, the Poincaré map on Σ shows attracting invariant curves surrounding points of an unstable period-six orbit $\{\gamma_0, \ldots \gamma_5\}$; see Figure 11. As can be seen from the panels of this figure, the number of invariant curves, labeled $\mathbb{T}_\Sigma$, changes with K.

One might be tempted to think that the change in the number of invariant curves is due to a bifurcation of the vector field (4.5), but this is not the case. Instead, there are several saddle transitions involving an underlying attracting invariant torus $\mathbb{T}$; see Figure 12. The torus $\mathbb{T}$ has been obtained by integrating from a suitable initial condition (after omitting transients), because the dynamics on it appears to be quasi-periodic (or of very high period); Figure 12 shows $\mathbb{T}$ (mauve) computed as a single orbit over 60,000 time steps of size 0.01; the repelling

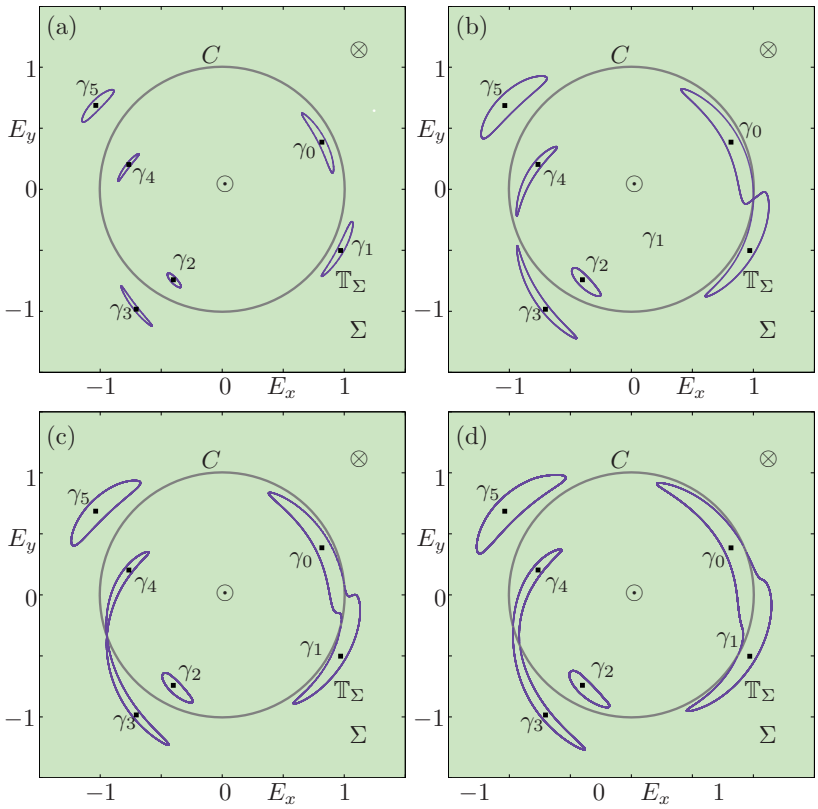


Figure 11. *The Poincaré map of* (4.5) *in the fixed section* Σ *defined by* (4.6) *has a number of attracting invariant curves, labeled* $\mathbb{T}_\Sigma$*, that surround points of an unstable period-six orbit* $\Gamma_\Sigma = \{\gamma_0, \dots \gamma_5\}$*. The number of invariant curves in* $\mathbb{T}_\Sigma$ *depends on the value of the injection strength* K*. Also shown is the circular tangency locus* C *given by* (4.7) *that divides* Σ *into two regions with upward* ($\odot$) *and downward* ($\otimes$) *flow. From* (a)–(d) K *takes the values* 0.1139, 0.11392468, 0.1139281, *and* 0.11395*; compare with Figure* 12.

periodic orbit Γ inside $\mathbb{T}$ is also shown; it was found as a fixed point of the sixth return map to the section Σ. Indeed, as is shown in Figure 12, there are exactly six intersections of Γ (black) and Σ (green).

To compare with the theory, note that the flow of (4.5) is tangent to Σ along the tangency locus

$$C = \left\{ (E, n) \in \Sigma \;\middle|\; |E|^2 = \Delta(-0.1) \right\}, \tag{4.7}$$

where

$$\Delta(n) := 1 - \frac{2\gamma n}{1 + 2Bn}. \tag{4.8}$$

Hence, C is a circle in Σ that is centered at $(0, 0, -0.1)$ with radius $\sqrt{\Delta(n)} \approx 1.0035$ for the given system parameters. We conclude from (4.5) that the flow points in the positive n-direction ($\odot$) inside C and in the negative n-direction ($\otimes$) outside C.

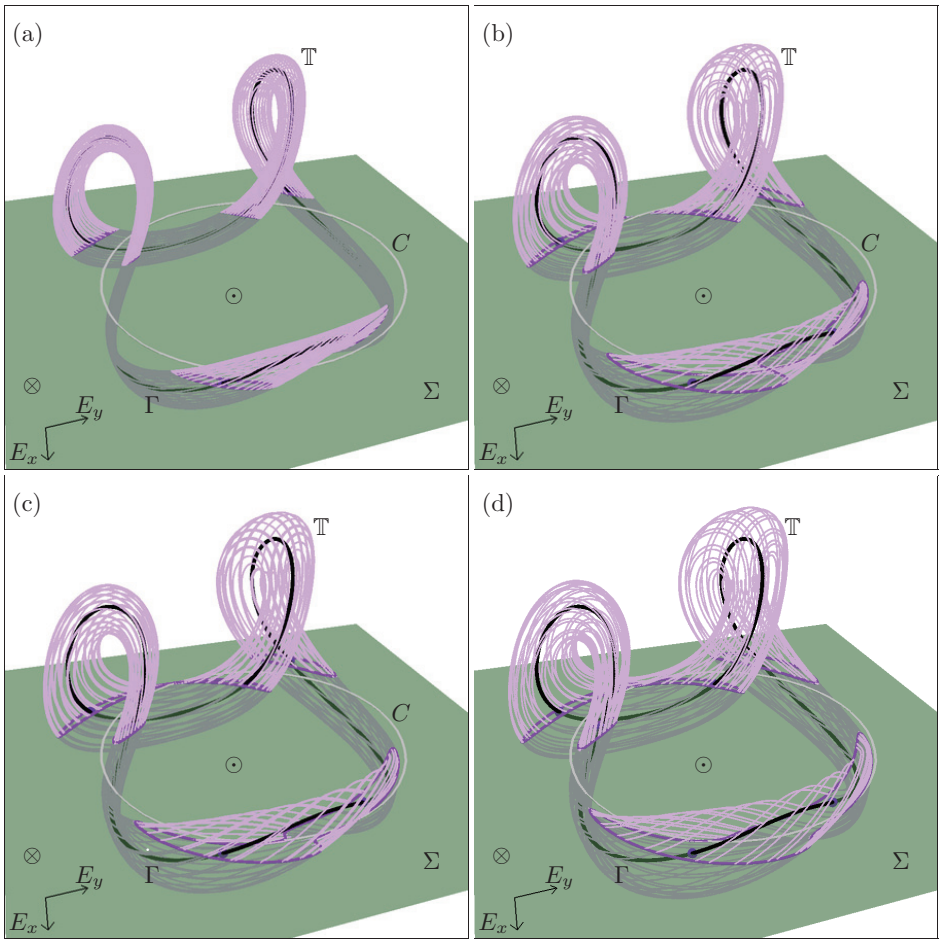


Figure 12. *The sets $\mathbb{T}_\Sigma$ and Γ_Σ in Figure 11 of the Poincaré map of (4.5) on the section Σ defined by (4.6) are intersections of an attracting invariant torus $\mathbb{T}$ and a repelling periodic orbit Γ of the flow. Also shown is the circular tangency locus C given by (4.7) that divides Σ into two regions with upward ($\odot$) and downward ($\otimes$) flow. From* (a)–(d) *K takes the values* 0.1139, 0.11392468, 0.1139281, *and* 0.11395.

In Figures 11 and 12 the parameter K is varied from (a) to (d) with the same values being used in the corresponding panels of the two figures. In panel (a) the intersection $\mathbb{T}_\Sigma$ consists of six disjoint invariant circles, each surrounding an intersection point γ_i. Under the first return map P on Σ each invariant circle is mapped to another invariant circle in $\mathbb{T}_\Sigma$. Note that a circle outside C is mapped to one inside C, and vice versa. Panels (b) of Figures 11 and 12 show the first saddle transition, where the invariant circles around γ_0 and γ_1 join to form a single invariant circle; locally the situation is topologically equivalent to Figures 10 and 9(b). While each component of $\mathbb{T}_\Sigma$ in Figure 11(a) maps to itself under the sixth return of P, this is no longer true after the saddle transition in Figure 11(b). After the bifurcation, as shown in Figure 11(c), the part inside C of the large component of $\mathbb{T}_\Sigma$ that surrounds γ_0 and γ_1 is mapped under P to the part of the same component that lies outside C. On the other hand, the outside part of this component is mapped under P^5 back to the inside part. Note

that panels (c) of Figures 11 and 12 also show the next saddle transition of the two invariant circles surrounding γ_3 and γ_4. Panels (d) of Figures 11 and 12 show the situation after the second saddle transition with the set of four circles in Σ that is invariant under P; two of the invariant circles now surround a pair of points of Γ_Σ.

As can be seen from the three-dimensional images of Figure 12, the invariant torus $\mathbb{T}$ does not appear to undergo any bifurcations (meaning that it remains a normally hyperbolic invariant manifold). The topological changes of the invariant set $\mathbb{T}_\Sigma$ of the Poincaré map shown in Figure 11 are entirely due to the fact that the torus changes its "thickness" with the injection strength K, which leads to saddle transitions. This is exactly the mechanism that was discussed with the simple geometric example in section 4.1 and in animation (69972_02.gif [3.85MB]).

5. Cubic tangency bifurcation in three-dimensional flows. In this section we discuss a different and final codimension-one tangency bifurcation of a two-dimensional invariant manifold of a three-dimensional flow. We identify and characterize this bifurcation in a planar section of the semiconductor laser system (4.5) in the next section. We then discuss the geometry with a simpler geometric model in section 5.2 and finally derive the normal form in section 5.3.

However, first we introduce a new geometric object—the *extended critical locus* denoted $\mathcal{C}$—to help understand the geometry of the flow. Suppose that we have as part of the setup a one-parameter family Σ_s of sections such that the bifurcation occurs at $\mathbf{x}^* \in \Sigma_{s^*}$. (If, as in the laser system, we are starting with only a single section Σ, then we define the family Σ_s in a natural way by moving it in the direction of the unit normal vector $\vec{n}_\Sigma(\mathbf{x})$, formally $\Sigma_s = \{\mathbf{x} + s\,\vec{n}_\Sigma(\mathbf{x}) \mid \mathbf{x} \in \Sigma\}$ and $\Sigma = \Sigma_{s^*}$; note that for sufficiently small s each Σ_s is a global section in the sense of (2.4).) The extended critical locus $\mathcal{C}$ is now defined as the union of the critical tangency loci $C(s)$ of the sections Σ_s, that is,

$$\mathcal{C} = \bigcup_s C(s). \tag{5.1}$$

We assume here that the dependence of Σ_s on s is smooth so that $\mathcal{C}$ is a smooth codimension-one submanifold of the phase space $\mathbb{R}^n$. Therefore, $\mathcal{C}$ is of the same dimension as the section Σ (i.e., it is a surface for $n = 3$), and $\mathcal{C} \cap \Sigma = C$. Generically, the extended critical locus is transverse to Σ_s, and, hence, knowing properties of the flow through $\mathcal{C}$ gives new geometric insight. We define the tangency locus D on $\mathcal{C}$ by

$$D := \{\mathbf{x} \in \mathcal{C} \mid f(\mathbf{x}) \cdot \vec{n}_\mathcal{C}(\mathbf{x}) = 0\}, \tag{5.2}$$

where $\vec{n}_\mathcal{C}(\mathbf{x})$ is the unit normal to $\mathcal{C}$ at the point $\mathbf{x}$. As with C on Σ the tangency locus D generically consists of codimension-one submanifolds (i.e., curves for $n = 3$) that divide $\mathcal{C}$ into regions with opposite directions of the flow.

5.1. Cubic tangency bifurcation in the semiconductor laser system. When the parameter K of the semiconductor laser system equation (4.5) is increased to values beyond those shown in Figure 11, then one encounters the codimension-one bifurcation of the invariant torus $\mathbb{T}$ that is illustrated in Figures 13 and 14.

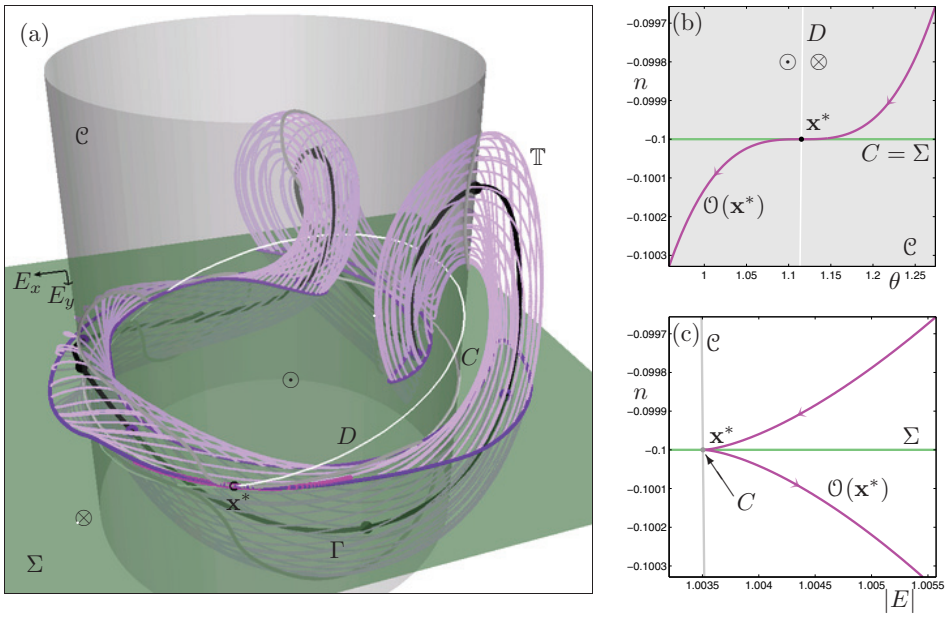


Figure 13. *Panel* (a) *shows the invariant torus* $\mathbb{T}$ *(mauve) of* (4.5) *at the moment of the cubic tangency bifurcation at* $\mathbf{x}^* \approx (0.44162, 0.90111, -0.1)$ *for* $K \approx 0.1140145$*; compare with Figure* 14(b)*. Also shown are the section* Σ *(green), the extended critical locus* $\mathcal{C}$ *(gray), and the respective intersection curves with* $\mathbb{T}$*. A segment of the orbit* $\mathcal{O}(\mathbf{x}^*)$ *(magenta curve) is shown in panel* (b) *in the* (θ, n)*-plane and in panel* (c) *in the* $(|E|, n)$*-plane (where* $E = |E|e^{i\theta}$*); notice the large difference in scales between* n *and* θ *in panel* (b)*.*

Figure 13(a) is a three-dimensional image of the surfaces $\mathbb{T}$, Σ, and $\mathcal{C}$ at the moment of bifurcation for $K \approx 0.1140145$. The extended tangency locus $\mathcal{C}$ was computed by considering the family of sections

$$\Sigma_s = \{(E, n) \in \mathbb{C} \times \mathbb{R} \mid n = s\},$$

where $\Sigma = \Sigma_{s^*}$ for $s^* = -0.1$. It follows from (4.7) and (4.8) that

$$\mathcal{C} = \left\{(E, n) \in \mathbb{C} \times \mathbb{R} \;\mid\; |E|^2 = \Delta(n)\right\}, \tag{5.3}$$

which is a cone. Furthermore, we find that

$$D = \left\{(E, n) \in \mathcal{C} \mid E_x = -\frac{n}{2K}\Delta(n),\; E_y = \sqrt{\Delta(n) - \frac{n^2}{4K^2}\Delta(n)^2}\right\}, \tag{5.4}$$

where n is chosen such that $E_y \in \mathbb{R}$, that is, $|n| \leq 2k/\sqrt{\Delta(n)}$. The cone $\mathcal{C}$ is the gray surface in Figure 13(a) that is divided by D (closed white curve) into two parts. The flow points from the inside of $\mathcal{C}$ to the outside above D and from the outside of $\mathcal{C}$ to the inside below D.

At the moment of bifurcation, illustrated in Figure 13, the torus $\mathbb{T}$ crosses the section Σ at the intersection point $\mathbf{x}^* \in C \cap D$ of C and D. Therefore, $\mathbb{T}_\Sigma$ is tangent to C in Σ and $\mathbb{T}_\mathcal{C}$ is tangent to C in $\mathcal{C}$; see row (b) of Figure 14. Figure 13(a) shows a relevant segment of $\mathcal{O}(\mathbf{x}^*)$ (magenta curve). Notice that this orbit segment remains very close to Σ and that it is

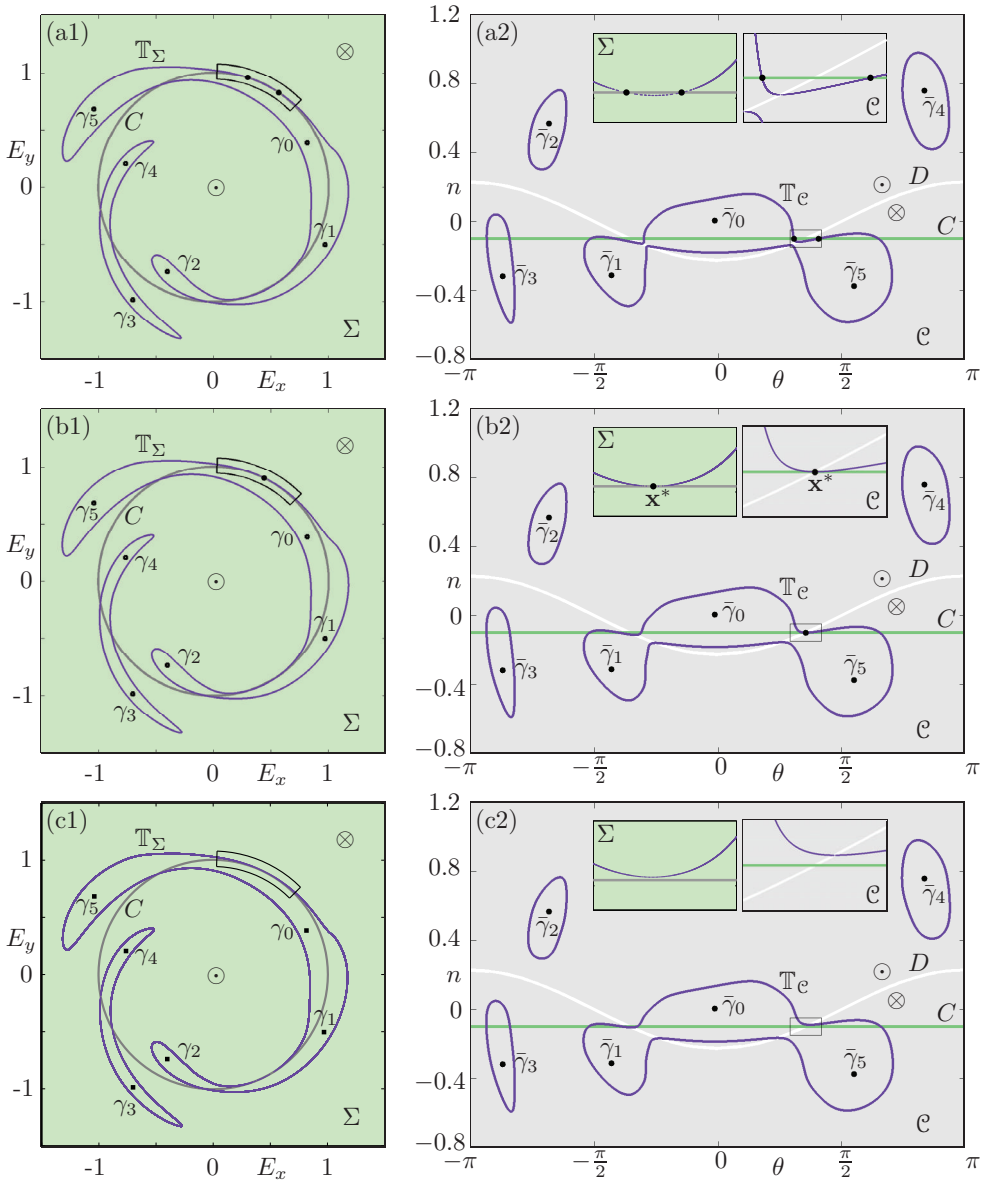


Figure 14. *Cubic tangency bifurcation for* (4.5). *The three rows show the situation before, at, and after the bifurcation. The left column shows the interaction of* $\mathbb{T}_\Sigma$ *with* C *in* Σ, *and the right column the interaction of* $\mathbb{T}_\mathcal{C}$ *with* D *in* $\mathcal{C}$; *the subpanels show enlargements near the tangency points in* Σ *(green subpanel) and* $\mathcal{C}$ *(gray subpanel) in the indicated regions. Rows* (a)–(c) *are for* $K = 0.1140105$, 0.1140145, *and* 0.1140185, *respectively; compare with Figure* 13(a).

very difficult to judge its position relative to other flow lines. Therefore, Figure 13(b) and (c) show the orbit through $\mathbf{x}^*$ in the radial and angular projections of $E = |E|e^{i\theta}$, that is, in the (θ, n)-plane and in the $(|E|, n)$-plane, respectively. In these projections one can clearly see the determining property of this orbit: it has a cubic tangency with the section Σ at $\mathbf{x}^*$, which is why we refer to this bifurcation as the *codimension-one cubic tangency bifurcation*.

What happens when one moves through the cubic tangency bifurcation for $K \approx 0.1140145$ is illustrated in Figure 14, where the green panels on the left show $\mathbb{T}_\Sigma$ in Σ and the gray panels on the left show $\mathbb{T}_\mathcal{C}$ in $\mathcal{C}$. Note that $\mathcal{C}$ has been "unrolled" and is shown in the (θ, n)-plane. Also shown are enlargements near the bifurcation point in Σ (green subpanel) and in $\mathcal{C}$ (gray subpanel), respectively; here $\mathbb{T}_\Sigma$ is plotted relative to the curve C, which appears as a straight line in the green subpanels. As is shown in the left column of Figure 14 and the associated green subpanels, the curve $\mathbb{T}_\Sigma$ moves relative to the curve C in the section Σ. At the same time we see in the right column of Figure 14 and the associated gray subpanels that $\mathbb{T}_\mathcal{C}$ moves relative to the curve $C = \Sigma \cap \mathcal{C}$ in the extended tangency locus $\mathcal{C}$. In Figure 14(a) there are two intersections between $\mathbb{T}_\Sigma$ and C and two between $\mathbb{T}_\mathcal{C}$ and C. Points on the part of $\mathbb{T}_\Sigma$ around γ_5 that lie outside C return to points on one of two separate arcs of $\mathbb{T}_\Sigma$, namely, the large arc inside C near γ_0 or the small piece between the two intersection points with C that is shown in the enlarged area of the section. Figure 14(b) is at the moment of cubic tangency bifurcation when the invariant manifold $\mathbb{T}_\Sigma$ has a quadratic tangency with the tangency locus C at the point $\mathbf{x}^*$. At the same time $\mathbb{T}_\mathcal{C}$ has a quadratic tangency with C at $\mathbf{x}^*$. Even though this bifurcation of the Poincaré map does not change the invariant curve $\mathbb{T}_\Sigma$, it does make a difference to the dynamics on the global section, because it changes the number of segments of $\mathbb{T}_\Sigma$ on either side of the tangency locus C. After the bifurcation, $\mathbb{T}_\Sigma$ and $\mathbb{T}_\mathcal{C}$ no longer intersect C; see Figure 14(c). This means that now the part of $\mathbb{T}_\Sigma$ around γ_5 that lies outside C returns to a single arc of $\mathbb{T}_\Sigma$ inside C.

It is generally quite difficult to find orbits with cubic tangencies with a section. Therefore, the insight that a cubic tangency bifurcation takes place when a two-dimensional invariant manifold $\mathbb{T}$ passes through a point in $C \cap D$ is very useful from a practical point of view (even though this is only a necessary and not a sufficient condition). Namely, the set $C \cap D$ consists generically of isolated points that can be calculated analytically. In particular, a cubic tangency bifurcation can occur only when $C \cap D \neq \emptyset$. One readily computes that for $|n| \leq 2k/\sqrt{\Delta(n)}$ the curves D and C intersect in two points, which are given by

$$C \cap D = \left(\frac{0.1}{2K}\Delta(n),\ \pm\sqrt{\Delta(n) - \frac{0.01}{4K^2}\Delta(n)^2},\ -0.1 \right) \approx \left(\frac{0.05035}{K},\ \pm\frac{1}{2}\sqrt{4.02808 - \frac{0.01014}{K^2}},\ -0.1 \right) \tag{5.5}$$

for the fixed values $\alpha = 2$, $B = 0.015$, $\gamma = 0.035$, and $\omega = 0.43$ used here, and for $s = s^* = -0.1$. Indeed, for $\Sigma_{-0.1}$ there are exactly two points in $C \cap D$. At the bifurcation for $K \approx 0.1140145$ we have $C \cap D \approx \{(0.44162, \pm 0.90111, -0.1)\}$, and $\mathbb{T}$ crosses the point $\mathbf{x}^* \approx (0.44162, 0.90111, -0.1)$; see Figures 13(a) and 14.

We finally remark that $C \equiv D$ is not generic in our context, but it occurs stably in Hamiltonian vector fields. In fact, this geometric situation was identified (but not in terms of C and D) by Birkhoff [3] as the one that allows the construction of a complete Poincaré section. Namely, for $C \equiv D$ the curve $C \subseteq \Sigma$ is invariant under the flow and orbits are spiraling around C.

5.2. A helical tube with a cubic tangency bifurcation. To get geometric insight into how a cubic tangency bifurcation can be "straightened out" to a normal form in the standard flowbox, we now consider a concrete geometric model of a simplified curved flow. To this end, we consider the family of sections

$$\Sigma_s = \{(x, y, z) \in \mathbb{R}^3 \mid z = s\} \tag{5.6}$$

in $\mathbb{R}^3$, where we further assume that the flow is such that

$$C = C(s) = \{(x, y, s) \in \Sigma_0 \mid y = 0\} \tag{5.7}$$

so that the flow points up for $y > 0$ and down for $y < 0$. Therefore,

$$\mathcal{C} = \{(x, y, z) \in \mathbb{R}^3 \mid y = 0\}. \tag{5.8}$$

To specify the flow further, we consider a helix Γ of radius 1 in $\mathbb{R}^3$ given by

$$\Gamma = \Gamma(\theta) = \{(x, y, z) \in \mathbb{R}^3 \mid x = \theta,\ y = \cos\theta,\ z = \sin\theta\}, \tag{5.9}$$

with period 2π as parameterized by $\theta \in \mathbb{R}$. We assume that the flow leaves Γ invariant. Notice that Γ spirals around C and intersects Σ in infinitely many points for $|s| < 1$.

The next step of the construction is to consider a one-parameter family of invariant tubes around Γ. To obtain a generic situation, the radius (around the respective points of Γ) of the tube must vary with the angular parameter θ. We consider here a standard tube that consists of circles of radius $k\,r(\theta)$ that lie in the plane spanned by the normal $N(\theta)$ and the binormal $B(\theta)$ of Γ at $x = \theta$. The parameter k is the "gross radius" in the sense that changing k changes the size of the tube but not its shape, which is given by $r(\theta)$; note that k plays the same role geometrically as the injection strength K in the laser system (4.5). We assume that $r(\theta)$ does not have extra periodicity or symmetries. Furthermore, $r(\theta)$ is such that there are no self-intersections or other bifurcations such as saddle transitions too near the bifurcation point. A function $r(\theta)$ that satisfies all conditions and the resulting formula of $\mathbb{T}$ can be found in Appendix A.

The helix Γ and the tube $\mathbb{T}$ are shown in Figure 15 for three different values of k. Also shown are the section $\Sigma = \Sigma_{-0.1804} = \{(x, y, z) \in \mathbb{R}^3 \mid z = -0.1804\}$ and the extended critical locus $\mathcal{C}$. The value $s = -0.1804$ for the height of the section Σ was determined so that there is a cubic tangency bifurcation for the fixed value $k = 0.2$; namely, it occurs at $\theta = 6.94193$. The intersection curves $\mathbb{T}_\Sigma$ and $\mathbb{T}_\mathcal{C}$ in Σ and $\mathcal{C}$, respectively, are shown in Figure 16. In Figures 15 and 16 the rows (a)–(c) show the situation before, at, and after the bifurcation. Figures 15 and 16 also show the critical locus D of $\mathcal{C}$. Since we do not specify an underlying flow, the curve D must also be constructed in a consistent way. In particular, D must go through the point of cubic tangency and be consistent with the positions of $\mathbb{T}_\mathcal{C}$. These requirements are met by D as defined in Appendix A, where the flow through $\mathcal{C}$ is in the direction of negative y above D and in the direction of positive y below D.

As can be seen from Figures 15 and 16, the planes Σ and $\mathcal{C}$, the tangency locus D, and the invariant objects Γ and $\mathbb{T}$ form a consistent geometrical model for the cubic tangency bifurcation. By consistency we mean here that the conditions we place on the unspecified

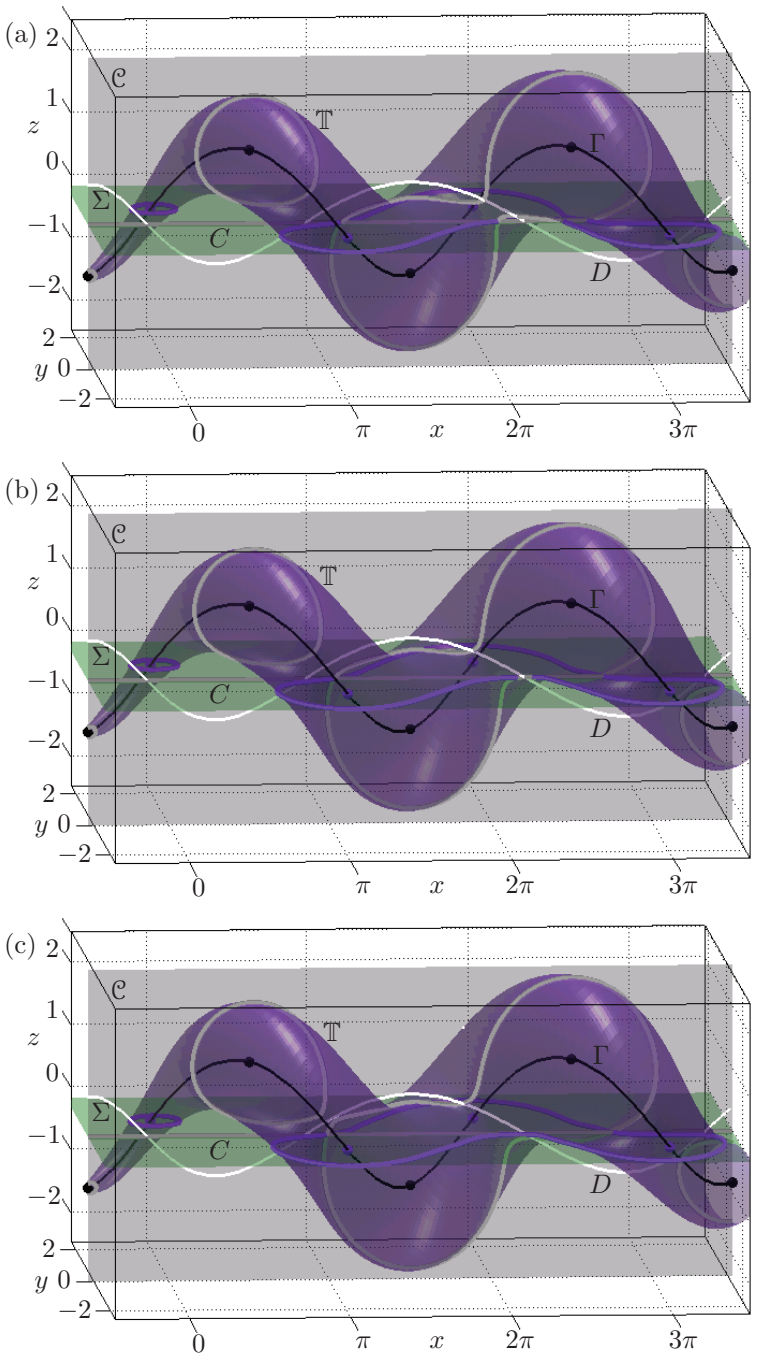


Figure 15. *Geometrical model of a cubic tangency bifurcation, consisting of a helical orbit Γ surrounded by a tube $\mathbb{T}$ of varying radius that spirals around the tangency locus C of a section Σ; also shown is the extended critical locus $\mathcal{C}$. Panels* (a)–(c) *are before, at, and after the bifurcation, namely, for $k = 0.19$, 0.2, and 0.21, while the section is given by $z = -0.1804$; see also Figure 16.*

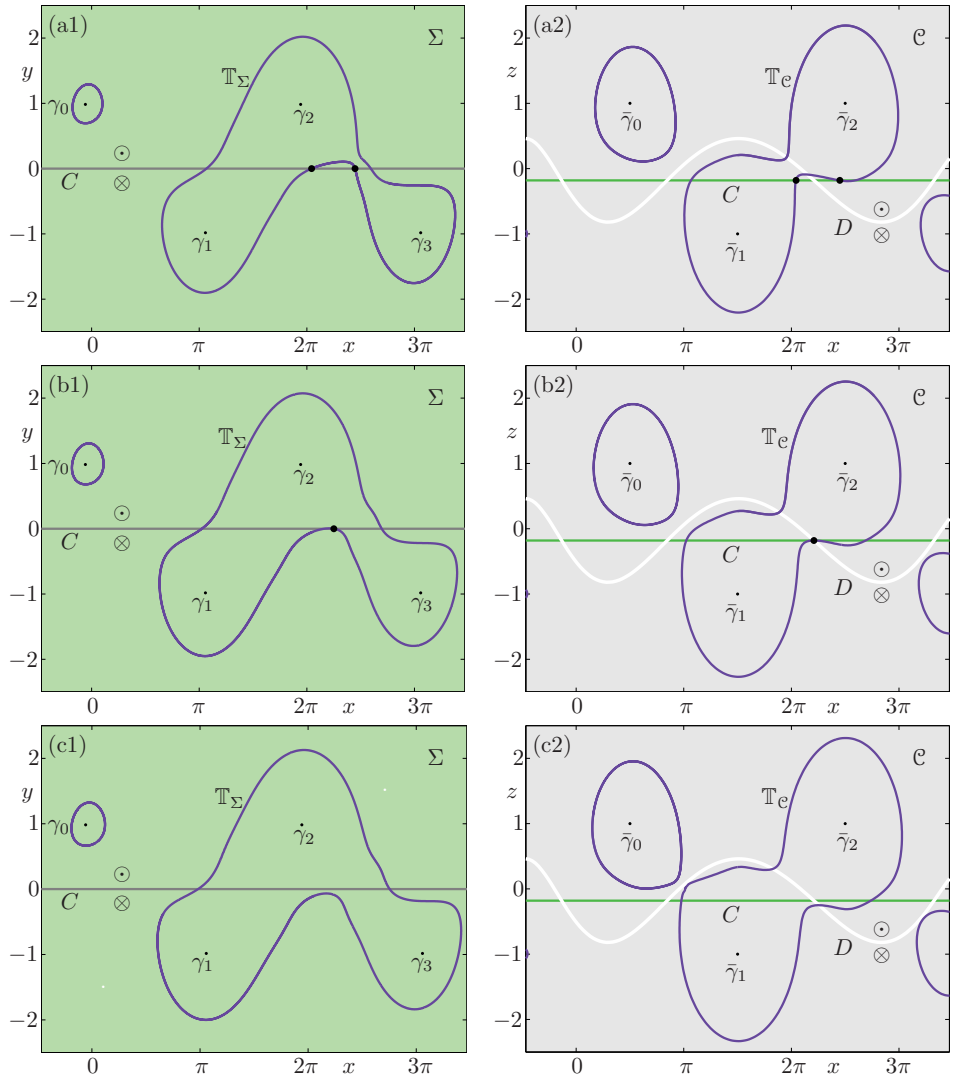


Figure 16. *The geometrical model of a cubic tangency bifurcation shown in the section* Σ *(left column) and in the plane* $\mathcal{C}$ *(right column). Panels* (a)–(c) *are before, at, and after the bifurcation, namely, for* $k = 0.19$, 0.2, *and* 0.21*; see also Figure* 15.

flow during the geometric construction are such that they can be realized by an actual flow; compare with Figure 14. As for the laser system in section 5.1, the bifurcation is unfolded by changing the gross radius of the invariant manifold $\mathbb{T}$, as specified by the parameter k in (A.1). (Note that the cubic tangency bifurcation could also be unfolded by moving the section, that is, by changing the parameter s in (5.6).) The key feature in the section Σ is the single closed component $\mathbb{T}_\Sigma$ that surrounds three points $\gamma_1, \gamma_2, \gamma_3 \in \Gamma_\Sigma$; see Figure 16 (left column). Before the bifurcation the closed component of $\mathbb{T}_\Sigma$ has four intersections with C; at the cubic tangency bifurcation the two innermost of them come together to a single point; and after the bifurcation there are only two intersection points (which is the minimal

number for a component of $\mathbb{T}_\Sigma$ that surrounds three points of Γ_Σ). In the plane $\mathcal{C}$ the object of interest is the single closed component $\mathbb{T}_\mathcal{C}$ that surrounds only two points $\overline{\gamma}_1, \overline{\gamma}_2 \in \Gamma_\mathcal{C}$; see Figure 16 (right column). Before the bifurcation the closed component of $\mathbb{T}_\mathcal{C}$ has four intersections with $\Sigma \cap \mathcal{C}$ (note that the situation in row (a) is close to a saddle transition), at the cubic tangency bifurcation the two innermost of them come together to a single point, and after the bifurcation there are only two intersection points (which is again the minimal number for a component of $\mathbb{T}_\mathcal{C}$ that surrounds two points of $\Gamma_\mathcal{C}$).

5.3. Normal form of the cubic tangency bifurcation. To obtain the normal form of the cubic tangency bifurcation in a three-dimensional flowbox we consider a small neighborhood of the bifurcation in phase space. The idea is then to "untwist" the helical structure of the flow, which means that the section Σ and the extended critical locus $\mathcal{C}$ deform to smooth surfaces. The key realization is that, owing to the cubic tangency of the orbit involved, in the normal form one needs to consider a section Σ that has a cusp singularity when projected along the orbits onto the in-set I (or the out-set O). The result in the absence of an invariant manifold was proved by Sotomayor and Teixeira [38, Figure 5.2] in the context of vector fields on a three-dimensional manifold with a codimension-one boundary. It can be phrased as follows in the present context.

Proposition 5.1. *In any sufficiently small flowbox the phase portrait near a cubic tangency of an orbit of a flow in $\mathbb{R}^3$ with a global section is topologically equivalent to the phase portrait in the standard flowbox (2.7) for $n = 3$ given by the section*

$$\Sigma = \{(u, v_1, v_2) \in \mathbb{R}^3 \mid v_1\, u - u^3 + v_2 = 0\}. \tag{5.10}$$

In order to understand the geometry of the flow in the standard flowbox and to help make the link with the previous sections, we now construct the extended tangency locus $\mathcal{C}$. Since $\vec{n}_\Sigma((0,0,0)) = (0,0,1)$, we embed the section Σ of (5.10) into the one-parameter family of sections

$$\Sigma_s = \{(u, v_1, v_2 + s) \in \mathbb{R}^3 \mid v_1\, u - u^3 + v_2 = 0\} \tag{5.11}$$

so that $\Sigma = \Sigma_0$. One readily finds that the tangency locus of Σ_s is

$$C = C(s) = \{(u, v_1, v_2 + s) \in \Sigma_s \mid v_1 = 3u^2,\ v_2 = -2u^3\}. \tag{5.12}$$

The projection $\widehat{C}$ of C onto the in-set I along the u-direction (the direction of the flow) has two branches that meet at a cusp point at $(v_1, v_2) = (0, s) \in I$. It follows that the extended critical locus $\mathcal{C}$ of the family Σ_s is the parabolic surface

$$\mathcal{C} = \{(u, v_1, v_2) \in \mathbb{R}^3 \mid v_1 = 3u^2\}. \tag{5.13}$$

Furthermore, the flow is tangent to $\mathcal{C}$ along the tangency locus

$$D = \{(u, v_1, v_2) \in \mathcal{C} \mid u = 0,\ v_1 = 0\},$$

where the flow is directed into the parabolic cylinder bounded by $\mathcal{C}$ for $u < 0$ and out of the parabolic cylinder for $u > 0$.

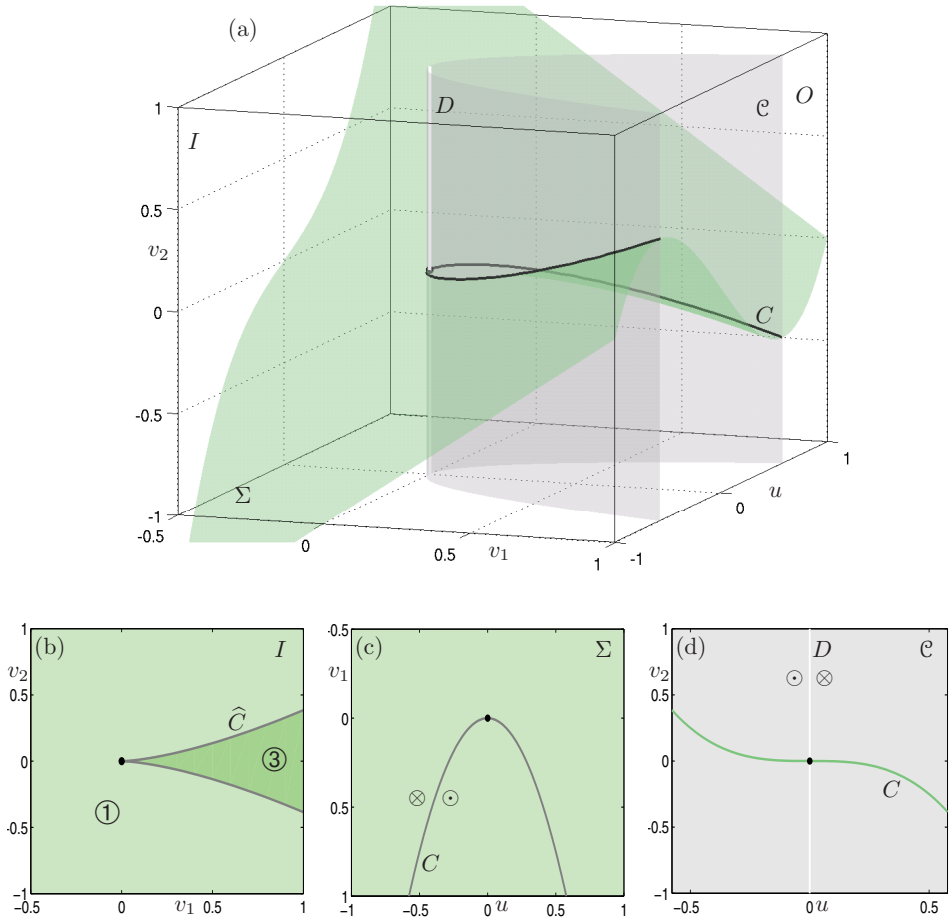


Figure 17. *Panel* (a) *depicts the cuspidal section* $\Sigma = \Sigma_0$ *(green) of* (5.11) *and the extended critical locus* $\mathcal{E}$ *inside the standard flowbox. Panel* (b) *shows the projection of* C *and* Σ *onto the in-set* I*; the symbols* ① *and* ③ *indicate the regions with one and three intersections with* Σ*, respectively. Panel* (c) *shows the curve* C *in* Σ*, and panel* (d) *the curves* C *and* D*. The direction of the flow is indicated by the symbols* $\odot$ *and* $\otimes$*.*

Figure 17(a) shows the section $\Sigma = \Sigma_0$ and the extended critical locus $\mathcal{E}$ inside the standard flowbox. Figures 17(b)–(d) show respective images on I, Σ, and $\mathcal{E}$. The projection curve $\widehat{C}$ divides the in-set I into two regions, labeled ① and ③ in Figure 17(b). Any orbit starting inside region ③ of I has three intersections with Σ while it "winds around" the curve C in (u, v_1, v_2)-space. (Rather, in the flowbox C winds around the straight orbit segment.) Similarly, orbits starting in region ① intersect Σ only once. Moreover, any orbit with $v_1 > 0$ intersects the parabolic surface $\mathcal{E}$ twice; compare with Figure 17(d). In other words, the geometry shown in Figure 17 is indeed topologically equivalent to that near a cubic tangency as discussed in sections 5.1 and 5.2.

We now present the normal form of the cubic tangency bifurcation. This bifurcation is of codimension one under the genericity conditions that the invariant manifold is in general position with respect to the section Σ at the moment of cubic tangency, the dependence on

the parameter is smooth, and the manifold crosses the section with positive speed.

Proposition 5.2. *In any sufficiently small flowbox, the unfolding of a cubic tangency of a two-dimensional invariant manifold of a flow in* $\mathbb{R}^3$ *with a global section is topologically equivalent to the unfolding in the standard flowbox* (2.7) *for* $n = 3$ *given by the one-parameter family of sections defined in* (5.11)*, where the invariant manifold is the plane*

$$M = \{(u, v_1, v_2) \mid v_1 = -v_2\}. \tag{5.14}$$

Proof. Suppose that the unfolding parameter of the quadratic tangency is η and the bifurcation takes place for $\eta = 0$. According to Proposition 5.1, any phase portrait of the unfolding in a flowbox near the cubic tangency point is topologically equivalent to that given by $\Sigma = \Sigma_0$ in the standard flowbox (2.7) for $n = 3$. Therefore, the invariant manifold is mapped to a surface $\widetilde{M}(\eta)$. The curve $\widetilde{M}(\eta) \cap I$ is (locally and for sufficiently small η) given by a function $\mu_\eta : v_1 \to v_2$ on the in-set I. Since $\widetilde{M}(0)$ is in general position, we have that $\frac{d\mu}{d\eta}(0) \neq 0$, meaning that the tangent vector to $\widetilde{M}(0) \cap I$ at the origin $(v_1, v_2) = (0, 0) \in I$ has both a v_1- and a v_2-component. Due to continuity on η, the same is true for $\widetilde{M}(\eta) \cap I$ for sufficiently small η. Therefore, the curve $\widetilde{M}(\eta) \cap I$ is locally a graph over the antidiagonal $(v_1, -v_1)$ of I, and it intersects the v_2-axis of I at a well-defined height $\mu_\eta(0) = s(\eta)$, where $s(0) = 0$. The u-independent coordinate change

$$(v_1, v_2) \mapsto (v_1, v_2 - v_1 - \mu(v_1))$$

maps $\widetilde{M}(\eta)$ to M as given by (5.14). Due to genericity of the dependence of $\widetilde{M}$ on η, we have that $\frac{d\mu}{d\eta}(0) = \frac{ds}{d\eta} \neq 0$, which implies that the image $\widetilde{\Sigma}(\eta)$ of Σ under this coordinate change is a surface with a generic cusp at $(u, v_1, v_2) = (0, 0, -s(\eta))$. As in the proof of Proposition 4.2, the surface $\widetilde{\Sigma}(\eta)$ can be brought to the required form (5.11) by a u-independent coordinate change that leaves M invariant. The overall η-family of coordinate changes is continuous by the genericity assumption on η. ∎

The important realization is that the manifold M of the normal form in the standard flowbox needs to be in general position relative to the cusp surface, which means that it must have nonnegative components in both the v_1- and the v_2-directions. This is ensured by the choice of the "diagonal" manifold M defined in (5.14). Note that a horizontal manifold (given by $v_2 = 0$) would always intersect the curve C in exactly one point, while a vertical manifold (given by $v_1 = 0$) would always intersect C at the cusp point. Both situations are not generic. Furthermore, for generic M as given in the normal form by (5.14), the cubic tangency bifurcation can alternatively be unfolded by moving the manifold M up and down. This is exactly the mechanism leading to a cubic tangency of an invariant torus when its gross radius changes as in sections 5.1 and 5.2.

In Figure 18 the unfolding given by Proposition 5.2 is presented in (u, v_1, v_2)-space, and in Figure 19 on the in-set I, the section Σ_s, and the extended critical locus $\mathcal{C}$; see also the accompanying animation (69972_03.gif [21.3MB]). Before the cubic tangency bifurcation, the invariant manifold M_Σ in Σ_s has two intersections with the curve C. Similarly, the intersection curve $M_\mathcal{C}$ in $\mathcal{C}$ intersects C twice. At the moment of bifurcation both M_Σ and $M_\mathcal{C}$ have quadratic tangencies with C in Σ_s and $\mathcal{C}$, respectively. Notice that these tangencies

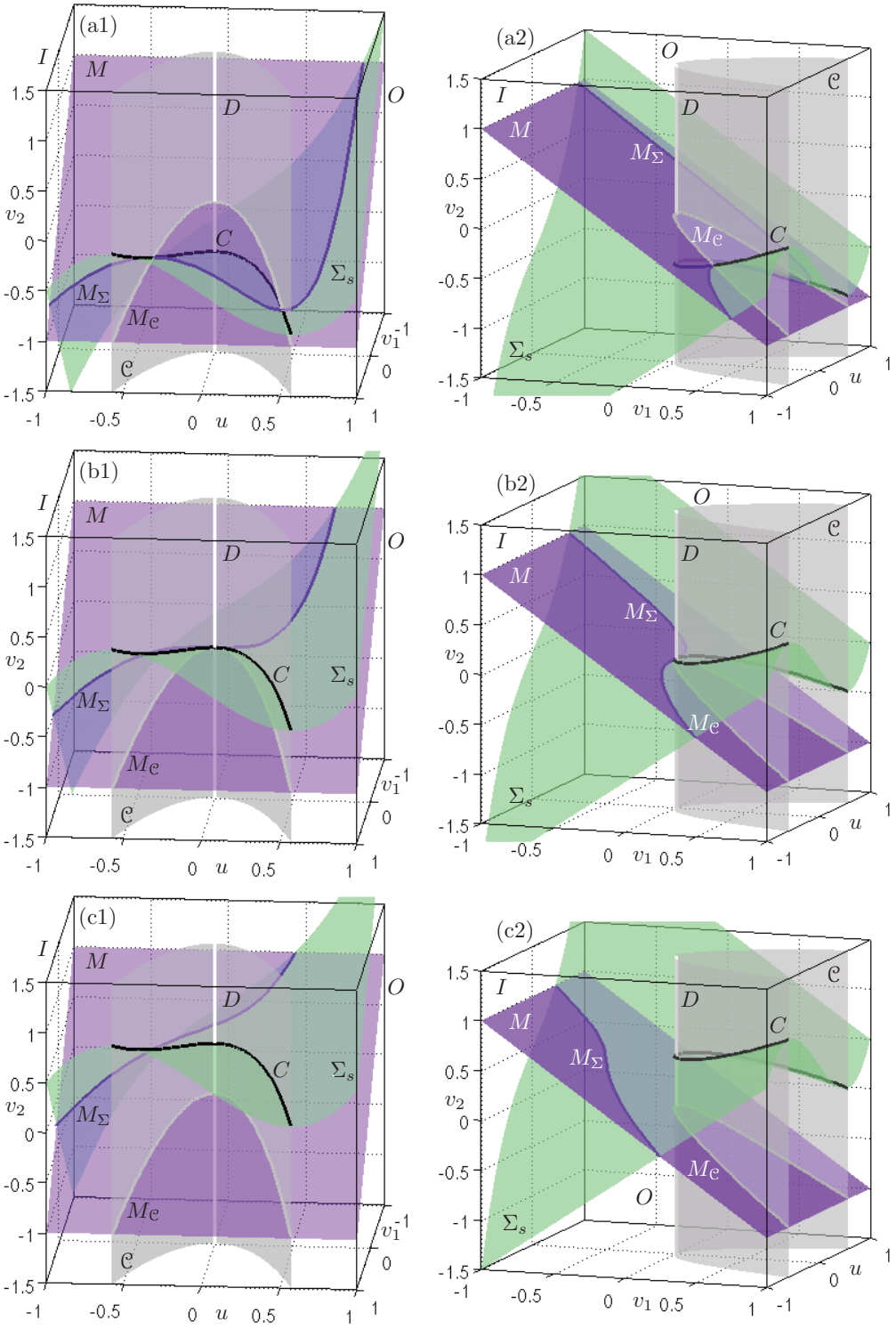


Figure 18. *Unfolding of the cubic tangency bifurcation in normal form in the standard flowbox, presented from two different viewpoints. Rows* (a)–(c) *show the invariant manifold* M *(purple), the section* Σ_s *(green), and the extended critical locus* $\mathcal{C}$ *(gray) before, at, and after the bifurcation. Also shown are the tangency loci* C *and* D*. From* (a)–(c) $s = -0.5$, $s = 0$, *and* $s = 0.5$*; see also Figure* 19 *and animation* (69972_03.gif [21.3MB]).

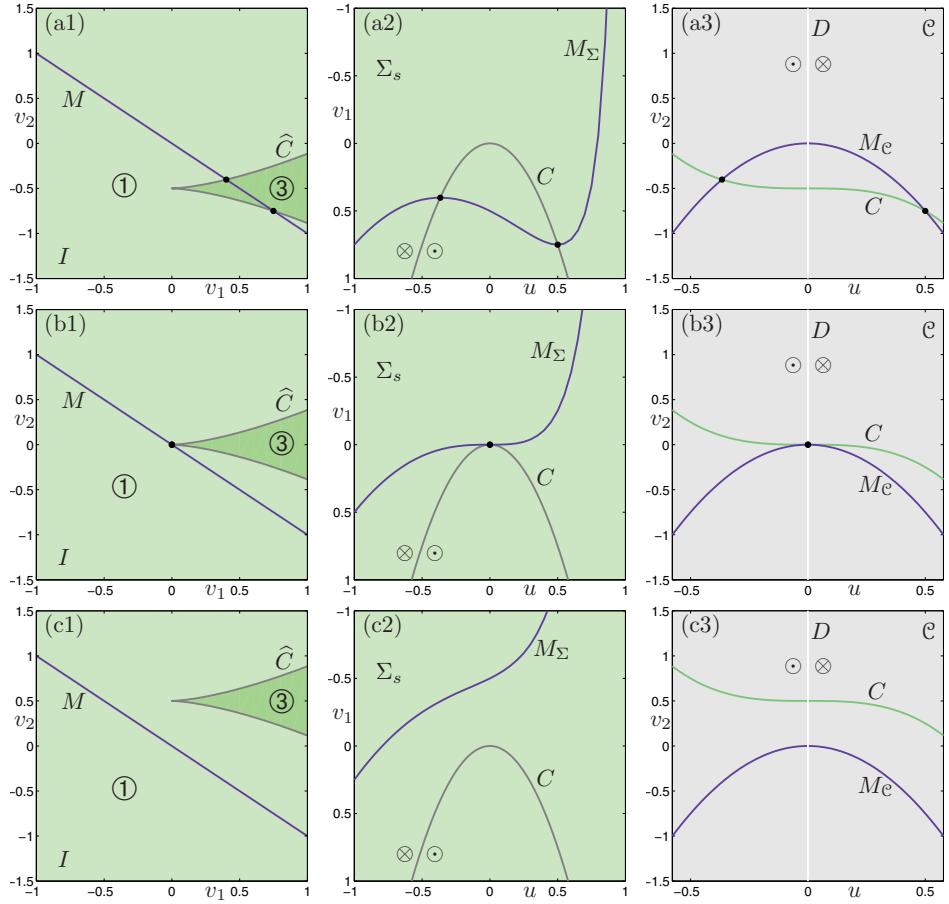


Figure 19. *Unfolding of the cubic tangency bifurcation projected onto the in-set I (left column), on the section Σ_s (middle column), and on the extended critical locus $\mathcal{C}$ (right column); compare with Figure 17. From* (a)–(c) *$s = -0.5$, $s = 0$, and $s = 0.5$; see also Figure 18 and animation* (69972_03.gif [21.3MB]).

indeed occur at the point $C \cap D$ as noted previously. After the bifurcation, neither M_Σ nor $M_{\mathcal{C}}$ intersect C. The comparison of Figure 19 with the corresponding Figure 14 of a cubic tangency bifurcation in the semiconductor laser system (4.5) demonstrates how the unfolding manifests itself in a concrete example.

6. Tangency bifurcations in higher dimension and of higher codimension. The unfoldings in the previous sections of the codimension-one tangency bifurcations of a manifold M with a global section Σ of flows in $\mathbb{R}^2$ and $\mathbb{R}^3$ give only a hint of the many possibilities for tangency bifurcations of a fixed codimension in $\mathbb{R}^n$ for any n. The key realization is that there are two "sources" of codimension: the order of contact with Σ of the orbit $\mathcal{O}(\mathbf{x}^*)$ of the tangency point $\mathbf{x}^*$ and possible codimension associated with tangencies of M and Σ at $\mathbf{x}^*$ in other directions. The first example of a tangency bifurcation with both types of tangencies is the codimension-one quadratic tangency in $\mathbb{R}^3$ in section 4.

Classifying and unfolding tangency bifurcations of higher codimension and for $n > 3$ is beyond the scope of this paper. However, we now give a general framework for this task, which

is based on a more detailed consideration of the intersection of the tangent spaces $T_M(\mathbf{x}^*)$ and $T_\Sigma(\mathbf{x}^*)$.

Definition 6.1. *Let M be an invariant manifold of dimension ℓ of a vector field f on $\mathbb{R}^n$ with a given planar $(n-1)$-dimensional global section Σ. Suppose that the following conditions are satisfied:*

(B1) *There is a point $\mathbf{x}^* \in M \cap \Sigma$ such that the orbit $\mathcal{O}(\mathbf{x}^*)$ has a tangency of degree $d \in \mathbb{N}$ with Σ at $\mathbf{x}^*$, where we assume that the tangency is at least quadratic, that is, $d \geq 2$.*

(B2) *The dimension of the orthogonal complement N of $f(\mathbf{x}^*)$ in $T_M(\mathbf{x}^*) \cap T_\Sigma(\mathbf{x}^*)$ is p.*

(B3) *The point $\mathbf{x}^*$ is a critical point of codimension q of the restriction $\vartheta|_N$ to N of the local chart $\vartheta : T_M(\mathbf{x}^*) \to \mathbb{R}^n$ of the manifold M at $\mathbf{x}^*$.*

Then we say that M and Σ have a d-tangency with singularity dimension p and singularity codimension q *at $\mathbf{x}^*$ (or d-p-q-tangency for short).*

The singularity codimension q is defined only for $p > 0$, meaning that $p = 0$ if and only if $q = 0$. Furthermore, by construction of N there must be a tangency (quadratic or of higher degree) along each of the base vectors of N so that $q \geq p$. From the data specified in Definition 6.1 one can determine the overall codimension.

Proposition 6.2. *The codimension c of a d-p-q-tangency of an ℓ-dimensional invariant manifold $M \subset \mathbb{R}^n$ with a global section is $c = d + q - \ell$, where $p < \ell < n$.*

Proof. For M and Σ in general position, $\dim(\Sigma \cap M) = \ell - 1$, independently of n. Furthermore, by (B1) and (B3) the point $\mathbf{x}^*$ is a critical point of codimension $(d-1)+q$ of the restriction of the chart ϑ to $f(\mathbf{x}^*) \oplus N = T_M(\mathbf{x}^*) \cap T_\Sigma(\mathbf{x}^*)$. Therefore, the codimension of the bifurcation is $c = (d-1) + q - (l-1)$. ■

Proposition 6.2 has some interesting immediate consequences. First, it shows that we indeed presented all codimension-one d-p-q-tangency bifurcations for $n \leq 3$; namely,

1. the 2-0-0-tangency for $\ell = 1$ is the quadratic tangency of a one-dimensional manifold in $\mathbb{R}^2$ and in $\mathbb{R}^3$ of Proposition 3.2 and Corollary 4.1, respectively;
2. the 2-1-1-tangency for $\ell = 2$ is the quadratic tangency of a two-dimensional manifold in $\mathbb{R}^3$ of Proposition 4.2; and
3. the 3-0-0-tangency for $n = 3$ and $\ell = 2$ is the cubic tangency of a two-dimensional manifold in $\mathbb{R}^3$ of Proposition 5.2.

Similarly, we can list all codimension-two tangency bifurcations for $n \leq 3$, which are as follows:

1. the 3-0-0-tangency for $\ell = 1$ is the cubic tangency of a one-dimensional manifold in $\mathbb{R}^2$ and in $\mathbb{R}^3$;
2. the 2-1-2-tangency for $\ell = 2$ is the quadratic tangency of a two-dimensional manifold with a cubic tangency in the one-dimensional N-direction;
3. the 3-1-1-tangency for $\ell = 2$ is the cubic tangency of a two-dimensional manifold with a quadratic tangency in the one-dimensional N-direction; and
4. the 4-0-0-tangency for $\ell = 2$ is a tangency of order four of an orbit on a two-dimensional manifold.

An important new element of codimension-two tangency bifurcations is that they cannot be unfolded solely by moving the section Σ (because Σ is of codimension one in $\mathbb{R}^n$). Of the cases above, only the unfolding for $\ell = 1$, that is, the codimension-two cubic 3-0-0-tangency, is straightforward.

Proposition 6.3. *In any sufficiently small flowbox, the unfolding of a cubic tangency of*

a one-dimensional invariant manifold of a flow in $\mathbb{R}^3$ *with a global section is topologically equivalent to the unfolding in the standard flowbox* (2.7) *for* $n = 3$ *given by the one-parameter family of sections* (5.11) *that interacts with the one-parameter family of manifolds*

$$M = \{(u, v_1, v_2) \mid v_1 = \lambda \text{ and } v_2 = 0\}.$$

This result follows from Proposition 5.1 in the same way that Proposition 3.2 follows from Proposition 3.1, namely, by the construction of a parameter-dependent coordinate change. Proposition 6.3 says that the unfolding of the codimension-two cubic tangency of a one-dimensional manifold in $\mathbb{R}^3$ (and, therefore, in all dimensions $n \geq 2$ and including $n = 2$) is given by the standard unfolding of a cusp bifurcation. Namely, the phase portrait is determined by the relative position of the point $M_I = (0, \lambda)$ in the in-set I relative to projection $\widehat{C}(s)$ of the fold curve; see Figure 17. Crossing one of the two branches of $\widehat{C}(s)$ corresponds to a codimension-one quadratic tangency that unfolds as given by Corollary 4.1. At the central codimension-two point the manifold passes exactly through the cusp point in I.

Suppose now that M is actually a segment of a periodic orbit Γ, meaning that there is a map R from the out-set O back to the in-set I that leaves M invariant. Then the number of intersections of Γ with Σ changes by two when Γ crosses C, even if this happens at the cusp point.

7. Conclusions and discussion. We considered the class of tangency bifurcations, which are bifurcations that one finds generically in global Poincaré maps but that are not due to bifurcations of the underlying vector field. Tangency bifurcations involve the interaction of an invariant manifold with the tangency locus of the section, which is nonempty unless the system is effectively periodically forced. Specifically, we presented a complete treatment of codimension-one tangency bifurcations of flows in $\mathbb{R}^2$ and $\mathbb{R}^3$, that is, for one-dimensional and two-dimensional global sections. We presented their normal forms in the standard flowbox by specifying suitable families of curved smooth sections that interact with straight invariant manifolds of appropriate dimension. Our approach is similar in spirit to that taken by Sotomayor and Teixeira, who considered flows on manifolds with one- and two-dimensional boundaries. The additional ingredients needed here are further coordinate transformations to "straighten out" the invariant manifold but such that the flowlines in the flowbox remain unchanged.

With the examples of the two-dimensional unforced Van der Pol oscillator and a three-dimensional model of a semiconductor laser with optical injection, we demonstrated how the codimension-one tangency bifurcations manifest themselves in a specific vector field. Namely, we studied the interactions of periodic orbits and invariant tori with the tangency locus of a given planar section. As is the case generically, the respective bifurcations of the Poincaré map can be brought about by either changing some system parameter or moving the section. The quadratic tangencies of one- or two-dimensional invariant manifolds with a planar section could be associated with the respective normal forms in the flowbox in a relatively straightforward manner. In the case of a cubic tangency, on the other hand, the operation of "straightening out" the flow to obtain the normal form in the flowbox is quite complicated. Therefore, a simplified geometric model at an intermediate step was constructed to help understand the normal-form transformation geometrically.

As we have shown for the concrete example of an attracting periodic orbit, a tangency bifurcation may give rise to a global topological change of the corresponding Poincaré map. Here, the properties of the map from the out-set back to the in-set become important. Similarly, one can study the consequences of tangency bifurcations of two-dimensional Poincaré maps in three-dimensional vector fields. However, there are more possibilities for the map from the out-set back to the in-set, because it is now two-dimensional as well. One example would be that the map has a repelling fixed point surrounded by an invariant curve, which would correspond to the case that the invariant manifold inside the flowbox is part of an invariant torus. Section 4.1 hints at bifurcations of the two-dimensional Poincaré map in the case that the flow on the torus is rational or irrational. However, the situation may be quite complicated if there are equilibria on the smooth torus. The classical example is the Cherry flow [8, 23], which gives rise to discontinuity of the Poincaré map on the invariant curve; in fact, the points where all iterates of this Poincaré map are defined form a Cantor set. A discussion of topological changes of two-dimensional Poincaré maps is left for future work.

The unfoldings presented here were shown to fit naturally into a general framework for the classification of tangency bifurcations of arbitrary codimension. The general idea is that a tangency of degree $d \geq 2$ of an orbit on the manifold is accompanied by other possible tangencies of the manifold in the directions normal to this orbit. This point of view provides a clear direction for the future study of tangency bifurcations in higher-dimensional spaces and of higher codimension. We already listed the codimension-two tangency bifurcations of a two-dimensional manifold in $\mathbb{R}^3$, and the construction of their unfoldings is an interesting challenge. Another important next step and the subject of our ongoing research is the study of tangency bifurcations in $\mathbb{R}^4$. This study starts with the tangency bifurcations of codimension one, as they are encountered naturally when a single parameter of the vector field is changed, or the section is moved. The normal forms that need to be developed involve intersections of hypersurfaces in $\mathbb{R}^4$, which is very hard to imagine and visualize. This difficulty can be overcome by considering the corresponding surfaces in the three-dimensional in-set of the flow box. Note that it will be a real challenge to identify and visualize tangency bifurcations in three-dimensional sections in concrete vector fields arising from applications.

Another topic of ongoing research is the classification of different mechanisms in which points or regions are created in the global section for which the flow never returns back to the section. In other words, the issue is to find bifurcations that create new sets of the section where the Poincaré map is not defined. As we already demonstrated in section 3.3, such bifurcations may be due to interactions between the global section and stable and unstable manifolds of equilibria and other invariant sets that do not lie in the section. Already for the case of three-dimensional flows, the study of these bifurcations in applications requires the use of numerical techniques for the computation of two-dimensional global invariant manifolds, such as those in [13, 24, 25].

Appendix A. Details of the helical tube construction. A standard tube of radius $k\,r(\theta)$ around the helix (5.9) is given by

$$\mathbb{T} = \{\Gamma(\theta) + k\,r(\theta)\,(N(\theta)\cos\phi + B(\theta)\sin\phi) \quad \text{for } \theta \in \mathbb{R},\ 0 \leq \phi < 2\pi\}\,, \tag{A.1}$$

where $N(\theta)$ is the normal and $B(\theta)$ the binormal at $x = \theta$. From (5.9) it follows that the

tangent, normal, and binormal are

$$T = \left(\frac{1}{\sqrt{2}}, -\frac{\sin\theta}{\sqrt{2}}, \frac{\cos\theta}{\sqrt{2}}\right), \quad N = (0, \cos\theta, \sin\theta), \quad B = \left(\frac{1}{\sqrt{2}}, \frac{\sin\theta}{\sqrt{2}}, -\frac{\cos\theta}{\sqrt{2}}\right).$$

Therefore, the helical tube (A.1) takes the form

$$\mathbb{T} = \left\{(x,y,z) \in \mathbb{R}^3 \mid x = \theta + k\, r(\theta)\frac{\sin\phi}{\sqrt{2}},\ y = \cos\theta + k\, r(\theta)\left(\cos\theta\cos\phi + \frac{\sin\theta\sin\phi}{\sqrt{2}}\right),\right.$$
$$\left. z = \sin\theta + k\, r(\theta)\left(\sin\theta\cos\phi - \frac{\cos\theta\sin\phi}{\sqrt{2}}\right) \text{ for } \theta \in \mathbb{R},\ 0 \le \phi < 2\pi\right\}.$$

A function $r(\theta)$ that satisfies the requirements mentioned in section 5.2 is

$$r(\theta) = \left(\sin^2\theta - \left(\frac{2\theta - 3\pi}{2\pi}\right)^2 - \frac{1}{2}\left(\frac{4\theta - 14\pi}{7\pi}\right)^2 + \frac{4}{5}\exp\left(\frac{11\pi}{4} - \theta\right) + 6\right).$$

For $K = 0.2$ we find a cubic tangency bifurcation with the plane $\Sigma = \Sigma_{-0.1804}$ at $\mathbf{x}^* \approx (6.94193, 0, -0.1804) \in C \subset \Sigma$. The curve $D \subset \mathcal{C}$ can now be constructed by observing the condition that it goes through $\mathbf{x}^*$ and is in general position with respect to the components of $\mathbb{T}_{\mathcal{C}}$. The definition

$$D = D(\theta) = \left\{(\theta, 0, z) \in \mathcal{C} \mid z = -0.64\sin\left(\frac{16\theta\pi^2 - 111.0709\pi^2}{-0.34\,(2\theta - 3\pi)^2 + 22.4\pi^2}\right) + 0.1804\right\}$$

meets these requirements, as can be seen from Figure 16. Note that the choice of $r(\theta)$ and D are by no means unique. The formulae presented here are indeed quite involved, but they allow one to compute all objects of interest while being consistent with the existence of an underlying flow.

Acknowledgments. The authors thank M. A. Teixeira and J. Sotomayor for helpful discussions.

REFERENCES

[1] H. D. I. Abarbanel, L. Korzinov, A. I. Mees, and N. F. Rulkov, *Small force control of nonlinear systems to given orbits*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44 (1997), pp. 1018–1023.
[2] V. I. Arnol′d, *Catastrophe Theory*, 3rd ed., Springer-Verlag, New York, 1992.
[3] G. D. Birkhoff, *Dynamical systems with two degrees of freedom*, Trans. Amer. Math. Soc., 18 (1917), pp. 199–300.
[4] A. Bolsinov, H. R. Dullin, and A. Wittek, *Topology of energy surfaces and existence of transversal Poincaré sections*, J. Phys. A, 29 (1996), pp. 4977–4985.
[5] C. Bonatti and M. A. Teixeira, *Topological equivalence of diffeomorphisms and curves*, J. Differential Equations, 118 (1995), pp. 371–379.
[6] H. W. Broer, I. Hoveijn, M. van Noort, C. Simó, and G. Vegter, *The parametrically forced pendulum: A case study in* 1 1/2 *degree of freedom*, J. Dynam. Differential Equations, 16 (2004), pp. 897–947.
[7] W.-C. C. Chan, *A note on diffeomorphism and periodic differential equations*, Math. Comput. Modelling, 36 (2002), pp. 1387–1392.

[8] T. M. Cherry, *Analytic quasi-periodic curves of discontinuous type on a torus*, Proc. London Math. Soc. (2), 44 (1938), pp. 175–215.
[9] D. R. J. Chillingworth, *Discontinuity geometry for an impact oscillator*, Dyn. Syst., 17 (2002), pp. 389–420.
[10] P. Collins and B. Krauskopf, *Entropy and bifurcations in a chaotic laser*, Phys. Rev. E (3), 66 (2002), 056201.
[11] H. R. Dullin and A. Wittek, *Complete Poincaré sections and tangent sets*, J. Phys. A, 28 (1995), pp. 7157–7180.
[12] J. P. England, B. Krauskopf, and H. M. Osinga, *Bifurcations of stable sets in noninvertible planar maps*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 891–904.
[13] J. P. England, B. Krauskopf, and H. M. Osinga, *Computing one-dimensional global manifolds of Poincaré maps by continuation*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 1008–1041.
[14] G. O. Fountain, D. V. Khakhar, and J. M. Ottino, *Visualization of three-dimensional chaos*, Science, 281 (1998), pp. 683–686.
[15] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory, Vol.* 1, Springer-Verlag, New York, 1985.
[16] K. Green and B. Krauskopf, *Global bifurcations and bistability at the locking boundaries of a semiconductor laser with phase-conjugate feedback*, Phys. Rev. E (3), 66 (2002), 016220.
[17] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[18] I. Gumowski and C. Mira, *Recurrences and Discrete Dynamic Systems*, Springer-Verlag, New York, 1980.
[19] J. Hale and H. Koçak, *Dynamics and Bifurcations*, Springer-Verlag, New York, 1991.
[20] M. W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.
[21] C. Jung, T. H. Seligman, and J. M. Torres, *Canonically transformed detectors applied to the classical inverse scattering problem*, J. Nonlinear Math. Phys., 12 (2005), pp. 404–411.
[22] W. Just and H. Kantz, *Some considerations on Poincaré maps for chaotic flows*, J. Phys. A, 33 (2000), pp. 163–170.
[23] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.
[24] B. Krauskopf and H. M. Osinga, *Computing geodesic level sets on global (un)stable manifolds of vector fields*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 546–569.
[25] B. Krauskopf, H. M. Osinga, E. J. Doedel, M. E. Henderson, J. Guckenheimer, A. Vladimirsky, M. Dellnitz, and O. Junge, *A survey of methods for computing (un)stable manifolds of vector fields*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 763–791.
[26] B. Krauskopf, H. M. Osinga, and B. B. Peckham, *Unfolding the cusp-cusp bifurcation of planar endomorphisms*, SIAM J. Appl. Dyn. Syst., 6 (2007), pp. 403–440.
[27] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 2004.
[28] E. H. Lockwood, *A Book on Curves*, Cambridge University Press, Cambridge, UK, 1961.
[29] S. Mancini, M. A. S. Ruas, and M. A. Teixeira, , *On divergent diagrams of finite codimension*, Port. Math. (N.S.), 59 (2002), pp. 179–194.
[30] Z. Nitecki, *Differentiable Dynamics*, MIT Press, Cambridge, MA, 1971.
[31] J. Palis and W. de Melo, *Geometric Theory of Dynamical Systems*, Springer-Verlag, New York, 1982.
[32] R. Peikert and F. Sadlo, *Visualization methods for vortex rings and vortex breakdown bubbles*, in Proceedings of the Eurographics/IEEE VGTC Symposium on Visualization, The Eurographics Association, Norrköping, Sweden, 2007, pp. 211–218.
[33] R. Peikert and F. Sadlo, *Flow topology beyond skeletons: Visualization of features in recirculating flow*, in Topology-Based Methods in Visualization, Springer-Verlag, New York, to appear.
[34] P. B. Percell, *Structural stability on manifolds with boundary*, Topology, 12 (1973), pp. 123–144.
[35] H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste*, Gauthier-Villars, Paris, 1892–1899.
[36] H. Poincaré, *New Methods of Celestial Mechanics*, American Institute of Physics, New York, 1993.
[37] M. G. Rosenblum, A. S. Pikovsky, and J. Kurths, *Phase synchronization of chaotic oscillators*, Phys. Rev. Lett., 76 (1996), pp. 1804–1807.

[38] J. Sotomayor and M. A. Teixeira, *Vector fields near the boundary of a* 3*-manifold*, in Dynamical Systems (Valparaiso, 1986), Lecture Notes in Math. 1331, Springer-Verlag, New York, 1988, pp. 169–195.
[39] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, Addison-Wesley, Reading, MA, 1994.
[40] M. A. Teixeira, *Generic bifurcation in manifolds with boundary*, J. Differential Equations, 25 (1977), pp. 65–89.
[41] M. A. Teixeira, *Generic bifurcation of sliding vector fields*, J. Math. Anal. Appl., 176 (1993), pp. 436–457.
[42] M. A. Teixeira, *Singularities of reversible vector fields*, Phys. D, 100 (1997), pp. 101–118.
[43] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics*, Chapman & Hall, London, 2002.
[44] S. Wieczorek, B. Krauskopf, T. B. Simpson, and D. Lenstra, *The dynamical complexity of optically injected semiconductor lasers*, Phys. Rep., 416 (2005), pp. 1–128.
[45] X.-S. Yang, *A remark on global Poincaré section and suspension manifold*, Chaos Solitons Fractals, 11 (2000), pp. 2157–2159.
[46] E. C. Zeeman, *Catastrophe Theory, Selected Papers* 1972–1977, Addison-Wesley, Reading, MA, 1977.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 755–765
© 2008 Society for Industrial and Applied Mathematics

Synchronous Chaos in Coupled Map Lattices with General Connectivity Topology*

Jonq Juang† and Yu-Hao Liang†

Abstract. The purpose of the paper is to address the synchronous chaos in coupled map lattices with general connectivity topology. Our main results contain the following. First, the master stability functions also hold for general connectivity topology with coupling through a nonlinear function that needs to be exactly the individual chaotic map. Second, the synchronization curve, composed of pieces of transverse Lyapunov exponent curves, is constructed. Third, necessary and sufficient conditions on coupling strength for yielding the synchronous chaos of the system are given. Moreover, the coupling strength d_c giving the fastest convergence rate of the initial values toward the synchronous state is explicitly obtained. It is also proved that such d_c is independent of the choice of the individual map. Finally, our results here can be applied to address questions of wavelength bifurcations and size instability.

Key words. stable synchronization, Lyapunov exponents, wavelength bifurcation, coupled map lattices

AMS subject classifications. 34C15, 37N35

DOI. 10.1137/070705179

1. Introduction.

A particularly interesting form of dynamical behavior occurs in networks of coupled systems or oscillators when all of the individual systems or oscillators acquire identical chaotic behavior. Such behavior of a network models many systems of interest in physics, biology, and engineering. A central dynamical question is: When is such synchronous behavior stable, especially in regard to coupling strengths in the network? Much progress in this direction has been made in lattices of coupled chaotic systems. Indeed, many results [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] give analytical criteria for determining the range of coupling strength to acquire locally or even globally stable synchronization. On the other hand, to the best of our knowledge, there are no general results for global synchronization in coupled map lattices (CMLs). There are, however, globally synchronous results for some special cases (see, e.g., [15]). As to the study of local synchronization in CMLs, the notion of master stability functions (MSFs) that allows one to isolate the contribution of the network structure in terms of the eigenvalues of the coupling matrix was introduced in [8], [16], [17], [18], [19] to determine the possible range of coupling strength. This function then defines a region of stably synchronous state in terms of the coupling strength and the eigenvalues of the coupling matrix. Most of the work done in finding such a region of stability of the synchronous state is numerical. In a few certain cases, such as when the coupling matrix is symmetric, the MSFs can be further reduced to a number of inequalities [20], [21], [22], [23]

$$h_{\max} + \ln|1 + d\lambda_i| < 0, \quad i = 2, \ldots, m. \tag{1.1}$$

*Received by the editors October 12, 2007; accepted for publication (in revised form) by B. Ermentrout April 1, 2008; published electronically July 23, 2008.
http://www.siam.org/journals/siads/7-3/70517.html

†Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan (jjuang@math.nctu.edu.tw, moonsea.am96g@g2.nctu.edu.tw).

Here $h_{\max}$ is the largest Lyapunov exponent of the individual map, λ_i are the nonzero eigenvalues of the $m \times m$ coupling matrix, and d is the coupling strength. The Gershgorin disk theory is then applied to obtain some sufficient conditions [23] on the coupling strength for local synchronization. The reason for the huge gap between the theory developed in the lattices of coupled chaotic systems and that of CMLs lies mostly in the fact that it is more natural to have a nonlinear coupling between oscillators in the CMLs. This is because a nonlinear coupling within suitable range of the coupling strength tends to yield an invariant region for the corresponding CMLs while linear coupling cannot. It should be noted that there is no such problem for the lattices of coupled chaotic systems. It should also be mentioned that all the analytical results of the lattices of the coupled chaotic systems stated above are linearly coupled.

The purpose of this paper is to give the best possible results for the local synchronization of the CMLs. Indeed, we first prove that (1.1) holds true for general connectivity topology with a limitation that the nonlinear coupling needs to be the individual chaotic map as well. Second, the synchronization curve, composed of pieces of transverse Lyapunov exponent curves, is derived. With the help of the synchronization curve, we give necessary and sufficient conditions on yielding synchronization of the CMLs. Such conditions then lead to the identification of the optimal coupling strength interval for acquiring synchronization of the CMLs. The optimal interval is to be termed the *synchronization interval* of the CMLs. Moreover, the coupling strength d_c, called the *center* of the *synchronization interval* and giving the fastest convergence rate of the initial values toward the synchronous state, can be identified. Such d_c is independent of the choice of the individual map. Like the applications, our work here can also be used to analytically quantify how the small-world scheme improves the synchronizability of the network [24], [25], [26], [27]. Furthermore, our results here can be applied to address questions of wavelength bifurcations [28], [29], [30], [31], [32] and size instability [32]. For CMLs or coupled chaotic systems, the following four scenarios are possible as the coupling varies: (i) no synchronization; (ii) the presence of short wavelength bifurcations (SWBs); (iii) the presence of intermediate wavelength bifurcations (IWBs); and (iv) the presence of long wavelength bifurcations (LWBs). Our main results give the following. First, if the coupling matrix has only real eigenvalues, then only (i) and (ii) are possible. Second, if the coupling matrix has complex eigenvalues, then all four scenarios are possible. Third, the critical values for which wavelength bifurcations occur as well as the exact number of oscillators capable of sustaining stably synchronous chaos can be explicitly computed. Finally, the minimum coupling value where all wavelength modes become de-excited enough to induce the stability of the synchronous state is also explicitly given.

We conclude this introductory section by mentioning the organization of the paper. The main results are contained in section 2. Three types of coupling matrices are provided in section 3 as illustrations and applications to our main results. Some concluding remarks about future research are addressed in section 4.

2. Main results. Consider a network of CMLs consisting of m oscillators. The equations of the motion then read

$$\boldsymbol{x}_i(n+1) = \boldsymbol{f}(\boldsymbol{x}_i(n)) + d\left(\sum_{k=1}^{m} g_{ik}\boldsymbol{h}(\boldsymbol{x}_k(n))\right), \quad i = 1, \ldots, m. \tag{2.1}$$

Here $\boldsymbol{f}:\mathbb{R}^l\to\mathbb{R}^l$, $l\geq 1$, represents the individual chaotic map, and $\boldsymbol{h}:\mathbb{R}^l\to\mathbb{R}^l$ is an arbitrary nonlinear function describing how each oscillator's variables are used in the coupling. The quantities g_{ij} are the coupling weights between the oscillators i and j. To consider the notion of synchronization, we assume that $\sum_{k=1}^m g_{ik}=0$ for each i, and 0 is the simple eigenvalue of the coupling matrix $\boldsymbol{G}=(g_{ij})$. The quantity d represents the coupling strength of the CMLs (2.1). To have an invariant region for CMLs (2.1), one usually chooses $\boldsymbol{h}$ as $\boldsymbol{f}$. Such nonlinear coupling between oscillators is what makes (2.1) harder to treat analytically. In vector-matrix form with $\boldsymbol{h}=\boldsymbol{f}$, (2.1) becomes

$$\boldsymbol{x}(n+1)=\boldsymbol{F}(\boldsymbol{x}(n))+d(\boldsymbol{G}\otimes\boldsymbol{I})\boldsymbol{F}(\boldsymbol{x}(n)), \tag{2.2}$$

where $\otimes$ denotes the Kronecker product, $\boldsymbol{x}(n)=(\boldsymbol{x}_1(n),\dots,\boldsymbol{x}_m(n))^T$, and $\boldsymbol{F}(\boldsymbol{x}(n))=(\boldsymbol{f}(\boldsymbol{x}_1(n)),\dots,\boldsymbol{f}(\boldsymbol{x}_m(n)))^T$.

To study the stability of the synchronous state $\{\boldsymbol{x}_i=\boldsymbol{s}\ \forall i\}$ of CML (2.2), we consider the variational equation of (2.2):

$$\begin{aligned}\boldsymbol{\xi}(n+1)&=D\boldsymbol{F}(\boldsymbol{s}(n))\boldsymbol{\xi}(n)+d(\boldsymbol{G}\otimes\boldsymbol{I})D\boldsymbol{F}(\boldsymbol{s}(n))\boldsymbol{\xi}(n)\\ &=\left[\boldsymbol{I}\otimes D\boldsymbol{f}(\boldsymbol{s}(n))+d\left(\boldsymbol{G}\otimes\boldsymbol{I}\right)\left(\boldsymbol{I}\otimes D\boldsymbol{f}\left(\boldsymbol{s}(n)\right)\right)\right]\boldsymbol{\xi}(n),\end{aligned} \tag{2.3a}$$

where $\boldsymbol{\xi}=(\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_m)$ and each $\boldsymbol{\xi}_i$ is the perturbation to the ith oscillator. Let $\boldsymbol{J}=\boldsymbol{P}^{-1}\boldsymbol{G}\boldsymbol{P}$, where $\boldsymbol{J}=[0]\oplus\boldsymbol{J}_1\oplus\cdots\oplus\boldsymbol{J}_p$ is the real Jordan canonical form of $\boldsymbol{G}$. Applying the change of variables $\boldsymbol{\eta}=(\boldsymbol{P}^{-1}\otimes\boldsymbol{I})\boldsymbol{\xi}$, we get

$$\boldsymbol{\eta}(n+1)=\left[(\boldsymbol{I}+d\boldsymbol{J})\otimes D\boldsymbol{f}(\boldsymbol{s}(n))\right]\boldsymbol{\eta}(n),$$

or, equivalently, in block diagonal form,

$$\begin{aligned}\boldsymbol{\eta}_i(n+1)&=\left[(\boldsymbol{I}+d\boldsymbol{J}_i)^n\otimes D\boldsymbol{f}^n(\boldsymbol{s}(1))\right]\boldsymbol{\eta}_i(1)\\ &=:\boldsymbol{A}_i(n)\boldsymbol{\eta}_i(1).\end{aligned} \tag{2.3b}$$

Let $\sigma(\boldsymbol{A})$ denote the spectrum of $\boldsymbol{A}$. Then $\sigma(\boldsymbol{A}_i(n)\boldsymbol{A}_i^*(n))$ equals

$$\begin{aligned}&\sigma([(\boldsymbol{I}+d\boldsymbol{J}_i)^n\otimes D\boldsymbol{f}^n(\boldsymbol{s}(1))]\,[(\boldsymbol{I}+d\boldsymbol{J}_i^*)^n\otimes(D\boldsymbol{f}^n(\boldsymbol{s}(1)))^*])\\ &=\sigma([(\boldsymbol{I}+d\boldsymbol{J}_i)^n\,(\boldsymbol{I}+d\boldsymbol{J}_i^*)^n]\otimes[D\boldsymbol{f}^n(\boldsymbol{s}(1))\cdot(D\boldsymbol{f}^n(\boldsymbol{s}(1)))^*])\\ &=\sigma((\boldsymbol{I}+d\boldsymbol{J}_i)^n\,(\boldsymbol{I}+d\boldsymbol{J}_i^*)^n)\cdot\sigma(D\boldsymbol{f}^n(\boldsymbol{s}(1))\cdot(D\boldsymbol{f}^n(\boldsymbol{s}(1)))^*)\\ &=\sigma((\boldsymbol{I}+d\bar{\boldsymbol{J}}_i)^n\,(\boldsymbol{I}+d\bar{\boldsymbol{J}}_i^*)^n)\cdot\sigma(D\boldsymbol{f}^n(\boldsymbol{s}(1))\cdot(D\boldsymbol{f}^n(\boldsymbol{s}(1)))^*),\end{aligned}$$

where $\bar{\boldsymbol{J}}=[0]\oplus\bar{\boldsymbol{J}}_1\oplus\cdots\oplus\bar{\boldsymbol{J}}_p$ is the Jordan canonical form of $\boldsymbol{G}$. Consequently, the Lyapunov exponents of (2.2) are

$$h_j+\lim_{n\to\infty}\frac{\ln\sqrt{\lambda_{k,i}}}{n}.$$

Here h_j are the Lyapunov exponents of the individual system $\boldsymbol{f}$, and $\lambda_{k,i}$ are the eigenvalues of $(\boldsymbol{I}+d\boldsymbol{J}_{\lambda_i})^n(\boldsymbol{I}+d\boldsymbol{J}_{\lambda_i}^*)^n$, where $\boldsymbol{J}_{\lambda_i}$ is a Jordan block of matrix $\boldsymbol{G}$ and λ_i is an eigenvalue of

$\boldsymbol{G}$. Let the size of matrix $\boldsymbol{J}_{\lambda_i}$ be $k_i \times k_i$, and let $\boldsymbol{N} = \boldsymbol{J}_{\lambda_i} - \lambda_i \boldsymbol{I}$. It should be noted that for sufficiently large n,

$$\begin{aligned}(\boldsymbol{I} + d\boldsymbol{J}_{\lambda_i})^n &= ((1 + d\lambda_i)\boldsymbol{I} + d\boldsymbol{N})^n = (1 + d\lambda_i)^n (\boldsymbol{I} + \alpha \boldsymbol{N})^n \\ &= (1 + d\lambda_i)^n \left(\boldsymbol{I} + \sum_{j=1}^{k_i - 1} \binom{n}{j} \alpha^j \boldsymbol{N}^j \right) \\ &=: (1 + d\lambda_i)^n \boldsymbol{T}_i,\end{aligned}$$

where $\alpha = d/(1 + d\lambda_i)$. Clearly, the order of the magnitude of each entry of $\boldsymbol{T}_i \boldsymbol{T}_i^*$ is at most $\mathrm{O}(n^{2k_i - 2})$. We conclude, via the Gershgorin disk theorem, that all eigenvalues of $\boldsymbol{T}_i \boldsymbol{T}_i^*$ are of the order $\mathrm{O}(n^{2k_i - 2})$. Consequently, the Lyapunov exponents of (2.2) are

$$h_j + \ln |1 + d\lambda_i|. \tag{2.4}$$

We summarize the above as follows.

Theorem 2.1. *Let $\boldsymbol{G} = (g_{ij})$ be the coupling matrix satisfying that all its row sums are zero and zero is a simple eigenvalue. Then the synchronous state of CML (2.2) is (locally) stable provided that*

$$h_{max} + \ln |1 + d\lambda_i| < 0, \quad i = 2, \ldots, m, \tag{2.5}$$

where $h_{\max}$ is the largest Lyapunov exponent of the individual map $\boldsymbol{f}$ and $\lambda_i \in \sigma(\boldsymbol{G}) - \{0\}$, $i = 2, \ldots, m$. That is to say, if d satisfies the inequalities in (2.5), then for any initial values of (2.2) that are sufficiently close to the synchronous state $\{\boldsymbol{x}_i = \boldsymbol{s}\ \forall i\}$, we have that each of the oscillators $\boldsymbol{x}_i(n)$ tends to the same state as n goes to infinity. Otherwise, CML (2.2) will not acquire local synchronization.

Remark. (i) The decoupling form (2.3b) of variational equation (2.3a) was first observed and proposed by Pecora and Carroll [8]. (ii) If the identity matrix $\boldsymbol{I}$ in (2.2) is replaced by a diagonal matrix $\boldsymbol{D}$ with some but not all diagonal elements being zero, then the corresponding system (2.2) is called a partial-state coupling. The partial-state coupling also finds applications in various fields. For instance, in self-pulsating laser diode equations (see, e.g., [33]), only the photon density can be coupled with the electron density of the active region. Moreover, in the case of coupled chaotic systems, the systems that are partial-state coupled may exhibit different dynamic behavior. For instance, it is well known (see, e.g., [7]) that for the coupled Lorentz systems, only if the x-component or y-component is coupled will the resulting system achieve synchronization.

We shall assume from here on that the real parts of the eigenvalues of $\mathbf{G}$ are nonpositive. To find the range of the coupling d so that (2.5) is fulfilled, we need to solve the following min max problem:

$$\begin{aligned}\min_{d \in \mathbb{R}} \max_{2 \le i \le m} |1 + d\lambda_i| &= \min_{d > 0} \max_{2 \le i \le m_1} |1 + d\lambda_i| \\ &=: \min_{d > 0} \max_{2 \le i \le m_1} r_i(d) =: \min_{d > 0} r(d).\end{aligned} \tag{2.6}$$

Here m_1 is the number of eigenvalues lying in upper complex plane or on the real axis. The curves $r_i(d)$ are termed the ith mode of the transverse Lyapunov exponent curves. The equalities above are due to the facts that $|1+d\lambda_i| = |1+d\bar{\lambda}_i|$, the real parts of the eigenvalues of $\mathbf{G}$ are nonpositive, and (2.5) is violated whenever $d \le 0$. Without loss of generality, we may assume that those distinct nonzero eigenvalues are λ_i, $i = 2, \ldots, m_1$, with $0 < |\lambda_2| \le \cdots \le |\lambda_{m_1}|$. The coupling value $d := d_c$ solving the min max problem (2.6) is the optimal choice of the coupling in the sense that it gives the fastest convergence rate of the initial values toward the synchronous state. To understand how $r(d)$ is formed, we need to know the ordering of $r_i(d)$. For $d > 0$, direct computations yield

$$\begin{aligned} r_i(d) &= \left[|\lambda_i|^2 d^2 + 2\,\mathrm{Re}(\lambda_i) d + 1\right]^{\frac{1}{2}} \\ &= \left[|\lambda_i|^2 \left(d - \frac{\mathrm{Re}(-\lambda_i)}{|\lambda_i|^2}\right)^2 + \frac{|\lambda_i|^2 - \mathrm{Re}^2(\lambda_i)}{|\lambda_i|^2}\right]^{\frac{1}{2}} \\ &=: \left[|\lambda_i|^2 (d - c_i)^2 + \tan^2\theta_i\right]^{\frac{1}{2}}. \end{aligned} \tag{2.7a}$$

Moreover, $r_i(d) \ge r_j(d)$ if and only if

$$\mathrm{Re}(\lambda_i) \ge \mathrm{Re}(\lambda_j) \quad \text{if } |\lambda_i| = |\lambda_j|, \tag{2.7b}$$

and

$$(|\lambda_i|^2 - |\lambda_j|^2)(d - d_{ij}) \ge 0 \quad \text{if } |\lambda_i| \ne |\lambda_j|, \tag{2.7c}$$

where

$$d_{ij} = \frac{2(\mathrm{Re}(-\lambda_i) - \mathrm{Re}(-\lambda_j))}{|\lambda_i|^2 - |\lambda_j|^2}. \tag{2.7d}$$

Let $A_i = \{j : 2 \le j \le m_1 \text{ and } |\lambda_i| = |\lambda_j|\}$. Then $\max_{j\in A_i} |1 + d\lambda_j| = |1 + d\lambda_k|$, where k is chosen so that $\mathrm{Re}(\lambda_k) \ge \mathrm{Re}(\lambda_j)\ \forall j \in A_i$. This gives that within each of the index set A_i, their corresponding quantities $|1+d\lambda_i|$ are well ordered for any $d > 0$. Consequently, to solve (2.6), we may assume, without loss of generality, that $0 < |\lambda_2| < \cdots < |\lambda_{m_1}|$ from here on. Using the terminology in [32], we see that the numbers 2 and m_1 correspond to the longest and shortest wavelength modes, respectively. The numbers in between 2 and m_1 are to be called intermediate wavelength modes. Since $d_{ij} = d_{ji}$ for any i and j we consider only d_{ij} with $i > j$. Our reduction process is now complete.

The following procedures are proposed to determine the "actual" node points of $r(d)$ from the candidate set $\{d_{ij} : i > j\}$.

(A) Set $k_0 = 0$, and $k_1 = \max\{l \mid \mathrm{Re}(\lambda_i) \le \mathrm{Re}(\lambda_l)\ \forall i = 2 \ldots, m_1\}$. Let k_2 be the largest index so that $0 < d_{k_2 k_1} \le d_{k k_1}\ \forall k_1 < k \le m_1$.

(B) Let k_3 be the largest index so that $d_{k_2 k_1} < d_{k_3 k_2} \le d_{k k_2}\ \forall k_2 < k$. The process can be continued until $k_p = m_1$ for some $p \le m_1$.

The next result shows that $\{k_i\}_{i=1}^{p}$ is the set of "actual" node points of $r(d)$.

Theorem 2.2. *Let $\boldsymbol{G}$ be given as in Theorem 2.1. Assume that the real parts of the eigenvalues of $\boldsymbol{G}$ are nonpositive. Then $r(d) = r_{k_i}(d)$ whenever $d_{k_i k_{i-1}} \le d \le d_{k_{i+1}k_i}$, $i = 1, \ldots, p$. Here $d_{k_1,k_0} = 0$ and $d_{k_{p+1}k_p} = \infty$.*

Proof. Denote $I_j = [d_{k_j k_{j-1}}, d_{k_{j+1}k_j}]$. It then follows from (2.7c) that if $i > j$ and $d_{ij} > 0$, then $r_i(d) > r_j(d)$ whenever $d > d_{ij}$ and $r_i(d) < r_j(d)$ whenever $0 < d < d_{ij}$. We then conclude that

(2.8a) (i) the ordering of $r_i(d)$ and $r_j(d)$ remains the same until both curves meet;

(2.8b) (ii) if $r_i(d^*) > r_j(d^*)$ for some $d^* > 0$ with $i > j$, then $r_i(d) > r_j(d)$ $\forall d \ge d^*$.

Using the first inequality in (2.7a), we have that $r(d) = r_{k_1}(d)$ for $\epsilon_1 > d \ge 0$. Here ϵ_1 is sufficiently small. It then follows from (2.8a), (2.8b), and procedure (A) that $r(d) = r_{k_1}(d)$ on I_1. Upon using (2.8a), we conclude that $r(d) = r_{k_2}(d)$ for $d \in (d_{k_2k_1}, d_{k_2k_1} + \epsilon_2)$. Here ϵ_2 is sufficiently small. Similarly, $r(d) = r_{k_2}(d)$ on I_2. We omit the proof of the remaining assertions of the theorem due to the similarity. ∎

Note that not all c_{k_i} given in (2.7a) could be critical points of $r(d)$. In fact, the critical points of $r(d)$ may not even come from the set $\{c_{k_i}\}$. We next identify the "actual" critical points of $r(d)$. Our next main result shows that $r(d)$ has exactly one critical point.

Theorem 2.3. *The curve $r(d)$ has a unique critical point d_c that solves the min max problem (2.6). Moreover, the optimal range of coupling d to sustain stably synchronous chaos of (2.2) is (d_l, d_r). Here d_l and d_r, $d_l < d_r$, are the intersection points (if any exist) of the straight line $y = e^{-h_{\max}}$ and the curve $y = r(d)$. Consequently, CML (2.2) acquires local synchronization if and only if $d \in (d_l, d_r)$.*

Proof. We break up the proof of the theorem into the following three steps.

Step I. We first claim that the number of c_{k_i} lying in the interior $\overset{\circ}{I}_i$ of I_i is at most one. Indeed, suppose there exist $c_{k_a} \in \overset{\circ}{I}_a$ and $c_{k_b} \in \overset{\circ}{I}_b$ with $c_{k_a} < c_{k_b}$. Then the following hold true: (i) $r_{k_a}(c_{k_b}) > r_{k_a}(c_{k_a})$. (ii) $r_{k_a}(c_{k_a}) > r_{k_b}(c_{k_a})$. (iii) $r_{k_b}(c_{k_a}) > r_{k_b}(c_{k_b})$. Inequalities (i) and (iii) hold true since c_{k_a} and c_{k_b} are, respectively, the minimum points of $r_{k_a}(d)$ and $r_{k_b}(d)$. The fact that $r_{k_a}(d)$ lies above all other curves on I_a leads to inequality (ii). Combining these inequalities, we have that $r_{k_a}(c_{k_b}) > r_{k_b}(c_{k_b})$, which is in contradiction to the fact that r_{k_b} is the maximum curve on I_{k_b}.

Step II. We next show that if $c_{k_i} \in \overset{\circ}{I}_i$, then $r(d)$ is decreasing on $(0, c_{k_i})$ and increasing on (c_{k_i}, ∞). Indeed, for $d \in I_{i+1}$, $r(d) = r_{k_{i+1}}(d) > r_{k_i}(d) > r_{k_i}(d_{k_{i+1}k_i}) = r_{k_{i+1}}(d_{k_{i+1}k_i})$. Using the conclusion in Step I and the fact that $r^2_{k_i}(d)$ is parabolic, we conclude that $r_{k_{i+1}}(d)$ must be increasing on I_{i+1}. On the other hand, $r_{k_{i-1}}(d)$ must be decreasing on I_{i-1} since $r_{k_{i-1}}(d) > r_k(d) > r_{k_i}(d_{k_ik_{i-1}}) = r_{k_{i-1}}(d_{k_ik_{i-1}})$. The monotonicity of $r(d)$ on each interval I_j, $1 \le j \le m_1$, can be similarly determined.

Step III. Since $r(d)$ is decreasing initially on I_1 and increasing eventually on I_p, there must be at least one critical point. If such points do not lie in the set of node points, then $r(d)$ has a unique critical point. Suppose $c_{k_i} \notin \overset{\circ}{I}_i$ $\forall i = 1, \ldots, p$. Then $r(d)$ is monotonic on each interval I_i. Suppose $r(d)$ first changes its monotonicity at $d_{k_{l+1}k_l}$ for some l. Then an argument similar to that given in Step II shows that once $r(d)$ becomes increasing on I_{l+1}, it will stay increasing the rest of the way. We have just completed the proof of the theorem. ∎

Remark. (i) If the straight line $y = e^{-h_{\max}}$ and the curve $y = r(d)$ do not intersect, then CML (2.2) will not achieve synchronization for any coupling strength. Suppose d_r and d_l exist. Then, as soon as d exceeds d_r, a certain wavelength mode is excited, which, in turn, causes the instability of the synchronous state. The illustration in Examples 2 and 3 shows that the excited wavelength mode could be either the shortest wavelength mode, the intermediate wavelength mode, or the longest wavelength mode. In any event, d_r is the exact critical value where wavelength bifurcation occurs. On the other hand, d_l is the exact critical value where all wavelength modes become de-excited enough to induce the stability of the synchronous state. (ii) Such $r(d)$ is called the *synchronization curve* of CML (2.2), and the interval (d_l, d_r), if it exists, is termed the *synchronization interval* of CML (2.2). Clearly, $d_c \in (d_l, d_r)$ and depends only on the connectivity topology.

Theorem 2.4. *Suppose the coupling matrix* $\mathbf{G}$ *has nonpositive real eigenvalues. Denote by* $\{\lambda_i\}_{i=2}^{m_1}$ *the distinct nonzero eigenvalues of* $\mathbf{G}$. *Then*

$$r(d) = \begin{cases} \lambda_2(d), & d \in [0, d_{m_1 2}] = I_1, \\ \lambda_{m_1}(d), & d \in (d_{m_1 2}, \infty) = I_2, \end{cases}$$

and $d_c = d_{m_1 2} = \frac{-2}{\lambda_2 + \lambda_{m_1}}$. *Consequently, depending on the quantity of* $h_{\max}$, *either CML* (2.2) *achieves no synchronization or SWB occurs as* d *varies. Furthermore, if* d_l *and* d_r *exist, then the* synchronization interval *of the corresponding CMLs is* $\left(\frac{1-e^{-h_{\max}}}{-\lambda_2}, \frac{1+e^{-h_{\max}}}{-\lambda_{m_1}}\right)$.

Proof. It is easily seen that $k_1 = 2$ and $k_2 = m_1$ since $d_{ij} = \frac{-2}{\lambda_i + \lambda_j}$. Thus, $r(d)$ is as asserted. The proof then follows from the facts that $c_{m_1} = -\frac{1}{\lambda_{m_1}} < \frac{-2}{\lambda_2 + \lambda_{m_1}} < -\frac{1}{\lambda_2} = c_2$ and $d_c = d_{m_1 2}$. Solving equations $y = r(d)$ and $y = e^{-h_{\max}}$, we have that d_l and d_r are as claimed. ∎

3. Illustrations and applications.

We illustrate our theorems with the following examples.

Example 1. Let the oscillators be diffusively coupled with periodic boundary conditions. For such $\mathbf{G}$, $m_1 = m$, $-\lambda_{m_1} = 4\sin^2 \frac{[\frac{m}{2}]\pi}{m}$, and $-\lambda_2 = 4\sin^2 \frac{\pi}{m}$.

Let $f(x) = 4x(1-x)$, $0 \le x \le 1$. Then $h_{\max} = \ln 2$, and the corresponding candidates for d_l and d_r are, respectively, $\frac{1}{8}\sin^{-2}\frac{\pi}{m}$ and $\frac{3}{8}\sin^{-2}\frac{[\frac{m}{2}]\pi}{m}$. However, $d_l \le d_r$ only if $m \le 5$. Hence, we conclude that the maximum number of oscillators to sustain synchronous chaos is 5.

We next compare our results with those obtained in [23], [34]. Their sufficient conditions on the coupling strength for obtaining stable synchronization are, respectively, given as follows: $\frac{1-e^{-h_{\max}}}{m} < d\, g_{ij} < \frac{1+e^{-h_{max}}}{m}$ and $\left(\sum_{k=1,\, k\neq i}^{m} |g_{ki} - g_{ji}|\right) + \left|\frac{1}{d} + g_{ii} - g_{ji}\right| < \frac{1}{d}e^{-h_{\max}}$ $\forall i, j$ with $i \neq j$. However, the first inequality above fails to find any suitable coupling strength provided that $\boldsymbol{G}$ has zero off-diagonal elements. If $\boldsymbol{G}$ is given as above with $m \ge 4$ and $f(x) = 4x(1-x)$, then the second inequality also fails to find any suitable coupling strength.

Example 2. Consider synchronization in a directed ring of $2K$ nearest neighbor coupled oscillators [19] with $K = 2$ and $m = 9$. Specifically, the coupling matrix $\boldsymbol{G}$ under consideration is a circulant matrix of the form

$$\boldsymbol{G} = \mathrm{circ}(-30,\ 13,\ 2,\ 0,\ \ldots,\ 0,\ 5.4,\ 9.6).$$

The spectrum of $\boldsymbol{G}$ is $\{-30+13e^{\frac{2(j-1)\pi}{9}i}+2e^{\frac{4(j-1)\pi}{9}i}+5.4e^{\frac{14(j-1)\pi}{9}i}+9.6e^{\frac{16(j-1)\pi}{9}i} : j = 1, \ldots, 9\}$. Here $\lambda_2 \approx -11.4024 + 1.1629i$, $\lambda_3 \approx -33.0293 + 2.1855i$, $\lambda_4 \approx -45 + 5.8890i$, and $\lambda_5 \approx -45.5683 + 3.3483i$. Direct computations yield that $d_{42} \approx 0.0348 < d_{52} < d_{32}$, $d_{54} \approx 0.0406$, and $c_5 < c_4 < d_{42} < d_{54} < c_2$ (see Figure 1). Consequently,

$$r(d) = \begin{cases} r_2(d), & d \in I_1 = [0, d_{42}], \\ r_4(d), & d \in I_2 = [d_{42}, d_{54}], \\ r_5(d), & d \in I_3 = [d_{54}, \infty], \end{cases}$$

the node points of $r(d)$ are d_{42} and d_{54}, and the critical point of $r(d)$ occurs at d_{42}. Let $f_\mu(x) = \mu x(1-x)$. For $\mu = 4$, since $e^{-h_{max}} = e^{-\ln 2} = 0.5 < r(d_{42})$, the *synchronization interval* does not exist. As μ varies from $\mu_\infty \approx 3.57$ to $\mu = 4$, scenarios (i), (ii), and (iii) described in the introductory section can be clearly observed from the figure.

On the other hand, the maximum number of oscillators on such connectivity topology to sustain stably synchronous chaos is 7. The claim above is done by checking the intersection of the equations $y = \frac{1}{2}$ and $y = r(d)$ $\forall m \leq 8$.

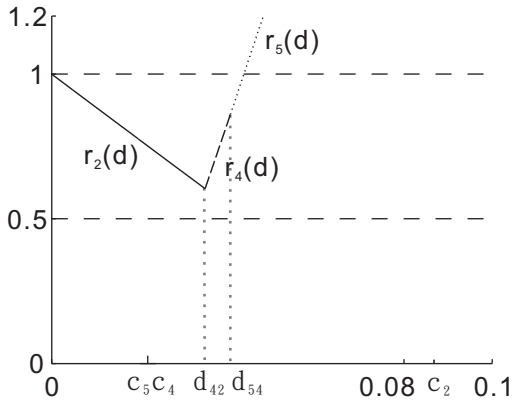


Figure 1. *Graph of $r(d)$ in Example 2. Here $d_{42} \approx 0.0348$, $d_{54} \approx 0.0406$, $r(d_{42}) \approx 0.6040$, $r(d_{54}) \approx 0.8604$, $c_5 \approx 0.02182$, $c_4 \approx 0.02185$, and $c_2 \approx 0.0868$. The critical point of $r(d)$ is d_{42}.*

Example 3. The following example shows that LWB is also possible. Let $\boldsymbol{G}$ be given as follows:

$$\begin{pmatrix} -30 & 3 & 12 & 5 & 10 \\ 10 & -30 & 3 & 12 & 5 \\ 5 & 10 & -30 & 3 & 12 \\ 12 & 5 & 10 & -30 & 3 \\ 3 & 12 & 5 & 10 & -30 \end{pmatrix}.$$

The spectrum of $\boldsymbol{G}$ is $\{0, -35.2639+10.7719i =: \lambda_2, -39.7361+2.5429i =: \lambda_3, \bar{\lambda}_2, \bar{\lambda}_3\}$. Then the graph of $r(d)$ is demonstrated in Figure 2. Consider $f(x) = 4x(1-x)$. Then the longest wavelength mode becomes excited to induce instability as d is increased beyond d_r.

Example 4. To illustrate the accuracy of our theorems, *synchronization intervals* established in Theorem 2.4 are compared with those obtained by the computer simulation. In particular, theoretically and numerically predicted *synchronization intervals* for three exam-

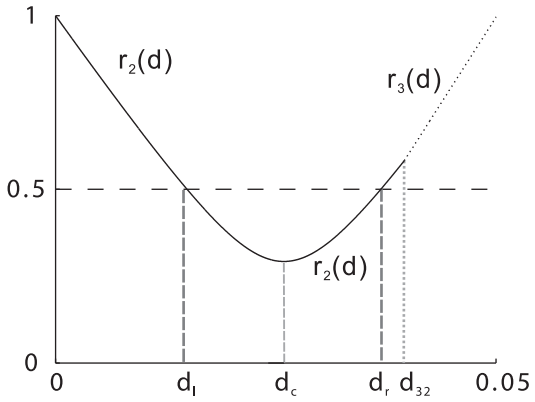


Figure 2. *Graph of $r(d)$ in Example 3. Here $d_{32} \approx 0.0396$, $c_2 \approx 0.0259$, $c_3 \approx 0.0251$, $d_l \approx 0.0149$, and $d_r \approx 0.0369$. The critical point of $r(d)$ is c_2.*

ples above are almost identical. Such comparisons are recorded in Figure 3. They are "almost" identical. This simulation is set up so that the differences between the initial values $\boldsymbol{x}_i$ are within 10^{-5}. Synchronization is achieved when their differences are within 10^{-15}.

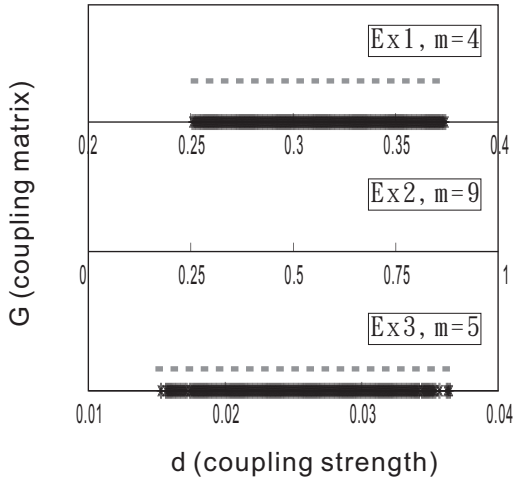


Figure 3. *Three typical* synchronization intervals *for coupled logistic map with various coupling matrices are shown. Solid (bold) lines are synchronization intervals obtained by computer simulation. Dotted (fine) lines are synchronization intervals predicted by our theorems. All are scaled for clear visualization.*

4. Conclusions. We conclude this paper by mentioning the difficulty one might face by applying our methods to more general cases, $\boldsymbol{D} \neq \boldsymbol{I}$ or $\boldsymbol{h} \neq \boldsymbol{f}$, and a possible approach to solving them.

Our main results in this paper are based on the study of the inequalities in (2.5). However, in the case that $\boldsymbol{D} \neq \boldsymbol{I}$ or $\boldsymbol{h} \neq \boldsymbol{f}$, it seems to be a nontrivial matter to find their corresponding inequalities such as (2.5). One possible approach is to find the lower and upper bounds of the Lyapunov exponents of (2.2), where both bounds have expressions similar to those in (2.4).

Acknowledgment. The authors would like to thank the referees for their helpful comments.

REFERENCES

[1] V. N. Belykh, N. N. Verichev, L. J. Kocarev, and L. O. Chua, *Chua's Circuit: A Paradigm for Chaos*, World Scientific, Singapore, 1993.

[2] V. N. Belykh, I. V. Belykh, K. V. Nevidin, and M. Hasler, *Hierarchy and stability of partially synchronous oscillations of diffusively coupled dynamical systems*, Phys. Rev. E (3), 62 (2000), pp. 6332–6345.

[3] V. N. Belykh, I. V. Belykh, K. V. Nevidin, and M. Hasler, *Persistent clusters in lattices of coupled nonidentical chaotic systems*, Chaos, 13 (2003), pp. 165–178.

[4] V. N. Belykh, I. V. Belykh, K. V. Nevidin, and M. Hasler, *Cluster synchronization in three-dimensional lattices of diffusively coupled oscillators*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 13 (2003), pp. 755–779.

[5] V. N. Belykh, I. V. Belykh, and M. Hasler, *Connection graph stability method for synchronized coupled chaotic systems*, Phys. D, 195 (2004), pp. 159–187.

[6] V. N. Belykh, I. V. Belykh, and M. Hasler, *Synchronization in asymmetrically coupled networks with node balance*, Chaos, 16 (2006), 015102.

[7] J. Juang, C. L. Li, and Y. H. Liang, *Global synchronization in lattices of coupled chaotic systems*, Chaos, 17 (2007), 033111.

[8] L. M. Pecora and T. L. Carroll, *Master stability functions for synchronized coupled systems*, Phys. Rev. Lett., 80 (1998), pp. 2109–2112.

[9] A. Pogromsky and H. Nijmeijer, *Cooperative oscillatory behavior of mutually coupled dynamical systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 48 (2001), pp. 152–162.

[10] W. Wang and J.-J. E. Slotine, *On partial contraction analysis for coupled nonlinear oscillators*, Biol. Cybernet., 92 (2005), pp. 38–53.

[11] C. W. Wu and L. O. Chua, *Synchronization in an array of linearly coupled dynamical systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 42 (1995), pp. 430–447.

[12] C. W. Wu, *Cooperative oscillatory behavior of mutually coupled dynamical systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 48 (2001), pp. 152–162.

[13] C. W. Wu, *Synchronization in coupled arrays of chaotic oscillators with nonreciprocal coupling*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 50 (2003), pp. 294–297.

[14] C. W. Wu, *Synchronization in Coupled Chaotic Circuits and Systems*, World Scientific Series on Nonlinear Science Series A 41, World Scientific, Singapore, 2002.

[15] W.-W. Lin and Y.-Q. Wang, *Chaotic synchronization in coupled map lattices with periodic boundary conditions*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 175–189.

[16] G. Hu, J. Yang, and W. Liu, *Instability and controllability of linearly coupled oscillators: Eigenvalue analysis*, Phys. Rev. E (3), 58 (1998), pp. 4440–4453.

[17] M. Zhan, G. Hu, and J. Yang, *Synchronization of chaos in coupled systems*, Phys. Rev. E (3), 62 (2000), pp. 2963–2966.

[18] K. S. Fink, G. Johnson, T. Carroll, D. Mar, and L. Pecora, *Three coupled oscillators as a universal probe of synchronization stability in coupled oscillator arrays*, Phys. Rev. E (3), 61 (2000), pp. 5080–5090.

[19] M. Barahona and L. M. Pecora, *Synchronization in small-world systems*, Phys. Rev. Lett., 89 (2002), 054101.

[20] P. M. Gade and R. E. Amritkar, *Spatially periodic orbits in coupled-map lattices*, Phys. Rev. E (3), 47 (1993), pp. 143–154.

[21] P. M. Gade, H. A. Cerdeira, and R. Ramaswamy, *Coupled maps on trees*, Phys. Rev. E (3), 52 (1995), pp. 2478–2485.

[22] P. M. Gade, *Synchronization of oscillators with random nonlocal connectivity*, Phys. Rev. E (3), 54 (1996), pp. 64–70.

[23] Y. Chen, G. Rangarajan, and M. Ding, *General stability analysis of synchronized dynamics in coupled systems*, Phys. Rev. E (3), 67 (2003), 026209.
[24] D. J. Watts and S. H. Strogatz, *Collective dynamics of "small-world" networks*, Nature, 393 (1998), pp. 440–442.
[25] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, NJ, 1999.
[26] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, *The large-scale organization of metabolic networks*, Nature, 407 (2000), pp. 651–654.
[27] S. H. Strogatz, *Exploring complex networks*, Nature, 268 (2001), pp. 268–276.
[28] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*, Springer-Verlag, New York, 1984.
[29] T. Bohr and O. B. Christensen, *Size dependence, coherence, and scaling in turbulent coupled-map lattices*, Phys. Rev. Lett., 63 (1989), pp. 2161–2164.
[30] L. A. Bunimovich, A. Lambert, and R. Lima, *The emergence of coherent structures in coupled map lattices*, J. Statist. Phys., 61 (1990), pp. 253–262.
[31] O. Cardoso, H. Willaime, and P. Tabeling, *Short-wavelength instability in a linear array of vortices*, Phys. Rev. Lett., 65 (1990), pp. 1869–1872.
[32] J. F. Heagy, L. M. Pecora, and T. L. Carroll, *Short wavelength bifurcations and size instabilities in coupled oscillator systems*, Phys. Rev. Lett., 74 (1995), pp. 4185–4188.
[33] C. Juang, T. M. Huang, J. Juang, and W. W. Lin, *A synchronization scheme using self-pulsating laser diodes in optical communication*, IEEE J. Quantum Electron., 36 (2000), pp. 300–304.
[34] G. Rangarajan and M. Ding, *Stability of synchronized chaos in coupled dynamical systems*, Phys. Lett. A, 296 (2002), pp. 204–209.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 766–794
© 2008 Society for Industrial and Applied Mathematics

On Synchronization and Traveling Waves in Chains of Relaxation Oscillators with an Application to Lamprey CPG*

Péter L. Várkonyi† and Philip Holmes‡

Abstract. We study chains of relaxation-type neural oscillators with local excitatory coupling. Phase reductions suggest that such networks typically exhibit traveling waves, but relaxation oscillators often synchronize. We examine these behaviors using the phase response and fast threshold modulation (FTM) theories, which respectively describe network behavior for infinitesimally weak and moderate coupling. Surprisingly, the two different approximations yield quantitatively consistent predictions for chains with one-way coupling. Specifically, approaching the relaxation limit, such chains can exhibit waves with vanishing phase differences (i.e., synchrony) propagating in the coupling direction, *or* waves with persistent phase differences traveling against the coupling direction. These results provide novel support for the finding that caudo-rostral coupling dominates in the lamprey central pattern generator (CPG), and they suggest that recent models may underestimate the role of network effects in burst generation.

Key words. fast threshold modulation, phase reduction, relaxation oscillators, synchronization, traveling waves

AMS subject classifications. 34C26, 34C15, 34C29

DOI. 10.1137/070710329

1. Introduction.

Phase reduction theory, originally developed by Malkin [30, 31] and independently rediscovered by Winfree [43] (cf. [14]), provides a method for the simplification and analysis of networks of coupled oscillators, including those composed of spontaneously oscillatory spiking or bursting neurons. Augmented by the averaging theorem [15] for weakly coupled systems, it allows one to reduce N sets of M ordinary differential equations (ODEs), each set describing an oscillator having a hyperbolic (attracting) limit cycle, to a system of N ODEs approximating the phases of each oscillator along its limit cycle. See [18] and [19] for more recent statements of Malkin's theorem. Phase reduction always applies for sufficiently weak coupling, but it often extends to stronger coupling [11].

According to this theory, chains of oscillators with local coupling generically exhibit traveling waves, except for symmetrical bidirectional coupling or special types of interactions such as coupling of neural oscillators via gap junctions [7, 22, 23]. However, the interaction of relaxation oscillators seems to be exceptional: they phase-lock with zero phase-difference,

*Received by the editors December 6, 2007; accepted for publication (in revised form) by D. Terman April 3, 2008; published electronically July 23, 2008. This work was partially supported by NSF EF-0425878 and NIH NS054271.

http://www.siam.org/journals/siads/7-3/71032.html

†Department of Mechanics, Materials and Structures, Budapest University of Technology and Economics, Müegyetem rkp. 3-9, H-1111 Budapest, Hungary (vpeter@mit.bme.hu). This author was supported by Imre Korányi and Zoltán Magyary fellowships as well as by OTKA-72368 and was hosted by the Program in Applied and Computational Mathematics of Princeton University.

‡Program in Applied and Computational Mathematics, and Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 (pholmes@math.princeton.edu).

(i.e., they synchronize) in cases where one would expect traveling waves [35, 36]. This synchrony is robust against perturbations: while phase oscillators compensate for perturbations by changing their phase relations, relaxation oscillators typically compensate via changes in waveforms.

There are at least two explanations for this behavior. Phase reduction neglects the effects of nonlinearities in coupling: it requires that orbits perturbed by coupling remain sufficiently close to unperturbed limit cycles at all times, which holds for *sufficiently weak* coupling $\epsilon \ll 1$. Relaxation oscillators combine fast and slow dynamics (i.e., two characteristic time-scales with ratio $\mu \ll 1$), and in this case phase reduction requires *extremely weak* coupling: $\epsilon \ll \mu$ [19]. In relevant ranges of μ, the oscillators' interactions are typically dominated by higher order effects that are not captured by phase theory. Fast threshold modulation (FTM) theory [35, 36] was introduced to explain this behavior. Motivated by synaptic coupling of neural oscillators, it applies to moderate or strong coupling: $\mu \ll \epsilon$.

Despite the apparent contrast between relaxation and phase oscillators, and the limited applicability of phase reduction to the former, phase theory can also account for synchronization of relaxation oscillators [19], and its predictions agree qualitatively with those of FTM theory. The reason for this unexpected behavior is that the function $H(\psi)$ describing the effect of coupling between two oscillators is discontinuous at certain points with respect to their phase difference ψ.

In this paper we apply phase reduction for weak coupling ($\epsilon \ll \mu \ll 1$) and a combination of FTM and phase theory for relaxation-type oscillators ($\mu \ll \epsilon \ll 1$). We ask if a given system exhibits traveling waves or synchrony and compare predictions of the two methods, thereby shedding light on behaviors expected under variations in coupling strength ϵ. Both approaches are required to obtain a global picture of the behavior of coupled chains, and we show that their predictions are *quantitatively* similar in chains with one-way coupling, despite the different mechanisms. In section 2 we analyze a pair of oscillators with one-way coupling in the phase reduction limit and outline a generalization to unidirectionally coupled oscillator chains, and section 3 is an analogous study of the FTM limit. In these sections we describe an interesting property of traveling wave solutions: in the limit $\mu \to 0$ waves propagating in the coupling direction approach synchronous dynamics, but counterpropagating waves persist (Theorems 2.1 and 3.1). These results provide quantitative conditions for traveling waves versus synchrony in arrays of unidirectionally coupled relaxation oscillators. The behavior of bidirectionally coupled chains is also briefly discussed at the end of each section. In section 4 we demonstrate that most, but not all, simple oscillators exhibit the first behavior: synchrony is more common in the relaxation limit than traveling waves, and we provide a sufficient condition for this in Theorem 4.1. Section 5 contains illustrative examples of both behaviors.

In section 6, we apply these results to the neural central pattern generator (CPG) of the lamprey. Recent lamprey CPG models [39, 27, 26] are double chains of relaxation-type bursters in which the burst frequency is adjusted by neuro-modulators that tune the slow time-scale μ so that the relaxation limit is approached as swimming speed decreases. Simulations indicate that these models exhibit synchrony in the relaxation limit and that phase lags between neighboring units depend strongly on swimming frequency, being small at low frequency and larger at high frequency. In contrast, the animal exhibits quasi–frequency-independent phase patterns. As we will show, this shortcoming could be eliminated if the model's parameters

were adjusted to generate traveling waves in the relaxation limit. The paper concludes with section 7. We relegate many technical details in the proofs of Theorems 2.1 and 3.1 to a series of appendices. Background on CPGs can be found in [8], and background on phase and relaxation oscillator models can be found in [21, 23].

2. Relaxation oscillators in the phase limit. We consider a pair of identical relaxation oscillators O_1, O_2, each with one slow variable x_j and one fast one y_j. The time-scale ratio is set by the parameter $0 < \mu \ll 1$, and $\mu \to 0$ is the relaxation limit. If O_1 receives weak coupling ($\epsilon \ll \mu$) from O_2 in the fast variable, the ODEs for O_1 are

$$\dot{x_1} = f(x_1, y_1), \tag{2.1}$$

$$\dot{y_1} = \frac{1}{\mu}\left[g(x_1, y_1) + \epsilon h(x_1, y_1, x_2, y_2)\right], \tag{2.2}$$

while those of O_2 are the same, but without the coupling term h. Henceforth we assume that the fast equation (2.2) has a cubic-shaped nullcline or *slow manifold* $g = 0$, and that for $\epsilon = 0$ a stable hyperbolic limit cycle Γ of period T exists, on which x_j slowly decreases (resp., increases) near the lower (resp., upper) branch of $g = 0$; see Figure 1. Henceforth, in describing a single oscillator, we typically drop the subscripts.

In section 2.1 we summarize the results of Izhikevich [19] and prepare for section 2.2. There we prove our first theorem, extending the leading order phase response curve (PRC) expressions of [19] to the next order and providing explicit estimates of the width and height of PRC peaks in terms of fractional powers of μ (Figure 5).

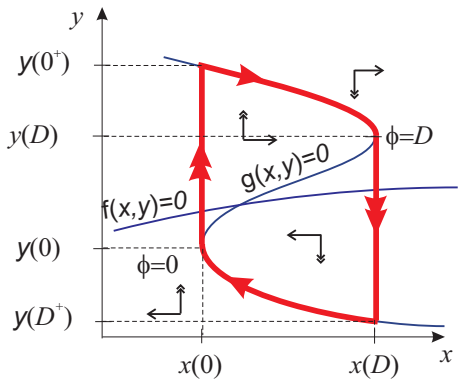


Figure 1. *Schematic phase portrait of a 2D relaxation oscillator with cubic-shaped nullcline (thin curve). Arrows show directions of the vectorfield; ϕ denotes the phase along the limit cycle (thick curve); $\phi = 0, D$ denote phase values at the instantaneous jumps; and $x(0)$, $y(0)$, $y(0^+)$, etc. denote the coordinates of the corresponding points in phase space. Oscillators are assumed to be active on the upper branch $\phi \in (0, D)$ and silent on the lower branch $\phi \in (D, 0)$.*

2.1. Phase reduction and previous results. Like any ODE with a stable hyperbolic limit cycle, (2.1)–(2.2) can be reduced to a phase description [18]. We define the phase $\phi = \phi(x, y)$ along Γ such that the periodic solution satisfies $\dot{\phi} = 2\pi/T \stackrel{\text{def}}{=} \omega$, and we let $x(\phi)$, $y(\phi)$, etc. denote coordinates of points on Γ. The system of four coupled ODEs may then be

reduced to the phase equations

$$\dot{\phi_1} = \omega + \frac{\epsilon}{\mu}\tilde{h}(\phi_1, \phi_2)z(\phi_1) + \mathcal{O}(\epsilon^2), \quad \dot{\phi_2} = \omega, \tag{2.3}$$

where $\tilde{h}(\phi_1, \phi_2) = h(x_1(\phi_1), y_1(\phi_2), x_2(\phi_2), y_2(\phi_2))$ denotes the coupling function evaluated on Γ. Here the PRC $z(\phi)$ represents the sensitivity of O_1 to perturbations from O_2, and $z(\phi) > 0$ (resp., $z(\phi) < 0$) means that an excitatory signal ($h > 0$) received at $(x(\phi), y(\phi))$ speeds up (resp., slows down) O_1. In deriving the PRC one expands about Γ in a Taylor series, thereby neglecting nonlinear ($O(\epsilon^2)$) coupling effects. See [10, 3] for recent examples of explicit PRC computations.

After introducing the phase difference $\psi = \phi_1 - \phi_2$, (2.3) can be averaged over the period T and subtracted to yield

$$\dot{\psi} = \frac{\epsilon}{2\pi\mu}\int_0^{2\pi} \tilde{h}(\varphi + \psi, \varphi)z(\varphi + \psi)\, d\varphi + O(\epsilon^2) \stackrel{\text{def}}{=} H(\psi) + O(\epsilon^2). \tag{2.4}$$

Zeros of H correspond to phase differences at which the oscillators phase-lock ($\psi = \text{const}$), and stable phase-locking occurs if $H(\psi) = 0$ and $dH(\psi)/d\psi < 0$. For details, see [18, 17].

While in general PRCs must be computed numerically, in [19, section 2] analytical formulae were derived for (2.1)–(2.2) in the limit $\mu \to 0$, as follows. Let $z^*(\phi) \stackrel{\text{def}}{=} z(\phi)/\mu$; subscripts x and y denote partial derivatives, $\phi^{(1)} = 0$ and $\phi^{(2)} = D$ denote phase values at the jumps, $y(\phi^{(j)}) = y(0), y(D)$ denote the value of y immediately before a jump, and $y(\phi^{(j)+}) = y(0^+), y(D^+)$ denotes its value immediately after it (Figure 1). If $\phi \neq \phi^{(j)}$, $j \in \{1, 2\}$, then

$$z^*(\phi) = -\frac{\omega f_y(x(\phi), y(\phi))}{f(x(\phi), y(\phi))g_y(x(\phi), y(\phi))}, \tag{2.5}$$

and near the jumps $\phi = \phi^{(j)}$,

$$z^*(\phi) = \frac{\delta(\phi - \phi^{(j)})\,\omega^2}{g_x(x(\phi^{(j)}), y(\phi^{(j)}))}\left[\frac{1}{f(x(\phi^{(j)}), y(\phi^{(j)}))} - \frac{1}{f(x(\phi^{(j)}), y(\phi^{(j)+}))}\right]. \tag{2.6}$$

Equation (2.5) follows from linearization in the neighborhood of the slow parts of Γ on the upper and lower branches of the $g = 0$ nullcline, and it may be derived from the adjoint formulation of the PRC [18], as in [19, (2.9)]. Equation (2.6) is found by considering the jumps from $y(0)$ to $y(0^+)$ and $y(D)$ to $y(D^+)$ (Figure 1; cf. [19, (2.4)]). The delta functions at $\phi^{(j)}$ play a central role in explaining the behavior of coupled relaxation oscillators; see Figure 2.

To formulate our results, we supplement (2.5)–(2.6) with the following conditions:

(i) The coupling term $h \equiv 1$ while O_2 is near the upper branch of its $g = 0$ nullcline (for $\mu \to 0$ this implies $\phi \in (0, D)$), and $h \equiv 0$ near the lower branch. During jumps, $0 \leq h \leq 1$.

(ii) $\partial f(x, y)/\partial y > 0$ at arbitrary points $(x(\phi), y(\phi))$ along the limit cycle.

(iii) The duty cycle $D/(2\pi) \leq 0.5$ (time spent on the upper (active) branch is not more than that spent on the lower branch).

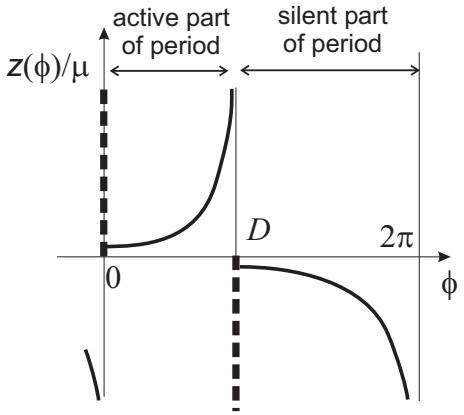


Figure 2. *Schematic PRC for a relaxation oscillator in the limit $\mu \to 0$ as derived in [19]. The function has singularities and delta functions (thick dashed lines) at $\phi = 0, D$. The sign of the delta functions is always as shown. The sign of the continuous parts is as shown if condition* (ii) *holds.*

Condition (i) simplifies the notion that oscillators have active (e.g., bursting) and silent (e.g., refractory) states; it has been used by other authors (e.g., [35, 36, 19]). Condition (ii) requires that $f(x, y)$ is strictly monotonic in y, which is true for most oscillator models, and it implies that the sign of z^* is as shown in Figure 2, due to the following facts:

(1) On the limit cycle of Figure 1, $f(x(0), y(0)) < 0$, $f(x(0), y(0^+)) > 0$, and $g_x(x(0), y(0)) < 0$, so (2.6) implies that the peak in z^* at $\phi = 0$ is positive. Similarly, the peak at $\phi = D$ is negative.

(2) Attractivity of the upper and lower branches of the $g = 0$ nullcline implies that $g_y < 0$; $f > 0$ on the upper branches and $f < 0$ on the lower branches, and $f_y > 0$, by condition (ii). Thus (2.5) gives positive and negative PRC values during the active and silent parts of the limit cycle, respectively.

Under condition (i), (2.4) simplifies to

$$H(\psi) = \frac{\epsilon}{2\pi} \int_{\psi}^{D+\psi} z^*(\phi)\, d\phi; \tag{2.7}$$

the resulting function $H(\psi)$ is shown in Figure 3, in which its key properties are also summarized. In particular, H is discontinuous: when $\phi^{(i)}$ enters or leaves the interval of integration $[\psi, D + \psi]$ in (2.7), the delta functions in (2.6) introduce step changes. If the decreasing step at $\psi = 0$ passes through 0, then $H(\psi)$ has a root at 0, corresponding to synchronization, which is robust against perturbations such as adding a constant to H. This fact was advanced in [19] to explain why weakly coupled relaxation oscillators tend to synchronize.

Note that the properties of $H(\psi)$ shown in Figure 3 require condition (iii); for $D > \pi$, H may have arbitrarily many roots or possibly none at all. Condition (iii) holds for the majority of important relaxation oscillators, and it appears elsewhere in the literature [24]. We also remark that, since $\lim_{\phi \to \phi^{(i)}} g_y = 0$ and $g_y \backsim (\phi^{(i)} - \phi)^{1/2}$, the rescaled PRC $z^*(\phi)$ in (2.5) has integrable singularities at $\phi = \phi^{(j)}$.

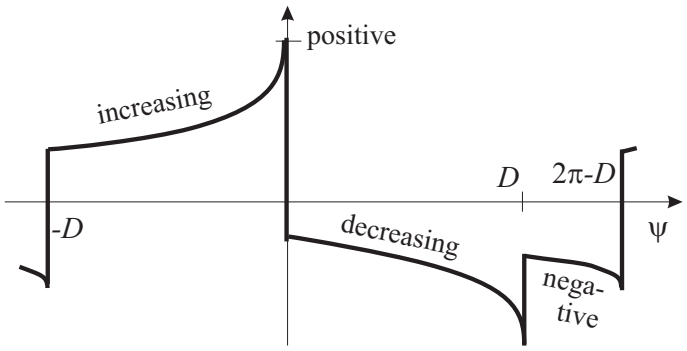


Figure 3. *Schematic coupling function $H(\psi)$ for a relaxation oscillator with weak one-way coupling in the limit $\epsilon \ll \mu \to 0$, as derived in* [19]. *If conditions* (i)–(iii) *hold, the qualitative form of H and the directions of the steps are as shown.*

2.2. Phase-locking and synchrony. While the discontinuity of H at $\psi = 0$ often yields synchronization, this is not the only possibility: if $\text{sign}(H(0^+)) = \text{sign}(H(0^-))$, the oscillators do not synchronize, although phase-locking can occur at $\psi \neq 0$ if H has nonzero roots. This latter case corresponds to traveling waves in a chain (see section 2.3). Here we state a result that shows that phase-locked solutions with $\psi = \Delta(\mu) \leq 0$ behave differently from those with $\Delta(\mu) > 0$, thereby distinguishing the two behaviors.

Theorem 2.1. *Let $H_\mu(\psi)$ denote the averaged coupling function at a given value of μ, and let $(x, y) = (\xi, \upsilon(\xi))$ denote the equation of the active branch of the slow manifold. Assume that the phase-shift $\Delta(\mu)$ satisfies $dH_\mu(\psi)/d\psi|_{\Delta(\mu)} < 0 = H_\mu(\Delta(\mu))$.*

(a) *If conditions* (i)–(iii) *hold and*

$$-\int_{x(0)}^{x(D)} \frac{f_y(\xi,\upsilon(\xi))}{f^2(\xi,\upsilon(\xi))g_y(\xi,\upsilon(\xi))}\,d\xi + \frac{1}{g_x(x(D),y(D))}\left[\frac{1}{f(x(D),y(D))} - \frac{1}{f(x(D),y(D+))}\right] > 0, \tag{2.8}$$

then $0 < \lim_{\mu\to 0}\Delta(\mu) < \pi$; i.e., O_1 leads O_2.

(b) *If conditions* (i)–(iii) *hold and* (2.8) *is false, then $\Delta(\mu) \backsim -\mu^{2/3}$ for small μ and $\lim_{\mu\to 0}\Delta(\mu) = 0$; i.e., the oscillators synchronize.*

Proof. We extend the results of [19] to the case $0 < \mu \ll 1$. The limit cycle consists of three parts, known in singular perturbed and boundary layer theory as the outer, inner, and intermediate limits [2]; see Figure 4.

1. Evolution along slow manifolds: These episodes occupy time $O(1)$, and the PRCs are well approximated by (2.5); thus $z^*(\phi) = O(1)$.
2. Transition from slow motion to jumps at the knees of the $g = 0$ nullcline: These take time $O(\mu^{2/3})$ [2], and, as shown below, the phase response converges to (2.6) in the relaxation limit; thus $z^*(\phi) = O(\mu^{-2/3})$.
3. Fast jumps of duration $O(\mu)$ between slow manifolds: Here $z^*(\phi) = O(1)$, as in case 1.

These statements follow from the arguments of [19], summarized in section 2.1, with the crucial additional fact, shown in Appendix A, that the delta function in (2.6) derives from

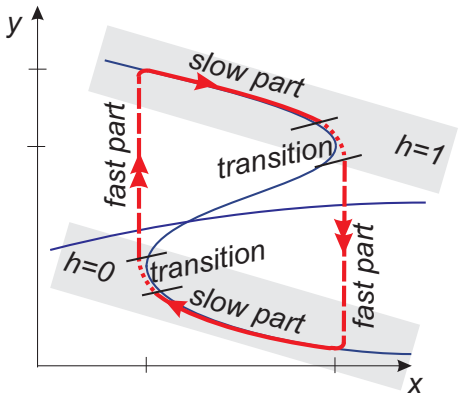


Figure 4. *Characteristic segments of the limit cycle of a relaxation oscillator: Slow motion along nullclines, fast jumps, and transition. Coupling h is assumed constant (≡ 0 or 1) near the slow nullclines, shown shaded.*

perturbations during transition and *not* during the fast jump. Intuitively, perturbations can advance the jump as solutions approach the fold, but perturbations in the fast variable have little effect during the jump ($z \backsim O(\mu)$, i.e., $z^* \backsim O(1)$), due to the $O(\mu^{-1})$ speed of the dynamics. The resulting PRC is shown in Figure 5.

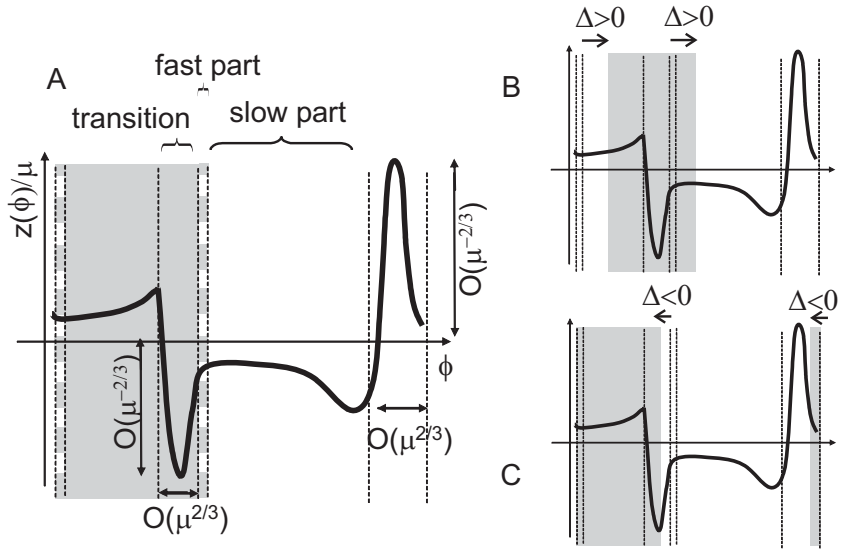


Figure 5. A: *PRCs for relaxation oscillators with* $0 \neq \mu \ll 1$. *Grey denotes the active part of the period* $h \equiv 1$, *white the silent part* $h \equiv 0$, *and striped, far from the nullclines, h not defined by condition* (i). *For synchrony,* $H(0)$ *is the integral of* z^* *over the grey interval.* B: *For small* $\psi > 0$, $H(\psi)$ *is the integral over a domain shifted rightward compared to the case* $\psi = 0$. C: *For small* $\psi < 0$, *the integration domain is shifted leftward and includes the peaks of* z^*.

If the two oscillators are in synchrony ($\psi = 0$), O_1 receives a coupling signal during both the slow and transitional parts near the *upper nullcline* according to condition (i); there is no input during slow and transition parts near the lower nullcline. Condition (i) does not

determine h during the fast jumps, since the state is far from both nullclines. However, the integral of z^* during these episodes is only $O(\mu)$ ($\sim$ the jump duration), so its contribution to H vanishes as $\mu \to 0$; cf. (2.7). Thus, if μ is sufficiently close to 0, $H_\mu(0)$ is well approximated by the integral of z^* over the slow and transition parts on the upper branch, and using (2.5)–(2.7), we obtain

$$H_\mu(0) = \frac{\epsilon\omega^2}{2\pi}\left[\int_0^D \frac{-f_y(x(\phi),y(\phi))}{\omega f(x(\phi),y(\phi))g_y(x(\phi),y(\phi))}\,d\phi \right. \\ \left. + \frac{1}{g_x(x(D),y(D))}\left(\frac{1}{f(x(D),y(D))} - \frac{1}{f(x(D),y(D^+))}\right)\right] + \mathcal{O}(\mu); \tag{2.9}$$

see also Figure 5(A).

Due to (2.1) and (2.3), we have $d\phi = dx \cdot \omega/f$; hence (2.9) is equal to the left-hand side of (2.8) modulo the $O(\mu)$ term and the $\epsilon\omega^2/2\pi$ factor. Thus, case (a) of Theorem 2.1 corresponds to $\lim_{\mu\to 0} H_\mu(0) > 0$. If $\psi = 0$ at $t = 0$, ψ increases according to (2.4), and the limits of integration in (2.7) must be moved rightward to locate a zero of $H(\psi)$, as in Figure 5(B). At that point the limits do not intersect the peaks of the PRC, implying that H has finite slope just above $\psi = 0$. The conclusion of part (a) follows from this fact.

On the other hand, if $H(0) < 0$, the domain of integration must be shifted to the left and will intersect the PRC's peaks; see Figure 5(C). A negative peak of width $O(\mu^{2/3})$ and slope $O(\mu^{-2/3})$, which shrinks to a vertical step in the relaxation limit of Figure 2, lies just below $\psi = 0$. $\Delta(\mu)$ cannot lie elsewhere than at this steep part, since conditions (i)–(iii) imply that at all other points $H(\psi)$ is either negative or increasing; see Figure 3. ■

We remark that if only condition (i) holds, we still have $\lim_{\mu\to 0}\Delta(\mu) \neq 0$ in case (a). In case (b), conditions (i)–(ii) without (iii) imply the existence of the synchronous solution but do not imply its uniqueness. Condition (i) without (ii)–(iii) means that the oscillators often but not always synchronize.

2.3. Oscillator chains in the phase reduction limit. The behavior of coupled pairs of phase oscillators generalizes to that of chains. Here we review basic results based on [22, 25] and outline some consequences of Theorem 2.1.

Consider a chain of n identical oscillators. For one-way nearest-neighbor coupling and if $H(\psi)$ crosses 0, phase differences between adjacent oscillators are equal to those between a coupled pair, being determined by the stable zeros of H. For two-way coupling ($H_1(\psi)$ and $H_2(\psi)$) in a long chain ($n >> 1$), one direction is typically dominant and phase relations are unaffected by connections in the other, except near boundaries. In special cases (e.g., $H_1 \approx H_2$) neither direction dominates and phase differences may be nonuniform. If the coupling is translation-symmetric and close but not necessarily adjacent oscillators are coupled, then the chain mimics the behavior of a reduced network with nearest-neighbor connections.

Thus, our analysis of a pair of units also explains the behavior of a wide class of chains. Case (a) of Theorem 2.1 ($0 < \lim_{\mu\to 0}\Delta(\mu) < \pi$) means that the unit that receives coupling is advanced in phase compared to the other. Analogously, a chain exhibits traveling waves propagating against the coupling direction (against the dominant direction for two-way coupling), and, according to the theorem, such traveling waves persist in the relaxation limit $\mu \to 0$. In contrast, case (b) corresponds to a phase lag of the unit receiving coupling that vanishes

in the relaxation limit. The corresponding phenomenon in chains is a traveling wave that propagates in the (dominant) coupling direction and approaches synchrony in the relaxation limit.

3. Relaxation oscillators in the FTM limit. We again consider the system (2.1)–(2.2), but now under the assumption $\mu \ll \epsilon \ll 1$. FTM theory describes the interaction of relaxation oscillators that exhibit sufficiently fast jumps simultaneously ($\mu \ll \epsilon$). It neglects interactions during the periods of slow dynamics, so most FTM results have been qualitative in nature. Here we augment these results by combining FTM with phase reduction theory, assuming $\epsilon \ll 1$ so that the latter applies except near jumps. We retain the notation of section 2 with phase ϕ along the unperturbed limit cycle Γ and $(x(\phi), y(\phi))$ denoting points in the phase plane.

3.1. FTM theory. FTM theory and the synchronization of relaxation oscillators are described in detail in [35] via the example of a mutually coupled pair. (Chains and other networks are considered in [36].) Here we perform a similar analysis of a pair with one-way coupling.

Since ϵ is large compared to μ, we consider an unperturbed limit cycle ($h \equiv 0$) and a separate, perturbed limit cycle for $h \equiv 1$; see Figure 6(A). Since O_2 receives no coupling signal it follows the former. Input to O_1 is either 1 or 0, depending on the state of O_2, so O_1 intermittently switches between the two cycles. Jumps are assumed to be instantaneous. (In phase reduction, one considers only the unperturbed limit cycle, but "jumps" are not instantaneous.)

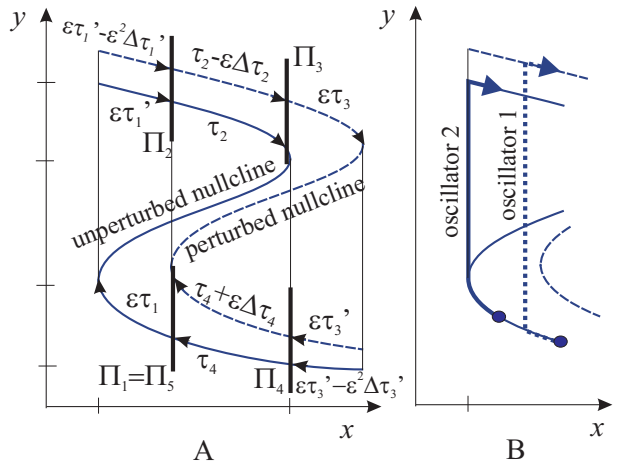


Figure 6. A: *Unperturbed (solid) and perturbed (dashed) nullclines.* Π_i *and* τ_i *represent Poincaré sections and times required to pass certain trajectories, respectively, as used in Appendix B.* B: *An example of FTM interaction:* O_2 *(solid) slightly leads* O_1 *(dashed); when* O_2 *jumps up,* O_1 *switches to the perturbed nullcline, leading to a synchronous jump.*

To illustrate FTM interaction, assume that the oscillators are almost synchronized and moving on the silent branch with O_1 slightly lagging behind O_2. When O_2 reaches the knee and jumps, it sends O_1 to the perturbed limit cycle, thereby causing a synchronous jump, after which O_1 takes the lead. See Figure 6(B). Similarly, synchronous jumps occur if O_1 slightly leads O_2 prior to jumping down. This typically results in rapid and robust synchronization.

The example shows that oscillators with FTM interaction can compensate for deviations from perfect synchrony by keeping fast jumps synchronous and modulating their locations. However, while FTM interactions at jumps act to synchronize the units, accumulating interactions during slow phases may shift them apart. The relative strength of the two effects determines whether synchronization occurs.

The fact that the phase equation (2.3) is not applicable near jumps is illustrated by the following example. Consider two impulsive perturbations, each of strength and duration $\mathcal{O}(\epsilon)$, delivered when O_1 is on the orbit segment of length $\epsilon\tau_1$ in Figure 6(A). Either one alone immediately moves the state to the upper branch, causing an $\mathcal{O}(\epsilon)$ phase change, large compared to its size of $\mathcal{O}(\epsilon^2)$ (strength $\times$ duration). However, if the impulses act successively, the second has an effect of only $\mathcal{O}(\epsilon^2)$. The implicit assumption of phase reduction theory, that successive perturbations are additive, is violated.

3.2. Phase-locking and synchrony. Examples like those above show that the notion of phase difference is unclear, so we introduce the following definitions to categorize different types of T-periodic interactions of pairs of oscillators under FTM:

(a) *Oscillator* O_j *leads* O_i if O_i is silent when O_j jumps up and O_i is active when O_j jumps down.

(b) *The oscillators alternate* if when either jumps up or down, the other is silent, or, alternatively, if when either jumps up or down, the other is active. Only the first case is possible under condition (iii).

(c) *The oscillators synchronize* if the time intervals that O_i spends in its active state are a subset of those spent by O_j in its active state or vice versa. This includes the case when jumps up and/or down are synchronous.

We can now state an analogue of Theorem 2.1.

Theorem 3.1. *Assume that two oscillators in the FTM limit ($\epsilon \to 0$, $\mu/\epsilon \to 0$) each have stable T-periodic solutions. Then*

(a) *if conditions* (i)–(iii) *and inequality* (2.8) *hold, O_1 leads O_2; and*

(b) *if conditions* (i)–(iii) *hold but inequality* (2.8) *does not, the oscillators synchronize.*

Proof. The lengthy proof is given in Appendix B. It relies primarily on defining a function H_{FTM}, analogous to H of section 2, which predicts the relative dynamics, and showing that H_{FTM} has the same shape in the limit $\epsilon \to 0$, $\mu/\epsilon \to 0$ as H does in the limit $\mu \to 0$, $\epsilon/\mu \to 0$. ■

In case (b) of Theorem 3.1 synchrony implies either that the jump in O_2 initiates an immediate jump in O_1 or that one of the oscillators jumps up earlier and jumps down later than the other. If there are synchronous jumps, it is intuitively clear that for small but nonzero μ this results in a small lag of the driven oscillator. Thus cases (b) of Theorems 2.1 and 3.1 are closely related. In contrast, if there is synchrony as defined above but no synchronous jumps, relations between the two theorems are less clear. However, in Appendix C we illustrate that the latter scenario is nongeneric for $\epsilon \to 0$.

There are some differences between the phase and FTM limits. For phase oscillators, $\Delta \backsim \mu^{2/3}$, as shown in Theorem 2.1(b). In the FTM case the $O(\mu)$ duration of fast jumps implies that $\Delta \backsim \mu$. We illustrate this by a numerical example in section 5.

3.3. Oscillator chains in the FTM case. As in phase response theory, the behavior of a unidirectionally coupled pair has implications for chains with one-way, nearest-neighbor coupling; i.e., in cases (a) and (b), chains typically exhibit traveling waves against and in the coupling direction, respectively, and in the relaxation limit synchrony results in case (b) but not in case (a). We suspect that more diffuse localized couplings can be reduced to the nearest-neighbor case, although we are unaware of specific studies of this type.

The behavior of chains FTM-coupled in both directions is less transparent than in case of phase-coupling. Oscillator pairs and arrays with *symmetrical* bidirectional FTM-coupling typically exhibit synchrony, and in contrast to phase-interaction, this is robust against perturbations of the coupling symmetry [35, 36]. These results indicate that asymmetrically coupled arrays (significantly stronger in one direction than in the other) are probably more likely to synchronize than those with unidirectional connections. Quantitative conditions for synchrony versus traveling waves for two-way coupling appear to be unknown. As we show in section 4, traveling wave behavior is much rarer than synchrony in one-way arrays. The above facts suggest that it is even rarer when connections in both directions are present.

4. Why do most oscillators synchronize? It was shown in sections 2 and 3 that oscillators can, but need not, synchronize in the relaxation limit $\mu \to 0$. As described in this section, we examined several simple two-dimensional relaxation oscillator models of neural oscillators and found that all exhibited synchrony. This suggests that inequality (2.8) of Theorem 2.1 is false for many examples. The following theorem provides a sufficient condition for this and hence for synchronization. Two oscillators satisfying conditions (i)–(iii) above and (iv) and (v) below always synchronize in the relaxation limit by Theorems 2.1 and 3.1.

Theorem 4.1. *Suppose that* (2.1)–(2.2) *satisfy condition* (ii)*, and additionally, at all points* (x, y) *on the "active" branch of the slow manifold* $g = 0$*, the following conditions hold.*

- *Condition* (iv): $f_x(x, y) \le 0$.
- *Condition* (v): $g_x(x, y) \le g_x(x(D), y(D))$.

Then inequality (2.8) *is false.*

Proof (by contradiction). The integral in (2.8) (or (2.9)) is evaluated along $g = 0$, on which an infinitesimal displacement $(d\xi, d\upsilon)$ satisfies

$$g_x(\xi, \upsilon)\, d\xi + g_y(\xi, \upsilon)\, d\upsilon = 0 \quad \text{or} \quad d\xi = -\frac{g_y(\xi, \upsilon)\, d\upsilon}{g_x(\xi, \upsilon)}. \tag{4.1}$$

We use (4.1) to change the variable of integration in (2.8) from ξ to υ, replacing $(\xi, \upsilon(\xi))$ by $(\xi(\upsilon), \upsilon)$, where $\xi(\upsilon)$ denotes the inverse of $\upsilon(\xi)$. Since $g_x < 0$ on the active nullcline, due to $g_x(x(D), y(D)) < 0$ and condition (v), this inverse is well defined in the case of interest. Inequality (2.8) becomes

$$\begin{aligned} &-\int_{y(D)}^{y(0+)} \frac{f_y(\xi(\upsilon), \upsilon)}{f^2(\xi(\upsilon), \upsilon) g_x(\xi(\upsilon), \upsilon)}\, d\upsilon \\ &+ \frac{1}{g_x(x(D), y(D))} \left[\frac{1}{f(x(D), y(D))} - \frac{1}{f(x(D), y(D^+))} \right] > 0. \end{aligned} \tag{4.2}$$

To show that (4.2) is false we first replace $g_x(x, y)$ in the integrand by the constant term $g_x(x(D), y(D))$, using the facts that $g_x(x, y) \le g_x(x(D), y(D)) < 0$ and $f_y < 0$ (condition (ii)),

which together imply that

$$\int_{y(D)}^{y(0+)} \frac{f_y(\xi(\upsilon),\upsilon)}{f^2(\xi(\upsilon),\upsilon)g_x(\xi(\upsilon),\upsilon)}\,d\upsilon \le \int_{y(D)}^{y(0+)} \frac{f_y(\xi(\upsilon),\upsilon)}{f^2(\xi(\upsilon),\upsilon)g_x(x(D),y(D))}\,d\upsilon.$$

We then multiply the resulting expression by the strictly positive quantity $-g_x(x(D),y(D))$ to deduce that, if (4.2) holds, then also

$$\int_{y(D)}^{y(0+)} \frac{f_y(\xi(\upsilon),\upsilon)}{f^2(\xi(\upsilon),\upsilon)}\,d\upsilon - \frac{1}{f(x(D),y(D))} + \frac{1}{f(x(D),y(D^+))} > 0,$$

which in turn implies that

$$\int_{y(D)}^{y(0+)} \frac{f_y(\xi(\upsilon),\upsilon)}{f^2(\xi(\upsilon),\upsilon)}\,d\upsilon - \frac{1}{f(x(D),y(D))} > 0, \tag{4.3}$$

where we use $1/f(x(D),y(D^+)) < 0$, since $f < 0$ on the "silent" branch of the slow manifold.

Next, using the chain rule and appealing to condition (iv) and the fact that $\xi(\upsilon)$ is a decreasing function, we have

$$\frac{df(\xi(\upsilon),\upsilon)}{d\upsilon} = f_y + f_x\frac{d\xi(\upsilon)}{d\upsilon} \ge f_y. \tag{4.4}$$

Equation (4.4) allows us to replace the partial derivative in the integrand of (4.3) by the total derivative and further to express it as an exact differential,

$$\int_{y(D)}^{y(0+)} \frac{df(\xi(\upsilon),\upsilon)/dy}{f^2(\xi(\upsilon),\upsilon)}\,d\upsilon = -f^{-1}(x_a(\upsilon),\upsilon)\big|_{y(D)}^{y(0+)}, \tag{4.5}$$

where $x_a(\upsilon)$ denotes points on the active branch and we use the fact that $(f^{-1})' = -f'/f^2$. Our inequality now reads

$$-f^{-1}(x_a(\upsilon),\upsilon)\big|_{y(D)}^{y(0+)} - \frac{1}{f(x(D),y(D))} = \frac{-1}{f(x(0),y(0+))} > 0. \tag{4.6}$$

But this is false, since $f(x,y) > 0$ on the active branch, providing our contradiction. ■

Theorem 4.1 applies to many oscillator models with unidirectional coupling that satisfies condition (i). We analyzed the van der Pol oscillator [37] and neuron models of FitzHugh–Nagumo [12, 33], Hindmarsh–Rose [16], Morris–Lecar [32], and Rinzel [34], as well as a two-dimensional spike-rate description of the bursting half-center in the lamprey from [42, p. 209], obtaining the results summarized in Table 1. We also checked inequality (2.8) (in some cases numerically) and found that it was false in every case, in four of which Theorem 4.1 applies. This suggests that it is not easy to find relaxation oscillators that do not synchronize. Nonetheless, in the next section we provide an example of this apparently rare behavior.

Table 1

Evaluation of the applicability of conditions (ii)–(v) *and Theorem* 4.1 *for some simple oscillators. The * means yes for low values of input current, and no for high values.*

Model ODE	(ii)	(iii)	(iv)	(v)	Thm. 4.1 applies?	Thm. 2.1 predicts synchrony?
van der Pol	yes	yes	yes	yes	**yes**	**yes**
FitzHugh–Nagumo	yes	*	yes	yes	*****	**yes**
Hindmarsh–Rose	yes	yes	yes	yes	**yes**	**yes**
Morris–Lecar	yes	yes	yes	yes	**yes**	**yes**
Rinzel	yes	yes	yes	no	**no**	**yes**
lamprey halfcenter	yes	yes	yes	no	**no**	**yes**

5. A numerical example. To demonstrate the above results, we now analyze a pair of van der Pol oscillators in Lienard variables [28, Chap. XI], [29], with one-way excitatory coupling. At the end of the section simulation results of chains are also shown. The uncoupled units include a parameter p that can be varied to produce two characteristic behaviors. $p = 0$ corresponds to the classical van der Pol oscillator. Oscillator 1 is described by

$$\begin{aligned} \dot{x_1} &= f(x_1, y_1) = y_1, \\ \mu \dot{y_1} &= g(x_1, y_1) + \epsilon h(y_2) \end{aligned} \tag{5.1}$$

$$= \left(y_1 - y_1^3/3 - x_1\right) \cdot \left(1 + (p \cdot x_1)^4\right) + \epsilon \left\{ \begin{array}{lll} 1 & \text{if} & y_2 > 0 \\ 0 & \text{if} & y_2 \le 0 \end{array} \right\}, \tag{5.2}$$

and the equations of oscillator 2 are the same but lack the coupling term $\epsilon\{\ldots\}$. The coupling obeys condition (i) and the uncoupled oscillators satisfy conditions (ii), (iii), and (iv) since $f_y \equiv 1$, the duty cycle is exactly 1/2 due to the symmetry of the vectorfield, and $f_x \equiv 0$. For $p = 0$, they also satisfy (v), because $g_x \equiv -1$. Thus they synchronize in the relaxation limit $\mu \to 0$ with extremely weak coupling (by Theorems 2.1 and 4.1) as well as with moderate coupling (by Theorems 3.1 and 4.1).

If $p \neq 0$, the S-shaped fast nullclines of (5.2) remain the same, but $g(x, y)$ becomes steeper as $|x_j|$ increases so that condition (v) does not hold and Theorem 4.1 cannot be applied. Numerical evaluations show that inequality (2.8) fails for $p < p_{cr} \approx 2.36$ but holds for $p \geq p_{cr}$. In the latter case, Theorems 2.1 and 3.1 predict persistent phase-shifts as $\mu \to 0$.

The oscillator pair was simulated with $\epsilon = 0.5$ and various values of μ and p. In every case, the system converged to a stable periodic orbit with period T equal to that of an uncoupled unit. The jth increasing zero-crossing of $y_1(t)$ and $y_2(t)$ (i.e., times t_{ij} when $y_i(t_{ij}) = 0$, $\dot{y}_i t_{ij} > 0$) were detected and the phase-shift Δ was determined according to

$$\Delta = \lim_{j \to \infty} \frac{t_{2j} - t_{1j}}{T}. \tag{5.3}$$

For weakly coupled relaxation oscillators ($\epsilon, \mu \to 0$), the respective meanings of $\pi > \Delta > 0$, $\Delta = 0$, and $-\pi < \Delta < 0$ are that the driven oscillator leads, is synchronous with, and lags behind the driver. Here ϵ is not very small, so $\Delta \approx 0$, but $\neq 0$ might also correspond

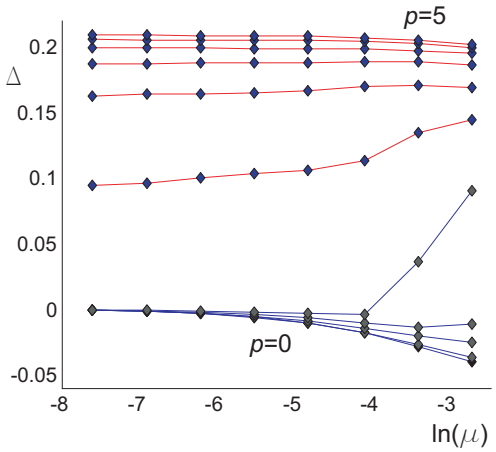


Figure 7. *Dependence of phase-shift between the coupled pair of oscillators* Δ *on* μ*. From bottom to top,* $p = 0, 0.5, 1, \ldots, 5$*. For small p, the shift is negative (the driven oscillator lags) and vanishes at the relaxation limit* $\mu \to 0$*; for large p the shift is positive and persists in the relaxation limit. Note that* $p = 2$ *is small in this regard, although it shows different behavior from the* $p < 2$ *cases for larger* μ *(fifth curve from the bottom).*

to synchrony (cf. the definition of synchrony in section 3.2). Nevertheless Δ is still a good indicator of the phase difference.

Figure 7 illustrates the dependence of Δ on μ for different values of p, showing that $\Delta(\mu)$ curves below 0, for small μ converge to 0 as $\mu \to 0$. Curves above 0, however, converge to a strictly positive limit. This corresponds to one of our main findings: if the driven oscillator leads the driver, the phase difference persists in the relaxation limit, but if the driver leads, the difference vanishes.

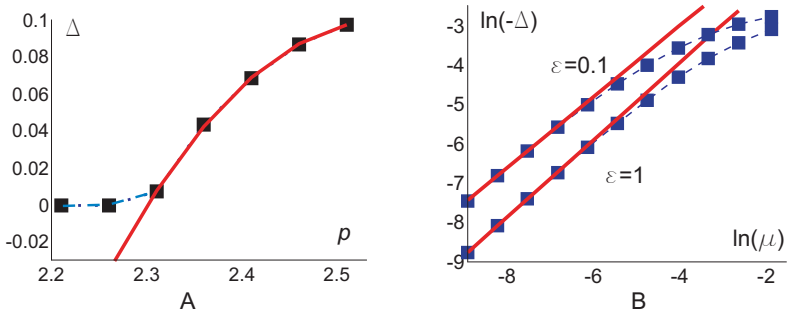


Figure 8. A: *Phase-shift* Δ *as function of p for* $\mu = 10^{-3}$*,* $\epsilon = 0.5$ *(squares), and a fitted quadratic curve* $\bar{\Delta}(p)$ *(solid line).* B: *Dependence of* Δ *on* μ *for* $p = 0$ *plotted on logarithmic scales for* $\epsilon = 1$ *and at* $\epsilon = 0.1$ *(squares). Linear regression in the range* $\mu = 10^{-4} \ldots 10^{-3}$ *reveals* $ln(-\Delta) \approx 0.98 \ln(\mu) + 0.04$ *and* $ln(-\Delta) \approx 0.90 \ln(\mu) + 0.67$*, respectively (solid lines).*

Figure 8(A) shows the numerically derived function $\Delta(p)$ for $\mu = 10^{-3}$ and $\epsilon = 0.5$. A quadratic fit for $\Delta > 0$ yields $\bar{\Delta}(p) = -1.646p^2 + 8.382p - 10.572$, whose root at $\bar{p}_{cr} \approx 2.301$ lies within 3% of $p_{cr} \approx 2.36$ predicted by inequality (2.8). The difference is primarily due to

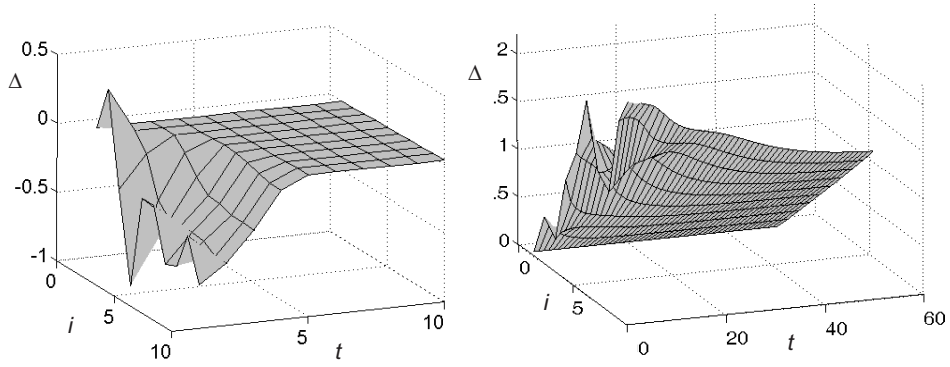


Figure 9. *Dynamics of a chain of* 10 *oscillators at* $p = 0$ *(left) and* 3 *(right). Each oscillator* i *is driven by its neighbor* $i-1$*;* t *denotes time normalized by the period of an uncoupled unit.* Δ *denotes the phase difference between units* i *and* 1*. Other parameter values:* $\epsilon = 0.5$*,* $\mu = 2 \cdot 10^{-3}$*.*

the relatively strong coupling.

We also determine numerically how Δ scales with μ for small p to examine $\lim_{\mu \to 0} \Delta = 0$. The predictions of sections 2.2 and 3.2 are $\Delta \backsim \mu^{2/3}$ for phase-oscillator interactions and $\Delta \backsim \mu^{1}$ for FTM interactions. According to simulations with $p = 0$, the exponent is approximately 0.97 in the range $\mu = 10^{-4} \ldots 10^{-3}$ if $\epsilon = 1$ and 0.90 if $\epsilon = 0.1$ (Figure 8(B)). These results reflect the fact that our parameter values are appropriate for FTM ($\mu << \epsilon << 1$); but they also show that the exponent decreases as ϵ decreases, moving toward the phase-approximation value, which holds for $\epsilon << \mu << 1$.

We close this section by illustrating the two types of behavior for chains. The two panels of Figure 9 show the spatio-temporal dynamics of phase-shifts along a chain of 10 unidirectionally coupled units. For $p = 0$ (left panel), the network rapidly synchronizes, while for $p = 3$ uniform phase-shifts develop on a longer time-scale. The final states agree with the predictions of Theorem 2.1. (The reason for the radically different decay times of transients is explained in [35].)

6. Applications to the lamprey CPG. The central pattern generator (CPG) of the lamprey has been a focus of research for over thirty years. Fictive swimming experiments [9] show that the CPG without muscles or afferent (feedback) inputs produces rhythms similar to real swimming: traveling waves of activation (motoneuron bursts) propagate from head to tail on both sides in antiphase. The wavelength remains approximately constant and equal to body length over a considerable speed range.

The components of the CPG and their interconnectivities have been partially determined [5], and a reduced network with three classes of neurons, representing one or a few segments of the animals' CPG, has been proposed; see Figure 10(A). Each segment has bilateral symmetry, with mutual inhibition between hemisegments. The entire CPG is modeled as a chain of such units [21, 23], intersegmental connections being of the same type as intrasegmental ones, but with strengths decreasing rapidly with distance. For simplicity here we assume only nearest-neighbor connections, but our results can be extended to more

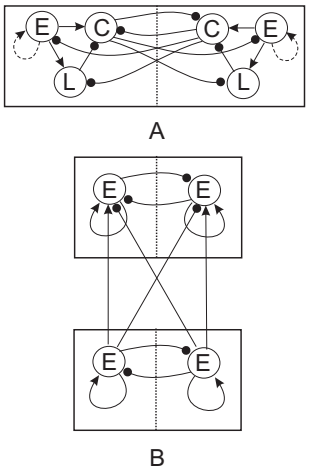


Figure 10. A: *Simplified structure of a segmental unit of the lamprey CPG first proposed by* [6]. *E, L, and C represent small groups of excitatory, crossed inhibitory, and lateral inhibitory neurons, respectively, and arrows and circles denote excitatory and inhibitory synapses. Bilaterally symmetric halves of the network are coupled by inhibition. The dashed self-excitatory connection occurs in some but not all models.* B: *A simplified network with E cells alone approximates the dynamics of cell-based rhythm generation (cf.* [38]*); two segmental units are shown.*

widespread connections, provided that they are short relative to the size of the full network.

Several models have implemented the network architecture of Figure 10(A), and two different pattern-generating mechanisms have been proposed [13]: rhythms being generated by network connections, or by small groups of bursting cells. Simple *network-based* models [4, 40] were able to reproduce the constant wavelength-swimming speed behavior, albeit over a limited frequency range. The *cell-based* mechanism inspired a series of detailed model studies [26, 39, 27] which encompass a wider frequency range, but with frequency-dependent wavelengths (small at low frequencies, large at high frequencies).

The core of the cell-based networks is a double chain of relaxation-type oscillators (Figure 10(B); see also [38]), each representing a small group of intrinsic bursters. The fast and the slow variables represent average activity (firing rate) and spike-rate adaptation due to slow calcium currents, respectively. In these models, swimming frequency is adjusted through serotonin concentration determining the speed of the slow dynamics, i.e., μ. (Other, less important, parameters also change with frequency, but we ignore these effects.) Thus, when swimming speed decreases, the speeds of the slow and the fast dynamics separate, approaching a relaxation limit. The tendency of chains of relaxation oscillators to synchronize offers a straightforward explanation for the wavelength-frequency behavior of these networks, and, as we now show, the results of sections 2–4 yield more precise predictions on cell-based CPG models.

The double chains of Figure 10 differ from the single chains studied in sections 2.3 and 3.3, but their behavior can be predicted in a similar manner by considering a pair of segments comprising four oscillators, as in Figure 10(B). Specifically, we assume condition (i) above, that the hemisegments remain out of phase, and we denote the strengths of intersegmental excitatory and inhibitory connections by ϵ_e and ϵ_i, respectively. Phase response theory

then yields the averaged coupling function $H_{\mu\,pair}$ governing intersegmental phase differences (cf. (2.4)) at given values of μ by superposing the contributions of the excitatory and inhibitory connections. In the relaxation limit $\lim_{\mu\to 0} H_{\mu\,pair} \stackrel{\text{def}}{=} H_{0\,pair}$ we obtain

$$\lim_{\mu\to 0} H_{\mu\,pair} \stackrel{\text{def}}{=} H_{0\,pair}(\psi) = \frac{1}{2\pi}\left[\epsilon_e \int_{\psi}^{D+\psi} z^*(\phi)\,d\phi - \epsilon_i \int_{\psi+\pi}^{D+\psi+\pi} z^*(\phi)\,d\phi\right]. \tag{6.1}$$

Equation (6.1) is the analogue of (2.7) with a second term $-\epsilon_i \int \ldots$ due to inhibition between the segments. Its consequence is a somewhat weaker and (in part (a)) more technical analogue of Theorem 2.1.

Theorem 6.1. *Let $(x, y) = (\xi, \upsilon(\xi))$ and $(x, y) = (\xi, \zeta(\xi))$, respectively, denote points on the active and silent branches of the $\dot{y} = 0$ nullcline. Assume that $\Delta(\mu)$ satisfies $dH_{\mu\,pair}(\psi)/d\psi|_{\Delta(\mu)} < 0 = H_{\mu\,pair}(\Delta(\mu))$.*

(a) *If conditions* (i)–(iii) *hold, and*

$$\begin{aligned} &-\epsilon_e \int_{x(0)}^{x(D)} \frac{f_y(\xi,\upsilon(\xi))}{f^2(\xi,\upsilon(\xi))g_y(\xi,\upsilon(\xi))}\,d\xi \\ &+\epsilon_i \int_{x(\pi)}^{x(\pi+D)} \frac{f_y(\xi,\zeta(\xi))}{f^2(\xi,\zeta(\xi))g_y(\xi,\zeta(\xi))}\,d\xi \\ &+\epsilon_e \frac{1}{g_x(x(D),y(D))}\left[\frac{1}{f(x(D),y(D))} - \frac{1}{f(x(D),y(D+))}\right] > 0, \end{aligned} \tag{6.2}$$

then for a coupling function of the form $H_{0\,pair}(\phi) = \Xi(\phi) + c$ (where Ξ is an arbitrary 2π periodic function and c is an arbitrary constant), there exists $\delta > 0$ independent of c such that if $|\Delta(0)| < \delta$, then $\Delta(0) > 0$.

(b) *If conditions* (i)–(iii) *hold, Δ is unique modulo 2π, and inequality* (6.2) *is false, then $\Delta(\mu) \backsim -\mu^{2/3}$ for small μ, and $\Delta(0) = 0$; i.e., the driven segment lags the driver and the segments synchronize in the relaxation limit.*

Proof (sketch). The proof of Theorem 6.1 is similar to that of Theorem 2.1, so we outline only the main points and differences between them.

Referring to (6.1) and the proof of Theorem 2.1, and noting that (6.2) has no "boundary" term due to inhibitory connections ϵ_i, since for duty cycle $D/(2\pi) < 0.5$ and $\phi_{ij} \approx 0$ the driven oscillator jumps when the inhibitory driver is inactive, we see that the left-hand side of inequality (6.2) is equal to $H_{0\,pair}(0)$. Hence, cases (a) and (b) of the theorem respectively correspond to $H_{0\,pair}(0) > 0$ and ≤ 0. The first term $\epsilon_e \int \ldots$ in (6.1) has a jump at $\phi = 0$, at the top of which it is positive, and this term is continuous on the left side of the jump provided that $-D < \phi < 0$ (see the proof of Theorem 2.1). The second term $-\epsilon_i \int \ldots$ is positive and continuous in ϕ if $|\phi| < (\pi - D)$ because z^* is negative and integrable in the interval $D < \phi < 2\pi$; cf. Figure 3. Hence, $H_{0\,pair}(\phi)$ itself has a jump at 0, is positive at the top of the jump ($H_{0\,pair}(0^-) > 0$), and is continuous in the interval $(-\min\{D, \pi - D\}, 0)$.

In case (b), at the bottom of the jump $H_{0\,pair}(0^+)$ is negative as in Theorem 2.1, so it has a stable root at 0, and because of uniqueness, $\Delta(0) = 0$. For μ small but positive, $\Delta(\mu) \backsim -\mu^{2/3}$ for the same reason as in Theorem 2.1. In case (a) at the bottom of the jump $H_{0\,pair}(0^+)$ is already positive, and thus $H_{0\,pair}(0^-) > s > 0$, where s denotes the magnitude

of the jump, which is independent of the constant c in the definition of $H_{0\,pair}$. On the negative side of 0, there is an interval in which $H_{0\,pair}(\phi)$ is continuous (see above). Thus by definition there exists δ (again independent of c) such that for arbitrary $0 < 0 - \phi < \delta$, we have $|H_{0\,pair}(0^-) - H_{0\,pair}(\phi)| < s$, yielding $H_{0\,pair}(\phi) > 0$. Thus if $|\Delta(0)| < \delta$, $\Delta(0)$ must be positive. ■

The implication of part (a) is that if $|\Delta(0)| << 1$, then $\Delta(0) > 0$; the driven segment leads the driver. Hence the statement in part (a) is similar to, and more specific than, part (a) of Theorem 2.1. Theorem 6.1 has two restrictions in addition to those of Theorem 2.1—the uniqueness of Δ in part (b) and its closeness to zero in part (a)—but neither affects its applicability to lamprey CPGs. Uniqueness means that the CPG has a unique stable traveling wave solution in agreement with the observation that the lamprey exhibits a single robust pattern of motion. It is also reasonable that $|\Delta(0)| << 1$ since the lamprey's notocord has $\mathcal{O}(100)$ segments, so intersegmental phase differences must be $\mathcal{O}(0.01 \times 2\pi)$ if wavelength is to equal body length.

Phase response theory does not always apply to the lamprey CPG, because intersegmental coupling is not necessarily weak, so FTM interactions may be more appropriate. Much as Theorem 2.1 has an analogous statement in Theorem 3.1, an analogue of Theorem 6.1 holds under the assumption of FTM interactions. Here we give an informative but inexact version, without proof.

Theorem 6.2. *Assume that two pairs of oscillators in the FTM limit ($\epsilon_{e,i} \to 0$, $\mu/\epsilon_{e,i} \to 0$) each have stable T-periodic solutions. Let Δ_{FTM} denote the time difference between activation of the ipsilateral driven and driver oscillators (positive if the driven activates first). Then,*

(a) *if conditions* (i)–(iii) *and inequality* (6.2) *hold and $|\Delta_{FTM}| << 1$, then the driven oscillators lead the drivers; and*

(b) *if conditions* (i)–(iii) *hold but inequality* (6.2) *fails and the network has only one stable T-periodic solution, then the oscillators synchronize (away from the relaxation limit, the driven segment lags the driver).*

Provided that the CPG has unidirectional intersegmental coupling, the consequences of Theorems 6.1–6.2 for the lamprey are as follows.

1. If inequality (6.2) fails, neighboring segments display a negative phase difference that vanishes in the relaxation limit. A chain will therefore exhibit waves that propagate in the direction of the intersegmental coupling, with wavelength approaching zero in the relaxation limit. Hence, in such models the wavelength is usually an increasing function of μ, instead of a constant with frequency. This behavior was seen in previous numerical simulations.

2. If inequality (6.2) holds, segments have positive phase differences that persist in the relaxation limit and traveling waves will propagate against the coupling direction, with approximately constant wavelength as μ is varied. Hence, intersegmental connections must be directed from tail to head to obtain head-to-tail traveling waves.

Our first important finding is that, if (6.2) holds, a double chain of oscillators can combine the advantages of previous cell- and network-based CPG models, namely, wide frequency-range and constant wavelengths. Theorem 4.1 implies that single chains rarely satisfy the analogous inequality (2.8), but the presence of an additional positive term $\epsilon_i \int \dots$ in (6.2), due to cross-inhibitory connections, provides more flexibility. On the other hand, as shown in section 3.3, the bidirectional coupling in the real network promotes synchrony under the

assumption of FTM interactions. Hence finding "well-behaved" models is difficult, but probably not hopeless, since the high number of ad hoc parameters in such models allows wide freedom for improvement. The failure of such an attempt would suggest that the cell-based mechanism is an oversimplification and that more sophisticated models, perhaps combining both rhythm-generating mechanisms, are required.

It is also worth noting that caudo-rostral (tail-to-head) coupling is required to produce waves that travel from head to tail. This confirms previous studies [41, 20] that used completely different arguments to show that ascending is stronger than descending coupling in the lamprey notocord.

7. Conclusions. This paper concerns coupled sets of planar relaxation oscillators. We focus on pairs of oscillators with unidirectional coupling but draw conclusions for two-way coupling and linear chains of oscillators. Our main theorems, Theorems 2.1 and 3.1, provide sufficient conditions for persistent phase lags and for synchrony in the limits of weak coupling and of large time-scale separation, using phase response theory and FTM theory, respectively. The key step involves estimation of an inequality (2.8) arising from the averaged coupling function.

Theorem 4.1 provides a sufficient condition for synchrony, and in section 4 we show that several models of bursting neurons satisfy this condition, which we conjecture to be the typical case. However, counterexamples can be found, as demonstrated in section 5. Finally, in section 6 we extend these results to the double chains featured in models of CPGs for swimming in lamprey, providing analogues of Theorems 2.1 and 3.1 in Theorems 6.1 and 6.2. These results partially explain why cell-based models of relaxation type approach synchrony as swimming speed decreases, violating the experimental observation of near-constant phase lags over a wide speed range, but they also offer hope that parameterizations that permit the observed behavior may be found.

More generally, the results in this paper reveal interesting relations between phase response and FTM theory, which apply in the distinctly different limits of weak coupling ($1 \gg \mu \gg \epsilon \to 0$) and strong time-scale separation ($1 \gg \epsilon \gg \mu \to 0$). In particular, we construct a composed Poincaré return map in the latter relaxation limit that is the analogue of the averaged coupling function in the former limit. This map is used to demonstrate that the tendency of unidirectionally coupled pairs or arrays of oscillators to synchronize is unaffected by extreme changes of the ratio ϵ/μ despite evident differences between the resultant coupling mechanisms. However, we also find that the rates of convergence to synchrony scale differently as perfect time-scale separation ($\mu \to 0$) is approached. In case of FTM interaction, our study raises further questions regarding the behavior of bidirectionally but asymmetrically coupled arrays as well as that of arrays with multiple (non–nearest-neighbor) coupling. A similar approach to the present one may be helpful in studying synchronization properties of such networks.

Appendix A. Theorem 2.1: The PRC at jumps. Here we locally approximate the PRC near jumps, showing that relaxation oscillators have large phase response values during transition, shortly before (but not during) jumps. This fact is central in the proof of Theorem 2.1.

We assume small but nonzero μ, in which case the stable limit cycle Γ of the ODEs (2.1)–(2.2) is close to but not exactly the same as that shown in Figure 1. Since notation like $x(\phi)$

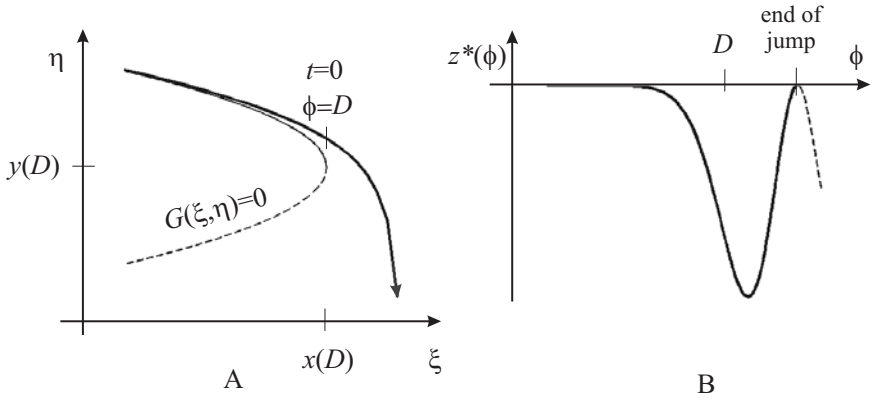


Figure 11. A: *An orbit in the neighborhood of a downward jump.* B: *The PRC near the jump.*

has been used to denote coordinates on Γ as $\mu \to 0$, we now use different notation ξ and η for the slow and the fast variables, respectively. We define time t and phase ϕ such that $t = 0$ and $\phi = D$ when $\xi = x(D)$; see Figure 11(A).

Leading order terms in the PRC near jumps at nondegenerate (quadratic) turning points are determined by the local approximation

$$\dot{x} \approx f(x(D), y(D)), \tag{A.1}$$

$$\mu\dot{y} \approx g_x(x(D), y(D))\,[x - x(D)] + \frac{g_{yy}(x(D), y(D))}{2}\,[y - y(D)]^2\,. \tag{A.2}$$

These ODEs have an explicit solution in terms of Airy functions and their derivatives [1], denoted below by Ai, Bi, Ai', and Bi':

$$\xi(t) = x(D) + ft, \tag{A.3}$$

$$\eta(\xi) = y(D) + \mu^{1/3}\left(\frac{4fg_x}{g_{yy}^2}\right)^{1/3}\frac{Ai'(\zeta) + a \cdot Bi'(\zeta)}{Ai(\zeta) + a \cdot Bi(\zeta)}, \tag{A.4}$$

where a is an arbitrary constant, the arguments $(x(D), y(D))$ have been suppressed (i.e., $f = f(x(D), y(D))$, etc.), and ζ is a rescaled version of ξ,

$$\zeta = -\mu^{-2/3}\left(\frac{g_x g_{yy}}{2f^2}\right)^{1/3}(\xi - x(D)). \tag{A.5}$$

The parameter a is determined by the asymptotic boundary condition that for $\xi \to -\infty$ the orbit follows the upper branch of the $g = 0$ nullcline, implying that $\eta(\xi) > y(D)$; see Figure 11(A). We note that $\xi \to -\infty$ corresponds to $\zeta \to \infty$ by (A.5), the limiting values of $Ai(\zeta)$, $Ai'(\zeta)$, $Bi(\zeta)$, and $Bi'(\zeta)$ are $+0$, -0, $+\infty$, and $+\infty$, respectively, and, from the nullcline geometry, $f > 0 > g_x$. Using these facts in (A.4), $\lim_{\xi\to-\infty}\eta(\xi) > y(D)$ implies that $a = 0$. The orbit gradually leaves the nullcline and η goes to minus infinity at $\zeta = A \approx -2.3381$, which is a zero of the function Ai. The term $\mu^{1/3}$ in (A.4) implies that, for small μ, η remains close to zero for most values of ζ and suddenly grows just before ζ reaches its critical value.

By definition, the PRC represents the (linear) advancing or retarding effect of a small, instantaneous perturbation $\Delta\eta$ in the fast variable. Such a perturbation results in a switch to another trajectory with $a \approx \Delta\eta \cdot da/d\eta$. The η-coordinate of the perturbed trajectory goes to infinity at $\zeta = \zeta_{end}(a)$. We assume that $\zeta_{end}(a)$ corresponds, via (A.5), to the rescaled slow coordinate ξ_{end} when the jump is finished and slow motion begins on the lower branch of the nullcline; as already noted, $\zeta_{end}(0) = A$. The perturbation changes this by

$$\Delta\xi_{end} = \Delta\eta \frac{da}{d\eta}\frac{d\zeta_{end}}{da}\frac{d\xi}{d\zeta} + \mathcal{O}(\Delta\eta^2). \tag{A.6}$$

To compute the effect of the perturbation, note that during the jump the slow variable evolves according to $\dot{\xi} \approx f(x(D), y(D))$, so that the jump duration increases by $\Delta\xi_{end}/f(x(D), y(D))$. After the jump, the orbit reverses direction in ξ: $\dot{\xi} \approx f(x(D), y(D^+)) < 0$. Thus positive $\Delta\xi$ has a further retarding effect of duration $\Delta\xi_{end}/f(x(D), y(D^+))$. The sum of these two terms represents the "time-response" to the perturbation, and a final scaling factor $d\phi/dt = \omega$ yields the local PRC

$$\begin{aligned}\Delta\phi = -\Delta\eta \frac{da}{d\eta}\frac{d\zeta_{end}}{da}\frac{d\xi}{d\zeta}\omega \left[f^{-1}(x(D), y(D)) - f^{-1}(x(D), y(D^+))\right] \\ + \mathcal{O}(\Delta\mu^2),\end{aligned} \tag{A.7}$$

where the minus sign implies that positive values correspond to shortening of the period.

We compute the components of the product in (A.7) one by one. First, $d\xi/d\zeta$ comes directly from (A.5):

$$\frac{d\xi}{d\zeta} = -\mu^{2/3}\left(\frac{2f^2}{g_x g_{yy}}\right)^{1/3}. \tag{A.8}$$

We find $da/d\eta = (d\eta/da)^{-1}$ using (A.4):

$$\begin{aligned}\frac{da}{d\eta} &= \left[\left(\frac{4\mu f g_x}{g_{yy}^2}\right)^{1/3} \frac{Bi'(\zeta)(Ai(\zeta) + a\,Bi(\zeta)) - Bi(\zeta)(Ai'(\zeta) + a\,Bi'(\zeta))}{(Ai(\zeta) + a\,Bi(\zeta))^2}\bigg|_{a=0}\right]^{-1} \\ &= \left[\left(\frac{4\mu f g_x}{g_{yy}^2}\right)^{1/3} \frac{Bi'(\zeta)\,Ai(\zeta) - Bi(\zeta)\,Ai'(\zeta)}{Ai(\zeta)^2}\right]^{-1}.\end{aligned} \tag{A.9}$$

The term $Bi'(\zeta)\,Ai(\zeta) - Bi(\zeta)\,Ai'(\zeta)$ is constant. The fact that its derivative is 0 follows from the definition of Airy functions: $v\,Ai(v) = Ai''(v)$, $v\,Bi(v) = Bi''(v)$. This straightforward calculation is omitted. Hence we may replace ζ by the constant $A \approx -2.3381$ defined above and use $Ai(A) = 0$ to obtain

$$\frac{da}{d\eta} = -\left[\left(\frac{4\mu f g_x}{g_{yy}^2}\right)^{1/3} \frac{Bi(A)\,Ai'(A)}{Ai(\zeta)^2}\right]^{-1}. \tag{A.10}$$

Finally, we determine $d\zeta_{end}/da$. $\zeta_{end}(a)$ denotes the location of the singularity in (A.4): it is the solution of

$$Ai(\zeta) + a\,Bi(\zeta) = 0. \tag{A.11}$$

Thus we have

$$\frac{d\zeta_{end}}{da} = \left[\frac{da}{d\zeta_{end}}\right]^{-1} = \left[-\frac{d\left[Ai(\zeta)/Bi(\zeta)\right]}{d\zeta}\bigg|_{\zeta=A}\right]^{-1}$$
$$= \frac{Bi^2(A)}{Ai(A)\,Bi'(A) - Ai'(A)\,Bi(A)} = -\frac{Bi(A)}{Ai'(A)}. \tag{A.12}$$

Substituting (A.3), (A.5), (A.8), (A.10), and (A.12) into (A.7), we obtain the PRC in terms of t:

$$\frac{\Delta\phi}{\Delta\eta} \approx \left[\frac{\mu f(x(D),y(D))g_{yy}(x(D),y(D))}{2g_x^2(x(D),y(D))}\right]^{1/3} Ai'^{-2}(A)$$
$$\times\, Ai^2\left(-\mu^{-2/3}\left(\frac{g_x(x(D),y(D))g_{yy}(x(D),y(D))f(x(D),y(D))}{2}\right)^{1/3} t\right)$$
$$\times\, \omega\left[\frac{1}{f(x(D),y(D))} - \frac{1}{f(x(D),y(D^+))}\right]. \tag{A.13}$$

Finally, replacing t by $\phi = \phi_0 + \omega t$, we obtain the approximate PRC $z(\phi)$ during the transition and jump. Note that in spite of its apparent complexity, the formula (A.13) contains only constants and a scaled Ai^2 function, so that we may write

$$z^*(\phi) = \frac{z(\phi)}{\mu} \approx B\mu^{-2/3}\, Ai^2\left(C\mu^{-2/3}(\phi - \phi_0)\right), \tag{A.14}$$

where B, C are $O(1)$ constants. See also Figure 11(B).

This formula demonstrates that large PRC values occur during transition, while the oscillator state is near the upper nullcline. (These correspond to the delta function in the relaxation limit; see (2.6).) The analogous result for the upward jump can be derived in the same way. In that case, the large values occur during transition at the lower nullcline. We remark, without explicit computations, that the integral of this approximation of the phase-response function from $-\infty$ to the end of the jump is equal to the coefficient of the delta function in (2.6).

Appendix B. Proof of Theorem 3.1. The proof is divided into five parts. In the first, notation and concepts are introduced; these include four mappings H_i representing the interactions of the oscillators during the four segments of their limit cycles (slow motions on the upper and lower nullcline branches, and jumps). In the second part we analyze the functions H_i. These results are used in the third part to demonstrate that the only possible forms of T-periodic interactions are synchrony or O_1 leading O_2. The fourth part contains the proof that if O_1 leads, then inequality (2.8) holds, and in the last part we show the converse in the case of synchrony.

B.1. Notation. Let $t_j^{(1)}, \ldots, t_j^{(5)}$ denote the times at which the state of O_j successively crosses the Poincaré sections $\Pi_1, \Pi_2, \ldots, \Pi_5 = \Pi_1$ defined in Figure 6(A). We shall construct a return map

$$\epsilon H_{FTM}(t_2^{(1)} - t_1^{(1)}) = \left[t_2^{(5)} - t_1^{(5)}\right] - \left[t_2^{(1)} - t_1^{(1)}\right], \tag{B.1}$$

which describes the time-shift during one cycle due to coupling, normalized by the coupling strength ϵ. Phase-locked solutions correspond to zeros of H_{FTM}, and their stability requires $0 \leq \epsilon\, dH_{FTM}(t)/dt \leq 2$. Thus, H_{FTM} is a similar predictor of dynamics to the coupling function $H(\psi)$ in the phase limit, although the stability condition differs. We assemble the maps H_i, $i = 1, 2, 3, 4$,

$$\epsilon H_i(t_2^{(i)} - t_1^{(i)}) = \left[t_2^{(i+1)} - t_1^{(i+1)}\right] - \left[t_2^{(i)} - t_1^{(i)}\right], \tag{B.2}$$

the composition of which defines

$$\begin{aligned} H_{FTM}(t) = {} & H_1(t) + H_2(t + \epsilon H_1(t)) \\ & + H_3(t + \epsilon H_2(t + \epsilon H_1(t))) \\ & + H_4(t + \epsilon H_3(t + \epsilon H_2(t + \epsilon H_1(t)))). \end{aligned} \tag{B.3}$$

In Figure 6(A), we introduce notation for the times required to travel along certain trajectories in the phase space. We use these to express the functions H_i for small ϵ in the next subsection. The notation reflects the scaling of these lengths; e.g., $\epsilon\Delta\tau_2$ in Figure 6(A) is $\mathcal{O}(\epsilon)$ because it represents the effect of a perturbation of strength $\mathcal{O}(\epsilon)$ and duration $\mathcal{O}(1)$.

B.2. The functions H_i. We shall use condition (iii), which implies that $\tau_4 > \tau_2$ if ϵ is sufficiently small. We also note that the functions H_i are invariant under translation by the period of the unperturbed limit cycle: $T = \epsilon(\tau_1 + \tau_1') + \tau_2 + \tau_4$.

To construct $H_1(t)$ we exploit the nature of FTM interactions. For $\epsilon\tau_1' \leq t \leq \tau_4$, O_1 receives no input between Π_1 and Π_2, so it travels on its unperturbed limit cycle, yielding $H_1(t) \equiv 0$. At $t = 0$, $H_1(0) = \epsilon\Delta\tau_1'$. If $-\epsilon\tau_1 - \tau_2 \leq t \leq -\epsilon\tau_1$, O_1 follows the perturbed limit cycle, so $H_1(t) \equiv (\tau_1 + \tau_1')$. For $-\epsilon\tau_1 \leq t \leq 0$, O_1 switches to the perturbed limit cycle from the curve of length $\epsilon\tau_1$, as shown in Figure 6(B). On the intervals $(-\epsilon\tau_1, 0)$ and $(0, \epsilon\tau_1')$, $H_1(t)$ is approximately linear (for $\epsilon \ll 1$). See Figure 12(A). The mapping $H_3(t)$ is generated in much the same way (Figure 12(C)).

To approximate $H_2(t)$ we use condition (ii), which implies that orbits move faster on the upper branch of the perturbed nullcline than on the unperturbed one: $\Delta\tau_2$ is positive. Thus for $t = 0$, O_1 receives input along the upper nullcline and arrives at Π_3 before O_2, so $H_2(0) = \Delta\tau_2$. The same argument holds if t is slightly negative, but if t is further decreased, the coupling signal turns off before O_1 reaches Π_3, until at $t = -\tau_2$ O_1 reaches Π_3 entirely on the unperturbed limit cycle. Thus $H_2(t) \equiv 0$ for $t < -\tau_2$, and it increases monotonically for $-\tau_2 < t$. See Figure 12(B).

O_1 receives input while traveling between Π_2 and Π_3, if $t \in (0, \epsilon\tau_1')$, so $H_2(t) \equiv \Delta\tau_2$ in this interval. If, however, t is increased further, O_1 has no input when crossing Π_2, and it first follows the unperturbed nullcline and jumps to the perturbed one if $t \in (\epsilon\tau_1', \epsilon\tau_1' + \tau_2)$; the bigger t is, the later this jump occurs. Thus H_2 decreases in this interval. If $t > \epsilon\tau_1' + \tau_2$, again $H_2(t) \equiv 0$.

Similar arguments lead to $H_4(t)$ (Figure 12(D)). $H_4(t) \equiv 0$ for $t \in (-\epsilon\tau_1, 0)$, because O_1 travels on the unperturbed nullcline. Because of condition (ii), traveling on the perturbed nullcline is always slower; thus $H_4(t)$ is nonpositive. It is monotonically decreasing in $(0, \tau_2 + \epsilon\tau_1')$, at which point O_1 starts on the perturbed limit cycle but switches along the

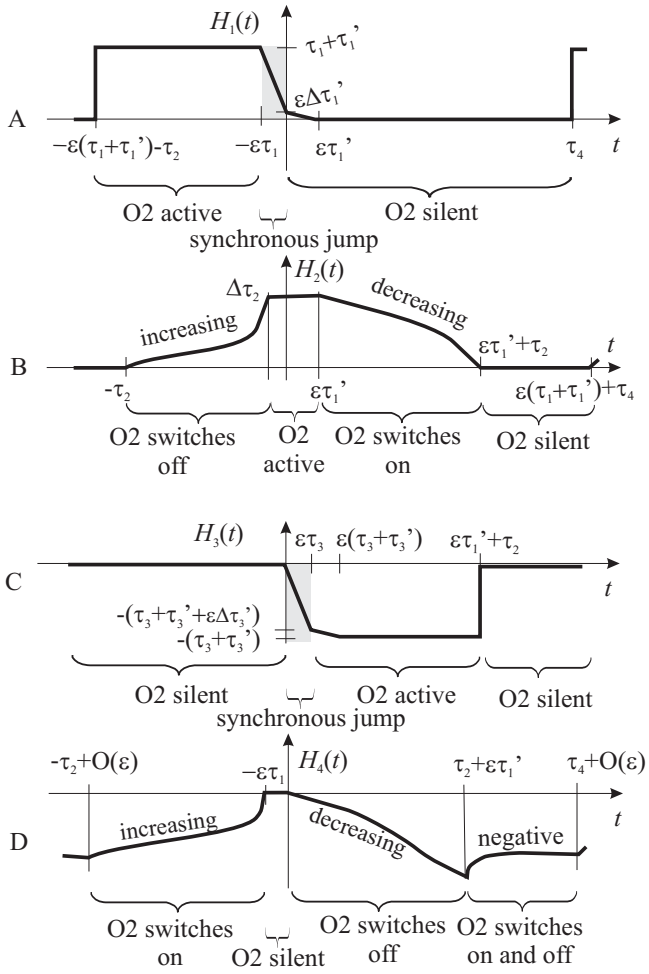


Figure 12. *The maps H_i.* A: O_2 active/silent *implies that O_2 is active/silent when O_1 jumps up;* synchronous jump *means that the upward jump of O_2 initiates an immediate jump in O_1.* B: O_2 active/silent *applies for the time interval in which O_1 travels from Π_2 to Π_3; O_2* switches on/off *means that O_2 is silent/active when O_1 crosses Π_2 but switches on/off before O_1 reaches Π_3. Analogous notation is used in* C *and* D.

way. The bigger t is, the longer it travels before switching to the unperturbed nullcline. H_4 has another, increasing part, corresponding to traveling on the unperturbed nullcline initially and switching to the perturbed one at some point (the bigger t, the later this happens), and a fourth region, marked "negative" in Figure 12(D), corresponding to starting and arriving on the unperturbed nullcline and spending an interval of length $\tau_2 + \epsilon\tau_1'$ on the other nullcline in between.

B.3. Possible forms of T-periodic dynamics. Here we show that on any stable T-periodic solution of the coupled oscillator pair, either O_1 and O_2 are in synchrony or O_1 leads O_2. We thereby exclude alternating dynamics or leading by O_2.

First assume that the oscillators alternate, so that, while O_2 is active, O_1 moves between Π_3 and Π_4. Thus, $t + \epsilon H_3(t + \epsilon H_2(t + \epsilon H_1(t))) \in (\tau_2 + O(\epsilon), \tau_4 + O(\epsilon))$: the interval marked

negative in Figure 12(D). This implies that

$$t \in (\tau_2 + O(\epsilon),\, \tau_4 + O(\epsilon)), \tag{B.4}$$

and while (B.4) holds, we can deduce the following.

1. From (B.4) and the fact that O_2 is silent when O_1 jumps up, $H_1(t) \equiv 0$ for sufficiently small ϵ.
2. Because O_2 is silent when O_1 is on its upper nullcline, $H_2(t + \epsilon H_1(t)) \equiv 0$.
3. Because O_2 is silent when O_1 jumps down, $H_3(t + \epsilon H_2(t + \epsilon H_1(t))) \equiv 0$.
4. From (B.4), $H_4(t + \epsilon H_3(t + \epsilon H_2(t + \epsilon H_1(t)))) < 0$.

Thus, $H_{FTM} < 0$ by (B.3). Since T-periodic solutions require $H_{FTM}(t) = 0$, this is a contradiction.

Assume now that O_2 leads O_1. In this case, the following hold:

1. O_2 is active when O_1 jumps up, corresponding to the $H_1(t) = \text{const} > 0$ part of H_1.
2. O_2 turns off before O_1 reaches Π_3, corresponding to the increasing part of H_2.
3. O_2 is silent when O_1 is between Π_3 and Π_4, implying that $H_3(t+\epsilon H_2(t+\epsilon H_1(t))) \equiv 0$.
4. O_2 turns on before O_1 reaches Π_1, corresponding to the increasing part of H_4.

In the interval of t that satisfies these requirements, all four functions are constant or increasing, implying that $H_{FTM}(t)$ itself is constant or increasing and hence that T-periodic solutions, if any exist, are unstable, from section B.1. This is again a contradiction.

Thus, we have proven that the only possible stable T-periodic solutions are synchrony or leading of O_1. In the next two subsections, we show that, if ϵ is sufficiently small, inequality (2.8) determines which case occurs.

B.4. If O_1 leads, then inequality (2.8) holds. Assume now that $H_{FTM}(t) = 0$ and O_1 leads. In this case O_2 is silent when O_1 is between Π_1 and Π_2, so that $H_1(t) < \epsilon\,\Delta\tau_1'$; see Figure 12(A). Similarly, O_2 is active when O_1 lies between Π_3 and Π_4, implying that $H_3(t + \epsilon H_2(t + \epsilon H_1(t))) < -(\tau_3 + \tau_3') + \epsilon\Delta\tau_3'$; see Figure 12(C). Combining these with the global inequalities $H_2 \le \Delta\tau_2$ and $H_4 \le 0$, (B.3) yields

$$0 = H_{FTM}(t) < \epsilon\Delta\tau_1' - (\tau_3 + \tau_3') + \epsilon\Delta\tau_3' + \Delta\tau_2. \tag{B.5}$$

The $\mathcal{O}(\epsilon)$ terms vanish as $\epsilon \to 0$; limiting values of the $\mathcal{O}(1)$ terms are found below.

For $\epsilon \ll 1$ (weak coupling) $\epsilon\Delta\tau_2$ is equal to the (appropriately scaled) linear phase-response to continuous ϵ perturbation during slow motion on the upper branch ($0 < \phi < D$); i.e., as shown in section 2,

$$\lim_{\epsilon\to 0} \Delta\tau_2 = \int_{x(0)}^{x(D)} \frac{f_y(\chi, y_a(\chi))}{f^2(\chi, y_a(\chi)) g_y(\chi, y_a(\chi))}\, d\chi. \tag{B.6}$$

We obtain $(\tau_3 + \tau_3')$ in the $\epsilon \to 0$ limit by approximating the ODEs defining O_1 at the upper right knee $[x(D), y(D)]$ by (A.1)–(A.2), derived in Appendix A. Substituting $h = 1$ and solving

$$g(x, y) + \epsilon \cdot 1 = 0 \quad \text{and} \quad g_y(x, y) = 0 \tag{B.7}$$

shows that the perturbation ϵh shifts the knee to the right by $\epsilon/g_x(x(D), y(D))$. Thus,

$$\epsilon\tau_3 \approx -\frac{\epsilon}{f(x(D), y(D))g_x(x(D), y(D))}, \tag{B.8}$$

$$\text{and} \quad \epsilon\tau_3' \approx \frac{\epsilon}{f(x(D), y(D^+))g_x(x(D), y(D))}, \tag{B.9}$$

yielding

$$\tau_3 + \tau_3' \approx \frac{1}{g_x(x(D), y(D))}\left[\frac{1}{f(x(D), y(D))} - \frac{1}{f(x(D), y(D^+))}\right]. \tag{B.10}$$

Substituting (B.6) and (B.10) into (B.5), we find that (2.8) holds in the $\epsilon \to 0$ limit. ■

Note that (B.10) represents the effect of FTM interactions but that it also agrees with predictions of phase reduction theory; cf. (2.6). This is the main reason that condition (2.6) holds in both the phase and the FTM limits.

B.5. Inequality (2.8) is false in case of synchrony. First we substitute the inequalities $H_1 \geq 0$ and $H_3 \geq -(\tau_3 + \tau_3')$ (cf. Figures 12(A,C)) into (B.3) to obtain

$$0 = H_{FTM}(t) \geq H_2(t + \epsilon H_1(t)) - (\tau_3 + \tau_3') + H_4(t + \epsilon H_3(t + \epsilon H_2(t + \epsilon H_1(t)))). \tag{B.11}$$

$H_2(t + H_1(t))$ is estimated by noting that, in case of synchrony, at least one of the following holds:

1. Upward jumps are synchronous.
2. Downward jumps are synchronous.
3. O_1 jumps up earlier and jumps down later than O_2.
4. O_2 jumps up earlier and jumps down later than O_1.

The first case yields $|t| \leq \epsilon\tau_1$; cf. Figure 12(A). Since $|H_1| \leq \tau_1 + \tau_1'$, we also have $|t + \epsilon H_1(t)| \leq \epsilon(\tau_1 + 2\tau_1')$; i.e., for arbitrary $\delta_1 > 0$, sufficiently small ϵ guarantees $|t + \epsilon H_1(t)| \leq \delta_1$. In the limit $\epsilon \to 0$, H_2 is determined exactly by phase theory. Moreover, H_2 is continuous because it describes interactions that occur only during slow dynamics and not during jumps. Appealing to continuity, we see that for arbitrary $\delta_2 > 0$, we may pick δ_1 sufficiently small that $|H_2(v) - H_2(0)| \leq \delta_2$ for all $|v| \leq \delta_1$. Thus, for ϵ sufficiently small, we conclude that

$$H_2(t + \epsilon H_1(t)) \geq \Delta\tau_2 - \delta_2. \tag{B.12}$$

Analogous arguments can be applied in the other three cases, and one can show in the same way that for sufficiently small ϵ, $|H_4(t + \epsilon H_3(t + \epsilon H_2(t + \epsilon H_1(t)))) - H_4(0)| \leq \delta_2$ also holds, implying that

$$H_4(t + \epsilon H_3(t + \epsilon H_2(t + \epsilon H_1(t)))) \geq -\delta_2. \tag{B.13}$$

Substituting inequalities (B.12)–(B.13) into (B.11), we obtain

$$0 = H_{FTM}(t) \geq \Delta\tau_2 - (\tau_3 + \tau_3') - 2\delta_2, \tag{B.14}$$

which holds for arbitrarily small δ_2 and so yields

$$0 \geq \Delta\tau_2 - (\tau_3 + \tau_3'). \tag{B.15}$$

Finally, using (B.6) and (B.10) in (B.15), we conclude that (2.8) cannot hold. ■

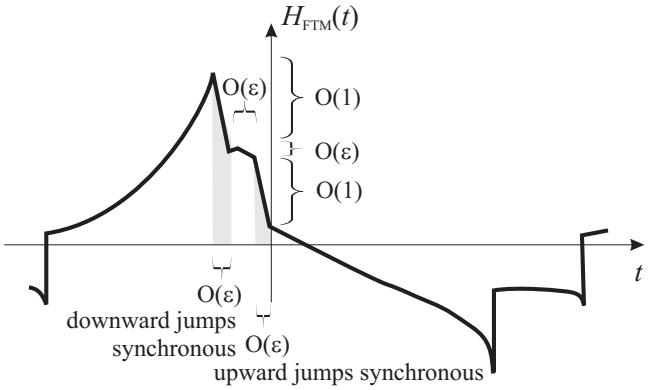


Figure 13. *An example of the composed map H_{FTM}. The proof of Theorem 3.1 relies on the fact that for $\epsilon \to 0$, the shape of H_{FTM} is similar to that of H in Figure 3. H_{FTM} has one or two steep, decreasing steps (two are shown here), inherited from H_1 and H_3. If there are two, they are separated by a plateau of width $O(\epsilon)$. A root of H_{FTM} in either steep part means that at least one jump is synchronous; a root in the plateau corresponds to synchronous activity in which neither jump is synchronous (cf. section 3.2). The latter is atypical in the limit $\epsilon \to 0$, as shown in Appendix C.*

Appendix C. Synchrony under weak coupling. The mappings H_2 and H_4 introduced in Appendix B have finite steepness, but H_1 and H_3 have $O(\epsilon^{-1})$-steep decreasing steps, corresponding to synchronous jumps of the two oscillators. The exact shape of the mapping H_{FTM} is model-specific, but in all cases it will also have one or two inherent steep parts due to (B.3), as shown in the example of Figure 13. According to the definition of section 3.2, the T-periodic orbits of the two oscillators are synchronous if one or both jumps coincide, or if one oscillator jumps up earlier but jumps down later than the other. These cases respectively correspond to two steep segments of H_{FTM} and the small plateau between them (if such a plateau exists). The width and height of the plateau is $O(\epsilon)$, and it vanishes in the limit $\epsilon \to 0$. Thus, synchronous activity typically means that either upward or downward jumps are synchronous.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Wiley Interscience, New York, 1984.

[2] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw Hill, New York, 1978.

[3] E. Brown, J. Moehlis, and P. Holmes, *On the phase reduction and response dynamics of neural oscillator populations*, Neural Comp., 16 (2004), pp. 673–715.

[4] J. T. Buchanan, *Neural network simulations of coupled locomotor oscillators in the lamprey spinal cord*, Biol. Cybernet., 66 (1992), pp. 367–374.

[5] J. T. Buchanan, *Contributions of identifiable neurons and neuron classes to lamprey vertebrate neurobiology*, Progress in Neurobiol., 63 (2001), pp. 441–466.

[6] J. T. Buchanan and S. Grillner, *Newly identified glutamate interneurons and their role in locomotion in the lamprey spinal cord*, Science, 236 (1987), pp. 312–314.

[7] A. H. Cohen, P. J. Holmes, and R. H. Rand, *The nature of the coupling between segmental oscillators of the lamprey spinal generator for locomotion: A mathematical model*, J. Math. Biol., 13 (1982), pp. 345–369.

[8] A. H. Cohen, S. Rossignol, and S. Grillner, eds., *Neural Control of Rhythmic Movements in Vertebrates*, Wiley, New York, 1988.
[9] A. H. Cohen and P. Wallén, *The neuronal correlate of locomotion in fish. "Fictive swimming" induced in an in vitro preparation of the lamprey*, Exp. Brain Res., 41 (1980), pp. 11–18.
[10] S. Coombes, *Phase locking in networks of synaptically coupled McKean relaxation oscillators*, Phys. D, 160 (2001), pp. 173–188.
[11] G. B. Ermentrout and N. Kopell, *Multiple pulse interactions and averaging in systems of coupled neural oscillators*, J. Math. Biol., 29 (1991), pp. 195–217.
[12] R. FitzHugh, *Impulses and physiological states in models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.
[13] S. Grillner, J. T. Buchanan, and A. Lansner, *Simulation of the segmental burst generating network for locomotion in lamprey*, Neuroscience Letters, 89 (1988), pp. 31–35.
[14] J. Guckenheimer, *Isochrons and phaseless sets*, J. Math. Biol., 1 (1975), pp. 259–273.
[15] J. Guckenheimer and P. J. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.
[16] J. L. Hindmarsh and R. M. Rose, *A model of the nerve impulse using two first-order differential equations*, Nature, 296 (1982), pp. 162–164.
[17] P. Holmes, R. J. Full, D. Koditschek, and J. Guckenheimer, *The dynamics of legged locomotion: Models, analyses, and challenges*, SIAM Rev., 48 (2006), pp. 207–304.
[18] F. C. Hoppensteadt and E. M. Izhikevich, *Weakly Connected Neural Networks*, Springer-Verlag, New York, 1997.
[19] E. M. Izhikevich, *Phase equations for relaxation oscillators*, SIAM J. Appl. Math., 60 (2000), pp. 1789–1804.
[20] T. Kiemel, K. M. Gormley, L. Guan, T. L. Williams, and A. H. Cohen, *Estimating the strength and direction of functional coupling in the lamprey spinal cord*, J. Comput. Neurosci., 15 (2003), pp. 233–245.
[21] N. Kopell, *Toward a theory of modelling central pattern generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. H. Cohen, S. Rossignol, and S. Grillner, eds., Wiley, New York, 1988, pp. 369–413.
[22] N. Kopell and G. B. Ermentrout, *Symmetry and phaselocking in chains of weakly coupled oscillators*, Comm. Pure Appl. Math., 39 (1986), pp. 623–660.
[23] N. Kopell and G. B. Ermentrout, *Chains of oscillators in motor and sensory systems*, in The Handbook of Brain Theory and Neural Networks, 2nd ed., M. A. Arbib, ed., MIT Press, Cambridge, MA, 2003, pp. 201–205.
[24] N. Kopell and D. Somers, *Anti-phase solutions in relaxation oscillators coupled through excitatory interactions*, J. Math. Biol., 33 (1995), pp. 261–280.
[25] N. Kopell, W. Zhang, and G. B. Ermentrout, *Multiple coupling in chains of oscillators*, SIAM J. Math. Anal., 21 (1990), pp. 935–953.
[26] J. H. Kotaleski, A. Lansner, and S. Grillner, *Neural mechanisms potentially contributing to the intersegmental phase lag in lamprey* ii. *Hemisegmental oscillations produced by mutually coupled excitatory neurons*, Biol. Cybernet., 81 (1999), pp. 299–315.
[27] A. Lansner, O. Ekeberg, and S. Grillner, *Realistic modeling of burst generation and swimming in the lamprey*, in Neurons, Networks, and Motor Behavior, P. S. G. Stein, S. Grillner, A. I. Selverston, and D. G. Stuart, eds., MIT Press, Cambridge, MA, 1997, pp. 165–171.
[28] S. Lefschetz, *Differential Equations: Geometric Theory*, 2nd ed., Interscience Publishers, New York, 1963.
[29] A. Liénard, *Étude des oscillations entretenues*, Revue Générale de l'Électricité, 23 (1928), pp. 901–912, 946–954.
[30] I. G. Malkin, *Methods of Poincaré and Linstedt in the Theory of Nonlinear Oscillations*, Gostexisdat, Moscow, 1949 (in Russian).
[31] I. G. Malkin, *Some Problems in Nonlinear Oscillation Theory*, Gostexisdat, Moscow, 1956 (in Russian).
[32] C. Morris and H. Lecar, *Voltage oscillations in the barnacle giant muscle fiber*, Biophys. J., 35 (1981), pp. 193–213.
[33] J. S. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating a nerve axon*, Proc. IRE, 50 (1962), pp. 2061–2070.

[34] J. Rinzel, *Excitation dynamics: Insights from simplified membrane models*, Federation Proc., 44 (1985), pp. 2944–2946.
[35] D. Somers and N. Kopell, *Rapid synchronization through fast threshold modulation*, Biol. Cybernet., 68 (1993), pp. 393–407.
[36] D. Somers and N. Kopell, *Waves and synchrony in networks of oscillators of relaxation and non-relaxation type*, Phys. D, 89 (1995), pp. 169–183.
[37] B. van der Pol, *On "relaxation-oscillations,"* The London, Edinburgh, and Dublin Phil. Magazine and J. of Sci., 7 (1926), pp. 978–992.
[38] P. L. Varkonyi, T. Kiemel, K. Hoffman, A. H. Cohen, and P. Holmes, *On the derivation and tuning of phase oscillator models for lamprey central pattern generators*, J. Comput. Neurosci., in press.
[39] T. Wadden, J. Hellgren, A. Lansner, and S. Grillner, *Intersegmental coordination in the lamprey: Simulations using a network model without segmental boundaries*, Biol. Cybernet., 76 (1997), pp. 1–9.
[40] T. L. Williams, *Phase coupling by synaptic spread in chains of coupled neuronal oscillators*, Science, 258 (1992), pp. 662–665.
[41] T. L. Williams and K. A. Sigvardt, *Intersegmental phase lags in the lamprey spinal cord: Experimental confirmation of the existence of a boundary region*, J. Comput. Neurosci., 1 (1994), pp. 61–67.
[42] H. Wilson, *Spikes, Decisions and Actions: The Dynamical Foundations of Neuroscience*, Oxford University Press, Oxford, UK, 1999.
[43] A. T. Winfree, *The Geometry of Biological Time*, 2nd ed., Springer-Verlag, New York, 2001.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 795–824
© 2008 Society for Industrial and Applied Mathematics

Neimark–Sacker Bifurcations in Planar, Piecewise-Smooth, Continuous Maps*

D. J. W. Simpson and J. D. Meiss†

Abstract. The multipliers of a fixed point of a piecewise-smooth, continuous map may change discontinuously as the fixed point crosses a discontinuity under smooth variation of parameters. We study the case when the multipliers "jump" from inside to outside the unit circle, and we assume the map is two-dimensional and piecewise-affine. The resulting dynamics is sometimes similar to the Neimark–Sacker bifurcation of a smooth map in which an attracting periodic or quasiperiodic orbit is created as the fixed point loses stability. However, the bifurcation is often much more complex, with multiple (chaotic) attractors, saddles, and repellors created or destroyed.

Key words. piecewise-smooth systems, resonance tongues, border-collision bifurcations

AMS subject classifications. 37G15, 37E30, 37E45

DOI. 10.1137/070704241

1. Introduction.

A dynamical system $F : M \to M$ is *piecewise-smooth* if it is everywhere smooth (i.e., C^k for some $k \in \mathbb{N}$) except on some codimension-one boundaries, called *switching manifolds*, that divide M into countably many regions. Such systems provide useful models for physical situations involving nonsmooth behavior such as impacts or rapid switching. Indeed, they are used in a wide variety of fields such as vibro-impacting systems and systems with friction [34, 5, 20, 25], circuitry including relay control systems [37, 4, 33], economics [27, 19], biology, and physiology [29, 17].

The theory of bifurcations in smooth dynamical systems is extensive and well grounded [18, 35]. However, the majority of this theory does not apply to piecewise-smooth systems. This paper is concerned with piecewise-smooth, continuous maps that are everywhere continuous but have a discontinuous Jacobian on the switching manifolds. As a fixed point of such a map crosses a switching manifold as a system parameter is varied, its associated multipliers may change discontinuously. The "jump" in multipliers may alter the stability of the fixed point. The resulting bifurcation is now called a *border-collision bifurcation*, a term coined by Nusse and Yorke [23] but originally called a C-bifurcation by Feigin whose work can be found in [9]. We study the two-dimensional case when complex multipliers jump from inside to outside the unit circle at the bifurcation. We show that this bifurcation exhibits diverse dynamical behavior, some of which is akin to that of a Neimark–Sacker bifurcation in a smooth map, some of which is like that of other border-collision bifurcations, and some of which are phenomena that to our knowledge have not been described previously.

*Received by the editors October 1, 2007; accepted for publication (in revised form) by T. Kaper April 7, 2008; published electronically July 23, 2008. This work was supported by the NSF under grants DMS-0202032 and DMS-0707695 and by the Mathematical Sciences Research Institute in Berkeley.
http://www.siam.org/journals/siads/7-3/70424.html
†Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526 (simpson@colorado.edu, jdm@colorado.edu).

Piecewise-smooth maps have two major applications. They appear as models of physical systems with both discrete and nonsmooth behavior, such as switching circuits and economics; second, they appear mathematically as Poincaré maps of piecewise-smooth flows with oscillatory behavior—see, for instance, work on *grazing bifurcations* and *sliding bifurcations* [8, 37, 16, 2]. They also provide a normal form for the so-called *corner-collision bifurcation* [7]. It is therefore of great interest to understand border-collision bifurcations in such maps.

To investigate local behavior near a border-collision bifurcation, it is first important to study the piecewise-linear approximation of the map at the bifurcation. The resulting piecewise-affine, continuous map provides a local approximation to a piecewise-smooth map near a differentiable point on a switching manifold. We can choose coordinates $z \in \mathbb{R}^n$ such that the local switching manifold becomes the plane $z_1 = 0$ and the piecewise-affine map takes the form

$$z' = \begin{cases} A_1 z + \mu c, & z_1 \le 0, \\ A_2 z + \mu c, & z_1 \ge 0, \end{cases} \tag{1}$$

where A_1 and A_2 are real-valued $n \times n$ matrices and $c \in \mathbb{R}^n$. In order for the map to be continuous, all columns of the two matrices A_i must be equal except for the first. It is assumed the fixed point crosses the switching manifold nontangentially, and thus $z^* = \mu(I - A_i)^{-1}c$ has $z_1^* = 0$ only when $\mu = 0$.

The work of Feigin [9] has provided an invaluable framework for the existence of fixed points and periodic solutions on either side of border-collision bifurcations for (1). Let σ_i^+ [σ_i^-] denote the number of real multipliers of A_i greater than 1 [less than -1] for $i = 1, 2$ (A_i may also have complex-valued multipliers). Feigin showed that

1. assuming A_1 and A_2 do not have a multiplier 1, then if $\sigma_1^+ + \sigma_2^+$ is even, (1) has a unique fixed point $\forall\ \mu \in \mathbb{R}$ continuously crossing the switching manifold at the origin when $\mu = 0$, and if $\sigma_1^+ + \sigma_2^+$ is odd, two fixed points exist for one sign of μ, colliding and annihilating at the origin when $\mu = 0$; and
2. assuming A_1 and A_2 do not have a multiplier -1, then if $\sigma_1^- + \sigma_2^-$ is odd, a period two cycle exists for one sign of μ colliding with the fixed point at the origin when $\mu = 0$, and if $\sigma_1^- + \sigma_2^-$ is even, the map (1) exhibits no 2-cycles.

The dynamics of one-dimensional, piecewise-affine, continuous maps with a single switching manifold are well understood [3, 8]. For example, it has been proved that in this case multiple attractors cannot coexist and every aperiodic attractor is chaotic [8].

For the two-dimensional case, we let $z = (x, y)$ and choose coordinates so that the locally smooth switching manifold corresponds to the y-axis. Following [23], the local map may be generically transformed by an affine change of coordinates to the canonical form $(x', y') = F_\mu(x, y; \tau_L, \tau_R, \delta_L, \delta_R)$, where F_μ is defined by

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{cases} A_L \begin{bmatrix} x \\ y \end{bmatrix} + \mu \begin{bmatrix} 1 \\ 0 \end{bmatrix}, & x \le 0, \\ A_R \begin{bmatrix} x \\ y \end{bmatrix} + \mu \begin{bmatrix} 1 \\ 0 \end{bmatrix}, & x \ge 0, \end{cases} \tag{2}$$

where

$$A_L = \begin{bmatrix} \tau_L & 1 \\ -\delta_L & 0 \end{bmatrix}, \quad A_R = \begin{bmatrix} \tau_R & 1 \\ -\delta_R & 0 \end{bmatrix}. \tag{3}$$

Again, the second columns of the two matrices are equal since the map is assumed continuous. The parameters of the normal form are simply the traces, $\tau_{L,R}$ and determinants, $\delta_{L,R}$, of the matrices $A_{L,R}$.

The normal form (2) has several simple properties that we will use below.

- The map F_μ is a homeomorphism of $\mathbb{R}^2$ if and only if $\delta_L\delta_R > 0$.
- If $\delta_L, \delta_R > 0$, the inverse is simply related to the original map by

$$F_\mu^{-1}(x, y; \tau_L, \tau_R, \delta_L, \delta_R) = F_{-\mu}\left(y, x; \frac{\tau_R}{\delta_R}, \frac{\tau_L}{\delta_L}, \frac{1}{\delta_R}, \frac{1}{\delta_L}\right). \tag{4}$$

  Thus the dynamical properties of F for positive μ are the same as those of F^{-1} for negative μ for a different combination of parameter values.
- The map has the scaling symmetry

$$F_{\lambda\mu}(\lambda x, \lambda y; \tau_L, \tau_R, \delta_L, \delta_R) = \lambda F_\mu(x, y; \tau_L, \tau_R, \delta_L, \delta_R) \ \forall\ \lambda > 0. \tag{5}$$

  Consequently, if $\mathcal{I}$ is an invariant set under F_μ, then $\lambda\mathcal{I}$ is an invariant set for $F_{\lambda\mu}$, $(\lambda > 0)$. Therefore, every bounded invariant set collapses onto the origin as $\mu \to 0$, and it is sufficient to consider only $\mu \in \{1, 0, -1\}$.

The canonical form (2) can have complex dynamics including multiple strange or quasi-periodic attractors, and the spectrum of its possible behaviors have not been completely classified. Partial classifications have been presented for the dissipative case, $|\delta_L|, |\delta_R| < 1$ [2], and for the case that the multipliers of the fixed point near the border-collision bifurcation remain real-valued [9].

In this paper, we will study the case in which the map has complex eigenvalues, in particular, when A_L corresponds to an attracting focus and A_R to a repelling focus. This is the case where one might expect an analogue of the Neimark–Sacker bifurcation to occur when a fixed point crosses the switching manifold.

In section 2 we describe basic properties of the canonical form map, (6), that we investigate. Periodic solutions are described by symbolic dynamics in section 2.1. Here we construct a linear system (10) to solve for periodic solutions and deduce their stability. We consider border-collision bifurcations of periodic solutions and then in section 2.2 define rotation numbers for general orbits.

Basic dynamics of the map (6) are described in section 3. We give a comparison to the smooth Neimark–Sacker bifurcation and show the effect of nonlinear terms. In section 3.1, via geometrical arguments, we prove the existence of an attracting set for some limiting cases and show that this set persists for nearby parameter values.

Particularly for the case of complex multipliers, it is of interest to compute regions in parameter space within which periodic solutions of a particular rotation number exist and are attracting. As with the case of circle maps, such regions are called resonance tongues (or Arnold tongues). In contrast to smooth systems, resonance tongues in piecewise-smooth

systems exhibit a distinctive lens-chain structure. This was first observed for a one-dimensional piecewise-linear circle map [36] and has been observed for the canonical form (2) in [38] and for other two-dimensional piecewise-linear maps [32, 15]. We will discuss these structures in detail in sections 3.2–3.5.

Complicated and unusual dynamics are described in section 4. In section 4.1 we provide an example where the fixed point is a saddle on the switching manifold and no invariant circle is created at the bifurcation. Multiple attractors are discussed in section 4.2; see also section 4.6 for additional complications.

The loss of stability of a periodic solution via an associated multiplier attaining the value -1 is detailed in section 4.3, and the loss of stability via a complex conjugate pair of associated multipliers crossing the unit circle is discussed in section 4.5. Period-doubled solutions appear far from basic periodic solutions due to the absence of quadratic and higher order terms in the canonical form. In a codimension-two situation we observe resonance tongues with reducible rotation numbers emanating from codimension-two points. Resonance tongues with reducible rotation numbers are discussed in section 4.4. These resonance tongues do not seem to form lens-chains. Conclusions are presented in section 5.

2. A two-dimensional map. In this paper we will study the map (2) when multipliers associated with the fixed point are complex-valued and jump from inside to outside the unit circle at the border-collision bifurcation. Thus we assume that A_L and A_R, (3), have eigenvalues $r_L e^{\pm 2\pi i \omega_L}$ and $\frac{1}{s_R} e^{\pm 2\pi i \omega_R}$, respectively, where $r_L, s_R \in (0,1)$ and without loss of generality $\omega_L, \omega_R \in (0, \frac{1}{2})$. Note that this corresponds to the case that $0 < \delta_L = r_L^2 < 1$ and $1 < \delta_R = 1/s_R^2$ so that the map (2) is an orientation preserving homeomorphism. With these new parameters, the normal form becomes

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{cases} \begin{bmatrix} 2r_L \cos(2\pi\omega_L) & 1 \\ -r_L^2 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \mu \begin{bmatrix} 1 \\ 0 \end{bmatrix}, & x \le 0, \\ \begin{bmatrix} \frac{2}{s_R} \cos(2\pi\omega_R) & 1 \\ -\frac{1}{s_R^2} & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \mu \begin{bmatrix} 1 \\ 0 \end{bmatrix}, & x \ge 0. \end{cases} \tag{6}$$

We will denote this map $(x', y') = f_\mu(x, y; r_L, s_R, \omega_L, \omega_R)$.

We may compute the existence of fixed points and 2-cycles by applying Feigin's analysis. Both A_L and A_R have no real-valued eigenvalues; hence, in the form (1), $\sigma_1^+ = \sigma_2^+ = \sigma_1^- = \sigma_2^- = 0$, and thus (6) has a unique fixed point $\forall\ \mu \in \mathbb{R}$, and the map has no period-two orbits for any values of the parameters. Explicitly, the fixed point is

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = \begin{cases} \dfrac{\mu}{r_L^2 - 2r_L \cos(2\pi\omega_L) + 1} \begin{bmatrix} 1 \\ -r_L^2 \end{bmatrix}, & \mu \le 0, \\ \dfrac{\mu}{s_R^2 - 2s_R \cos(2\pi\omega_R) + 1} \begin{bmatrix} s_R^2 \\ -1 \end{bmatrix}, & \mu \ge 0. \end{cases} \tag{7}$$

The fixed point moves from the left-half plane (LHP) when $\mu < 0$, where it is a stable focus, to the origin at $\mu = 0$, and then to the right-half plane (RHP) when $\mu > 0$, where it is an unstable focus.

Since (6) is a homeomorphism, its inverse is given by the symmetry (4). Explicitly, we have

$$f_\mu^{-1}(x,y) = \begin{cases} \begin{bmatrix} 0 & -s_R^2 \\ 1 & 2s_R\cos(2\pi\omega_R) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - \mu \begin{bmatrix} 0 \\ 1 \end{bmatrix}, & y \le 0, \\ \begin{bmatrix} 0 & -\frac{1}{r_L^2} \\ 1 & \frac{2}{r_L}\cos(2\pi\omega_L) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} - \mu \begin{bmatrix} 0 \\ 1 \end{bmatrix}, & y \ge 0. \end{cases} \tag{8}$$

2.1. Periodic solutions. Each orbit of (6) can be coded by a symbol sequence that gives its itinerary relative to the switching manifold. We will denote an orbit of (6) by a sequence $z^{(t)} = (x^{(t)}, y^{(t)})$, where $z^{(t+1)} = f_\mu(z^{(t)})$. The *symbol sequence* for an orbit is a sequence $\mathcal{S} = \{\mathcal{S}_t \in \{\mathsf{L}, \mathsf{R}\} : t \in \mathbb{Z}\}$, where

$$\mathcal{S}_t = \begin{cases} \mathsf{L}, & x^{(t)} \le 0, \\ \mathsf{R}, & x^{(t)} \ge 0. \end{cases} \tag{9}$$

Note that we allow the selection of either L or R when $x^{(t)} = 0$; however, since the map is continuous on the switching manifold, this ambiguity will not cause difficulties.

A period-n orbit, $\{z^{(0)}, z^{(1)}, \dots, z^{(n-1)}\}$, with $x^{(i)} \neq 0$ for each i, is defined by symbol sequence of length n that is unique up to a cyclic permutation. Given a periodic symbol sequence, the orbit is determined by the linear system

$$\begin{aligned} z^{(1)} &= A_{\mathcal{S}_0} z^{(0)} + b, \\ z^{(2)} &= A_{\mathcal{S}_1} z^{(1)} + b, \\ &\vdots \\ z^{(0)} &= A_{\mathcal{S}_{n-1}} z^{(n-1)} + b, \end{aligned}$$

where $b = (\mu, 0)^{\mathsf{T}}$. Elimination of the points $z^{(1)}, \dots, z^{(n-1)}$ gives

$$(I - M_{\mathcal{S}}) z^{(0)} = P_{\mathcal{S}} b, \tag{10}$$

where

$$M_{\mathcal{S}} = A_{\mathcal{S}_{n-1}} \dots A_{\mathcal{S}_0}, \tag{11}$$

$$P_{\mathcal{S}} = I + A_{\mathcal{S}_{n-1}} + A_{\mathcal{S}_{n-1}} A_{\mathcal{S}_{n-2}} + \dots + A_{\mathcal{S}_{n-1}} \dots A_{\mathcal{S}_1}. \tag{12}$$

We will call (10) the *n-cycle solution system*. If $(I - M_{\mathcal{S}})$ is nonsingular for a given period-n symbol sequence $\mathcal{S}$, (10) has the unique solution

$$z^{(0)} = (I - M_{\mathcal{S}})^{-1} P_{\mathcal{S}} b. \tag{13}$$

If, in addition, the consistency conditions (9) are satisfied, the period-n orbit exists and is said to be *admissible*; otherwise, it is *virtual* (terminology widely used in piecewise-smooth system theory [8]).

Stability of an admissible orbit that has no points on the switching manifold is easily determined. In this case, there exists a neighborhood of $z^{(0)}$ within which all points follow the symbol sequence $\mathcal{S}$ for at least n iterates. Moreover if $w^{(0)}$ is an element of this neighborhood, then

$$w^{(n)} = M_{\mathcal{S}} w^{(0)} + P_{\mathcal{S}} b. \tag{14}$$

Note that $z^{(0)}$ is the unique fixed point of this linear system; thus the stability of the n-cycle is determined by $M_{\mathcal{S}}$. For instance, the n-cycle is stable if and only if both multipliers of $M_{\mathcal{S}}$ lie inside the unit circle. For this reason we call $M_{\mathcal{S}}$ the *stability matrix* of $\mathcal{S}$.

Suppose now that an admissible n-cycle has one point on the switching manifold, say, $z^{(0)}$; that is, $x^{(0)} = 0$ and $x^{(i)} \neq 0 \; \forall \; i \neq 0$. Then there is a neighborhood of $z^{(0)}$ such that all points in this neighborhood located on one side of the switching manifold (say, the LHP) follow the symbol sequence $\mathcal{S}$ for n steps, whereas all points in the neighborhood on the other side follow the symbol sequence $\mathcal{S}^*$ for n steps, where $\mathcal{S}^*$ differs from $\mathcal{S}$ only in the first component. Thus for any $w^{(0)}$ in the neighborhood, we have

$$w^{(n)} = \begin{cases} M_{\mathcal{S}} w^{(0)} + P_{\mathcal{S}} b, & w_1^{(0)} \leq 0, \\ M_{\mathcal{S}^*} w^{(0)} + P_{\mathcal{S}^*} b, & w_1^{(0)} \geq 0. \end{cases} \tag{15}$$

Note that $P_{\mathcal{S}^*} = P_{\mathcal{S}}$ and $M_{\mathcal{S}^*}$ differs from $M_{\mathcal{S}}$ in only the first column; thus (15) is a piecewise-affine continuous map of the form (1). Hence, using Feigin's results, the problem of the existence of n-cycles and $2n$-cycles near the bifurcation is reduced to counting the number of real multipliers of $M_{\mathcal{S}}$ and $M_{\mathcal{S}^*}$ greater than 1 and less than -1.

We call $P_{\mathcal{S}}$ the *border-collision matrix* of $\mathcal{S}$ in view of the following lemma.

Lemma 1. *Suppose* $(I - M_{\mathcal{S}})$ *is nonsingular and* $\mu \neq 0$. *Then the point* $z^{(0)} = (I - M_{\mathcal{S}})^{-1} P_{\mathcal{S}} (\mu, 0)^{\mathsf{T}}$ *lies on the switching manifold if and only if* $P_{\mathcal{S}}$ *is singular.*

Proof. We introduce the map

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{cases} A_L \begin{bmatrix} x \\ y \end{bmatrix} + b, & x \leq 0, \\ A_R \begin{bmatrix} x \\ y \end{bmatrix} + b, & x \geq 0, \end{cases} \tag{16}$$

which is equivalent to (6) when $b = (b_1, b_2)^{\mathsf{T}} = (\mu, 0)^{\mathsf{T}}$. This map displays the same dynamics when $b = (\mathrm{sign}(b_1 + b_2), 0)^{\mathsf{T}}$ since the shift transformation $(x, y) \mapsto (x, y) + (0, -b_2)$ transforms $b = (b_1, b_2)^{\mathsf{T}}$ to $(b_1 + b_2, 0)^{\mathsf{T}}$, and a positive scaling (preserving the sign of x) transforms b to $(\mathrm{sign}(b_1 + b_2), 0)^{\mathsf{T}}$. Thus for any $b \in B = \{(b_1, b_2)^{\mathsf{T}} \mid \mu(b_1 + b_2) > 0\}$, $b_1 + b_2$ has the same sign as μ, and thus (16) exhibits the same dynamical structure as (6). In particular, if the periodic orbit with symbol sequence $\mathcal{S}$ for (6) has the point $z^{(0)}$ on the switching manifold, then the corresponding point for the map (16), $(I - M_{\mathcal{S}})^{-1} P_{\mathcal{S}} b$, will also lie on the switching manifold. Thus if $z^{(0)}$ lies on the switching manifold, then $(I - M_{\mathcal{S}})^{-1} P_{\mathcal{S}} B$ is a subset of the switching manifold. But B is a two-dimensional set; therefore, $P_{\mathcal{S}}$ is singular.

Let

$$K = \{(k, -k)^{\mathsf{T}} \mid k \in \mathbb{R}\}. \tag{17}$$

Note that $\forall b \in K$, the above shift transformation transforms b to the origin; hence in this case (16) has a unique fixed point z^*, which lies on the switching manifold. Clearly this point is a solution to the n-cycle solution system of $\mathcal{S}$; thus $z^* = (I - M_{\mathcal{S}})^{-1} P_{\mathcal{S}} b$. Now suppose that $P_{\mathcal{S}}$ is singular. If $b \in \text{null}(P_{\mathcal{S}}) \cap K$, $P_{\mathcal{S}} b = 0$, and hence $z^* = 0$. But for the origin to be a fixed point of (16) we must have $b = 0$. Thus $\text{null}(P_{\mathcal{S}}) \cap K = \{0\}$. It follows that $P_{\mathcal{S}}$ maps K onto $P_{\mathcal{S}}\mathbb{R}^2$. Thus $\exists b \in K$ such that $P_{\mathcal{S}} b = P_{\mathcal{S}}(\mu, 0)^{\mathsf{T}}$. The map (16) with this b has $z^* = (I - M_{\mathcal{S}})^{-1} P_{\mathcal{S}} b$ on the switching manifold. Thus $z^{(0)}$ for (6) also lies on the switching manifold. ■

The next lemma is useful for describing dynamics near where $M_{\mathcal{S}}$ has an eigenvalue 1.

Lemma 2. *Suppose $P_{\mathcal{S}}$ is nonsingular and $\mu \neq 0$. Then the n-cycle solution system (10) has a solution if and only if $(I - M_{\mathcal{S}})$ is nonsingular.*

Proof. Clearly if $(I - M_{\mathcal{S}})$ is nonsingular, (10) has the unique solution (13). Suppose $(I - M_{\mathcal{S}})$ is singular. Following the argument in Lemma 1, for any $b \in K$, (17), the map (16) has a unique fixed point z^* and $(I - M_{\mathcal{S}})z^* = P_{\mathcal{S}} b$. Thus $P_{\mathcal{S}} K \subset \text{rng}(I - M_{\mathcal{S}})$, the range of $(I - M_{\mathcal{S}})$, but $P_{\mathcal{S}}$ is nonsingular; thus we must have $P_{\mathcal{S}} K = \text{rng}(I - M_{\mathcal{S}})$. Finally, $(\mu, 0)^{\mathsf{T}} \notin K$; therefore, (10) has no solution. ■

The above lemmas and discussion can easily be extended to higher dimensional piecewise-affine maps.

2.2. Rotation numbers. The rotation number (or winding number) of an orbit of a map is a characterization of the average increase in angle per iteration. It is most easily defined for maps on circles or annuli. The rotation number for a map on $\mathbb{R}^2$ can be defined relative to a fixed point, z^*, if one exists, because removing z^* from the plane leaves an annulus. An alternative, intrinsic definition is called the "self-rotation number" [10]. The first definition is dependent upon the choice of z^*, but since (6) has a unique fixed point for any combination of parameter values, it is natural to define the rotation number about this point.

Since orbits of (6) rotate in a clockwise direction about z^*, we define the angle $\phi : \mathbb{R}^2 \setminus \{z^*\} \to (-\pi, \pi]$ as

$$\begin{aligned} \phi(z) &= \text{polar angle of } (x - x^*) - \mathrm{i}(y - y^*) \\ &= \text{atan2}(y^* - y, x - x^*), \end{aligned} \tag{18}$$

where atan2 is the two argument arctangent. To compute the rotation number we simply average the changes in ϕ over the iterations of the map. Let $\Delta\phi : \mathbb{R}^2 \setminus \{z^*\} \to [0, 2\pi)$ be

$$\Delta\phi(z) = \phi(f(z)) - \phi(z) \mod 2\pi; \tag{19}$$

then the rotation number of an orbit is

$$\rho(z^{(0)}) = \lim_{n\to\infty} \frac{1}{2\pi n} \sum_{i=0}^{n-1} \Delta\phi(z^{(i)}) \tag{20}$$

if this limit exists. It can be shown [30] that, for any point in $\mathbb{R}^2 \setminus \{z^*\}$ the limit (20) exists even if the orbit is unbounded and that

$$0 < \rho(z) < \frac{1}{2}. \tag{21}$$

Clearly if $z^{(0)}$ belongs to a period-n orbit, the rotation number is rational: $\rho = m/n$. If (6) has an invariant circle, then since f_μ is a homeomorphism, so is its restriction to the circle. Since any invariant circle bounds an invariant disk, the Brouwer fixed point theorem implies that the disk must contain a fixed point, which must be z^*, since the fixed point of (6) is unique. This is consistent with (21), which implies that the rotation number of the circle is nonzero. The usual definition for the rotation number on a circle [18, 6] coincides with (20). Furthermore, all points on the invariant circle have the same rotation number [18, 6].

3. Resonance tongues and basic dynamics. Recall that a Neimark–Sacker bifurcation in a sufficiently smooth map generically occurs when an attracting fixed point has a complex pair of multipliers, say, $\lambda(\mu) = r(\mu)e^{2\pi i\omega(\mu)}$ and $\bar{\lambda}(\mu)$, that cross the unit circle, at, say, $\mu = 0$ [18]. If the map satisfies a nondegeneracy assumption and $\lambda^n(0) \neq 1$ for $n = 1$, 2, 3, or 4, then an invariant circle is created or destroyed as μ crosses zero. The circle emerges from the fixed point with a size $O(|\mu|^{\frac{1}{2}})$. The criticality of the bifurcation is determined by cubic terms in the normal form; when it is *supercritical* the invariant circle is stable and exists when $|\lambda(\mu)| > 1$, and when it is *subcritical* the opposite is true. For small μ, dynamical behavior on the circle is determined by the rotation number $\omega(\mu)$. If $\omega(\mu) = \frac{m}{n}$ with $\gcd(m, n) = 1$ and $n > 4$, the motion is called *weakly resonant* or *mode locked* and there generically exist two or more period-n orbits on the invariant circle. When $n \leq 4$, the dynamics is *strongly resonant* and an invariant circle need not exist [18]. When $\omega(\mu)$ is irrational, all orbits on the circle are dense and quasiperiodic.

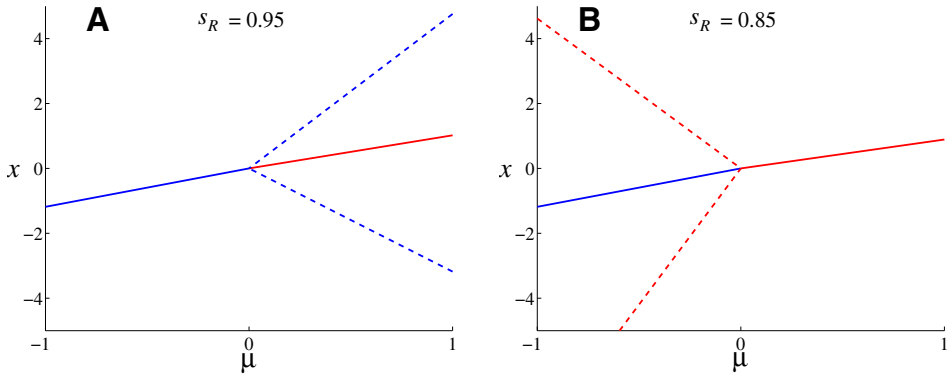


Figure 1. *Bifurcation diagrams of (6) when $\omega_L = \omega_R = 0.16$, $r_L = 0.9$. In panel A, $s_R = 0.95$; in panel B, $s_R = 0.85$. Blue (red) lines denote stable (unstable) solutions. Solid lines correspond to the fixed point, and dashed lines correspond to the maximum and minimum values of a periodic solution. Phase portraits corresponding to panel A are shown in Figure 2.*

The behavior of the piecewise-smooth map (6) can be much more complicated when μ crosses zero. However, for some parameter values, its behavior is similar to the classical Neimark–Sacker bifurcation; two example bifurcation diagrams for (6) are shown in Figure 1. For panel A of the figure, when $\mu < 0$, the stable fixed point is a global attractor; see Figure 2, panel A. When $\mu > 0$, the fixed point is unstable and is encircled by a stable invariant circle with rotation number $\rho \approx 0.1601$ whose basin appears to be the entire phase space except for the fixed point (see Figure 2, panel B). Consequently, this bifurcation is analogous to a

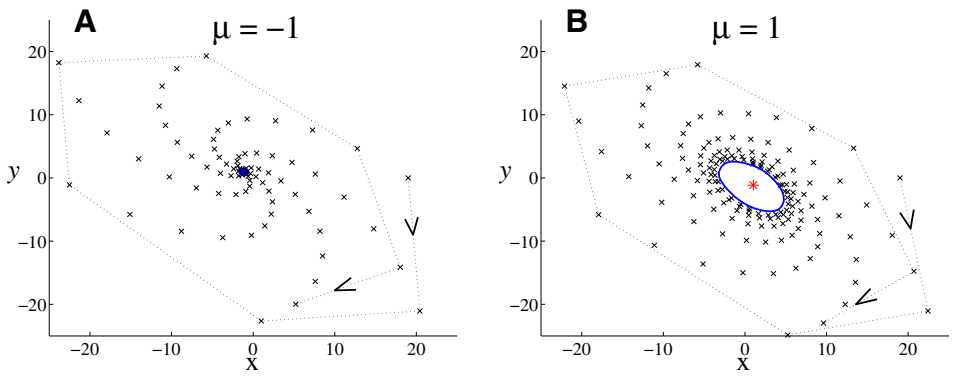


Figure 2. *Phase portraits of (6) with the same parameter values as those in Figure 1, panel A. In both panels one orbit is shown and the first few iterates are connected to illustrate the direction of motion. The fixed point is indicated by an asterisk.*

supercritical Neimark–Sacker bifurcation. An important difference is that, unlike in smooth systems, the size of the invariant circle grows linearly with respect to μ. This is a simple consequence of the scaling symmetry (5) and is common in other bifurcations in piecewise-smooth systems [20, 8]—for example, the discontinuous Andronov–Hopf bifurcation [31].

By contrast, the bifurcation in panel B of Figure 1 is analogous to a subcritical Neimark–Sacker bifurcation. When $\mu < 0$, there is an unstable invariant circle whose radius shrinks linearly to zero as $\mu \to 0^-$. Points inside the circle are attracted to the stable fixed point, whereas the forward orbits of points outside the circle are unbounded. When $\mu > 0$, the orbit of every initial condition, except for the unstable fixed point, is unbounded.

A fundamental question regarding the map (6) is: What governs the criticality of the Neimark–Sacker-like bifurcation? As will be shown later, the answer is not simple. For instance, there may exist no invariant circles for any value of μ (see section 4.1), or both supercritical and subcritical behavior may be observed together (see section 4.6).

Numerical simulations suggest that (6) does not exhibit strong resonance. We have not observed significantly different dynamical phenomena when $\omega_L = \omega_R = 1/3$ or $1/4$ in (6).

The addition of nonlinear terms to (6) does not affect structurally stable dynamics (e.g., when the created orbits are hyperbolic) for sufficiently small values of μ. For instance, bifurcation diagrams are shown in Figure 3 for the map

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = f_\mu(x, y) - \begin{bmatrix} x^2 \\ 0 \end{bmatrix}. \tag{22}$$

In panel A of Figure 3, the parameter values of panel A of Figure 1 are used. For small $\mu > 0$, we find an attracting invariant circle which grows in amplitude approximately linearly with respect to μ. For larger values of μ, the nonlinear map develops a stable 6-cycle.

In panel B, parameters are chosen such that the piecewise-affine map, (6), has a stable 3-cycle for $\mu > 0$. For small $\mu > 0$, the nonlinear map exhibits the same 3-cycle, which initially grows at the same linear rate as for the piecewise-affine map. However, as μ is increased, the 3-cycle undergoes a period-doubling cascade. The cascade is comprised of usual smooth

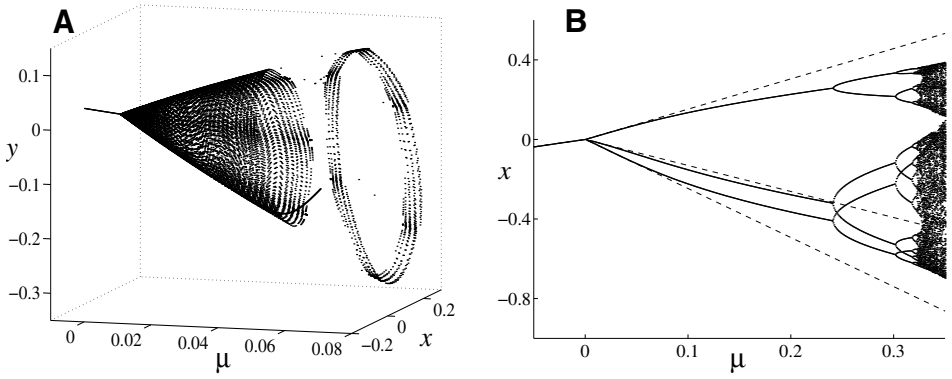


Figure 3. *Bifurcation diagrams of the nonlinear map, (22), computed numerically by plotting forward orbits after transients have decayed. In panel A the parameter values are the same as those in Figure 1, panel A; in panel B, $\omega_L = \omega_R = 0.4$, $r_L = 0.2$, and $s_R = 0.7$. In panel B, the corresponding bifurcation diagram of (6) is superimposed (shown as dashed lines).*

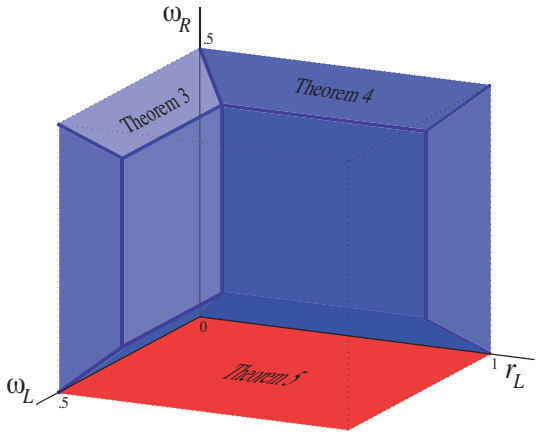


Figure 4. *Schematic summary of the results of section 3.1. Regions in parameter space are shaded blue where it is known that there exists an asymptotically stable invariant set and red where no attractor exists. The regions are labeled by their corresponding theorems in section 3.1.*

period-doubling bifurcations, where the attracting set is bounded away from the switching manifold.

The rest of this section describes the dynamics of (6) when $\mu = 1$ under variation of the remaining four parameters.

3.1. Limiting parameter values. Here we prove the existence or absence of an attractor for some limiting values of the parameters. This is summarized in Figure 4. There are two limiting domains in this figure, small r_L and small ω_L, for which we prove there exists an asymptotically stable invariant set in Theorems 3 and 4. The third limit in Figure 4 corresponds to $\omega_R = 0$ for which we prove in Theorem 5 almost all orbits are unbounded. Numerical results for regions in the figure where the theorems do not apply are given in the

remaining subsections. Throughout this section we let $f = f_1$ denote the map (6), with $\mu = 1$.

Theorem 3. *Consider* (6) *with* $\mu = 1$ *and assume* $0 < s_R < 1$, $0 < \omega_L$, $\omega_R < 1/2$ *as usual. Then there is an* $\varepsilon > 0$ *such that whenever* $0 \le r_L < \varepsilon$, *the map has an asymptotically stable invariant set.*

Proof. First suppose $r_L = 0$. Since $\mu = 1$, the unique fixed point of (6) lies in the RHP. The fixed point is an unstable focus; thus any point in the RHP other than the fixed point maps into the LHP in finitely many steps. Moreover, if $z = (x, y)$ is the first point in the LHP, then it is easy to see that $y < 0$. The image of this point is $f(z) = (y + 1, 0)$, with an x component less than one. If this image is still in the LHP, then the second iterate $f^2(z) = (1, 0) \in$ RHP. Consequently, any point in the LHP maps into the RHP in at most two steps.

Thus the forward orbit of every point other than the fixed point intersects the segment $\mathcal{I} = \{(x, 0) \mid x \in [0, 1]\}$. Since $\mathcal{I}$ is compact, there is an $N \in \mathbb{N}$ such every point in $\mathcal{I}$ maps back into $\mathcal{I}$ in at most N steps. Hence

$$\Omega = \bigcup_{i=0}^{N} f^i(\mathcal{I})$$

is a compact, bounded, forward invariant set of f. Note that $z^* \notin \Omega$. Let Σ be any compact neighborhood of Ω, $\Omega \subset \mathrm{int}(\Sigma)$, that does not contain the fixed point. Since Σ is compact, every point in Σ maps to $\mathcal{I}$; thus there is an $\hat{N} \in \mathbb{N}$ such that $f^{\hat{N}}(\Sigma) \subset \Omega$, i.e., Σ is a trapping set for $f^{\hat{N}}$.

When r_L is small, a similar construction must hold: since (6) depends continuously on r_L, there is an $\varepsilon > 0$ such that whenever $0 \le r_L < \varepsilon$, Σ is still a trapping set for $f^{\hat{N}}$. Let

$$\Lambda = \bigcap_{i=0}^{\infty} f^i(\Sigma). \tag{23}$$

Then Λ is an attracting set for (6); it is necessarily asymptotically stable and invariant. ∎

The attracting set (23) may not be an "attractor" as it need not be minimal or chain-transitive [24, 28]. Indeed, we will see in section 3.2 that f can have multiple attractors.

Theorem 4. *Consider* (6) *with* $\mu = 1$ *and* $0 < r_L$, $s_R < 1$, $0 < \omega_R < 1/2$ *as usual. Then there is an* $\varepsilon > 0$ *such that whenever* $0 \le \omega_L < \varepsilon$, *the map has an asymptotically stable invariant set.*

Proof. When $\omega_L = 0$, the left half of (6) becomes

$$z' = A_L z + \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad A_L = \begin{bmatrix} 2r_L & 1 \\ -r_L^2 & 0 \end{bmatrix}.$$

This affine map has a unique (virtual) fixed point $z_L^* = \frac{1}{(1-r_L)^2}(1, -r_L^2)^\mathsf{T}$ in the RHP; see panel A of Figure 5. A_L has a repeated eigenvalue r_L and a single eigenvector $(1, -r_L)^\mathsf{T}$. The one-dimensional invariant manifold E^{r_L} of the virtual fixed point intersects the y-axis at $\frac{r_L}{1-r_L} > 0$. Observe that E^{r_L} separates orbits of the left-half map (i.e., if z is below (above) E^{r_L}, then so is z'). Let $\Xi = \{(x, y) \mid x, y \ge 0,\ y \le r_L(\frac{1}{1-r_L} - x)\}$; see Figure 5, panel A.

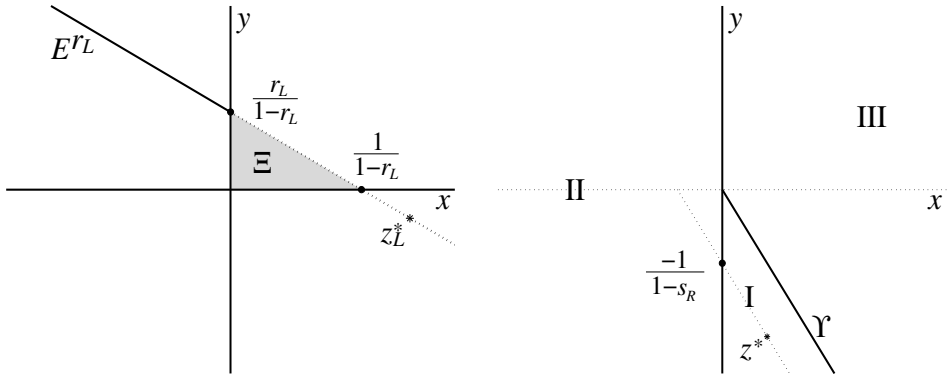


Figure 5. *Partitions of phase space when $\mu = 1$, applied in Theorems 4 and 5. In panel A, $\omega_L = 0$; in panel B, $\omega_R = 0$.*

Consider the forward orbit of any $z \in \Xi$ under the full map (6). Since the map in the RHP corresponds to an unstable focus, z maps to a point $\hat{z}$ in the LHP in finitely many steps, and since for any $x > 0$, $y' = -\frac{1}{s_R^2}x < 0$, $\hat{z}$ lies below the x-axis. Every point in the LHP below E^{r_L} maps into the RHP below E^{r_L} in finitely many steps. Since for any $x < 0$, $y' = -r_L^2 x > 0$, the first iterate in the RHP will lie above the x-axis and thus in Ξ. Since Ξ is compact, there is an $N \in \mathbb{N}$ such that $\forall\ z \in \Xi$, $f^n(z) \in \Xi$ for some $n \le N$. Hence,

$$\Omega = \bigcup_{i=0}^{N} f^i(\Xi)$$

is a compact, bounded, forward invariant set for f.

Following the argument in Theorem 3, this implies that for $0 \le \omega_L < \epsilon$ there is an attracting set Λ, (23) for f, where Σ is a neighborhood of Ω. ■

Theorem 5. *Consider (6) with $\mu = 1$ and the limiting case $\omega_R = 0$. Assume that the remainder of the parameter values satisfy $0 < r_L$, $s_R < 1$, $0 < \omega_L < 1/2$ as usual. Then the forward orbit of any point other than the fixed point is unbounded.*

Proof. When $\omega_R = 0$, the right half of (6) becomes

$$z' = A_R z + \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad A_R = \begin{bmatrix} \frac{2}{s_R} & 1 \\ -\frac{1}{s_R^2} & 0 \end{bmatrix}.$$

This affine map has a unique fixed point $z^* = -\frac{1}{(1-s_R)^2}(s_R^2, -1)^\mathsf{T}$ in the RHP. A_R has a repeated eigenvalue $\frac{1}{s_R}$ and a single eigenvector $(s_R, -1)^\mathsf{T}$ whose associated one-dimensional invariant manifold Σ intersects the y-axis at $-\frac{1}{1-s_R} < 0$. Let Υ be the line parallel to Σ which passes through the origin. We use Υ to partition phase space into three regions, as shown in panel B of Figure 5. Region III, the set $\{(x, y) \mid x \ge 0,\ x > -s_R y\}$, is forward invariant for (6) because for any z in region III, $z' = (\frac{2}{s_R}x + y + 1, -\frac{1}{s_R^2}x)^\mathsf{T}$; that is,

$$x' = \frac{2}{s_R}x + y + 1 > \frac{2}{s_R}x - \frac{1}{s_R}x + 1 \Rightarrow x' > \frac{1}{s_R}x > 0.$$

Using $x = -s_R^2 y'$, we see that $x' > -s_R y'$, and hence z' is in region III. Since the map in the LHP corresponds to a virtual stable focus, and since for any $x < 0$, $y' = -r_L^2 x > 0$, every point in region II is mapped into region III in finitely many iterations. Thus the forward orbit of any point either starts in the RHP and remains there or enters region III via the LHP in finitely many steps and remains in region III. In any case, the tail of the forward orbit is contained in the RHP, and since dynamics in the RHP are governed by an unstable linear map, the forward orbit of every point except z^* is unbounded. ■

The limiting case, $s_R = 1$, has been described previously in [32, 26, 13]; we will discuss it further in section 3.3.

3.2. Resonance tongues when $\mu = 1$. Figures 6, 7, and 8 show numerically computed regions of existence of stable periodic orbits with periods shown in the color bar at the bottom of Figure 6. As is commonly observed in piecewise smooth systems, these "resonance tongues" usually have the form of a chain of lens-shaped regions [37, 27]. The rotation number of each lens-chain is fixed, and, as we will discuss below, within a given chain, the symbol sequence changes from lens to lens. The tongues emanate from $s_R = 1$, and their sizes are ordered by the Farey sequence, as indicated along the top of Figure 6. In some places tongues overlap corresponding to the coexistence of multiple stable periodic solutions.

The figures are computed on a grid of 1024 frequency values and 128 values of s_R. For each set of parameters, we numerically check the admissibility conditions (9) and the stability conditions (that the multipliers of the stability matrix (11) lie inside the unit circle) for periodic orbits up to period 30 with certain symbol sequences described below. If no stable n-cycle was found, we compute N iterates along the forward orbit of the point $z = (M, 0)$ for N up to 10^4. If $\max_{i=0}^{n} |x^{(i)}|$ appeared to grow steadily as $n \to N$, this orbit is declared to be unbounded, and the corresponding point is shaded white. Otherwise the point is shaded black and presumably corresponds to bounded motion with a period larger than 30. Because the orbits for small ω appear to range over a large domain, we used $M = 10^{12}$.

In some cases, multiple attracting periodic solutions may exist, and our algorithm simply assigns a color based on the first periodic orbit that it finds; see section 4.2.

In Figure 6, notice that when $\omega_R = 0$, there are no stable solutions for any value of s_R as foreseen by Theorem 5. Similarly, when $\omega_L = 0$ (as seen in the lower two diagrams when $\omega_R = -\phi$), there is a stable solution for all values of s_R in accordance with Theorem 4. In Figure 8, notice that the diagram corresponding to the smaller value of r_L has stable orbits over a larger range of parameter space, as would be expected from Theorem 3.

3.3. Dynamics near $s_R = 1$. We first describe the dynamics for the case $s_R = 1$, corresponding to the top edge of the figures. When $\mu = 1$, the fixed point z^*, (7), lies in the RHP and is a center. Points sufficiently near z^* rotate around it with rotation number ω_R on invariant ellipses contained in the RHP. The largest of this family of nested ellipses has one point on the switching manifold. The boundary of the region in phase space within which this rigid rotation occurs has a geometry dependent upon the rationality of ω_R [32]. When ω_R is irrational, the boundary is the largest ellipse. When $\omega_R = m/n$ is rational, the boundary is an invariant n-sided polygon $\mathcal{P}$. In this case there are points in the region bounded by the largest ellipse and $\mathcal{P}$. These points simply belong to m/n-cycles contained in the RHP.

Numerically we have observed that the boundary attracts nearby points outside the region,

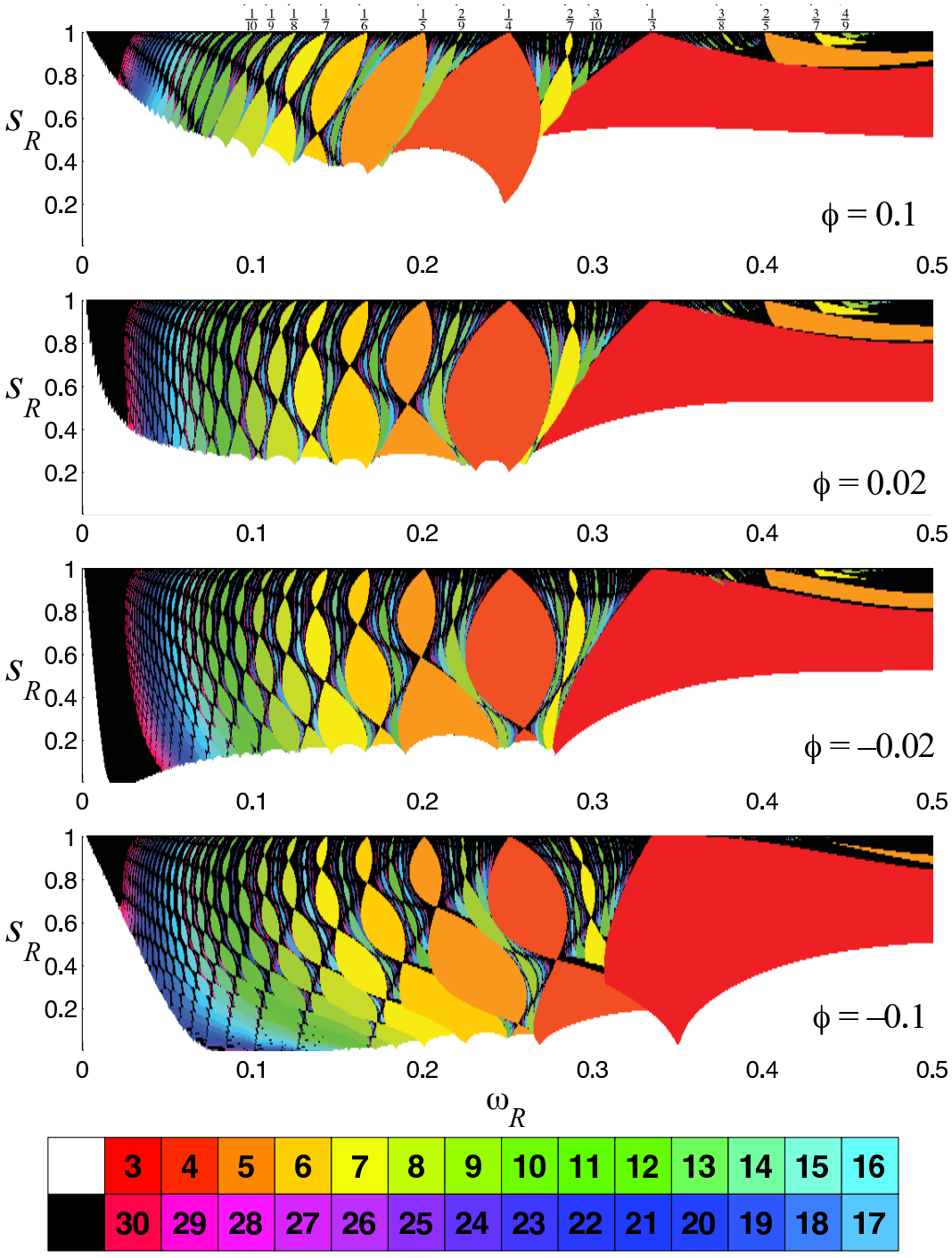


Figure 6. *Resonance tongues for $\mu = 1$, $r_L = 0.2$, $\omega_L = \omega_R + \phi$ for four different values of ϕ. Each colored region corresponds to the existence of a stable periodic orbit with period shown in the color bar at the bottom. White regions correspond to unbounded orbits and black regions to orbits with period larger than 30.*

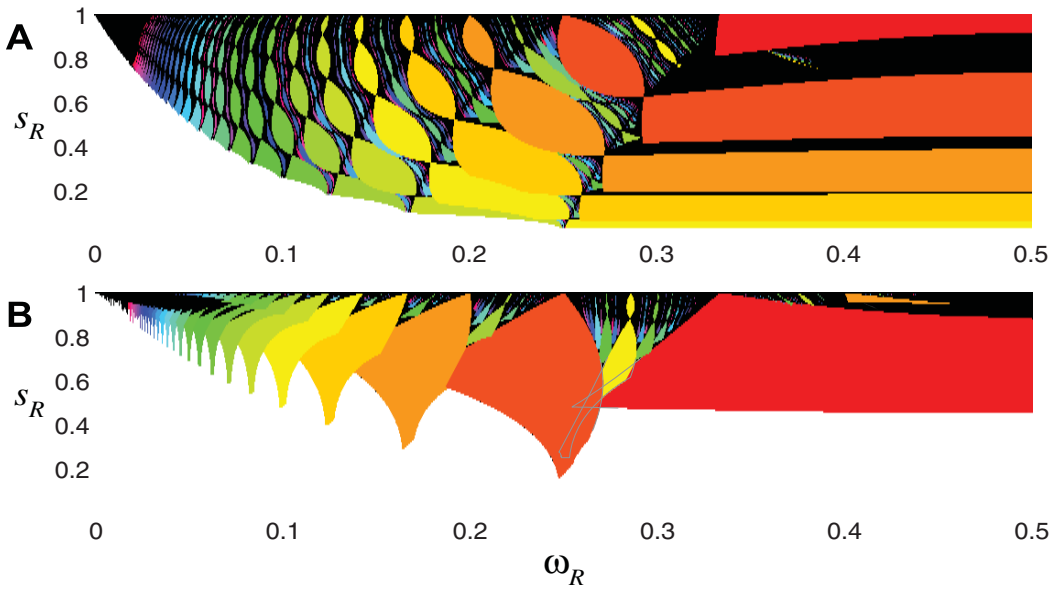


Figure 7. *Resonance tongues of* (6) *for* $\mu = 1$. *In panel A,* $r_L = 0.3$, $\omega_L = 0.09$*; in panel B,* $r_L = 0.16$, $\omega_L = 0.38$. *In panel B, the overlapping of the* 1/3, 1/4, *and* 2/7 *tongues is emphasized. The color scheme is the same as that in Figure* 6.

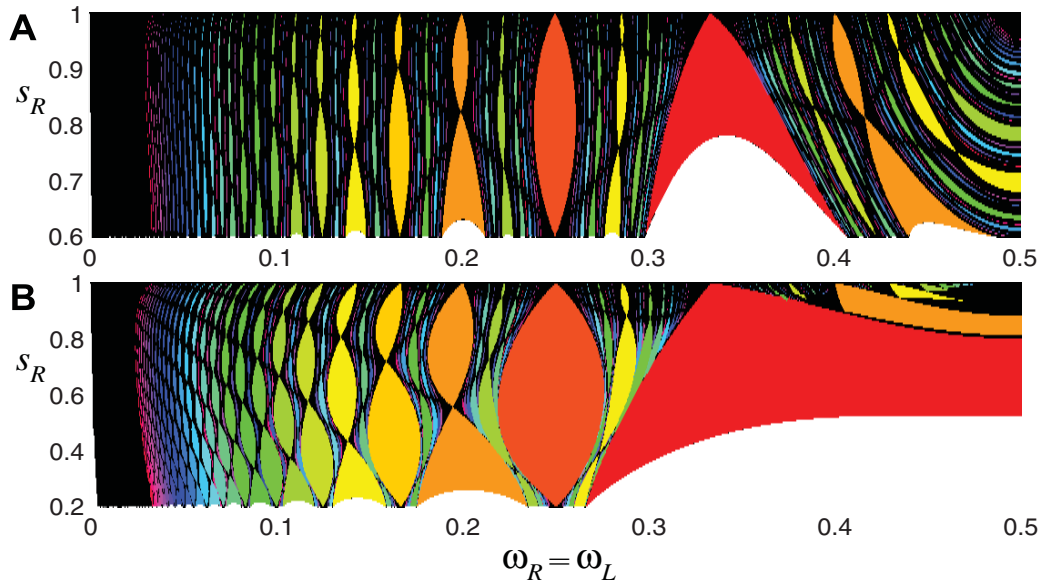


Figure 8. *Resonance tongues of* (6) *for equal polar angles, i.e.,* $\omega_L = \omega_R$, *when* $\mu = 1$. *In panel A,* $r_L = 0.6$*; in panel B,* $r_L = 0.2$. *The color scheme is the same as that in Figure* 6.

prompting the following conjecture.

Conjecture 1. *Consider* (6) *with* $\mu = 1$ *and assume that* $0 < r_L < 1$, $0 < \omega_L$, $\omega_R < 1/2$

as usual. Then there is an $\varepsilon > 0$ such that whenever $1 - \varepsilon < s_R \leq 1$, the map has an asymptotically stable invariant set.

Note that other attractors may exist when $s_R = 1$. For example, in panel A of Figure 7, the period-three cycle exists $\forall \omega_R > \frac{1}{3}$ when $s_R = 1$. For these values there will also be an invariant polygon or ellipse with rotation number ω_R.

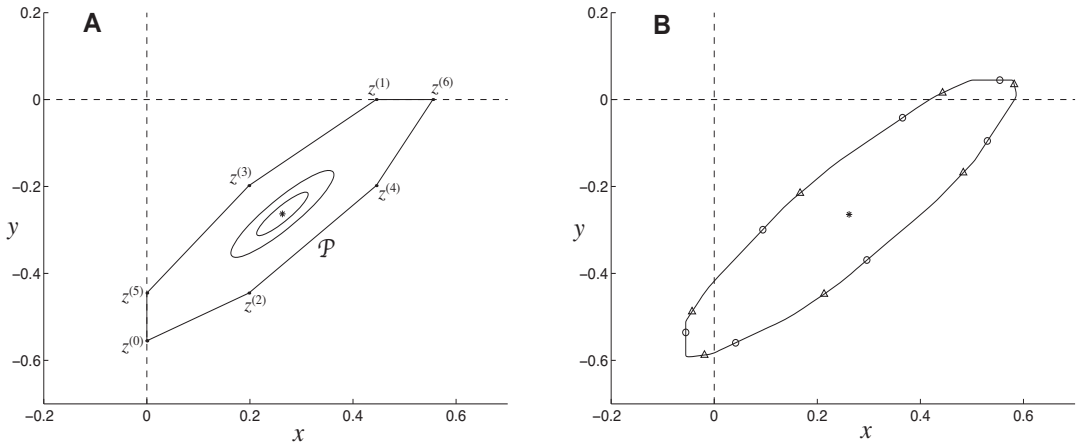


Figure 9. *Phase portraits of (6) when $\mu = 1$, $\omega_L = \omega_R = 3/7$, $r_L = 0.9$, and $s_R = 1$ in panel A and $s_R = 0.995$ in panel B. The fixed point z^* is indicated by an asterisk. In panel A, we show the invariant heptagon $\mathcal{P}$ and two of the uncountably many invariant ellipses near the center z^*, each consisting of infinitely many 3/7-cycles. In panel B, the right-half map is no longer area preserving; $\mathcal{P}$ appears to become an invariant circle attracting from both sides. The stable (saddle) 3/7-cycle is indicated by triangles (circles).*

When $\omega_R = m/n$ is rational, the bounding polygon $\mathcal{P}$ has one side on the switching manifold; see Figure 9. Denote the vertices of $\mathcal{P}$ by $\{z^{(i)} \mid i = 0, \ldots, n-1\}$, so that each $z^{(i)}$ maps to $z^{(i+1)}$. Suppose that $z^{(0)} = (0, y^{(0)})$ and $z^{(d)} = (0, y^{(d)})$ are the two points on the switching manifold and without loss of generality that $y^{(0)} < y^{(d)}$. Since the rotation number is m/n, $z^{(1)}$ is the mth vertex clockwise from $z^{(0)}$; thus $z^{(i)}$ is the $(im \bmod n)$th vertex clockwise from $z^{(0)}$. Since $z^{(d)}$ is adjacent to $z^{(0)}$, $dm \bmod n = 1$. In other words, d is the multiplicative inverse of m modulo n, which—as is well known—exists if and only if m/n is an irreducible fraction [14]. For example, when $m/n = 3/7$, $d = 5$ since $5 \times 3 \bmod 7 = 15 \bmod 7 = 1$; see Figure 9. Alternatively d may be computed via the Farey tree. If m_1/n_1 and m_2/n_2 are the Farey neighbors of m/n ($m_1 + m_2 = m$ and $n_1 + n_2 = n$), then $mn_i \bmod n = \pm 1$ for $i = 1, 2$; thus d is either n_1 or n_2.

The n-cycle $\{z^{(i)}\}$ may be thought of as having a symbol sequence $\mathcal{S}$, where $\mathcal{S}_0 = \mathcal{S}_d = \mathsf{L}$ and the remaining $(n-2)$ elements equal R. Each of the conditions $x^{(0)} = 0$ and $x^{(d)} = 0$ defines a codimension-one manifold in parameter space that forms a boundary for the lens that emanates from $s_R = 1$. Within the lens there exist both stable and saddle orbits with rotation number m/n. These orbits collide and annihilate on the boundaries in a border-collision bifurcation. The stable orbits have symbol sequence $\mathcal{S}$, whereas the saddle orbits have one point in the LHP and $(n-1)$ points in the RHP. An example of a phase portrait showing these two orbits and the invariant circle formed from the unstable manifolds of the saddle is shown in panel B of Figure 9.

3.4. Resonance tongue boundaries. More generally, the codimension-one boundaries of resonance tongues correspond to either a border-collision bifurcation or a loss of stability. In the first case, one point on the attracting n-cycle collides with the switching manifold. The resulting border-collision bifurcation may be classified by Feigin's analysis, as discussed in section 2.1. Typically we find that the stable n-cycle annihilates with a saddle-type n-cycle that coexists within the lens, but this is not always the case (see section 4.4). A second possible codimension-one boundary corresponds to the loss of stability of the attracting cycle when one or more multipliers of its stability matrix $M_{\mathcal{S}}$ leaves the unit circle. In codimension-one situations this can happen in exactly three ways: a real multiplier may pass through 1 or -1 or a complex conjugate multiplier pair may cross the unit circle.

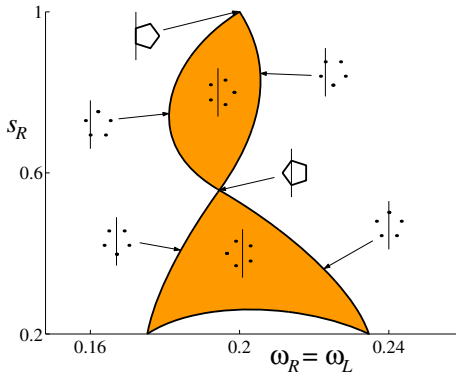


Figure 10. *The 2/5-resonance tongue of (6) with $\mu = 1$, $r_L = 0.2$, as seen in Figure 8, panel B. Shown are schematic phase portraits of the stable 2/5-cycle in relation to the switching manifold.*

As an example, Figure 10 illustrates the lens-chain for the 2/5 orbit and also shows phase portraits of the stable cycle in the two lenses and on their boundaries. In the upper lens, two points of the stable 2/5-cycle lie in the LHP; in the lower lens three points lie in the LHP. The left and right boundaries correspond to border-collision bifurcations along which the stable cycle collides and annihilates with a saddle cycle. Along the bottom boundary a multiplier associated with the stable cycle passes through 1.

Interestingly, this last bifurcation does not resemble a saddle-node bifurcation in a smooth system. If the stability matrix $M_{\mathcal{S}}$ has a multiplier $\lambda^* = 1$, then $I - M_{\mathcal{S}}$ is singular. By Lemma 2, the system (10) has no solution provided that the border-collision matrix $P_{\mathcal{S}}$ (12) is nonsingular. In this case, as $\lambda^* \to 1^-$, the n-cycle becomes unbounded. Of course, this is a somewhat artificial consequence of the lack of nonlinear terms in (6). Note that the saddle cycle exists on both sides of the $\lambda^* = 1$ bifurcation curve.

If Conjecture 1 holds, the bounding invariant polygon $\mathcal{P}$ for $s_R = 1$ is an attracting invariant set, and so it persists as such a set, typically as an invariant circle for $s_R < 1$. There are many ways in which an invariant circle of planar, piecewise-affine continuous maps can break up [32, 39]. For instance, the stable and unstable manifolds of the saddle cycle may transversely intersect, replacing the invariant circle with a homoclinic tangle [32, 38, 39]. The invariant circle consisting of a stable and saddle cycle and the unstable invariant manifolds of the saddle may disappear upon collision and annihilation of the cycles in a border-collision

bifurcation. Also, a loss of stability of the stable cycle may lead to the destruction of the invariant circle [32, 21]. These mechanisms are similar to those that occur in smooth maps [1].

The symbol sequence of an m/n periodic orbit $\{z^{(i)} \mid i = 0, \ldots, n-1\}$ that lies on an invariant circle may be determined by the same process as when $s_R = 1$. Assume the invariant circle intersects the switching manifold twice and that the m/n-cycle has l points in the LHP and $(n-l)$ points in the RHP. If $z^{(0)}$ is the first point in the cycle along the invariant circle to the left of its lower intersection with the switching manifold, then the points in the LHP are $z^{(0)}, z^{(d)}, z^{(2d \bmod n)}, \ldots, z^{((l-1)d \bmod n)}$, where $dm \bmod n = 1$ as before. Therefore, $\mathcal{S}_0 = \mathcal{S}_d = \cdots = \mathcal{S}_{(l-1)d \bmod n} = \mathsf{L}$ and $\mathcal{S}_i = \mathsf{R}$ otherwise.

Stable n-cycles associated with the topmost lens in a tongue are comprised of two points in the LHP and $(n-2)$ points in the RHP. Numerically we have observed that other lenses correspond to stable n-cycles with one more point in the LHP than the connected lens above. The saddle n-cycle has one fewer point in the LHP and one more point in the RHP than its stable counterpart.

Figure 8 shows resonance tongues when $\omega_L = \omega_R$ for two different fixed values of r_L. Lens-chains emanate from $s_R = 1$ and may extend to $s_R = r_L$. If n is odd, we have observed that the bottom-most lens corresponds to stable n-cycles with $\frac{n+1}{2}$ points in the LHP and the boundary of this lens intersects $s_R = r_L$ at one point. Alternatively, if n is even, the bottom-most lens corresponds to stable n-cycles with $\frac{n}{2}$ points in the LHP and there are possibly two intersection points of the boundary of this lens with $s_R = r_L$. The boundary curve connecting these two points bends above $s_R = r_L$ as in Figure 10 and corresponds to the associated stability matrix having a multiplier 1. We were unable to find any stable solutions when $s_R < r_L$, prompting the following conjecture.

Conjecture 2. *Suppose $\mu = 1$, $0 < \omega_L = \omega_R < 1/2$, and $0 < s_R < r_L < 1$. Then (6) has no stable solutions.*

3.5. Shrinking points. As in [36], we call points where resonance tongues have zero width *shrinking points*. Since this corresponds to the intersection of two border-collision-bifurcations, there is a period-n orbit, $\{z^{(i)}\}$, that has two points on the switching manifold. As a result of the piecewise-affine structure of the map, line segments connecting $z^{(i)}$ and $z^{(i+d)}$ map to one another and hence form an invariant n-gon comprised of uncountably many m/n-cycles; see Figure 11 for an example. If $\mathcal{S}$ is the symbol sequence of these cycles, then $(I - M_{\mathcal{S}})$ is singular at the shrinking point. Note that Lemma 2 does not apply here since $P_{\mathcal{S}}$ is singular; the n-cycle solution system (10) has uncountably many solutions.

Denote the two vertices of the n-gon on the switching manifold by $w_1 = (0, y_1)$ and $w_2 = (0, y_2)$ and assume without loss of generality that $y_1 < y_2$. Since these points are on a periodic orbit, $f^s(w_1) = w_2$ for some s. For example, in Figure 11, $s = 5$. The polygon persists as an attracting invariant set as parameters are continuously varied, though it will no longer necessarily contain only periodic orbits. When the polygon persists as an invariant circle, the points of intersection of the invariant circle with the switching manifold, w_1 and w_2, also vary continuously. Moreover, it is a codimension-one phenomenon for w_1 to map into w_2 in s iterations. Thus there exists a curve in two-dimensional parameter space along which $f^s(w_1) = w_2$. We call such a curve a *shrinking point curve*. These curves also exist for one-dimensional piecewise-linear circle maps, and in [36], the authors were able to obtain

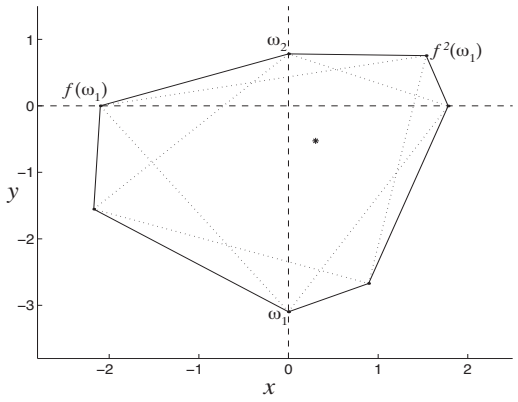


Figure 11. *A phase portrait of (6) for parameter values ($r_L = 0.6$, $s_R \approx 0.7583$, $\omega_L = \omega_R \approx 0.2841$) corresponding to a shrinking point on the 2/7-resonance tongue. The invariant circle is a heptagon that consists entirely of 2/7-cycles. As when $s_R = 1$, the vertices map to one another as do the sides. The stable fixed point is indicated by an asterisk. The dotted lines connect the vertices with their images.*

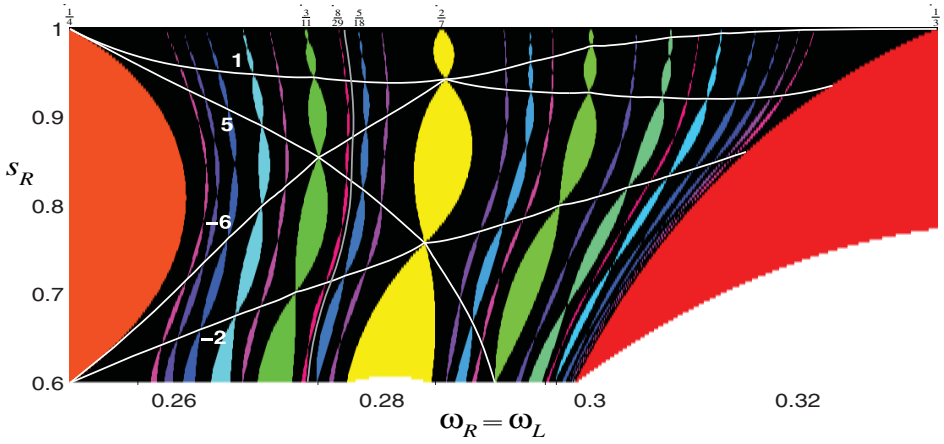


Figure 12. *A magnification of panel A of Figure 8. The white curves are shrinking point curves for $s = -6$, -2, 1, and 5. The gray curve corresponds to stable solutions for which the rotation number is $\frac{1}{2+\gamma} \approx 0.2764$, where γ is the golden ratio.*

analytical expressions.

Shrinking point curves can be computed numerically by first finding an approximate invariant circle by the algorithm described in [12] and then estimating the points w_1 and w_2 by interpolation. We then vary the parameters to minimize $|f^s(w_1) - w_2|$ for some fixed s; an example is shown in Figure 12.

For instance, when $s = 5$, we obtain a curve that extends from the top of the 1/4-resonance region to the bottom of the 3/10-resonance region, as shown in Figure 12. This curve intersects the 2/7-resonance region at which point the associated 2/7-polygon appears as in Figure 11. As seen in this figure, w_1 maps into w_2 upon five iterations of (6). At this point w_2 maps into w_1 in two iterations; hence the shrinking point curve for which $s = -2$ also crosses this point.

Similarly the curves for $s = 1$ and $s = -6$ intersect at the upper shrinking point of the 2/7 lens-chain.

In addition to invariant circles with rational rotation numbers, we can approximately find irrational circles by following a sequence of rational lens-chains whose rotation numbers limit on a given irrational number. For example, in Figure 12, the gray one-dimensional curve corresponds to the existence of an invariant circle whose rotation number is computed to be $\frac{1}{2+\gamma}$, where γ is the golden ratio, within an error of 10^{-10}. Note that this curve is sandwiched between resonance tongues whose rotation numbers correspond to the Farey sequence

$$\frac{1}{3}, \frac{1}{4}, \frac{2}{7}, \frac{3}{11}, \frac{5}{18}, \frac{8}{29} \to \frac{1}{2+\gamma} \approx 0.2764.$$

Resonance tongues associated with the first six rotation numbers in this sequence are shown in Figure 12.

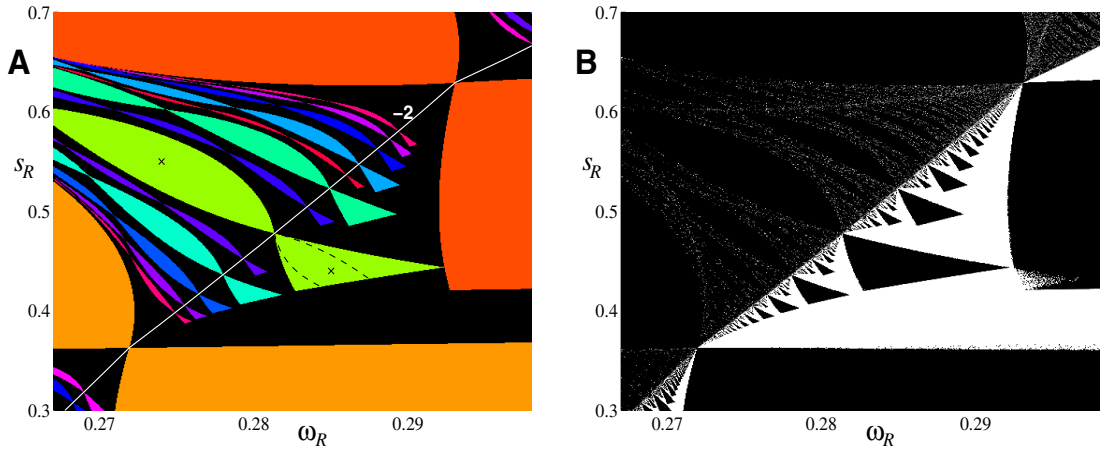


Figure 13. *Panel A shows a magnification of panel A of Figure 7 ($r_L = 0.3$, $\omega_L = 0.09$, and $\mu = 1$). The white curve is the shrinking point curve with $s = -2$; the dashed black curves correspond to first homoclinic tangencies. Phase portraits for parameter values indicated by crosses are shown in Figure 14. Panel B shows numerically calculated Lyapunov exponents over the same parameter range. Black (white) areas correspond to negative (positive) Lyapunov exponents. Gray areas correspond to numerically computed Lyapunov exponents with a magnitude less than* 0.0005.

To distinguish between regular and chaotic orbits, we numerically computed Lyapunov exponents for initial conditions on a 512×512 grid in parameter space; see Figure 13. To create panel B, for each choice of parameter values, we iterated a randomly chosen initial condition for 10^4 steps after removing transients. When there are multiple attractors, the value we compute for the Lyapunov exponent depends upon which basin of attraction the initial random point is located. For example, the black and white area centered at $(\omega_R, s_R) = (0.294, 0.43)$ corresponds to the coexistence of a stable 1/4-cycle and a chaotic attractor born in a flip bifurcation of a 2/9-cycle.

If there exists an attracting invariant circle with an irrational rotation number, the map restricted to the circle is semiconjugate to rigid rotation [6]. Thus, in this case, orbits on the circle are quasiperiodic and have zero Lyapunov exponent. For this reason the upper-left

half of Figure 13, panel B, contains many data points for which the numerically computed Lyapunov exponent has a value within 0.0005 of zero. These grey curves also seem to fall between the lens-chains corresponding to rational rotation numbers, which is consistent with their rotation number being irrational.

The line that runs diagonally through Figure 13 is the $s = -2$ shrinking point curve. To the left of this curve, there appear to be no chaotic solutions, or at least they are very much less common. Figure 14 shows stable and unstable invariant manifolds of the saddle 2/9-cycle on either side of the shrinking point curve. In panel A the associated eigenvalues of both the stable and saddle 2/9-cycles are positive and the unstable manifolds form an attracting invariant circle. In panel B the associated eigenvalues of the stable 2/9-cycle are negative and the unstable manifolds spiral into the stable 2/9-cycle. We have observed a similar situation near the shrinking point curve for all lens-chains shown in Figure 13, panel A; thus we believe Figure 14 illustrates a typical scenario.

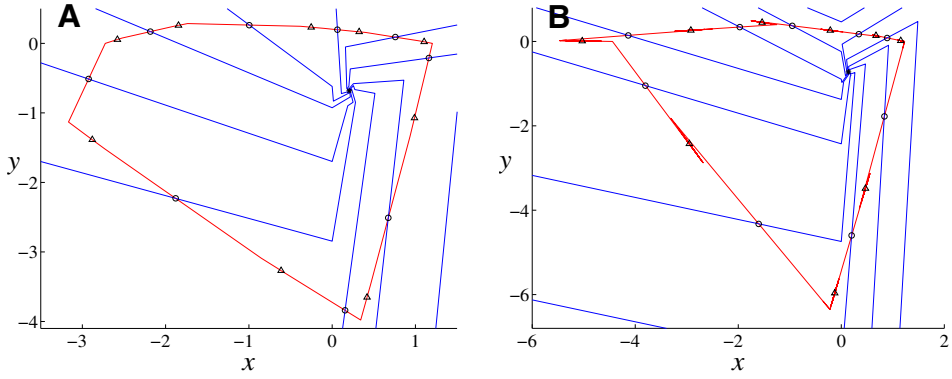


Figure 14. *Stable (blue) and unstable (red) invariant manifolds of a saddle* 2/9*-cycle of* (6) *when* $r_L = 0.3$, $\omega_L = 0.09$, *and* $\mu = 1$. *In panel A,* $\omega_R = 0.274$ *and* $s_R = 0.55$. *In panel B,* $\omega_R = 0.285$ *and* $s_R = 0.44$. *Triangles (circles) denote stable (saddle)* 2/9*-cycles.*

Variation of the parameters of panel B toward a border-collision bifurcation leads to a collision of the stable and unstable manifolds of the saddle cycle. Beyond curves of first tangency (shown in Figure 13) the invariant circle no longer exists. The stable and unstable manifolds intersect transversely forming a homoclinic tangle.

4. More complex phenomena. Despite its simple appearance, the map (6) exhibits an extremely rich array of behavior beyond that described in section 3. Via an example, in section 4.1 we examine the border-collision bifurcation when no invariant circle is created. In section 4.2 we look at multiple attractors. As mentioned in section 3.4, there are three paths by which stable periodic orbits generically lose stability. The first is via an associated real-valued multiplier crossing 1, as detailed in section 3.4. In section 4.3 we detail the case when the crossing occurs at -1. We observe non–lens-chain structures, and these are expounded in section 4.4. The third scenario, that of a complex conjugate pair of multipliers crossing the unit circle, is looked at in section 4.5. Section 4.6 introduces more exotic border-collision bifurcations occurring at $\mu = 0$.

4.1. Saddle fixed points for $\mu = 0$. The white regions in Figure 8 correspond to parameter values where no attractor exists when $\mu = 1$. Here we describe the corresponding border-collision bifurcation that occurs at $\mu = 0$. We find that no invariant circle is created; thus the bifurcation bears little relation to a Neimark–Sacker bifurcation in a smooth system.

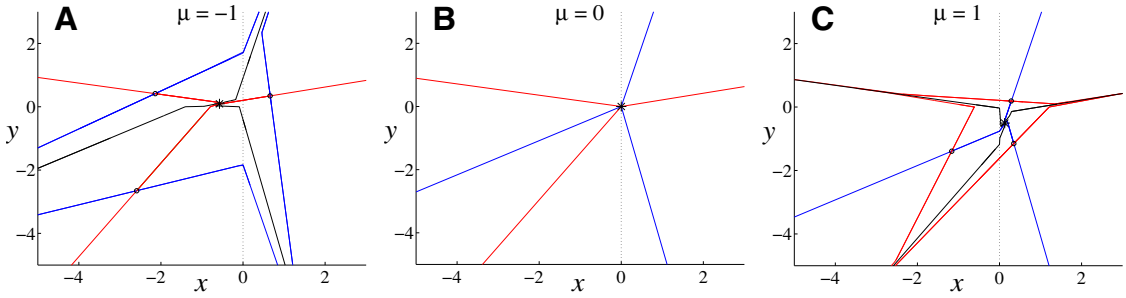


Figure 15. *Stable (blue) and unstable (red) manifolds of a period-3 saddle (indicated by circles) for (6) when $\omega_L = \omega_R = 0.38$, $r_L = 0.4$, $s_R = 0.5$ for $\mu = -1, 0, 1$. The fixed point is indicated with an asterisk.*

As an example, in Figure 15 we show phase portraits for $\mu = -1, 0, 1$. When $\mu = -1$, the unique fixed point of (6) lies in the LHP and is stable. However, its basin of attraction is not the entire plane: there is a period-three saddle with the symbol sequence LLR. The one-dimensional piecewise-linear stable manifolds of this orbit (the blue curves) form the boundary of the basin of attraction for the fixed point. Orbits of points in the interior of the complement of this region are unbounded. When $\mu = 1$, the opposite situation occurs: the fixed point lies in the RHP and is a repellor. The unstable manifolds (red curves) of a period-three orbit with symbol sequence RRL form the boundary for the basin of repulsion of the fixed point.

The LLR orbit that exists when $\mu < 0$ is destroyed at $\mu = 0$; however, its unstable manifolds persist as an invariant set of piecewise-linear curves contained in the basin of repulsion; these are the black curves in panel C of Figure 15. These manifolds extend to infinity and connect to the fixed point. Similarly, the stable manifolds of the RRL 3-cycle become an invariant set of piecewise-linear curves in the basin of attraction of the fixed point when $\mu = 1$.

When $\mu = 0$, the two period-three orbits collide with the fixed point at the origin. The stable manifolds of the RRL orbit and the unstable manifolds of the LLR orbit become stable and unstable manifolds of the origin, which is now a saddle with six hyperbolic sectors.

We believe the above dynamical behavior is generic. When $\omega_L = \omega_R$ and the origin is of saddle type for $\mu = 0$, we expect n stable invariant rays and n unstable invariant rays to emanate from the origin, as we have observed above for $n = 3$. In general, when $\omega_L \neq \omega_R$, more complications may arise.

4.2. Multiple attractors. The overlapping of resonance tongues corresponds to the coexistence of multiple stable periodic cycles. Multiple attractors in piecewise-smooth maps have been described previously; see, for instance, [38, 32, 11, 22, 21]. An overlap is shown, for example, in panel B of Figure 7 near $\omega_R = 0.27$ and $s_R = 0.5$. Two examples of the phase space for this situation are shown in Figure 16. For example, for the parameters of panel A, there exist simultaneous stable period-three and period-four cycles. For this case there are also saddle period-three and period-four cycles, and the stable manifold of the latter saddle

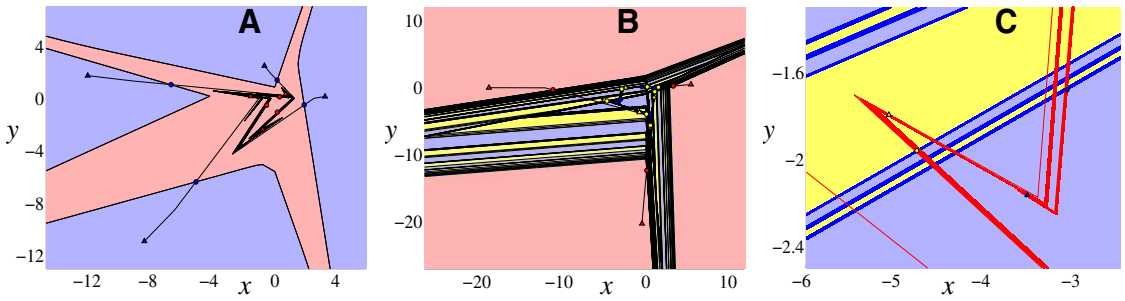


Figure 16. *Attractors and their basins of attraction for (6) $\mu = 1$ and $(r_L, s_R, \omega_L, \omega_R) = (0.45, 0.55, 0.2, 0.4)$ in panel A and $(0.16, 0.52, 0.38, 0.265)$ in panel B. Stable (saddle) periodic orbits are indicated by triangles (circles) and shaded red for period 3, blue for period 4, and yellow for period 7. Panel C is a magnification of panel B; here stable (unstable) invariant manifolds are colored blue (red).*

forms the boundary between the basins of attraction of the two stable orbits.

The coexistence of three attractors is illustrated in panel B of Figure 16; these correspond to the mutual intersection of the 1/3, 1/4, and 2/7 resonance tongues as seen in Figure 7, panel B. The one-dimensional stable manifold of the period-three saddle forms the boundary of the basin of attraction of the stable 3-cycle. However, the boundary between the remaining two basins is more complicated. The stable manifolds of the period-four and period-seven saddles transversely intersect the unstable manifolds of all three saddles forming a collection of homoclinic and heteroclinic orbits. The stable manifolds of the period-four and period-seven saddles appear to accumulate and form a fractal boundary between the two basins. We note that we have been able to find up to six coexisting attractors for the map (6) (specifically when $(\mu, r_L, s_R, \omega_L, \omega_R) = (1, 0.68, 0.8, 0.38, 0.27)$ there exist attractors with symbolic sequences $\mathsf{LRLL}(\mathsf{RLL})^k$ for $k = 7, \ldots, 12$); we believe that arbitrarily many distinct stable solutions can coexist.

4.3. Flip bifurcations of periodic solutions. An attracting periodic orbit of (6) can lose stability by a period-doubling or flip bifurcation. Suppose that an n-cycle with symbol sequence $\mathcal{S}$ has a stability matrix $M_{\mathcal{S}}$, with one multiplier inside the unit circle and one multiplier near -1; call it λ_{PD}. As described in section 2.1, if no points of the n-cycle lie on the switching manifold, there is a neighborhood of its initial condition with the same symbol sequence for the first n iterates; recall (14). Let $v_{\mathsf{PD}}^{(0)}$ denote the eigenvector of $M_{\mathcal{S}}$ at $z^{(0)}$ associated with λ_{PD}. Since $v_{\mathsf{PD}}^{(0)}$ corresponds to the dynamically slow direction of the periodic orbit, orbits of (6) that start sufficiently close to the periodic orbit approach the image vectors $v_{\mathsf{PD}}^{(i)}$ associated with λ_{PD} for each $z^{(i)}$. This is illustrated in Figure 17.

When $\lambda_{\mathsf{PD}} = -1$, points on the slow eigenvectors that are sufficiently close to the period-n orbit return to themselves after $2n$ iterations; therefore, there exists a segment of $2n$-cycles with a symbol sequence equal to the concatenation of $\mathcal{S}$ with itself, which we will denote $\mathcal{S}_{\mathsf{PD}}$.

This family of $2n$-cycles has a "first" point of intersection with the switching manifold. Let $q_\delta = z^{(0)} + \delta v_{\mathsf{PD}}$; then for $\lambda_{\mathsf{PD}} = -1$ the orbits of q_δ have the symbol sequence $\mathcal{S}_{\mathsf{PD}}$ providing $\delta \in [0, \delta_{\max}]$ for some $\delta_{\max} > 0$. Moreover, a single point on the $2n$-cycle will generically touch the switching manifold at $\delta_{\max}$. Without loss of generality, we can assume that a permutation

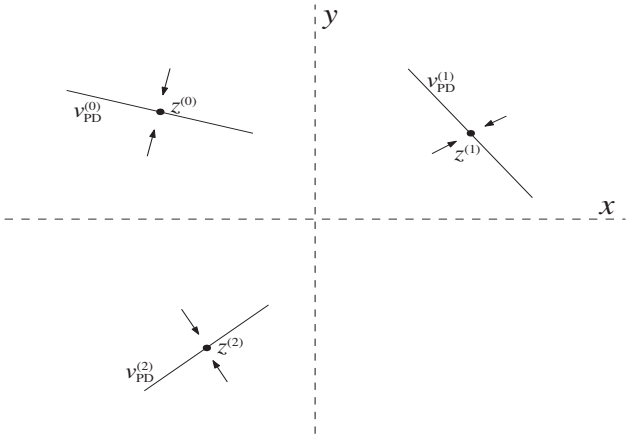


Figure 17. *Schematic showing an n-cycle and its associated slow eigenvectors in a neighborhood of the n-cycle.*

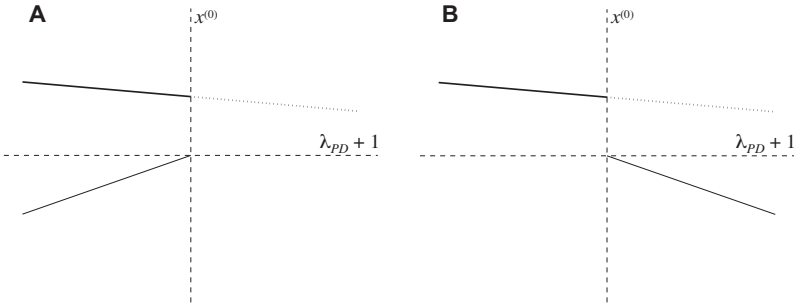


Figure 18. *Schematics of bifurcation diagrams about a flip bifurcation of (6). The solid (dotted) line corresponds to a stable (unstable) n-cycle. The thin line corresponds to a $2n$-cycle of unknown stability.*

of the orbit is selected so that the point that hits the switching manifold first is $q_{\delta_{\max}}$. Thus, when $\lambda_{\mathsf{PD}} = -1$, there is a period-$2n$ orbit with symbol sequence $\mathcal{S}_{\mathsf{PD}}$ that has its first point on the switching manifold.

Since the first symbol is ambiguous, we can also declare that this orbit has a switching manifold with its first symbol flipped; call this sequence $\mathcal{S}_{\mathsf{PD}^*}$. This orbit is an admissible solution of the system (10) with sequence $\mathcal{S}_{\mathsf{PD}^*}$.

The x-component of this solution will typically change and will have the correct sign, corresponding to $\mathcal{S}_{\mathsf{PD}^*}$ for only one sign of $\lambda_{\mathsf{PD}} + 1$: this will correspond to an admissible, isolated period-$2n$ orbit. This $2n$-cycle will coexist with either the stable n-cycle, as sketched in panel A, or the unstable n-cycle, as sketched in panel B of Figure 18. We have not been able to derive a general result regarding the stability of the period-doubled cycle; however, it appears most often to be unstable (see section 4.4 for a stable example).

For the case that the doubled cycle is unstable and exists for $\lambda_{\mathsf{PD}} < -1$, we have observed that it is typically embedded in a complicated attractor. In some cases, this attractor coincides

with the doubled cycle when it is created and grows in size as $\lambda_{\mathsf{PD}}+1$ decreases; see panel A of Figure 19. Alternatively, the attractor can be large and contain the n-cycle when it is created; see panel B of Figure 19. In both cases the Lyapunov exponent for the attractor appears to be positive ($\gamma_L \approx 0.0747$ in panel A and $\gamma_L \approx 0.0129$ in panel B), suggesting that the attractor is chaotic. As the parameters are varied further, the multiple-piece attractor may undergo merging [21].

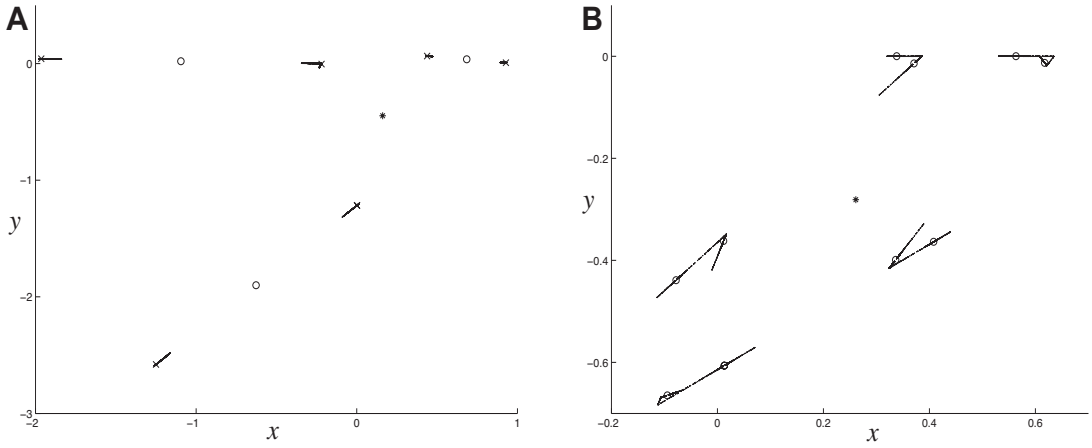


Figure 19. *Phase portraits of* (6) *when* $\lambda_{\mathsf{PD}} \approx -1.01$, $\mu = 1$. *In panel A,* $n = 3$*; in panel B,* $n = 10$. *An unstable fixed point is indicated by an asterisk. A saddle* n*-cycle is indicated by circles. In panel A, a saddle* $2n$*-cycle is indicated by crosses. In both panels, a complicated attractor is indicated by dots. The parameter values are* $(r_L, s_R, \omega_L, \omega_R) = (0.18202, 0.6, 0.09, 0.38)$ *in panel A and* $(r_L, s_R, \omega_L, \omega_R) = (0.0193, 0.964, 0.38, 0.41)$ *in panel B.*

So far we have assumed that a single point, $q_{\delta_{\max}}$, on the period-$2n$ orbit with $\lambda_{\mathsf{PD}} = -1$ hits the switching manifold. By varying another parameter, it is possible to have two such points on the switching manifold, say, $q^{(i)}_{\delta_{\max}}$ and $q^{(j)}_{\delta_{\max}}$, when $\lambda_{\mathsf{PD}} = -1$; this corresponds to a codimension-two bifurcation.

In this case there will be two curves that cross at the codimension-two point; these correspond to the vanishing of the x-components of each $q_{\delta_{\max}}$ individually. These two curves divide the neighborhood of the codimension-two point into four quadrants. In one quadrant there will be no admissible period-$2n$ orbit, and in two quadrants exactly one of the two new orbits will be admissible; one will have a symbol sequence with the ith symbol in PD flipped and the other with the jth symbol flipped. In the final quadrant, a new period-$2n$ switching manifold will be admissible—that corresponding to flipping *both* the ith and jth symbols.

In Figure 20 we show two examples where the doubly flipped period-$2n$ orbits are stable. In these cases the primary orbits, 1/3 and 2/5, respectively, are seen to lose stability with $\lambda_{\mathsf{PD}} = -1$, and there is a narrow tongue corresponding to stable orbits with rotation numbers 2/6 and 4/10, respectively, that emanates from a codimension-two point on the period-doubling curve. Interestingly, there also exist additional resonance tongues in the neighborhood of this codimension-two point. These appear in sequences with rotation numbers $2k/6k$ for k up to 7 in panel A and rotation numbers $4k/10k$ for k up to 3 in panel B. We have been unable to locate resonance tongues corresponding to larger values of k. In panel A the doubled

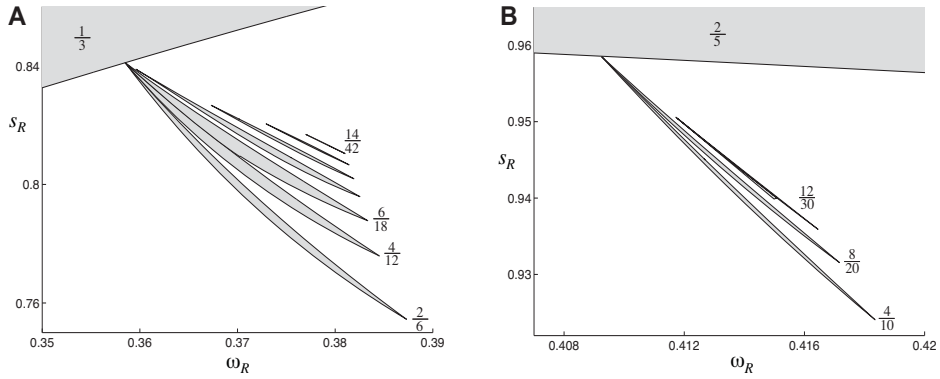


Figure 20. *Magnifications of panels A and B of Figure 7 near codimension-two flip bifurcations of (6). For panel A, the original orbit has rotation number 1/3 and symbol sequence* LLR*, and for panel B it has rotation number 2/5 and symbol sequence* LRLRR*. The shaded regions correspond to the existence of a stable orbit with the rotation numbers shown.*

orbits have symbol sequences $(\mathsf{LLR})^{2k-1}\mathsf{RRR}$; note that only the first three tongues in the sequence appear to emanate from the codimension-two point. In panel B the doubled orbits have symbol sequences $(\mathsf{LRLRR})^{2k-1}\mathsf{RRRRR}$, and only the first two tongues emanate from the codimension-two point.

4.4. Reducible rotation numbers. The rotation number of an n-cycle of the map (6) is a fraction m/n, where m is the number of times the orbit revolves around the fixed point in n steps. In this section we discuss the case in which m/n is a reducible fraction and describe new codimension-one bifurcations that do not seem to exist for the irreducible case. Furthermore, we will see that resonance regions associated with reducible rotation numbers do not appear to exhibit the familiar lens-chain structure.

An example with reducible rotation numbers is shown in Figure 21. Here there are three separate, adjoining resonance tongues. The tongue (abe) corresponds to parameters for which there exists a stable 3/9-cycle with symbol sequence LRLLRRRRL. The right-hand boundary of this region, (a-e), corresponds to a flip bifurcation. To the right of this boundary the 3/9-cycle exists but is not attracting. Throughout this tongue there also exists a saddle 3/9-cycle that collides and annihilates with the stable cycle in a usual border-collision bifurcation on the boundary (a-b).

The boundary (b-e) also corresponds to a border-collision bifurcation; however, here the 3/9-cycle persists above the boundary. Above this boundary the 3/9-cycle has symbol sequence RRLLRRRRL and is stable to the right of the boundary (d-i), which corresponds to a flip bifurcation. Along (b-d) a 6/18-cycle is created that exists above the boundary and is stable to the right of (c-g), which also corresponds to a flip bifurcation. This period-doubled cycle has a symbol sequence that is the concatenation of the symbol sequences of the two stable 3/9-cycles. Finally, the boundaries (g-h) and (f-i) correspond to border-collision bifurcations beyond which saddles of the same rotation number continue to exist.

Of all the border-collision boundaries that we have so far discussed, the boundary (d-e) is the only one for which stable n-cycles exist on both sides of the bifurcation. Also, of all the

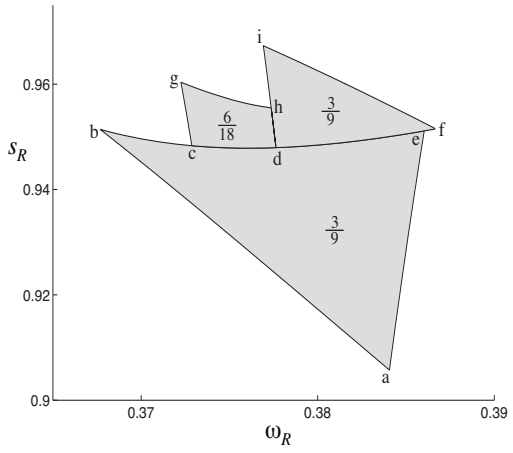


Figure 21. *Resonance tongues of* (6) *for* $\mu = 1$, $r_L = 0.475$, $\omega_L = 0.09$. *The bottom and upper right tongues correspond to a stable* 3/9*-cycle. The upper left tongue corresponds to a stable* 6/18*-cycle.*

codimension-one flip bifurcation boundaries so far discussed, the boundary (d-h) is the only one along which the period-doubled solution is stable. We have observed scenarios similar to Figure 21 for different rotation numbers. We conclude that periodic solutions for which the rotation number is a reducible fraction appear to exhibit a wider variety of codimension-one bifurcations than those with a rotation number that is irreducible. In particular, we do not see the familiar lens-chain structure.

4.5. Neimark–Sacker bifurcations of periodic solutions. Just as the fixed point can undergo a Neimark–Sacker bifurcation, so can a periodic orbit. When this bifurcation does not coincide with a border collision, the stability matrix $M_{\mathcal{S}}$ will have a pair of complex eigenvalues on the unit circle so that $\det(M_{\mathcal{S}}) = 1$. By (11) this corresponds to $r_L^l = s_R^{n-l}$, where l is the number of L's in $\mathcal{S}$. As an example, this bifurcation occurs for the period-three orbit LLR, when $(\mu, r_L, s_R, \omega_L) = (1, 0.5, 0.25, 0.25)$ and any $\omega_R \in (\hat{\omega}, \frac{1}{2})$, where $\hat{\omega} = \frac{1}{2\pi}\cos^{-1}(\frac{19}{32}) \approx 0.3512$. If the complex multiplier pair of an m/n orbit crosses the unit circle at $e^{2\pi i \frac{p}{q}}$ for an irreducible fraction p/q, there will exist uncountably many period-qn orbits with rotation numbers qm/qn at the bifurcation. Generically one of these cycles has two points on the switching manifold. In a similar manner as for the codimension-two flip bifurcation described in section 4.3, a qm/qn-resonance tongue emanates from this codimension-two Neimark–Sacker point in parameter space. However, we have not found an example with (6) for which the qn-cycle is stable.

4.6. Further complications. We have mostly considered the case $\mu > 0$ for the map (6). However, using the symmetry property (4), each stable (unstable) solution for $\mu > 0$ corresponds to an unstable (stable) solution of the same period and rotation number for $\mu < 0$ if the R and L parameters are exchanged. For some choices of parameter values, nontrivial periodic orbits can be observed for both signs of μ. For example, in panel A of Figure 22, we see an unstable 4-cycle for $\mu < 0$ and a stable 3-cycle for $\mu > 0$. Alternatively, a stable n-cycle can coexist with the stable fixed point. For example, in panel B of Figure 22, a stable 5-cycle coexists with the stable fixed point for $\mu < 0$, and a stable period-six orbit is created

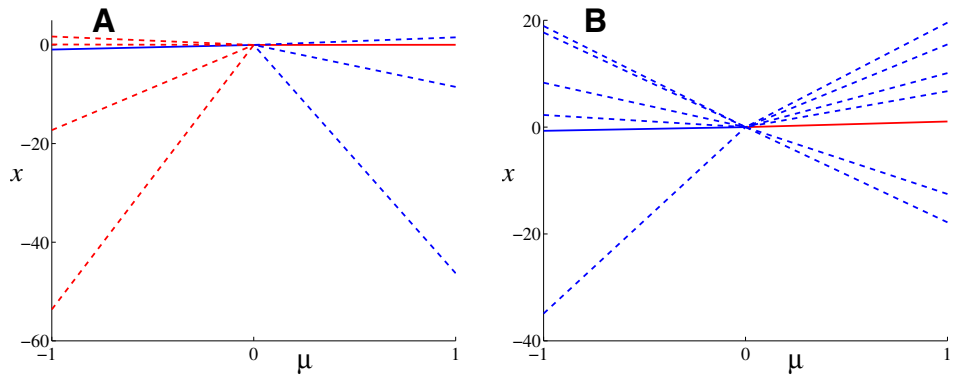


Figure 22. *Bifurcation diagrams of* (6) *for* $(r_L, s_R, \omega_L, \omega_R) = (0.2, 0.7, 0.49, 0.12)$ *in panel A and* $(0.25, 0.18, 0.25, 0.38)$ *in panel B. Blue (red) lines denote stable (unstable) solutions, solid lines correspond to the fixed point, and dashed lines correspond to periodic orbits.*

for $\mu > 0$.

Recall also from section 4.2 that multiple attractors may coexist. Similarly stable and unstable solutions with different periods may coexist, and various combinations of these phenomena may occur. The bifurcation at $\mu = 0$ may be extremely complex.

5. Conclusion. In this paper we have studied border-collision bifurcations of fixed points in planar, piecewise-smooth, continuous maps for the case that the multipliers are complex and "jump" from inside to outside the unit circle.

We investigated a piecewise-linear approximation (6), which we believe preserves local dynamics under the addition of nonlinear terms, except at higher codimension points such as at the boundaries of resonance tongues. For example, at shrinking points nonlinear terms may dramatically affect dynamics. The map (6) is a homeomorphism with a single fixed point that is stable for $\mu < 0$ and unstable for $\mu > 0$. We have found a large variety of periodic, quasiperiodic, and chaotic attractors for this system. Typically these attractors are created at the bifurcation, exist for one sign of μ, and grow in size linearly with respect to μ. These features are commonly observed in piecewise-smooth systems [8, 20] but are not typical in smooth systems. We have seen that, unlike the one-dimensional case, [8], multiple attractors may coexist in our two-dimensional map. Also, attractors may coexist with repellors; the bifurcation at $\mu = 0$ may be very complicated.

For some parameter values the border-collision bifurcation is analogous to that of a Neimark–Sacker bifurcation in a smooth map: an invariant circle is created that may be stable or unstable. As the parameters of the map are varied, the invariant circle may be destroyed. Though we investigated some of the ways in which this may occur, we have not given a complete classification.

Periodic orbits are classified by their symbol sequence $\mathcal{S}$ and rotation number m/n. Stable m/n-cycles exist in parameter regions (resonance tongues) that are bounded by curves of border-collision bifurcations or loss of stability. When m/n is an irreducible fraction, the resonance tongue has the form of a chain of lenses (also seen by [37, 27]); each lens corresponds to cycles with a particular symbol sequence. When m/n is reducible, lenses of the

associated resonance tongue may connect along intervals rather than at points. We observe a wider variety of border-collision bifurcations at boundaries of tongues with reducible rotation numbers.

There are many directions by which our investigations may be extended. For example, if we do not require that the multipliers of the fixed point are complex, just that they both jump from inside the unit circle to outside, we expect more complications to arise. In particular, for this case the map need not be a homeomorphism. We would like to extend our results here to higher dimensional maps; however, a center-manifold analysis cannot be applied to this bifurcation problem in the usual manner, and we are unaware of a theory of dimension reduction in piecewise-smooth systems.

Acknowledgment. We would like to thank Holger Dullin for many helpful discussions.

REFERENCES

[1] D. G. Aronson, M. A. Chory, G. R. Hall, and R. P. McGehee, *Bifurcations from an invariant circle for two-parameter families of maps of the plane: A computer-assisted study*, Comm. Math. Phys., 83 (1982), pp. 303–354.
[2] S. Banerjee and C. Grebogi, *Border collision bifurcations in two-dimensional piecewise smooth maps*, Phys. Rev. E (3), 59 (1999), pp. 4052–4061.
[3] S. Banerjee, M. S. Karthik, G. Yuan, and J. A. Yorke, *Bifurcations in one-dimensional piecewise smooth maps: Theory and applications in switching circuits*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 47 (2000), pp. 389–394.
[4] S. Banerjee and G. C. Verghese, eds., *Nonlinear Phenomena in Power Electronics*, IEEE Press, New York, 2001.
[5] B. Brogliato, *Nonsmooth Mechanics: Models, Dynamics and Control*, Springer-Verlag, New York, 1999.
[6] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai, *Ergodic Theory*, Springer-Verlag, New York, 1982.
[7] M. di Bernardo, C. J. Budd, and A. R. Champneys, *Normal form maps for grazing bifurcations in n-dimensional piecewise-smooth dynamical systems*, Phys. D, 160 (2001), pp. 222–254.
[8] M. di Bernardo, C. J. Budd, A. R. Champneys, and P. Kowalczyk, *Piecewise-Smooth Dynamical Systems. Theory and Applications*, Springer-Verlag, New York, 2008.
[9] M. di Bernardo, M. I. Feigin, S. J. Hogan, and M. E. Homer, *Local analysis of C-bifurcations in n-dimensional piecewise-smooth dynamical systems*, Chaos Solitons Fractals, 10 (1999), pp. 1881–1908.
[10] H. R. Dullin, D. Sterling, and J. D. Meiss, *Self-rotation number using the turning angle*, Phys. D, 145 (2000), pp. 25–46.
[11] M. Dutta, H. E. Nusse, E. Ott, J. A. Yorke, and G. Yuan, *Multiple attractor bifurcations: A source of unpredictability in piecewise smooth systems*, Phys. Rev. Lett., 83 (1999), pp. 4281–4284.
[12] K. D. Edoh and J. Lorenz, *Numerical approximation of rough invariant curves of planar maps*, SIAM J. Sci. Comput., 25 (2003), pp. 213–223.
[13] M. Gallegati, L. Gardini, T. Puu, and I. Sushko, *Hicks' trade cycle revisited: Cycles and bifurcations*, Math. Comput. Simulation, 63 (2003), pp. 505–527.
[14] J. A. Gallian, *Contemporary Abstract Algebra*, Houghton Mifflin, Boston, 1998.
[15] L. Gardini, T. Puu, and I. Sushko, *The Hicksian model with investment floor and income ceiling*, in Business Cycle Dynamics: Models and Tools, T. Puu and I. Sushko, eds., Springer-Verlag, New York, 2006, pp. 179–191.
[16] C. Halse, M. Homer, and M. di Bernardo, *C-bifurcations and period-adding in one-dimensional piecewise-smooth maps*, Chaos Solitons Fractals, 18 (2003), pp. 953–976.
[17] J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.
[18] Yu. A. Kuznetsov, *Elements of Bifurcation Theory*, 3rd ed., Appl. Math. Sci. 112, Springer-Verlag, New York, 2004.

[19] J. Laugesen and E. Mosekilde, *Border-collision bifurcations in a dynamic management game*, Comput. Oper. Res., 33 (2006), pp. 464–478.
[20] R. I. Leine and H. Nijmeijer, *Dynamics and Bifurcations of Non-Smooth Mechanical Systems*, Lecture Notes in Applied and Computational Mathematics 18, Springer-Verlag, Berlin, 2004.
[21] Y. Maistrenko, I. Sushko, and L. Gardini, *About two mechanisms of reunion of chaotic attractors*, Chaos Solitons Fractals, 9 (1998), pp. 1373–1390.
[22] H. E. Nusse, E. Ott, and J. A. Yorke, *Border-collision bifurcations: An explanation for observed bifurcation phenomena*, Phys. Rev. E (3), 49 (1994), pp. 1073–1076.
[23] H. E. Nusse and J. A. Yorke, *Border-collision bifurcations including "period two to period three" for piecewise smooth systems*, Phys. D, 57 (1992), pp. 39–57.
[24] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 2001.
[25] K. Popp, *Non-smooth mechanical systems*, J. Appl. Math. Mech., 64 (2000), pp. 765–772.
[26] T. Puu, L. Gardini, and I. Sushko, *On the change of periodicities in the Hicksian multiplier-accelerator model with a consumption floor*, Chaos Solitons Fractals, 29 (2006), pp. 681–696.
[27] T. Puu and I. Sushko, eds., *Business Cycle Dynamics: Models and Tools*, Springer-Verlag, New York, 2006.
[28] C. Robinson, *Dynamical Systems. Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, FL, 1999.
[29] R. Rosen, *Dynamical System Theory in Biology*, Wiley-Interscience, New York, 1970.
[30] D. J. W. Simpson, *Bifurcations in Piecewise-Smooth, Continuous Systems*, Ph.D. thesis, in progress.
[31] D. J. W. Simpson and J. D. Meiss, *Andronov-Hopf bifurcations in planar, piecewise-smooth, continuous flows*, Phys. Lett. A, 371 (2007), pp. 213–220.
[32] I. Sushko and L. Gardini, *Center bifurcation for a two-dimensional piecewise linear map*, in Business Cycle Dynamics: Models and Tools, T. Puu and I. Sushko, eds., Springer-Verlag, New York, 2006, pp. 49–78.
[33] C. K. Tse, *Complex Behavior of Switching Power Converters*, CRC Press, Boca Raton, FL, 2003.
[34] M. Wiercigroch and B. De Kraker, eds., *Applied Nonlinear Dynamics and Chaos of Mechanical Systems with Discontinuities*, World Scientific, Singapore, 2000.
[35] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Texts Appl. Math. 2, Springer-Verlag, New York, 2003.
[36] W.-M. Yang and B.-L. Hao, *How the Arnol'd tongues become sausages in a piecewise linear circle map*, Comm. Theoret. Phys., 8 (1987), pp. 1–15.
[37] Z. T. Zhusubaliyev and E. Mosekilde, *Bifurcations and Chaos in Piecewise-Smooth Dynamical Systems*, World Scientific, Singapore, 2003.
[38] Z. T. Zhusubaliyev, E. Mosekilde, S. Maity, S. Mohanan, and S. Banerjee, *Border collision route to quasiperiodicity: Numerical investigation and experimental confirmation*, Chaos, 16 (2006), 023122.
[39] Z. T. Zhusubaliyev, E. Soukhoterin, and E. Mosekilde, *Quasiperiodicity and torus breakdown in a power electronic dc/dc converter*, Math. Comput. Simulation, 73 (2007), pp. 364–377.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 825–842
© 2008 Society for Industrial and Applied Mathematics

Master-Slave Global Stochastic Synchronization of Chaotic Oscillators*

Maurizio Porfiri† and Roberta Pigliacampo†

Abstract. We study global synchronization of coupled chaotic systems with random intermittent coupling. We use stochastic Lyapunov stability theory and partial averaging techniques to show that global synchronization is possible if the switching period is sufficiently small and if the oscillators globally synchronize under a time-averaged coupling. We study mean square and almost sure global synchronization, and we determine quantitative bounds for the exponential rate of decay of the synchronization error. We focus on master-slave synchronization, where two dynamical systems are coupled via a directed feedback that randomly switches among a finite set of given constant functions at a prescribed time rate. We apply the proposed approach to the synchronization of Chua circuits.

Key words. master-slave synchronization, global synchronization, stochastic synchronization, chaos, Chua circuit, exponential stability

AMS subject classifications. 34C15, 34C29, 34D23, 74H65, 93C10, 93E15

DOI. 10.1137/070688973

1. Introduction.

Chaos synchronization is a topic of great interest, due to its observation in a huge variety of phenomena of different natures. In many biological systems, synchronization plays an important role in self-organization of organisms' groups [11]. Examples of synchronization include communication among fireflies [10, 37], locomotion of animals [14], molecular and cellular activity [26], and cardiac stimulation [23, 29, 46]. The study of neural activity [48, 58, 68] and brain disorders [3, 57] is a correlated issue as well. Other examples of synchronization can be found in ecological systems [6], meteorology [18], chemistry [26, 39], gas-liquid bubbling dynamics [61, 66], and optics [59, 67]. Many reviews on chaos synchronization are currently available; see, for example, [2, 7, 13, 25, 47, 52, 54].

In the literature, different paradigms have been proposed to enforce synchronization of two or more coupled chaotic oscillators. We mention, among the others, peer-to-peer coupling [4, 24, 56, 60, 64], back-stepping [8], generalized synchronization [71, 72], phase synchronization [7], and master-slave synchronization [12, 22, 28, 33, 34, 43, 49, 50, 51, 65, 72, 73]. In this work, we focus on master-slave synchronization. In this case, one system acts as a "master" by driving the other system that behaves consequently as a "slave."

Most of the research efforts on chaos synchronization focus on time-invariant coupling [2, 7, 13, 25, 47, 52, 54]. Nevertheless, experimental and numerical evidence on master-slave synchronization shows that synchronization can also be achieved using time-varying intermittent feedback coupling [22, 33, 72, 73]. In [22], experimental results on synchronization

*Received by the editors April 21, 2007; accepted for publication (in revised form) by M. Silber March 5, 2008; published electronically July 25, 2008. This work was supported by the National Science Foundation via grant CMMI-0745753.
http://www.siam.org/journals/siads/7-3/68897.html
†Mechanical and Aerospace Engineering Department, Polytechnic University, Brooklyn, NY 11201 (mporfiri@poly.edu, rpigli01@utopia.poly.edu).

of two periodically coupled chaotic circuits are presented. In [33, 72], the slave system is driven by a sequence of samples of the master's state. In [73], the signal transmission from the master to the slave system is adaptively controlled. That is, the driving signal is transmitted only when it is expected to reduce the synchronization error.

In this work, we establish sufficient conditions for global exponential synchronization of coupled chaotic systems with random intermittent coupling. We associate to the stochastic system a partially averaged system describing synchronization under time-constant coupling. We transform the synchronization problem into a nonlinear stochastic stability problem by describing the system's dynamics through the synchronization error. We present a general framework for assessing global exponential stability of switched stochastic systems from global exponential stability of their time-averaged counterparts. Under certain regularity conditions, we show that global exponential synchronization of intermittently coupled oscillators is possible if the oscillators exponentially synchronize under a time-constant average coupling and if the coupling is switching sufficiently quickly. In addition, we provide a rigorous estimate for the slowest fast-switching rate that guarantees global synchronization. We particularize our findings to linearly coupled oscillators, where intermittent coupling is made possible through a switching linear state feedback. The switching feedback gain changes randomly over time while assuming values among a finite set of constant values. For this configuration, global exponential synchronization of the partially averaged system can be studied using well-established and manageable techniques based on Lyapunov stability theory and Gerschgorin's theorem, such as those presented in [34].

Synchronization of oscillator networks under intermittent fast-switching coupling is also studied in [4, 56, 64] in case of peer-to-peer synchronization. In [56, 64], only local asymptotic synchronization results based on linearized dynamics are presented, while this work and [4] focus on global synchronization by retaining the nonlinear nature of the coupled systems. In [4], a new type of small-world network of cells with chaotic oscillators is investigated. Cells are coupled through a time-varying network that consists of a fixed, so-called *pristine network* and intermittent links between any pair of cells that are used to describe small-world dynamical effects. As shown by the authors, intermittent links facilitate synchronization of the oscillator network by reducing the synchronization threshold. Time-varying interconnections are considered as binary independent identically distributed random variables. That is, they do not influence each other, they have the same probability to be present, and when they are on, they share the same strength. In addition, time-varying interconnections are allowed to change only simultaneously and at equally spaced instants of time, thereby fixing the switching rate of the time-varying network to a constant value. The authors determine rigorous bounds for the strength of the intermittent coupling and the switching rate to guarantee global asymptotic synchronization of the oscillator network for almost all the switching sequences, that is, global asymptotic almost sure synchronization. The claims are proved by showing that, under some general hypotheses, there exists an autonomous quadratic Lyapunov function for the synchronization error dynamics that asymptotically goes to zero for almost all the switching sequences. The autonomous quadratic Lyapunov function is constructed from a thorough analysis of the synchronization problem over the time-averaged network topology based on the connection graph stability method [5].

In this paper, we extend the mathematical tools presented in [4] to global exponential

synchronization of coupled oscillators. Beyond analyzing almost sure synchronization, we also focus on mean square synchronization. That is, we derive conditions for the second moment of the error dynamics to converge to zero. Mean square and almost sure convergence of stochastic processes are in general not equivalent [30, 69]. For example, if a stochastic process takes on increasingly large values with decreasing probability and the rate of increase is sufficiently fast, then it may converge almost surely while its moments diverge; see, for example, [42]. Derived results on mean square synchronization can be potentially useful in assessing the effects of perturbations and unmodeled dynamics on synchronizability of oscillator networks; see, for example, the approaches outlined in [17, 70]. We provide rigorous bounds for the exponential rate of decay of the synchronization error dynamics. In particular, we analyze the exponential rate of convergence to zero of the error second moment and of the probability of the error norm to be larger than a threshold value. Estimating these convergence rates can be potentially useful in assessing the coupling performance and in optimizing the design of time-varying coupling strategies. In addition, unlike [4], we consider time-varying interconnections that can in principle assume a variety of values and change at a nonconstant but bounded switching rate. One of the main contributions of our work is Theorem 2.4, which extends recent results on Lyapunov stability theory [1] and averaging methods [27, 53] for deterministic dynamical systems to stochastic systems. Theorem 2.4 applies to a large class of nonautonomous candidate Lyapunov functions, including the autonomous quadratic candidate Lyapunov functions considered in [4]. In addition, it yields general conditions for exponential stability of stochastic nonautonomous nonlinear systems that include as a special case the results in [4].

The system studied in this paper finds many practical applications. For example, in communication and signal processing, chaotic behavior can be used for message encryption and secure communication [16, 19, 20, 32, 35, 45]. Higher communication efficiency can be potentially achieved through sporadic transmission of the driving signal. This is particularly useful when the available resources are shared and the amount of information that can be transmitted is limited.

We organize the paper as follows. In section 2, we present our general results on stability of nonlinear stochastic systems. In section 3, we apply these results to the master-slave synchronization problem. As a sample case, in section 4 we consider the case of two stochastically coupled Chua circuits. Section 5 is left for conclusions.

2. Global exponential stability through fast-switching. We consider the integral equation in $\mathbb{R}^n$

$$x(t) = x(\sigma_k) + \int_{\sigma_k}^{t} f(x(\xi), \xi, \Omega) d\xi, \tag{2.1}$$

where $t \in [\sigma_k, \sigma_{k+1})$, $\sigma_k = k\varepsilon$, $\varepsilon > 0$, $k \in \mathbb{Z}^+$, and n is a positive integer. The function f is defined in $\mathbb{R}^n \times \mathbb{R}^+ \times \Theta$ and is piecewise continuous with respect to t. Here, Ω is a discrete random variable taking values in the finite set $\Theta = \{\omega_1, \ldots, \omega_N\}$, with N a positive integer. We assume that the origin is an equilibrium of every sample system; that is, $f(0, t, \omega) = 0$ for every $t \in \mathbb{R}^+$ and every $\omega \in \Theta$. We further assume that for every $\omega \in \Theta$ the function $f_\omega(\bullet, \bullet) = f(\bullet, \bullet, \omega)$ is globally Lipschitz in $\mathbb{R}^+$, with Lipschitz constant $L_{\omega,\varepsilon}$. In addition, we

require that $L_{\omega,\varepsilon} \leq L$, where L is a constant independent of ω and ε. We note that (2.1) describes a Markovian nonlinear nonhomogeneous jump system; see, for example, [15]. We study the solutions of (2.1) for $t \geq t_0 \in \mathbb{R}^+$ and initial conditions $x(t_0) = x_0$.

In what follows, we use $\mathsf{E}[\bullet]$ to indicate expectation, we denote probability with $P\{\bullet\}$, and we use "a.s." for abbreviating "almost sure" or "with probability one"; see, for example, [30]. We refer to $\| \bullet \|$ as the Euclidean norm in $\mathbb{R}^n$ or the corresponding induced norm in $\mathbb{R}^{n\times n}$. We use the superscript T for matrix transposition. For brevity, we refer to the stability of the origin as the system stability.

In this section, we establish sufficient conditions for global stability of the stochastic system (2.1). To this aim, we recall the definitions of global mean square exponential stability (see, for example, [21, 44]) and global almost sure exponential stability (see, for example, [44]). As stated in section 1, mean square stability and almost sure stability are in general not equivalent. The relationship between these concepts in stochastic observer design has been studied in [69], whereas a comparison between them in the analysis of asynchronous systems with Poisson transitions can be found in [42].

Definition 2.1. *The system* (2.1) *is globally mean square exponentially stable if there exist* $\alpha \geq 0$ *and* $\beta > 0$ *such that for any* $x_0 \in \mathbb{R}^n$ *and* $t_0 \in \mathbb{R}^+$

$$\mathsf{E}[\|x(t)\|^2] \leq \alpha \|x_0\|^2 e^{-\beta(t-t_0)}$$

$\forall t \geq t_0$.

Definition 2.2. *The system* (2.1) *is globally almost surely exponentially stable if there exist a constant* $\zeta \geq 0$ *and a positive bounded random variable* ϱ *such that for any* $x_0 \in \mathbb{R}^n$ *and* $t_0 \in \mathbb{R}^+$

$$\|x(t)\| \leq \varrho e^{-\zeta(t-t_0)}$$

a.s. $\forall t \geq t_0$.

From classical Lyapunov stability theory, it is well known that a deterministic dynamical system is uniformly asymptotically stable if there exists a positive definite decrescent candidate Lyapunov function whose time derivative along the solutions of the system is strictly negative definite; see, for example, [36]. In [1], this condition is relaxed, and it is shown that if the candidate Lyapunov function decreases when evaluated at a discrete sequence of time instants, the system is uniformly asymptotically stable. In this case, the time derivative of the Lyapunov function can assume positive and negative values. The following theorem extends the results of [1] from the deterministic to the stochastic case and serves as a preliminary result to establish our main claim, that is, Theorem 2.4.

Theorem 2.3. *Consider the system* (2.1)*, and suppose that there exists a function* $V : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}$ *such that* $\forall (x,t) \in \mathbb{R}^n \times \mathbb{R}^+$

$$\lambda_{min}\|x\|^2 \leq V(x,t) \leq \lambda_{max}\|x\|^2 \tag{2.2}$$

with λ_{min} *and* λ_{max} *positive nonzero real constants. Assume also that there exists* ν*, with* $0 < \nu \leq 1$*, such that*

$$\mathsf{E}[V(x(\sigma_{k+1}),\sigma_{k+1})|x(\sigma_k)] - V(x(\sigma_k),\sigma_k) \leq -\nu V(x(\sigma_k),\sigma_k) \tag{2.3}$$

for every $k \in \mathbb{Z}^+$. Then, (2.1) is globally mean square exponentially stable and globally almost surely exponentially stable.

Proof. See Appendix A. ∎

Remark 1. *In the literature, different definitions of almost sure exponential stability for stochastic differential equations of the kind of (2.1) have been proposed; see, for example, [44, 40]. In this work we use the definition of [44], which, loosely speaking, enforces the exponential stability of almost all the sample systems in the deterministic sense. On the other hand, the exponential stability notion introduced in [40] implies that the probability that $\|x(t)\|$ is greater than or equal to a given quantity decreases with an exponential decay. A detailed discussion on the differences among used stability notions for stochastic systems can be found in [38]. For completeness, in Appendix A we also show that under the hypotheses of Theorem 2.3, (2.1) is globally almost surely exponentially stable in the sense of [40].*

We associate to (2.1) the partially averaged system

$$\dot{x}(t) = \overline{f}(x(t), t) = \mathsf{E}[f(x(t), t, \Omega)]. \tag{2.4}$$

Equation (2.4) represents a deterministic nonautonomous nonlinear system. We notice that the origin is an equilibrium of the partially averaged system; that is, $\overline{f}(0, t) = 0$ for every $t \in \mathbb{R}^+$. If (2.4) is globally exponentially stable, by the converse theorem of Lyapunov (see [36, Theorem 3.12]) we know that there exists a function V that is bounded by quadratic forms of x and whose time derivative is strictly negative definite along the system trajectories; see, for example, (2.5) and (2.6) below. In the following theorem, we show that if V satisfies further regularity conditions and the switching period is sufficiently small, the original system (2.1) is globally mean square exponentially stable and globally almost surely exponentially stable.

Theorem 2.4. *Consider the system (2.1) and the associated partially averaged system (2.4), and suppose that there exists a function $V(x, t)$ which satisfies the following conditions:*

1. *There exist positive real numbers λ_{min} and λ_{max} such that for every $(x, t) \in \mathbb{R}^n \times \mathbb{R}^+$*

$$\lambda_{min}\|x\|^2 \leq V(x, t) \leq \lambda_{max}\|x\|^2. \tag{2.5}$$

2. *There exists $w > 0$ such that for every $(x, t) \in \mathbb{R}^n \times \mathbb{R}^+$*

$$\frac{\partial V}{\partial t}(x, t) + \frac{\partial V}{\partial x}(x, t)\overline{f}(x, t) \leq -w\|x\|^2. \tag{2.6}$$

3. *$\forall t \in \mathbb{R}^+$, $\frac{\partial V}{\partial x}(0, t) = 0$ and $\frac{\partial V}{\partial x}$ is globally Lipschitz with Lipschitz constant C_v. Moreover, for every $t \in \mathbb{R}^+$, $\frac{\partial^2 V}{\partial x \partial t}(0, t) = 0$ and $\frac{\partial^2 V}{\partial x \partial t}$ is globally Lipschitz with Lipschitz constant C_{vt}.*

There exists an $\varepsilon^* > 0$ such that $\forall \varepsilon < \varepsilon^*$ system (2.1) is globally mean square exponentially stable and globally almost surely exponentially stable. The function $V(x, t)$ is called a Lyapunov function.

Proof. The derivative of V along the solution of (2.1) is

$$\dot{V}(x(t), t) = \frac{\partial V}{\partial t}(x(t), t) + \frac{\partial V}{\partial x}(x(t), t) f(x(t), t, \Omega). \tag{2.7}$$

For every nonnegative integer k, we define

$$\Delta V(\sigma_{k+1},\sigma_k)=\mathsf{E}[V(x(\sigma_{k+1}),\sigma_{k+1})|x(\sigma_k)]-V(x(\sigma_k),\sigma_k). \tag{2.8}$$

From (2.1), (2.7), and (2.8) we have

$$\begin{aligned}\Delta V(\sigma_{k+1},\sigma_k)&=\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\dot{V}(x(t),t)dt\right]\\&=\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial t}(x(t),t)+\frac{\partial V}{\partial x}(x(t),t)f(x(t),t,\Omega)dt\right]\\&=\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial x}(x(t),t)f(x(t),t,\Omega)-\frac{\partial V}{\partial x}(x(\sigma_k),t)f(x(\sigma_k),t,\Omega)dt\right]\\&\quad+\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial t}(x(t),t)-\frac{\partial V}{\partial t}(x(\sigma_k),t)dt\right]\\&\quad+\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial t}(x(\sigma_k),t)+\frac{\partial V}{\partial x}(x(\sigma_k),t)f(x(\sigma_k),t,\Omega)dt\right].\end{aligned} \tag{2.9}$$

We seek an upper bound for the absolute values of the three terms in the summation above. We start our analysis by considering the first and second terms. Using the Lipschitz conditions on f_ω and on the first and second derivatives of V and following the argument of [53, parts II and III of the proof of Theorem 2] for each realization ω of Ω, we have

$$\begin{aligned}&\left|\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial x}(x(t),t)f(x(t),t,\omega)\right.\\&\qquad\left.-\frac{\partial V}{\partial x}(x(\sigma_k),t)f(x(\sigma_k),t,\omega)dt\right|\le 2L_{\omega,\varepsilon}^2C_ve^{2\varepsilon L_{\omega,\varepsilon}}\varepsilon^2\|x(\sigma_k)\|^2,\\&\left|\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial t}(x(t),t)-\frac{\partial V}{\partial t}(x(\sigma_k),t)dt\right|\le L_{\omega,\varepsilon}C_{vt}e^{2\varepsilon L_{\omega,\varepsilon}}\varepsilon^2\|x(\sigma_k)\|^2.\end{aligned}$$

Since $\sum_{i=1}^N P\{\Omega=\omega_i\}=1$ and $L_{\omega,\varepsilon}\le L$ for each ω and ε, the absolute value of the first term of the summation (2.9) is less than or equal to

$$2L^2C_ve^{2\varepsilon L}\varepsilon^2\|x(\sigma_k)\|^2.$$

In addition, the absolute value of the second term is less than or equal to

$$LC_{vt}e^{2\varepsilon L}\varepsilon^2\|x(\sigma_k)\|^2.$$

Now, we consider the third term on the right side of (2.9):

$$\begin{aligned}&\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial t}(x(\sigma_k),t)+\frac{\partial V}{\partial x}(x(\sigma_k),t)f(x(\sigma_k),t,\Omega)dt\right]\\&\qquad=\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial t}(x(\sigma_k),t)+\frac{\partial V}{\partial x}(x(\sigma_k),t)\overline{f}(x(\sigma_k),t)dt\right]\\&\qquad\quad+\mathsf{E}\left[\int_{\sigma_k}^{\sigma_{k+1}}\frac{\partial V}{\partial x}(x(\sigma_k),t)\{f(x(\sigma_k),t,\Omega)-\overline{f}(x(\sigma_k),t)\}dt\right].\end{aligned}$$

By hypothesis, (2.6) provides a bound for the integrand in the first term in the summation above, while the second term is equal to zero. Using the above computed bounds in (2.9), we find

$$\Delta V(\sigma_{k+1}, \sigma_k) \leq [g(\varepsilon) - w\varepsilon]\|x(\sigma_k)\|^2, \tag{2.10}$$

where the function $g(\varepsilon)$ is defined by

$$g(\varepsilon) = (2L^2 C_v e^{2\varepsilon L} + LC_{vt} e^{2\varepsilon L})\varepsilon^2. \tag{2.11}$$

Noticing that $g(0) = 0$ and $g'(0) = 0$, we have that there exists $\varepsilon^* > 0$ such that

$$\Delta V(\sigma_{k+1}, \sigma_k) \leq -\overline{w}\|x(\sigma_k)\|^2, \tag{2.12}$$

where $\overline{w} = [w\varepsilon - g(\varepsilon)] > 0$ for every $\varepsilon < \varepsilon^*$. In conclusion, if the switching period ε is sufficiently small, (2.5), (2.8), and (2.12) imply that the hypotheses of Theorem 2.3 are satisfied. Thus the claim follows. ■

Remark 2. *We assumed that ε is a fixed period of time. This hypothesis can be relaxed. In fact, Theorem 2.4 can be generalized considering not equally spaced switching instants. If the maximum time distance between two adjacent switching events is ε_{max}, (A.9) in Theorem 2.3 still holds with $\gamma = e^{L\varepsilon_{max}}$, while Theorem 2.4 holds with $\varepsilon_{max} \leq \varepsilon^*$.*

Remark 3. *Theorem 2.4 applies to a large class of candidate Lyapunov functions, including autonomous quadratic functions of the type $V(x,t) = x^T Px$, where P is a symmetric positive definite constant matrix. Autonomous quadratic candidate Lyapunov functions are considered in [4]. In this case, conditions 1 and 3 of Theorem 2.4 are automatically satisfied. In fact, (2.5) holds for $\lambda_{min} = \min\{\lambda(P)\}$ and $\lambda_{max} = \max\{\lambda(P)\}$, since P is a constant matrix (here, $\lambda(\bullet)$ indicates the spectrum of the matrix). Furthermore, condition 3 of Theorem 2.4 is satisfied with $C_v = 2\|P\|$ and $C_{vt} = 0$. The claims of Theorem 2.4 imply as a special case the global almost sure asymptotic stability results proved in [4] for the class of small-world networks considered therein.*

3. Master-slave synchronization.

3.1. Problem statement.

We consider the master system

$$\dot{x}(t) = Ax(t) + g(x(t)) + u(t), \tag{3.1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^n$ is the input vector, $A \in \mathbb{R}^{n\times n}$ is a constant matrix, g is a nonlinear function, n is a positive integer, and $t \in \mathbb{R}^+$ indicates the time variable. We construct a slave system for (3.1):

$$\dot{\widetilde{x}}(t) = A\widetilde{x}(t) + g(\widetilde{x}(t)) + u(t) + K(t)(x(t) - \widetilde{x}(t)). \tag{3.2}$$

System (3.2) is unidirectionally coupled to the master system (3.1) through the feedback gain matrix function $K(t)$. We consider the case where $K(t)$ is a piecewise constant signal that, in every time interval $[\sigma_k, \sigma_{k+1})$, with $\sigma_k = k\varepsilon$, $\varepsilon > 0$, and $k \in \mathbb{Z}^+$, equals the random variable $\mathfrak{K}_k$. We assume that the random variables $\mathfrak{K}_k$ are independent and identically distributed discrete

random variables that take values in the finite set $\{K_1, K_2, \ldots, K_N\}$, with N a positive integer. We refer to $\mathfrak{K}$ as the common random variable describing the full set of random variables $\{\mathfrak{K}_k\}_{k=0}^{\infty}$. Following [34], we assume that

$$g(x) - g(\widetilde{x}) = M_{x,\widetilde{x}}(x - \widetilde{x})$$

for some bounded matrix $M_{x,\widetilde{x}}$, whose elements depend on x and $\widetilde{x}$. As discussed in [34], this condition applies to a large variety of chaotic systems. We note that assuming the matrix $M_{x,\widetilde{x}}$ is bounded does not imply that the oscillators' states are bounded.

We express the system of equations (3.1) and (3.2) in terms of the error function $e = x - \widetilde{x}$:

$$\begin{aligned} \dot{e}(t) &= Ae(t) + g(x(t)) - g(\widetilde{x}(t)) - K(t)e(t) \\ &= (A - K(t))e(t) + M_{x(t),x(t)-e(t)}e(t). \end{aligned} \tag{3.3}$$

The stochastic nonautonomous nonlinear system in (3.3) can be written in the form (2.1), where f is defined by

$$f(e, t, \mathfrak{K}) = (A - \mathfrak{K} + M_{x(t),x(t)-e})e. \tag{3.4}$$

We say that the two oscillators in (3.1) and (3.2) globally mean square synchronize if the error system in (3.3) is globally mean square exponentially stable; see Definition 2.1. Similarly, we say that the two oscillators in (3.1) and (3.2) globally almost surely synchronize if the error system in (3.3) is globally almost surely exponentially stable; see Definition 2.2.

We note that, for $i = 1, \ldots, N$, the function $f_i = f(\bullet, \bullet, K_i)$ is globally Lipschitz in $\mathbb{R}^+$ with Lipschitz constant $L_i = \|A\| + m + \|K_i\|$, where $\|M\| \leq m$. In addition, the Lipschitz constants L_i are bounded by $L = \|A\| + m + \max_{1\leq i\leq N}\{\|K_i\|\}$. We further notice that $f(0, t, K_i) = 0\ \forall t \in \mathbb{R}^+$.

3.2. Global stochastic synchronization. In this section, we combine the general findings of section 2 on stochastic stability of nonlinear systems with available results on synchronizability of deterministic master-slave systems to provide sufficient conditions for global synchronization of the master-slave system described by (3.1) and (3.2) under fast-switching conditions. In particular, we make use of the results of [34], where a criterion for assessing global exponential stability of (3.3) is given in the case of constant feedback gain.

We associate to the system (3.3) the partially averaged system

$$\dot{e}(t) = (A + M_{x(t),x(t)-e(t)})e(t) - \overline{K}e(t), \tag{3.5}$$

where $\overline{K} = \mathsf{E}[K(t)] = \sum_{i=1}^{N} p_i K_i$ is the time-averaged constant feedback gain. Here, p_i indicates the probability of $K(t)$ assuming value K_i, that is, $p_i = P\{\mathfrak{K} = K_i\}$.

Global exponential stability of (3.5), that is, global exponential synchronization of the master-slave system under constant feedback coupling $\overline{K}$, can be enforced using the results of [34]. For clarity, we restate here the main theorem of [34] adapted to the present notation.

Theorem 3.1. *The system* (3.5) *is globally exponentially stable if the feedback gain matrix* $\overline{K}$ *is chosen such that*

$$\bar{l}_i(\xi, t) \leq -w < 0, \quad i = 1, 2, \ldots, n,$$

for every $\xi \in \mathbb{R}^n$ *and* $t \in \mathbb{R}^+$, *where the* $\bar{l}_i(\xi, t)$*'s are the eigenvalues of the matrix*

$$Q(\xi, t) = (A - \overline{K} + M_{x(t),x(t)-\xi})^T P + P(A - \overline{K} + M_{x(t),x(t)-\xi})$$

and P *is a positive definite symmetric constant matrix. A Lyapunov function for* (3.5) *can be constructed as*

$$V(e) = e^T P e \tag{3.6}$$

with

$$\dot{V}(e(t)) = e^T(t) Q(e(t), t) e(t) \leq -w\|e(t)\|^2. \tag{3.7}$$

The Lyapunov function (3.6) constructed for the partially averaged system can be used to assess the stability of the stochastic system. In fact, (3.7) is equivalent to (2.6), and conditions 1 and 3 of Theorem 2.4 are automatically satisfied as observed in Remark 3. Equation (2.12), specified for the case at hand, reads

$$2L^2 C_v e^{2L\varepsilon}\varepsilon^2 - w\varepsilon = 0, \tag{3.8}$$

and it yields the sought value of ε^*. By applying Theorem 2.4, we claim that the system (3.3) is globally mean square exponentially stable and globally almost surely exponentially stable $\forall \varepsilon < \varepsilon^*$. We summarize the above arguments in the following corollary.

Corollary 3.2. *Consider the system* (3.3) *and the corresponding partially averaged system* (3.5). *If the feedback gain matrix* $K(t)$ *is chosen such that*

$$\bar{l}_i(\xi, t) \leq -w < 0, \quad i = 1, 2, \ldots, n, \tag{3.9}$$

for every $\xi \in \mathbb{R}^n$ *and* $t \in \mathbb{R}^+$, *where the* $\bar{l}_i(\xi, t)$*'s are the eigenvalues of the matrix*

$$Q(\xi, t) = (A - \overline{K} + M_{x(t),x(t)-\xi})^T P + P(A - \overline{K} + M_{x(t),x(t)-\xi}) \tag{3.10}$$

and P *is a positive definite symmetric constant matrix, then there exists an* $\varepsilon^* > 0$ *such that* $\forall \varepsilon < \varepsilon^*$ *the system* (3.3) *is globally mean square exponentially stable and globally almost surely exponentially stable. The time duration* ε^* *is the nonzero solution of* (3.8).

Remark 4. *The conditions of Corollary* 3.2 *do not generally constrain the structure of the feedback gain matrices* $\{K_1, \ldots, K_N\}$. *In the special case where the feedback gains* $\{K_1, \ldots, K_N\}$ *and the probabilities* $\{p_1, \ldots, p_N\}$ *lead to a diagonal time-averaged feedback gain matrix* $\overline{K}$ *and the matrix* P *is diagonal, inequalities* (3.9) *may be directly enforced using Gerschgorin's theorem, as illustrated in* [34].

Remark 5. *The type of intermittent coupling considered in this paper has been also analyzed in the framework of consensus theory* [31, 55]. *However, in consensus theory, the individual systems' dynamics is linear, while in the present case the coupled systems are strongly nonlinear.*

4. Case study: Synchronization of two chaotic Chua circuits. As an example, we apply our results to synchronization of Chua circuits; see, for example, [62]. A Chua circuit system is described by

$$\begin{cases} \dot{x_1} &= a(x_2 - x_1 - h(x_1)), \\ \dot{x_2} &= x_1 - x_2 + x_3, \\ \dot{x_3} &= -bx_2, \end{cases} \tag{4.1}$$

where $a > 0$, $b > 0$, and the nonlinear function h has the form

$$h(x_1) = m_1 x_1 + \frac{1}{2}(m_0 - m_1)\{|x_1 + 1| - |x_1 - 1|\} \tag{4.2}$$

with $m_0 < 0$ and $m_1 < 0$. We define

$$h(x_1) - h(\widetilde{x}_1) = w_{x_1,\widetilde{x}_1}(x_1 - \widetilde{x}_1), \tag{4.3}$$

where $w_{x_1,\widetilde{x}_1}$ depends on x_1 and $\widetilde{x}_1$ and is bounded by $m_0 \leq w_{x_1,\widetilde{x}_1} \leq m_1$; see, for example, [34].

We consider the case where $K(t)$ is a diagonal matrix. Following (3.2), the slave system of (4.1) is constructed as follows:

$$\begin{cases} \dot{\widetilde{x}}_1 &= a(\widetilde{x}_2 - \widetilde{x}_1 - h(\widetilde{x}_1)) + k_1(t)(x_1 - \widetilde{x}_1), \\ \dot{\widetilde{x}}_2 &= \widetilde{x}_1 - \widetilde{x}_2 + \widetilde{x}_3 + k_2(t)(x_2 - \widetilde{x}_2), \\ \dot{\widetilde{x}}_3 &= -bx_2 + k_3(t)(x_3 - \widetilde{x}_3). \end{cases} \tag{4.4}$$

Combining (4.1) and (4.4), we obtain (3.3) with

$$A = \begin{bmatrix} -a & a & 0 \\ 1 & -1 & 1 \\ 0 & -b & 0 \end{bmatrix}, \quad K(t) = \begin{bmatrix} k_1(t) & 0 & 0 \\ 0 & k_2(t) & 0 \\ 0 & 0 & k_3(t) \end{bmatrix}, \quad g(x) = \begin{bmatrix} -ah(x) \\ 0 \\ 0 \end{bmatrix}.$$

We observe that $g(x) - g(\widetilde{x}) = M_{x,x-e}e$ with

$$M_{x,x-e} = \begin{bmatrix} -aw_{x_1,x_1-e_1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{4.5}$$

and $\|M\| \leq a|m_0|$.

We associate to the system (3.3) the partially averaged system

$$\dot{e} = Ae + M_{x,x-e}e - \overline{K}e, \tag{4.6}$$

where

$$\overline{K} = \begin{bmatrix} \overline{k}_1 & 0 & 0 \\ 0 & \overline{k}_2 & 0 \\ 0 & 0 & \overline{k}_3 \end{bmatrix}. \tag{4.7}$$

By choosing $P = I$ and by setting

$$\begin{aligned} \overline{k}_1 &\geq \frac{1}{2}(1 - a - 2am_0 + w), \\ \overline{k}_2 &\geq \frac{1}{2}(a - 1 + |1 - b| + w), \\ \overline{k}_3 &\geq \frac{1}{2}(|1 - b| + w), \end{aligned} \tag{4.8}$$

the partially averaged system is globally exponentially stable [34]. This follows directly from Gerschgorin's theorem, as anticipated in Remark 4. We also have $C_v = 2$, and $L = \|A\| + a|m_0| + \max_{1\leq i\leq N}\{\|K_i\|\}$. Equation (3.8) gives the value of ε^* that ensures the global mean square exponential stability and the global almost sure exponential stability of the stochastic system $\forall \varepsilon < \varepsilon^*$.

Here, we present a few numerical results that illustrate how two stochastically coupled Chua circuits synchronize for a sufficiently fast switching rate. We select $a = 9.78$, $b = 14.97$, $m_0 = -1.31$, and $m_1 = -0.75$ in order to have chaotic behavior of the system [34]. We let $K(t)$ switch randomly between the two constant matrices K_1 and K_2, where K_1 is the zero matrix and

$$K_2 = \begin{bmatrix} 20 & 0 & 0 \\ 0 & 27.5 & 0 \\ 0 & 0 & 20 \end{bmatrix}. \tag{4.9}$$

For these parameters we have $\|A\| = 18.89$, $\max_{1\leq i\leq N}\{\|K_i\|\} = \|K_2\| = 27.5$, and $L = 59.20$. Selecting $p_1 = 0.6$ and $p_2 = 0.4$, w can be chosen from (4.8) to be equal to 0.5. From (3.8) we have that for $\varepsilon \leq \varepsilon^* = 3.56 \times 10^{-5}$ the system synchronizes globally mean square exponentially and globally almost surely exponentially. Figure 1 depicts the trajectories of the master and slave systems on the x_1-x_2 and x_1-x_3 planes for $\varepsilon = 10^{-5}$. This figure shows that the two systems synchronize even if the initial conditions are significantly different.

5. Conclusions. In this paper, we presented a general criterion for global synchronization of randomly coupled chaotic oscillators. We focus on the case of two oscillators in master-slave configuration. The two systems are coupled through a stochastic unidirectional feedback that is realized through a switching function that switches randomly among a finite set of constant values. Using tools based on Lyapunov stability and partial averaging we showed that, under suitable regularity conditions, the synchronization characteristics of the partially averaged system are inherited by the stochastic system. Our findings are illustrated through numerical simulations on Chua circuits. The proposed approach can be applied to global synchronization of complex networks and seems particularly promising for analyzing pinning-controllability (see, for example, [63]) with intermittent control.

The main claim on stochastic stability of this work, Theorem 2.4, applies to a large class of stochastic nonautonomous nonlinear systems and Lyapunov functions. In particular, it can be applied to autonomous quadratic Lyapunov functions, such as those studied in [4]. In addition, our main claim yields strong conditions on global exponential stability that include, as a special case, the asymptotic stability results derived in [4].

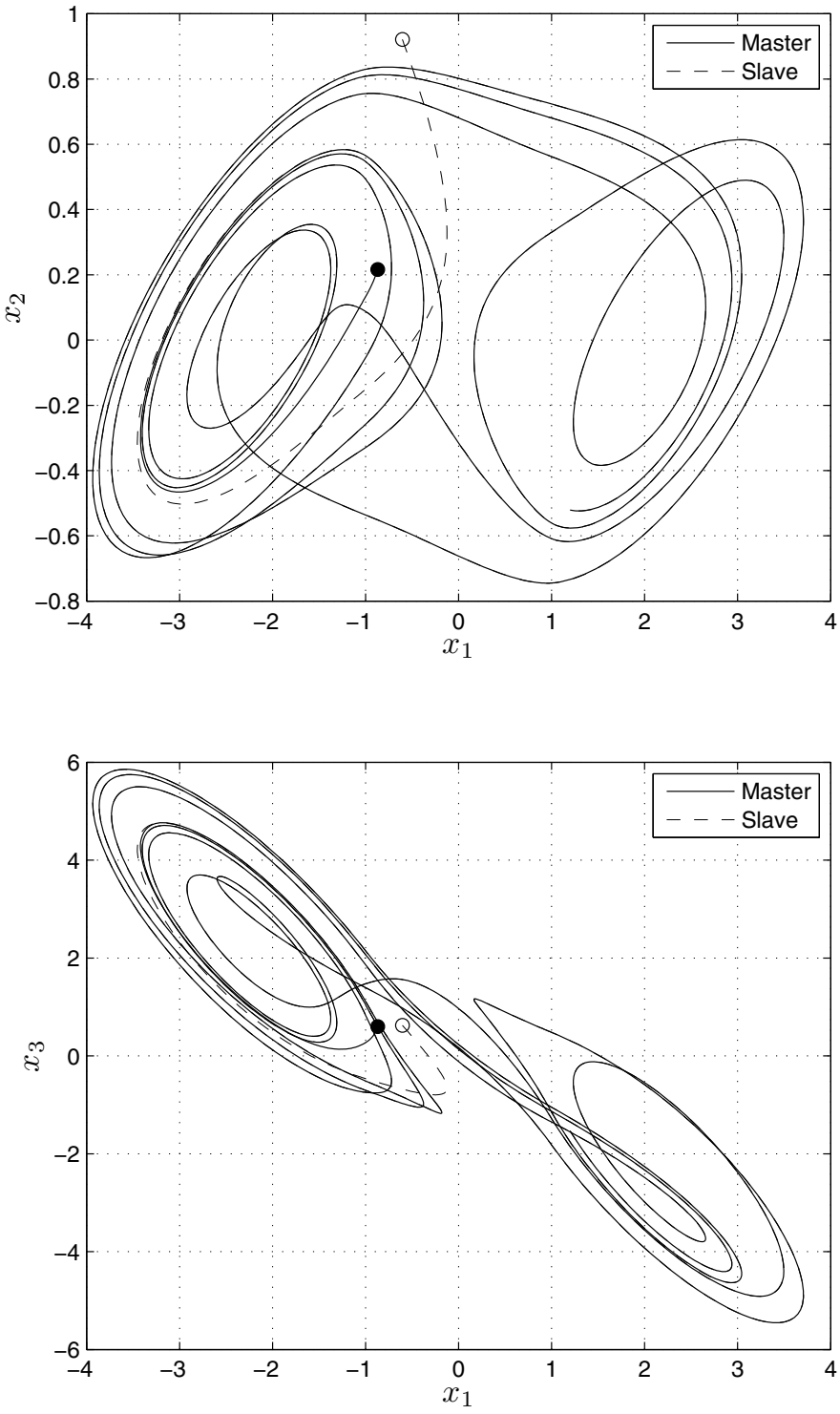


Figure 1. *Trajectories of the master and slave systems in the x_1-x_2 and x_1-x_3 planes.*

Appendix A. Proof of Theorem 2.3.

Proof. We start by proving the global mean square exponential stability. Consider arbitrary initial time $t_0 \in \mathbb{R}^+$ and initial condition $x_0 \in \mathbb{R}^n$. We define the index $\hat{k}$ so that $t_0 \in [\sigma_{\hat{k}-1}, \sigma_{\hat{k}})$. Specifying (2.3) at the $\hat{k}$th and $(\hat{k}+1)$th switching instants, we have

$$\mathsf{E}[V(x(\sigma_{\hat{k}+1}), \sigma_{\hat{k}+1})|x(\sigma_{\hat{k}})] \leq (1-\nu)V(x(\sigma_{\hat{k}}), \sigma_{\hat{k}}), \tag{A.1}$$

$$\mathsf{E}[V(x(\sigma_{\hat{k}+2}), \sigma_{\hat{k}+2})|x(\sigma_{\hat{k}+1})] \leq (1-\nu)V(x(\sigma_{\hat{k}+1}), \sigma_{\hat{k}+1}). \tag{A.2}$$

By taking the conditional expected value of (A.2) we obtain

$$\mathsf{E}[\mathsf{E}[V(x(\sigma_{\hat{k}+2}), \sigma_{\hat{k}+2})|x(\sigma_{\hat{k}+1})]|x(\sigma_{\hat{k}})] \leq (1-\nu)\mathsf{E}[V(x(\sigma_{\hat{k}+1}), \sigma_{\hat{k}+1})|x(\sigma_{\hat{k}})]. \tag{A.3}$$

Using the smoothing lemma (see, for example, Lemma 1.1 on page 474 in [30]) and inequality (A.1) in (A.3), we find

$$\mathsf{E}[V(x(\sigma_{\hat{k}+2}), \sigma_{\hat{k}+2})|x(\sigma_{\hat{k}})] \leq (1-\nu)^2 V(x(\sigma_{\hat{k}}), \sigma_{\hat{k}}).$$

We explicitly note that the hypotheses of the smoothing lemma are verified since (2.1) defines a Markov process.

Iterating the argument above for any positive integer $n > \hat{k}$, we obtain

$$\mathsf{E}[V(x(\sigma_n), \sigma_n)|x(\sigma_{\hat{k}})] \leq (1-\nu)^{n-\hat{k}} V(x(\sigma_{\hat{k}}), \sigma_{\hat{k}}). \tag{A.4}$$

By using the bounds in (2.2), (A.4) gives

$$\mathsf{E}[\|x(\sigma_n)\|^2|x(\sigma_{\hat{k}})] \leq \lambda_{max}/\lambda_{min}(1-\nu)^{n-\hat{k}}\|x(\sigma_{\hat{k}})\|^2. \tag{A.5}$$

Inequality (A.5) can be used to derive an upper bound for the unconditioned expected value that is needed to assess the global mean square exponential stability according to Definition 2.1. Since $\hat{k}$ defines a given instant of time and x_0 is a prescribed initial condition, $x(\sigma_{\hat{k}})$ is a finite-state random variable taking values in $\{x_1(\sigma_{\hat{k}}), \ldots, x_N(\sigma_{\hat{k}})\}$, where N is the cardinality of the event set Θ. From the definition of conditional expectation (see, for example, [9]), we have

$$\mathsf{E}[\|x(\sigma_n)\|^2] = \sum_{i=1}^{N} \mathsf{E}[\|x(\sigma_n)\|^2|x_i(\sigma_{\hat{k}})]P\{x_i(\sigma_{\hat{k}})\}, \tag{A.6}$$

where $P\{x_i(\sigma_{\hat{k}})\}$ is the probability that $x_i(\sigma_{\hat{k}})$ is the realization of the random variable $x(\sigma_{\hat{k}})$. Hence, using inequality (A.5), equation (A.6) yields

$$\mathsf{E}[\|x(\sigma_n)\|^2] \leq \sum_{i=1}^{N} \frac{\lambda_{max}}{\lambda_{min}}(1-\nu)^{n-\hat{k}}\|x_i(\sigma_{\hat{k}})\|^2 P\{x_i(\sigma_{\hat{k}})\}. \tag{A.7}$$

In order to assess the global mean square exponential stability, we need to analyze the system dynamics inside every switching interval. Given a generic switching interval $[\sigma_k, \sigma_{k+1})$

and an instant $\bar{t} \in [\sigma_k, \sigma_{k+1})$, using the triangle inequality $\forall t \geq \bar{t}$ in $[\sigma_k, \sigma_{k+1})$, equation (2.1) yields

$$\|x(t)\| \leq \|x(\bar{t})\| + \int_{\bar{t}}^{t} \|f(x(\xi), \xi, \Omega)\| d\xi. \tag{A.8}$$

Since the functions f_ω are globally Lipschitz in $\mathbb{R}^+$ and all the corresponding Lipschitz constants are bounded by a constant L, (A.8) yields

$$\|x(t)\| \leq \|x(\bar{t})\| + \int_{\bar{t}}^{t} L\|x(\xi)\| d\xi.$$

Using the Gronwall–Bellman inequality (see, for example, [36]), we have

$$\|x(t)\| \leq \gamma \|x(\bar{t})\| \tag{A.9}$$

with $\gamma = e^{L\varepsilon}$. Therefore, using (A.9) in (A.7), we find that $\forall t \in [\sigma_n, \sigma_{n+1})$

$$\mathsf{E}[\|x(t)\|^2] \leq \gamma^2 \mathsf{E}[\|x(\sigma_n)\|^2] \leq \gamma^2 \sum_{i=1}^{N} \frac{\lambda_{max}}{\lambda_{min}} (1-\nu)^{n-\hat{k}} \|x_i(\sigma_{\hat{k}})\|^2 P\{x_i(\sigma_{\hat{k}})\}. \tag{A.10}$$

Inequality (A.9) can also be used to find an upper bound for $\|x(\sigma_{\hat{k}})\|$ in terms of the initial conditions. In fact, since $t_0 \in [\sigma_{\hat{k}-1}, \sigma_{\hat{k}})$ for the definition of $\hat{k}$, from (A.9) we obtain

$$\|x(\sigma_{\hat{k}})\| \leq \gamma \|x(t_0)\|. \tag{A.11}$$

Finally, using (A.11) to bound the right side of (A.10), we obtain

$$\begin{aligned} \mathsf{E}[\|x(t)\|^2] &\leq \gamma^4 \sum_{i=1}^{N} \frac{\lambda_{max}}{\lambda_{min}} (1-\nu)^{n-\hat{k}} \|x(t_0)\|^2 P\{x_i(\sigma_{\hat{k}})\} \\ &\leq \gamma^4 \frac{\lambda_{max}}{\lambda_{min}} (1-\nu)^{n-\hat{k}} \|x(t_0)\|^2 \\ &\leq \alpha \|x(t_0)\|^2 e^{-\beta(t-t_0)}, \end{aligned} \tag{A.12}$$

where we defined

$$\alpha = \gamma^4 (1-\nu)^{-2} \lambda_{max}/\lambda_{min}, \quad \beta = -\ln(1-\nu)/\varepsilon. \tag{A.13}$$

Therefore, according to Definition 2.1, the system (2.1) is globally mean square exponentially stable.

In the second part of the proof, we establish the global almost sure exponential stability. Following [44], we let χ be a positive real number such that $\chi \leq \beta$, where β is defined in (A.13). From the Markov inequality (see, for example, [30, Theorem 1.1, p. 120]), we have

$$P\left\{ \sup_{t \in [\sigma_{n-1}, \sigma_n)} \|x(t)\|^2 > e^{-(\beta-\chi)(\sigma_{n-1}-t_0)} \right\} \leq e^{(\beta-\chi)(\sigma_{n-1}-t_0)} \mathsf{E}\left[\sup_{t \in [\sigma_{n-1}, \sigma_n)} \|x(t)\|^2 \right]. \tag{A.14}$$

Using (A.12) in (A.14), we find

$$P\left\{\sup_{t\in[\sigma_{n-1},\sigma_n)}\|x(t)\|^2 > e^{-(\beta-\chi)(\sigma_{n-1}-t_0)}\right\} \le \alpha\|x(t_0)\|^2 e^{-\chi(\sigma_{n-1}-t_0)}. \tag{A.15}$$

By taking the infinite summation with respect to n of both sides of (A.15) and by noticing that $e^{-\chi\varepsilon}$ is less than one, we find

$$\sum_{n=\hat{k}}^{\infty} P\left\{\sup_{t\in[\sigma_{n-1},\sigma_n)}\|x(t)\|^2 > e^{-(\beta-\chi)(\sigma_{n-1}-t_0)}\right\} < \infty. \tag{A.16}$$

Thus, by directly applying the Borel–Cantelli lemma (see, for example, [41, Lemma 1, p. 192]), we obtain that

$$\sup_{\sigma_n>k\ge\sigma_{n-1}}\|x(t)\| \le e^{-1/2(\beta-\chi)(\sigma_{n-1}-t_0)} \tag{A.17}$$

holds a.s. for all but finitely many n. Hence, for any sample system, there exists a positive integer n_0 such that (A.17) holds a.s. for any $n > n_0$. Therefore, for any $n > n_0$ and $t \in [\sigma_{n-1}, \sigma_n)$, the following inequality holds a.s.:

$$\|x(t)\| \le e^{-1/2(\beta-\chi)(\sigma_{n-1}-t_0)} \le e^{1/2(\beta-\chi)\varepsilon} e^{-1/2(\beta-\chi)(t-t_0)}. \tag{A.18}$$

By repetitively applying the Gronwall–Bellman inequality in (A.9) and by using (A.18), we finally find that $\forall t \ge t_0$ the following inequality holds a.s.:

$$\|x(t)\| \le \gamma^{n_0-\widehat{k}+1} e^{1/2(\beta-\chi)\varepsilon}\|x(t_0)\| e^{-1/2(\beta-\chi)(t-t_0)}, \tag{A.19}$$

which proves the global almost sure exponential stability of (2.1) according to Definition 2.2.

For completeness, we also show that (2.1) is globally almost surely exponentially stable in the sense of [40]. We notice that, since (2.3) holds and $V(x(\sigma_k), \sigma_k)$ is a positive quantity, the sequence of $V(x(\sigma_k), \sigma_k)$ is a supermartingale; see, for example, Definition 2.4 in [30]. Therefore, we can apply the supermartingale inequality (see, for example, [41, Proposition 1, p. 31]) and obtain that for every $\eta > 0$

$$P\left\{\sup_{\infty>k\ge n} V(x(\sigma_k), \sigma_k) \ge \eta\right\} \le \frac{\mathsf{E}[V(x(\sigma_n), \sigma_n)|x(\sigma_{\hat{k}})]}{\eta}. \tag{A.20}$$

Substituting (A.4) into (A.20) and using condition (2.5), we find

$$P\left\{\sup_{\infty>k\ge n}\|x(\sigma_k)\|^2 \ge \frac{\eta}{\lambda_{min}}\right\} \le \frac{\lambda_{max}}{\eta}(1-\nu)^{n-\hat{k}}\|x(\sigma_{\hat{k}})\|^2. \tag{A.21}$$

Using (A.9) in (A.21), we obtain the following inequality for the continuous time process $x(t)$:

$$P\left\{\sup_{\infty>(t-t_0)\ge T}\|x(t)\|^2 \ge \frac{\eta}{\lambda_{min}}\right\} \le \frac{\gamma^2\lambda_{max}}{\eta}(1-\nu)^{n-\hat{k}}\|x(t_0)\|^2, \tag{A.22}$$

where T is an arbitrary time duration and the index n satisfies $n = \hat{k} + \lfloor T/\varepsilon \rfloor$, where $\lfloor \bullet \rfloor$ refers to the integer division. By defining the positive quantity $\mu = \gamma\sqrt{\eta/\lambda_{min}}$ and expressing the left-hand side of (A.22) in terms of the supremum of $\|x(t)\|$, we obtain

$$P\left\{\sup_{\infty>(t-t_0)\geq T}\|x(t)\| \geq \mu\right\} \leq \frac{\lambda_{max}\gamma^4}{\lambda_{min}\mu^2}(1-\nu)^{n-\hat{k}}\|x(t_0)\|^2. \tag{A.23}$$

Replacing the constants α and β defined in (A.13) into (A.23), we finally derive the following bound on the exponential rate of decay of the probability that $\|x(t)\|$ is larger than μ:

$$P\left\{\sup_{\infty>(t-t_0)\geq T}\|x(t)\| \geq \mu\right\} \leq \frac{\alpha}{\mu^2}\|x(t_0)\|^2 e^{-\beta T}. \tag{A.24}$$ ■

Acknowledgments. The authors would like to thank Ms. Francesca Fiorilli and Dr. Davide Spinello for their careful review of the manuscript.

REFERENCES

[1] D. Aeyels and J. Peuteman, *On exponential stability of nonlinear time-varying differential equations*, Automatica J. IFAC, 35 (1999), pp. 1091–1100.
[2] B. R. Andrievskii and A. L. Fradkov, *Control of chaos: Methods and applications*, Autom. Remote Control, 64 (2003), pp. 673–713.
[3] L. Angelini, M. De Tommaso, M. Guido, K. Hu, P. Ch. Ivanov, D. Marinazzo, G. Nardulli, L. Nitti, M. Pellicoro, C. Pierro, and S. Stramaglia, *Steady-state visual evoked potentials and phase synchronization in migraine patients*, Phys. Rev. Lett., 93 (2004), 038103.
[4] I. V. Belykh, V. N. Belykh, and M. Hasler, *Blinking model and synchronization in small-world networks with a time-varying coupling*, Phys. D, 195 (2004), pp. 188–206.
[5] I. V. Belykh, V. N. Belykh, and M. Hasler, *Connection graph stability method for synchronized coupled chaotic systems*, Phys. D, 195 (2004), pp. 159–187.
[6] B. Blasius and L. Stone, *Chaos and phase synchronization in ecological systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 10 (2000), pp. 2361–2380.
[7] S. Boccaletti, J. Kurths, G. Osipov, D. L. Valladares, and C. S. Zhou, *The synchronization of chaotic systems*, Phys. Rep., 366 (2002), pp. 1–101.
[8] S. Bowong and F. M. M. Kakmeni, *Synchronization of uncertain chaotic system via backstepping approach*, Chaos Solitons Fractals, 21 (2004), pp. 999–1011.
[9] P. Bremaud, *Markov Chains, Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer-Verlag, New York, 1999.
[10] J. Buck and E. Buck, *Synchronous fireflies*, Scientific American, 234 (1976), p. 74.
[11] S. Camazine, W. Ristine, and M. E. Didion, *Self-Organization in Biological Systems*, Princeton University Press, Princeton, NJ, 2003.
[12] T. L. Carroll and L. M. Pecora, *Synchronizing chaotic circuits*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 38 (1991), pp. 453–456.
[13] G. Chen and X. Yu, eds., *Chaos Control Theory and Applications*, Lecture Notes in Control and Inform. Sci. 292, Springer-Verlag, Berlin, 2003.
[14] J. J. Collins and I. Stewart, *Coupled nonlinear oscillators and the symmetries of animal gaits*, J. Nonlinear Sci., 3 (1993), pp. 349–392.
[15] O. L. V. Costa, M. D. Fragoso, and R. P. Marques, *Discrete-Time Markov Jump Linear Systems*, Springer-Verlag, London, 2005.
[16] K. M. Cuomo, V. A. Oppenheim, and S. H. Strogatz, *Synchronization of Lorentz-based chaotic circuits with application to communications*, IEEE Trans. Circuits Systems II, 40 (1993), pp. 626–633.

[17] C. E. de Souza and D. F. Coutinho, *Robust stability of a class of uncertain Markov jump*, IEEE Trans. Automat. Control, 51 (2006), pp. 1825–1831.
[18] G. S. Duane, P. J. Webster, and J. B. Weiss, *Co-occurrence of northern and southern hemisphere blocks as partially synchronized chaos*, J. Atmospheric Sci., 56 (1999), pp. 4183–4205.
[19] M. Feki, *An adaptive chaos synchronization scheme applied to secure communication*, Chaos Solitons Fractals, 18 (2003), pp. 141–148.
[20] M. Feki, B. Robert, G. Gelle, and M. Colas, *Secure digital communication using discrete-time chaos synchronization*, Chaos Solitons Fractals, 18 (2003), pp. 881–890.
[21] X. Feng, K. A. Loparo, Y. Ji, and H. J. Chizeck, *Stochastic stability properties of jump linear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 38–53.
[22] L. Fortuna, M. Frasca, and A. Rizzo, *Experimental pulse synchronization of two chaotic circuits*, Chaos Solitons Fractals, 17 (2003), pp. 335–361.
[23] A. Garfinkel, M. L. Spano, W. L. Ditto, and J. N. Weiss, *Controlling cardiac chaos*, Science, 257 (1992), pp. 1230–1235.
[24] Z. Ge and Y. Chen, *Synchronization of mutual coupled chaotic systems via partial stability theory*, Chaos Solitons Fractals, 34 (2007), pp. 787–794.
[25] L. Glass and M. C. Mackey, *From Clocks to Chaos: The Rhythms of Life*, Princeton University Press, Princeton, NJ, 1988.
[26] A. Goldbeter, *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour*, Cambridge University Press, Cambridge, UK, 1996.
[27] G. Grammel and I. Maizurna, *Exponential stability and partial averaging*, J. Math. Anal. Appl., 283 (2003), pp. 276–286.
[28] G. Grassi and S. Mascolo, *Nonlinear observer design to synchronize hyperchaotic systems via a scalar signal*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 44 (1997), pp. 1011–1014.
[29] M. R. Guevara, A. Shirier, and L. Glass, *Phase-locked rhythms in periodically stimulated heart cell aggregates*, Amer. J. Physiology, 254 (1988), pp. H1–H10.
[30] A. Gut, *Probability: A Graduate Course*, Springer-Verlag, New York, 2005.
[31] Y. Hatano and M. Mesbahi, *Agreement over random networks*, IEEE Trans. Automat. Control, 50 (2005), pp. 1867–1872.
[32] S. Hayes, C. Grebogi, and E Ott, *Communicating with chaos*, Phys. Rev. Lett., 70 (1993), pp. 3031–3034.
[33] M. Itoh, T. Yang, and L. O. Chua, *Conditions for impulsive synchronization of chaotic and hyperchaotic systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 551–560.
[34] G. P. Jiang, W. K. S. Tang, and G. Chen, *A simple global synchronization criterion for coupled chaotic system*, Chaos Solitons Fractals, 15 (2003), pp. 925–935.
[35] B. Jovic, C. P. Unsworth, G. S. Sandhu, and S. M. Berber, *A robust sequence synchronization unit for multi-user DS-CDMA chaos-based communication systems*, Signal Processing, 87 (2007), pp. 1692–1708.
[36] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 1995.
[37] D. Kim, *A spiking neuron model for synchronous flashing of fireflies*, Bio Systems, 76 (2004), pp. 7–20.
[38] F. Kozin, *A survey of stability of stochastic systems*, Automatica J. IFAC, 5 (1969), pp. 95–112.
[39] Y Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, Berlin, 1984.
[40] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.
[41] H. J. Kushner, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, New York, 1971.
[42] R. P. Leland, *Stability of asynchronous systems with Poisson transitions*, IEEE Trans. Automat. Control, 39 (1994), pp. 182–185.
[43] A. I. Lerescu, N. Contandache, S. Oancea, and I. Grosu, *Collection of master-slave synchronized chaotic systems*, Chaos Solitons Fractals, 22 (2004), pp. 599–604.
[44] X. Mao, *Exponential Stability of Stochastic Differential Equations*, Marcel Dekker, New York, 1994.
[45] A. N. Miliou, I. P. Antoniades, S. G. Stavrinides, and A. N. Anagnostopulos, *Secure communication by chaotic synchronization: Robustness under noisy conditions*, Nonlinear Anal. Real World Appl., 8 (2007), pp. 1003–1012.
[46] R. E. Mirollo and S. H. Strogatz, *Synchronization of pulsed-coupled biological oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 1645–1662.

[47] E. Mosekilde, Y. Maistrenko, and D. Postnov, *Chaotic Synchronization: Applications to Living Systems*, World Scientific, River Edge, NJ, 2002.
[48] T. I. Netoff and S. J. Schiff, *Decreased neuronal synchronization during experimental seizures*, J. Neurosci., 22 (2002), pp. 7297–7307.
[49] J. H. Park, *Chaos synchronization of a chaotic system via nonlinear control*, Chaos Solitons Fractals, 25 (2005), pp. 579–584.
[50] L. M. Pecora and T. L. Carroll, *Synchronization in chaotic system*, Phys. Rev. Lett., 64 (1990), pp. 821–824.
[51] L. M. Pecora and T. L. Carroll, *Master stability functions for synchronized coupled systems*, Phys. Rev. Lett., 80 (1998), pp. 2109–2112.
[52] L. M. Pecora, T. L. Carroll, G. A. Johnson, and D. J. Mar, *Fundamentals of synchronization in chaotic systems, concepts, and applications*, Chaos, 7 (1997), pp. 520–543.
[53] J. Peuteman and D. Aeyels, *Exponential stability of nonlinear time-varying differential equations and partial averaging*, Math. Control Signals Systems, 15 (2002), pp. 42–70.
[54] A. Pikovsky, M. Rosemblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge University Press, Cambridge, UK, 2001.
[55] M. Porfiri and D. J. Stilwell, *Consensus seeking over random weighted directed graphs*, IEEE Trans. Automat. Control, 52 (2007), pp. 1767–1773.
[56] M. Porfiri, D. J. Stilwell, E. M. Bollt, and J. D. Skufca, *Random talk: Random walk and synchronizability in a moving neighborhood network*, Phys. D, 224 (2006), pp. 102–113.
[57] M. L. V. Quyen, J. Martinerie, C. Adam, and F. J. Varela, *Nonlinear analyses of interictal eeg map the brain interdependences in human focal epilepsy*, Phys. D, 127 (1999), pp. 250–266.
[58] P. E. Rapp, T. R. Bashore, J. M. Martinerie, A. M. Albano, I. D. Zimmerman, and A. I. Mees, *Dynamics of brain electrical activity*, Brain Topography, 2 (1989), pp. 99–118.
[59] R. Roy and K. S. Thornburg, *Experimental synchronization of chaotic lasers*, Phys. Rev. Lett., 72 (1994), pp. 2009–2012.
[60] N. F. Rulkov, *Images of synchronized chaos: Experiments with circuits*, Chaos, 6 (1996), pp. 262–279.
[61] S. U. Sarnobat, S. Rajput, D. D. Bruns, D. W. DePaoli, C. S. Daw, and K. Nguyen, *The impact of external electrostatic fields on gas-liquid bubbling dynamics*, Chemical Engineering Science, 59 (2004), pp. 247–258.
[62] L. P. Shil'nikov, *Chua's circuit: Rigorous results and future problems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1993), pp. 489–519.
[63] F. Sorrentino, M. di Bernardo, F. Garofalo, and G. Chen, *Controllability of complex networks via pinning*, Phys. Rev. E, 75 (2007), 046103.
[64] D. J. Stilwell, E. M. Bollt, and D. G. Roberson, *Sufficient conditions for fast switching synchronization in time-varying network topologies*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 140–156.
[65] J. Sun, Y. Zhang, F. Qiao, and Q. Wu, *Some impulsive synchronization criterions for coupled chaotic systems via unidirectional linear error feedback approach*, Chaos Solitons Fractals, 19 (2004), pp. 1049–1055.
[66] A. Tufaile and J. C. Sartorelli, *The circle map dynamics in air bubble formation*, Phys. Lett. A, 287 (2001), pp. 74–80.
[67] G. D. Van Wiggeren and R. Roy, *Communication with chaotic lasers*, Science, 279 (1998), pp. 1198–1200.
[68] T. Womelsdorf and P. Fries, *The role of neuronal synchronization in selective attention*, Current Opinion in Neurobiology, 17 (2007), pp. 1–7.
[69] E. Yaz, *On the almost sure and mean-square exponential convergence of somestochastic observers*, IEEE Trans. Automat. Control, 35 (1990), pp. 935–936.
[70] E. Yaz, *Robustness of discrete-time systems for unstructured stochastic perturbations*, IEEE Trans. Automat. Control, 36 (1991), pp. 867–869.
[71] G. Zhang, Z. Liu, and Z. Ma, *Generalized synchronization of different dimensional chaotic dynamical system*, Chaos Solitons Fractals, 32 (2007), pp. 773–779.
[72] R. Zhang, Z. Xu, S. X. Yang, and X. He, *Generalized synchronization via impulsive control*, Chaos Solitons Fractals, 38 (2008), pp. 97–105.
[73] M. Zochowski, *Intermittent dynamical control*, Phys. D, 145 (2000), pp. 181–190.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 843–867
© 2008 Society for Industrial and Applied Mathematics

A Computer-Assisted Proof of Σ_3-Chaos in the Forced Damped Pendulum Equation*

Balázs Bánhelyi†, Tibor Csendes†, Barnabas M. Garay‡, and László Hatvani§

Abstract. The present paper is devoted to studying Hubbard's pendulum equation

$$\ddot{x} + 10^{-1}\dot{x} + \sin(x) = \cos(t).$$

Using rigorous/interval methods of computation, the main assertion of Hubbard on chaos properties of the induced dynamics is raised from the level of experimentally observed facts to the level of a theorem completely proved. A special family of solutions is shown to be chaotic in the sense that, on consecutive time intervals $(2k\pi, 2(k+1)\pi)$ $(k \in \mathbb{Z})$, individual members of the family can freely "choose" between the following possibilities: the pendulum crosses the bottom position exactly once clockwise or does not cross the bottom position at all or crosses the bottom position exactly once counterclockwise. The proof follows the topological index/degree approach by Mischaikow, Mrozek, and Zgliczynski. The new feature of this paper is a definition of the transition graph for which the periodic orbit lemma—the key technical result of the approach mentioned above—turns out to be a consequence of Brouwer's fixed point theorem. The role of wholly automatic versus "trial-and-error with human overheads" computer procedures in detecting chaos is also discussed.

Key words. forced damped pendulum, Σ_3-chaos, computer-assisted proof, transition graph, interval arithmetic

AMS subject classifications. 34C28, 37D45, 70K40, 70K55, 65G30

DOI. 10.1137/070695599

1. Introduction and the main results.

The complexity of the solutions to the forced damped pendulum equation

$$m\ell\ddot{x} + b\dot{x} + mg\sin(x) = A\cos(\omega t)$$

and of related systems is one of the most frequently studied problems in dynamics. For certain values of the parameters, small perturbation theory can be applied to prove chaotic behavior.

However, a purely theoretical approach can hardly lead to a proof for chaos if small perturbation methods break down such as in the case where

$$\ddot{x} + 10^{-1}\dot{x} + \sin(x) = \cos(t) \tag{1.1}$$

*Received by the editors June 26, 2007; accepted for publication (in revised form) by T. Sauer March 21, 2008; published electronically July 25, 2008. This work was supported by the Hungarian National Science Foundation grants OTKA T 048377, T 046822, T037491, and T049516.
http://www.siam.org/journals/siads/7-3/69559.html
†Institute of Informatics, University of Szeged, H-6701 Szeged P.O. Box 652, Hungary (banhelyi@inf.u-szeged.hu, csendes@inf.u-szeged.hu). The first author is grateful for the Ferenc Deák Scholarship DFÖ 19/2007.
‡Department of Mathematics, Budapest University of Technology, H-1521 Budapest, Hungary, and Computer and Automation Institute (SZTAKI), Hungarian Academy of Sciences, H-1111 Budapest, Hungary (garay@math.bme.hu).
§Bolyai Institute of Mathematics, University of Szeged, H-6701 Szeged P.O. Box 428, Hungary, and Analysis and Stochastics Research Group of the Hungarian Academy of Sciences (hatvani@math.u-szeged.hu).

(i.e., for parameters $m\ell = mg = A = \omega = 1$ and $b = 10^{-1}$) investigated by Hubbard [24]. Based on numerical experiments and the accompanying abstract considerations mimicking Smale's geometric horseshoe construction, Hubbard [24] made the existence of Σ_3-chaos—both on Poincaré sections of the 2π-solution mapping and also in more natural terms of the dynamics—quite plausible. His main result can be stated as follows.

Theorem H (Hubbard [24]). *Suppose we are given a bi-infinite sequence* $\{\varepsilon_k\}_{k\in\mathbf{Z}} \in \{-1;0;1\}^{\mathbb{Z}}$, *arbitrarily chosen. Then the pendulum governed by* (1.1) *has at least one motion that corresponds to the bi-infinite sequence* $\{\varepsilon_k\}_{k\in\mathbb{Z}}$ *in the sense that, during the time interval* $(2k\pi, 2(k+1)\pi)$, *the pendulum bob*

- *crosses the bottom position exactly once clockwise if and only if* $\varepsilon_k = -1$,
- *does not cross the bottom position at all if and only if* $\varepsilon_k = 0$, *or*
- *crosses the bottom position exactly once counterclockwise if and only if* $\varepsilon_k = 1$

and does not point downward at the time instants $t = 2k\pi$, $k \in \mathbf{Z}$.

The first aim of this paper is to interpret Hubbard's observation within the Mischaikow–Mrozek framework of computer-assisted proofs for horseshoe-type chaos. We use the word "observation" because, as is written on page 755 of [24], "no statement is proved anywhere." Hubbard arranges numerical evidence according to the framework of symbolic dynamics. We complete his work by filling in the gaps via refinements of some of his theoretical arguments (in particular, by introducing the small quadrangles L_ℓ, M_ℓ, R_ℓ, $\ell \in \mathbb{Z}$) and performing the necessary rigorous interval arithmetic computations. We will show that Theorem H is a consequence of a technical result based on Figure 10 in Hubbard [24], which shows images and preimages of three large quadrangles, the convex hulls of the smaller sets $L_\ell \cup M_\ell \cup R_\ell$, $\ell = -1, 0, 1$. In short, the observation is turned into a theorem.

Theorem 1.1. *There exist compact pairwise disjoint quadrangles*

$$L_0, M_0, R_0 \subset \left\{(x, \dot{x}) \in \mathbf{R}^2 \;\middle|\; 0 < x < 2\pi\right\}$$

with the following properties. Given a bi-infinite sequence $\{\varepsilon_k\}_{k\in\mathbb{Z}} \in \{-1;0;1\}^{\mathbb{Z}}$, *there exists a solution* $x = x(\{\varepsilon_k\}_{k\in\mathbb{Z}}) : \mathbf{R} \to \mathbf{R}$ *to* (1.1) *such that*

$$(x(2k\pi), \dot{x}(2k\pi)) \in \begin{cases} L_{\sigma_k} & \text{if} \quad \varepsilon_k = -1, \\ M_{\sigma_k} & \text{if} \quad \varepsilon_k = 0, \\ R_{\sigma_k} & \text{if} \quad \varepsilon_k = 1, \end{cases} \tag{1.2}$$

where $\sigma_{k+1} = \sigma_k + \varepsilon_k$, $k \in \mathbb{Z}$ *with* $\sigma_0 = 0$, *and*

$$L_\ell = L_0 + (2\ell\pi, 0), \quad M_\ell = M_0 + (2\ell\pi, 0), \quad R_\ell = R_0 + (2\ell\pi, 0), \quad \ell \in \mathbb{Z}. \tag{1.3}$$

Quadrangles L_0, M_0, and R_0 are shown in Figure 2. Property (1.2) means that the horizontal $2\ell\pi$-translates L_ℓ, M_ℓ, R_ℓ of the carefully chosen quadrangles L_0, M_0, R_0 are visited by trajectories of the Poincaré mapping

$$\Pi : \mathbb{R}^2 \to \mathbb{R}^2, \quad (x(0), \dot{x}(0)) \to (x(2\pi), \dot{x}(2\pi))$$

in the given order prescribed by the bi-infinite sequence $\{\varepsilon_k\}_{k\in\mathbb{Z}}$. The underlying circle of abstract topological results on transition graphs and iterates of continuous mappings are the

key parts of the landmark paper by Mischaikow and Mrozek [30] and of the great number of contributions that followed. The essence of the Mischaikow–Mrozek approach is to prove the existence of an abundance of combinatorially different periodic orbits and then, by using the density of periodic orbits in the shift dynamics, to pass to the existence of horseshoe-type chaos. The main technical tool is represented by what we call Lemma 2.1 in section 2 below. Lemma 2.1 relates to transition graphs and periodic orbits in two dimensions and constitutes the main step in proving Theorem 1.1.

The second aim of this paper is to provide an elementary proof of a higher-dimensional generalization of Lemma 2.1. Higher-dimensional versions of Lemma 2.1 were given by Gidea and Zgliczynski [21] and Pireddu and Zanolin [37]. The underlying definitions of the transition graphs in [21] and [37] (the latter being motivated by [25]) are different. However, both proofs are based on Brouwer degree arguments. Here we will give a third definition of the transition graph in higher dimensions—the two-dimensional case having been settled by Papini and Zanolin [34]—where a simple application of Brouwer's fixed point theorem suffices. This implies, in particular, that in some of the earliest computer-assisted proofs for horseshoe-type chaos [30], [52], [53], [54], Conley index and/or Brouwer degree arguments can be replaced by applications of Brouwer's fixed point theorem. See also Remark 1.

The computer-assisted parts of the proofs of Theorems 1.1 and H were performed in the LINUX and Cygwin environments, on a typical modern PC. We used the PROFIL/BIAS [27] programming environment which supports interval arithmetics and the Validated Numerical ODE (VNODE) package by Nedialkov [32], [33]. Our basic references for rigorous/interval computation and set-valued numerics are [1] and [13], respectively.

The computer program used for the proof can be downloaded from the Web page http://www.inf.u-szeged.hu/~banhelyi/FDP together with a short introduction and screenshots of the installation procedure.

This paper is organized as follows. Section 2 begins with a definition of the transition graph in two dimensions, goes on to state Lemma 2.1, and ends with a proof of Theorem 1.1. Theorem H and a higher-dimensional generalization of Lemma 2.1 are proved in sections 4 and 5, respectively. Connections to a four-dimensional neural networks model are investigated in section 6. Section 3 is devoted to a discussion of the role of the computer in chaos detection.

The results on symbolic dynamics and various forms of the pendulum equation can be found in a variety of papers. Two early results in this direction concern the standard pendulum equation with damping and variable length (but without an external forcing term) $\ddot{x} + b\dot{x} + (1 + c\sin(\mu t))\sin(x) = 0$. They were obtained by applying Melnikov's approach [48] and a computer-assisted version of the shooting method [23], respectively. The concept of a chaotic oscillation for the case $b = 0$ was defined in [17]. For the singularly perturbed van der Pol equation $\varepsilon\ddot{x} + \varphi(x)\dot{x} + \varepsilon x = p(t)$, where φ and p are piecewise constant, the existence of embedded symbolic dynamics was proved by Levi [28] in 1981. He used Newhouse's abstract results on homoclinic bifurcations.

From the enormous (and still mathematically sound) literature on chaos in electrical circuits, we refer the reader to the computer-assisted proofs of Galias [18] for chaos in Chua's circuit as well as to the computer-assisted proof of Yang and Li [47] for chaos in Josephson junctions.

Chaos results for the time-periodic nonlinear Hill equation $\ddot{x} + q(t)g(x) = 0$ were obtained

by topological and variational methods. The slightly more general time-periodic equations $\ddot{x}+b\dot{x}+q(t)g(x)=0$ and $\ddot{x}+\partial W(t,x)/\partial x=h(t)$ were investigated in [7] and [6], respectively. For details, generalizations, and more references, see the survey by Papini and Zanolin [35]. Note that Hubbard's pendulum equation (1.1) is not included in their discussions of theoretical and computational results, however.

2. Transition graph and chaos associated. For $j\in\mathbb{Z}$, define

$$\begin{aligned} Q_j &= \{(x_1,x_2)\in\mathbb{R}^2 \mid 3j+1\le x_1\le 3j+2,\ 0\le x_2\le 1\},\\ \lambda_j &= \{x\in Q_j \mid x_1=3j+1\},\quad \rho_j=\{x\in Q_j \mid x_1=3j+2\},\\ E_j &= \{(x_1,x_2)\in\mathbb{R}^2 \mid 3j+1\le x_1\le 3j+2,\ |2x_2-1|>1\}. \end{aligned}$$

Let $X=\cup_{j\in\mathbb{Z}}Q_j\subset\mathbb{R}^2$, and consider a continuous mapping $\varphi: X\to\mathbb{R}^2$ with coordinate functions φ_1,φ_2. The *transition graph* $\mathcal{G}(\varphi)$ *of* φ is defined as a directed graph with vertex set $\mathbf{V}(\mathcal{G})=\mathbb{Z}$. For $j,\tilde{j}\in\mathbf{V}(\mathcal{G})$, the pair $(j,\tilde{j})$ belongs to the edge set $\mathbf{E}(\mathcal{G})$ of $\mathcal{G}(\varphi)$ if

$$\varphi(Q_j)\subset\mathbb{R}^2\setminus \mathrm{cl}(E_{\tilde{j}}) \tag{2.1}$$

and one of the following conditions holds true:

$$\varphi_1(x)<3\tilde{j}+1 \quad\text{for}\quad x\in\lambda_j \quad\text{and}\quad \varphi_1(x)>3\tilde{j}+2 \quad\text{for}\quad x\in\rho_j \tag{2.2}$$

or

$$\varphi_1(x)>3\tilde{j}+2 \quad\text{for}\quad x\in\lambda_j \quad\text{and}\quad \varphi_1(x)<3\tilde{j}+1 \quad\text{for}\quad x\in\rho_j. \tag{2.3}$$

Sets Q_j,λ_j,ρ_j,E_j $(j=0,1,2)$ as well as relation $(0,2)\in\mathbf{E}(\mathcal{G})$ are shown in Figure 1. We write $\mathbf{V}=\mathbf{V}(\mathcal{G})=\mathbb{Z}$ and $\mathbf{E}=\mathbf{E}(\mathcal{G})$ in the following. For $N\in\mathbb{N}$, the directed graph $\mathcal{C}=\mathcal{C}(j_0,j_1,\dots,j_N)$ is a *directed* $(N+1)$*-circle in* $\mathcal{G}(\varphi)$ if $\mathbf{V}(\mathcal{C})=\{j_0,j_1,\dots,j_N\}\subset\mathbb{Z}$ and, with the convention $j_{N+1}=j_0$, $\mathbf{E}(\mathcal{C})=\{(j_k,j_{k+1})\}_{k=0}^{N}\subset\mathbf{E}$. The directed graph $\mathcal{P}=\mathcal{P}(\{j_k\mid k\in\mathbb{Z}\})$ is a *directed bi-infinite path in* $\mathcal{G}(\varphi)$ if $\mathbf{V}(\mathcal{P})=\{j_k\mid k\in\mathbb{Z}\}\subset\mathbb{Z}$ and $\mathbf{E}(\mathcal{P})=\{(j_k,j_{k+1})\}_{k\in\mathbb{Z}}\subset\mathbf{E}$. The definition of directed finite and infinite paths (i.e., paths having a root vertex) in $\mathcal{G}(\varphi)$ follows a similar pattern and will not be included here.

Lemma 2.1. *Let $\mathcal{C}=\mathcal{C}(j_0,j_1,\dots,j_N)$ be a directed circle in the transition graph $\mathcal{G}(\varphi)$. Then there is a finite sequence of points $\{q_k\}_{k=0}^{N}\subset X$ such that, with the convention $q_{N+1}=q_0$,*

$$q_{k+1}=\varphi(q_k) \quad\textit{and}\quad q_k\in Q_{j_k},\quad k=0,1,\dots,N.$$

Actually, Lemma 2.1 comes from the paper by Mischaikow and Mrozek [30]. As stated above, it is a version of the main result in Zgliczynski [52]. The formulation and the proof of a higher-dimensional generalization of Lemma 2.1 will be postponed until section 5.

Corollary 2.2. *Let $\mathcal{P}=\mathcal{P}(\{j_k\}_{k\in\mathbb{Z}})$ be a directed bi-infinite path in the transition graph $\mathcal{G}(\varphi)$. Assume that either*

(A) *every directed infinite path in $\mathcal{P}$ has infinitely many different vertices*

or

(B) *$\mathcal{G}$ (as a directed graph) is connected.*

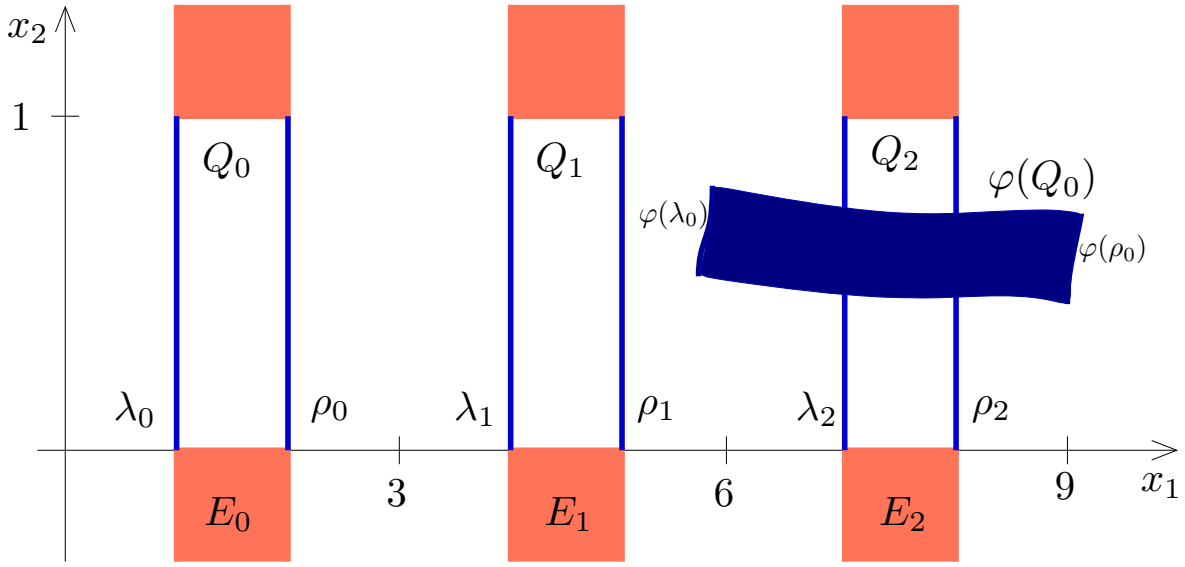


Figure 1. *Notation used to define the transition graph in two dimensions.*

Then there is a bi-infinite sequence of points $\{q_k\}_{k\in\mathbb{Z}} \subset X$ *with the property that*

$$q_{k+1} = \varphi(q_k) \quad \textit{and} \quad q_k \in Q_{j_k}, \quad k \in \mathbb{Z}.$$

Proof. Case (A). Choose a positive integer $\ell = L$ and consider the finite path with consecutive vertices $(j_{-L}, j_{-L+1}, \dots, j_L) \in \mathbb{Z}^{2L+1}$. Next, choose an integer $M > L$ such that $j_M \neq j_k$ for $k = -L, \dots, L$. Redefining φ on Q_{j_M}, we may assume that $(j_M, j_{-L}) \in \mathbf{E}$. Thus the extended finite sequence $(j_{-L}, \dots, j_L, j_{L+1}, \dots, j_M) \in \mathbf{V}^{L+1+M}$ forms the set of consecutive vertices of a directed circle in $\mathcal{G}(\varphi)$. Applying Lemma 2.1, we conclude there must exist a finite sequence of points $\{q_k^L\}_{|k|\le L} \subset X$ such that

$$q_{k+1}^L = \varphi(q_k^L) \quad \text{for} \quad k = -L, \dots, L-1 \quad \text{and} \quad q_k^L \in Q_{j_k} \quad \text{for} \quad k = -L, \dots, L.$$

Repeating the previous considerations for $\ell = L+1, L+2, \dots$, a standard Bolzano–Weierstrass subsequence argument in the limiting process $\ell \to \infty$ leads to the desired result.

Case (B). The connectedness of $\mathcal{G}$ is equivalent to the property that every directed finite path in $\mathcal{P}$ is contained in a directed circle of $\mathcal{G}(\varphi)$. Consequently, with some minor modifications, the argument we applied in proving case (A) can be repeated here. ■

Corollary 2.2 asserts the existence of a φ-trajectory which visits the Q_j's in the prescribed order: a directed bi-infinite path of type (A) or (B) of the transition graph is shadowed by a φ-trajectory. Directed $(N+1)$-circles in $\mathcal{G}(\varphi)$ are shadowed by $(N+1)$-periodic φ-trajectories. This is the essence of Lemma 2.1.

Remark 1. If $N = 0$, then Lemma 2.1 simplifies to the Colorado fixed point theorem in [3]. If the vertical coordinate is missing, then Lemma 2.1 simplifies to a well-known result in one-dimensional dynamics (see, for example, Lemma III.1.4 in [40]) whose proof is based solely on the intermediate value theorem. The proof of a higher-dimensional generalization

of Lemma 2.1 in section 5 mimics the standard derivation of the Miranda theorem from Brouwer's fixed point theorem [36]. Note that the Miranda theorem is nothing else but the higher-dimensional counterpart of the intermediate value theorem. It is known to be equivalent to Brouwer's fixed point theorem and to many other important results in topology [51]. Its history can be traced back to Poincaré and Bohl. Not long ago, the Miranda theorem appeared as a root test in numerical analysis and interval computation [16], [15], [42] as well as in chaos theory for two-dimensional mappings [34], [4]. The "rectangular nature" of the Miranda theorem fits in beautifully with the rectangles used to define the transition graph as well as the rectangles used in rigorous/interval computation.

Remark 2. Observe that Lemma 2.1 remains valid if the right-hand side of inclusion (2.1) is weakened to $\mathbb{R}^2 \setminus E_{\tilde{j}}$ and the strict inequalities in (2.2) and (2.3) are replaced by their nonstrict counterparts. (In fact, for $\ell = 1, 2, \dots$, it is elementary to construct a modified map $\varphi^\ell : X \to \mathbb{R}^2$ satisfying $|\varphi_\ell - \varphi| < 1/\ell$ for which Lemma 2.1 (as stated above) applies. Allowing $\ell \to \infty$, the existence of the desired φ-periodic trajectory follows from the Bolzano–Weierstrass argument.) The reason for stating Lemma 2.1 in the form presented above is to make the result stable with respect to small perturbations. Actually, if the conditions of Lemma 2.1 are met, and a continuous mapping $\tilde{\varphi} : \cup_{j\in\mathbb{Z}} Q_j \to \mathbb{R}^2$ satisfies $\max\{|\varphi(q) - \tilde{\varphi}(q)| \mid q \in \cup_{k=0}^{N} Q_{j_k}\} \le \eta$ with η sufficiently small, then the $(N+1)$-tuple $(j_0, j_1, \dots, j_N) \in \mathbb{Z}^{N+1}$ forms a directed circle in $\mathcal{G}(\tilde{\varphi})$ as well. As we shall see below, it is exactly this robustness property of the transition graph which makes Lemma 2.1 so suitable in computer-assisted proofs for horseshoe-type chaos. Stability in small perturbations in turn ensures stability in numerical approximations, including those with rounding errors.

Now we shall return to (1.1), which was studied by Hubbard [24].

In what follows we will demonstrate how Corollary 2.2 applies and how it leads to a complete proof of Theorem 1.1. The strategy is to find a bi-infinite sequence of pairwise disjoint compact sets $\{K_j\}_{j\in\mathbb{Z}}$ in the Poincaré plane $\{(x, \dot{x}) \in \mathbb{R}^2\}$ such that, up to a coordinate transformation h, Corollary 2.2 applies to the associated Poincaré mapping $\Pi : (x(0), \dot{x}(0)) \to (x(2\pi), \dot{x}(2\pi))$ of (1.1). We need a homeomorphism h of the Poincaré plane onto the standard plane $\{(x_1, x_2) \in \mathbb{R}^2\}$ such that, for

$$\varphi = h\Pi h^{-1}|X : X \to \mathbb{R}^2 \quad \text{with} \quad Q_j = h(K_j), \quad j \in \mathbb{Z},$$

Corollary 2.2 directly applies. Here, of course, $X = \cup_{j\in\mathbb{Z}} Q_j$, and $h\Pi h^{-1}|X$ means the restriction of $h\Pi h^{-1}$ to X. Since Π is 2π-periodic in the x variable and the number of different ε_k's is three, the bi-infinite sequence $\{K_j\}_{j\in\mathbb{Z}}$ is sought as a collection of the horizontal $2\ell\pi$-translates of the three specially chosen quadrangles L_0, M_0, R_0 (compare the notation in (1.3) and see Figure 2) with

$$K_{3\ell} = L_0 + (2\ell\pi, 0), \quad K_{3\ell+1} = M_0 + (2\ell\pi, 0), \quad K_{3\ell+2} = R_0 + (2\ell\pi, 0), \quad \ell \in \mathbb{Z}.$$

Given a bi-infinite sequence $\{\varepsilon_k\}_{k\in\mathbb{Z}} \in \{-1; 0; 1\}^{\mathbb{Z}}$, it is essential that the directed bi-infinite path $\mathcal{P} = \mathcal{P}(\{j_k\}_{k\in\mathbb{Z}})$ with $j_k = 3\sigma_k + 1 + \varepsilon_k$ (where—as defined in Theorem 1.1—$\sigma_0 = 0$ and $\sigma_{k+1} = \sigma_k + \varepsilon_k$ for $k \in \mathbb{Z}$) be a subgraph of $\mathcal{G}(\varphi)$. Applying Corollary 2.2, trajectories satisfying (1.2) correspond to the directed bi-infinite path $\mathcal{P} = \mathcal{P}(\{j_k\}_{k\in\mathbb{Z}})$ and vice versa.

Proof of Theorem 1.1. The successful realization of the strategy outlined above depends on the careful choice of the quadrangles L_0, M_0, R_0 and of the coordinate transformation

h. Noting the horizontal 2π-translation invariance property of the collection $\{K_j\}_{j\in\mathbb{Z}}$, the continuous mapping $\varphi = h\Pi h^{-1}|X$ is prescribed to be 9-periodic with respect to the x_1 variable. This can be guaranteed by requiring that the coordinate functions of homeomorphism $h : \{(x, \dot{x}) \in \mathbb{R}^2\} \to \{(x_1, x_2) \in \mathbb{R}^2\}$ satisfy

$$h_1(x + 2\pi, \dot{x}) = 9 + h_1(x, \dot{x}) \quad \text{and} \quad h_2(x + 2\pi, \dot{x}) = h_2(x, \dot{x}). \tag{2.4}$$

The existence of quadrangles L_0, M_0, R_0 that lead to a *transition graph suitably complex* depends on the inner structure of the Poincaré mapping.

Following Hubbard [24], define quadrangles $K_0 = L_0$, $K_1 = M_0$, $K_2 = R_0$ as

$$K_j = \operatorname{conv}\{V_{ul}^{K_j}, V_{ur}^{K_j}, V_{ll}^{K_j}, V_{lr}^{K_j}\}, \quad j = 0, 1, 2,$$

which are the closed convex hulls of their respective upper left, upper right, lower left, and lower right vertices. (The letters L, M, and R stand for left, middle, and right, respectively.) The coordinates of these vertices are

$$\begin{aligned} V_{ul}^{L_0} &= (1.000,\ -0.985), & V_{ur}^{L_0} &= (1.970,\ -0.208),\\ V_{ll}^{L_0} &= (1.226,\ -1.350), & V_{lr}^{L_0} &= (2.226,\ -0.516),\\ V_{ul}^{M_0} &= (2.436,\ 0.166), & V_{ur}^{M_0} &= (2.481,\ 0.201),\\ V_{ll}^{M_0} &= (2.758,\ -0.123), & V_{lr}^{M_0} &= (2.796,\ -0.092),\\ V_{ul}^{R_0} &= (3.197,\ 0.775), & V_{ur}^{R_0} &= (3.800,\ 1.258),\\ V_{ll}^{R_0} &= (3.398,\ 0.389), & V_{lr}^{R_0} &= (4.412,\ 1.202). \end{aligned}$$

See Figure 2. For details on how the individual vertices were found, see the third paragraph of section 3 below.

Now consider the broken line in Figure 2, namely,

$$\mathcal{L}_1 = \{\text{the vertical half-line below } W_1^1\} \cup [W_1^1, W_1^2] \cup \{\text{the vertical half-line above } W_1^2\},$$

where

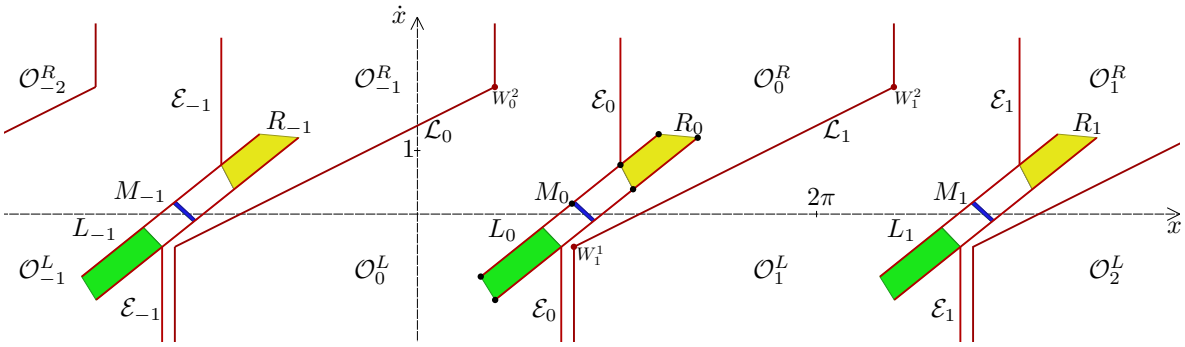
$$W_1^1 = (w_1^1, w_2^1) = V_{lr}^{L_0} + (0.2, 0), \qquad W_1^2 = (w_1^2, w_2^2) = (7.5, 2),$$

and $[W_1^1, W_1^2]$ stands for the closed line segment between W_1^1 and W_1^2. The open strip between $\mathcal{L}_1$ and the translated broken line $\mathcal{L}_0 = \mathcal{L}_1 + (-2\pi, 0)$ shall be denoted by $\mathcal{S}_0$. Now with "conv" standing for the closed convex hull of the points in braces, define

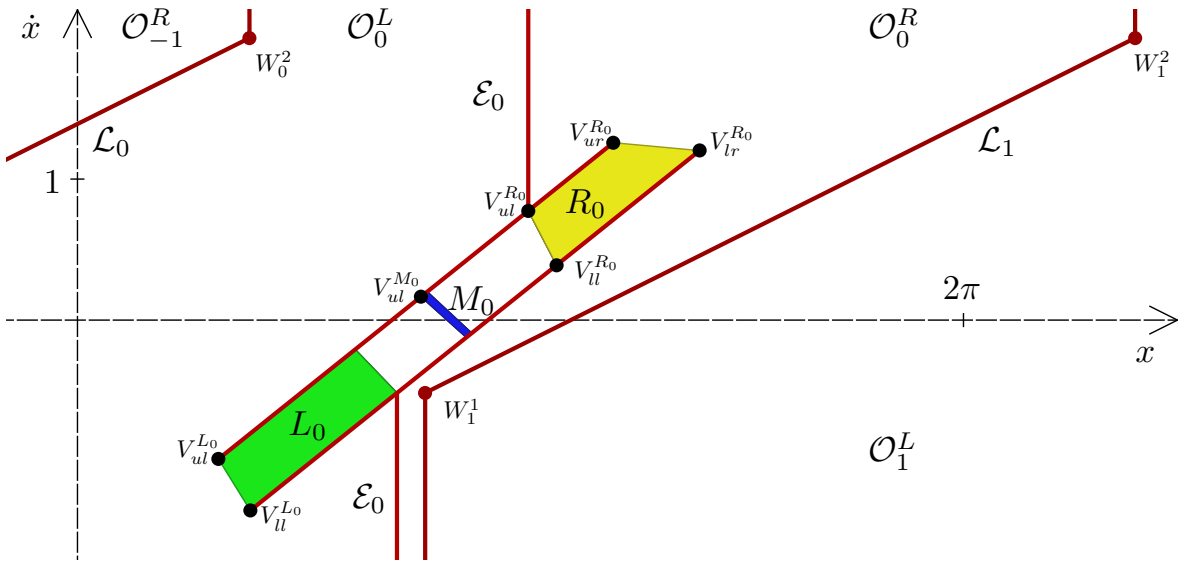
$$\begin{aligned} \mathcal{D}_0 = \{&\text{the vertical half-line below } V_{lr}^{L_0}\} \cup L_0 \cup \operatorname{conv}\{V_{ur}^{L_0}, V_{ul}^{M_0}, V_{ll}^{M_0}, V_{lr}^{L_0}\}\\ &\cup M_0 \cup \operatorname{conv}\{V_{ur}^{M_0}, V_{ul}^{R_0}, V_{ll}^{R_0}, V_{lr}^{M_0}\} \cup R_0 \cup \{\text{the vertical half-line above } V_{ul}^{R_0}\}. \end{aligned}$$

The open strips between $\mathcal{D}_0$ and $\mathcal{L}_0$ (resp., $\mathcal{L}_1$) will be denoted by $\mathcal{O}_0^L$ (resp., $\mathcal{O}_0^R$). The union of the right-hand side boundary of the strip $\mathcal{O}_0^L$ and the left-hand side boundary of the strip $\mathcal{O}_0^R$ will be denoted by $\mathcal{B}_0$. Finally, we let

$$\mathcal{E}_0 = \mathcal{B}_0 \setminus \{(V_{ul}^{L_0}, V_{ll}^{L_0}) \cup (V_{ur}^{R_0}, V_{lr}^{R_0})\},$$



(a) The entire region considered



(b) Focusing just on the middle three quadrangles

Figure 2. *Notation used in proving Theorem 1.1.*

where, for example, $(V_{ul}^{L_0}, V_{ll}^{L_0})$ stands for the open line segment connecting $V_{ul}^{L_0}$ and $V_{ll}^{L_0}$. (The closed line segment connecting $V_{ul}^{L_0}$ and $V_{ll}^{L_0}$, for example, will be denoted by $[V_{ul}^{L_0}, V_{ll}^{L_0}]$. Note that $\mathcal{E}_0$ is the union of ten closed line segments and two closed half-lines. See Figure 2 again.)

The crucial properties responsible for the edge structure of the transition graph are

$$\Pi(R_{-1}),\ \Pi(M_0),\ \Pi(L_1)\ \subset\ \mathcal{S}_0 \setminus \mathcal{E}_0, \tag{2.5}$$

$$\Pi([V_{ul}^{R_{-1}}, V_{ll}^{R_{-1}}]),\ \Pi([V_{ul}^{M_0}, V_{ll}^{M_0}]),\ \Pi([V_{ur}^{L_1}, V_{lr}^{L_1}])\ \subset\ \mathcal{O}_0^L, \tag{2.6}$$

$$\Pi([V_{ur}^{R_{-1}}, V_{lr}^{R_{-1}}]),\ \Pi([V_{ur}^{M_0}, V_{lr}^{M_0}]),\ \Pi([V_{ul}^{L_1}, V_{ll}^{L_1}])\ \subset\ \mathcal{O}_0^R. \tag{2.7}$$

See Figure 3, which shows the sets $\Pi(L_0)$ (a translated copy of $\Pi(L_1)$), $\Pi(M_0)$, and $\Pi(R_0)$ (a translated copy of $\Pi(R_{-1})$). The subset relations (2.5), (2.6), and (2.7) will be checked by

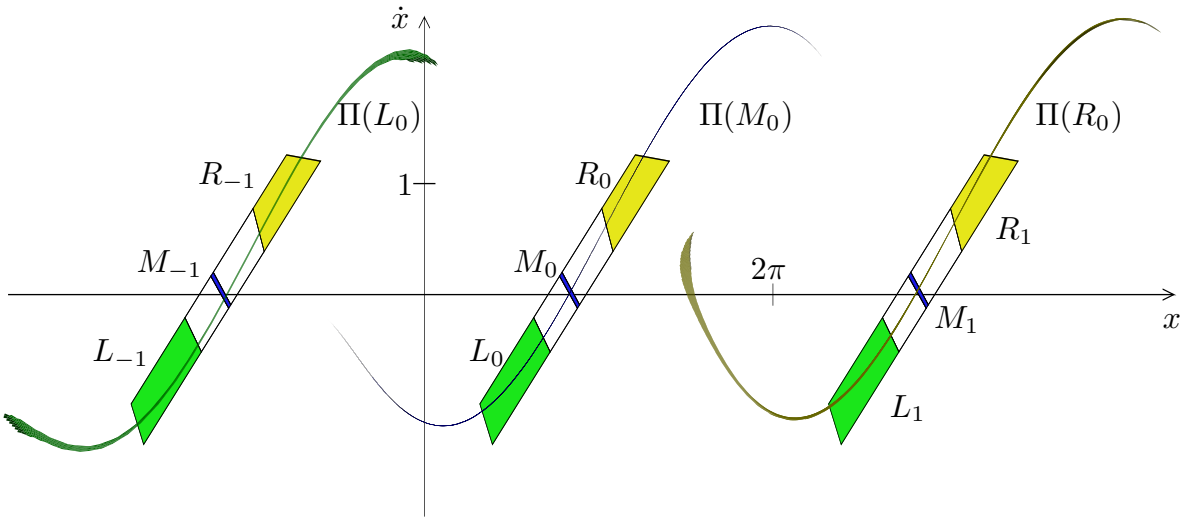


Figure 3. *Images of the specially chosen quadrangles under* Π.

computer. Note that the sets $\mathcal{S}_0 \setminus \mathcal{E}_0$, $\mathcal{O}_0^L$, and $\mathcal{O}_0^R$ are open and all nine sets $\Pi(R_{-1}), \dots, \Pi([V_{ul}^{L_1}, V_{ll}^{L_1}])$ on the respective left-hand sides are compact. Hence inclusions (2.5), (2.6), and (2.7) remain valid if the entire construction is repeated with the sets $\mathcal{D}_0$, $\mathcal{B}_0$, and $\mathcal{E}_0$ slightly thicker, that is, if $\mathcal{D}_0$, $\mathcal{B}_0$, and $\mathcal{E}_0$ are replaced by their closed neighborhoods $\mathcal{D}$, $\mathcal{B}$, and $\mathcal{E}$, suitably chosen.

Next we will start constructing a homeomorphism h subject to condition (2.4). We also require that $Q_j = h(K_j)$ with

$$(3j+1,1) = h(V_{ul}^{K_j}), \quad (3j+2,1) = h(V_{ur}^{K_j}), \quad j = 0,1,2,$$
$$(3j+1,0) = h(V_{ll}^{K_j}), \quad (3j+2,0) = h(V_{lr}^{K_j}), \quad j = 0,1,2$$

(i.e., the corresponding vertices are mapped to each other), and

$$\operatorname{cl}(E_0 \cup E_1 \cup E_2) \subset h(\mathcal{E}), \qquad \{(x_1,x_2) \in \mathbb{R}^2 \mid x_1 = 0\} = h(\mathcal{L}_0). \tag{2.8}$$

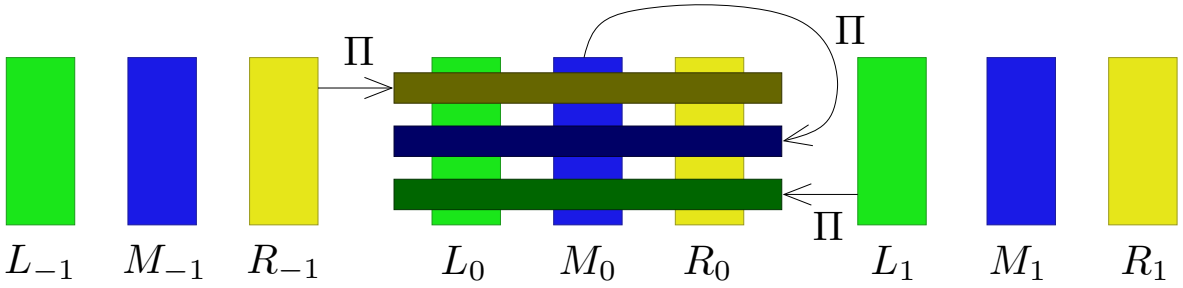
Due to the piecewise linear boundaries of the sets involved, the construction of h is elementary. We have a fair amount of freedom in choosing h. Advanced results of two-dimensional topology are not needed for this. Note that, by translation symmetry, the broken line $\mathcal{L}_1$ is mapped onto the line of the equation $x_1 = 9$.

Recall that $X = \cup_{j\in\mathbb{Z}} Q_j$. Then property (2.5) and the inclusion in (2.8) imply that

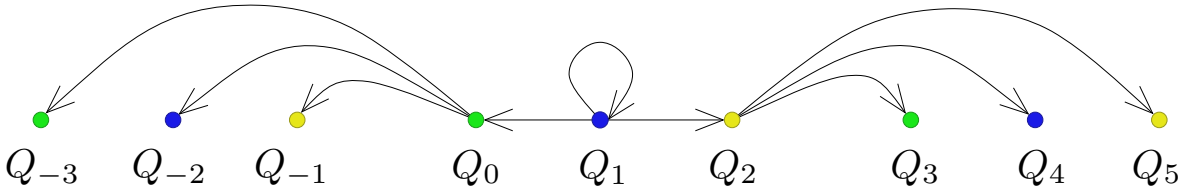
$$\varphi(X) \subset \mathbb{R}^2 \setminus \operatorname{cl}\left(\cup_{j\in\mathbb{Z}} E_j\right).$$

Using (2.6), (2.7), we conclude that the transition graph of φ is as follows. The vertex set of $\mathcal{G}(\varphi)$ is obviously $\mathbf{V} = \mathbb{Z}$, and $\mathcal{G}(\varphi)$ is 3-periodic in the sense that $(j,\tilde{j}) \in \mathbf{E}$ if and only if $(j+3, \tilde{j}+3) \in \mathbf{E}$. The edges starting from the vertex subset $\{0,1,2\}$ are like those shown in Figure 4(a):

$$(0,-3);(0,-2);(0,-1);(1,0);(1,1);(1,2);(2,3);(2,4);(2,5).$$



(a) A segment of the 3-periodic transition graph $\mathcal{G}(\varphi)$



(b) Σ_3-chaos in the quotient dynamics of Π

Figure 4. *Combinatorial complexity in Hubbard's forced damped pendulum equation.*

Thus we arrive at the schematic phase portrait of the Poincaré mapping depicted in Figure 4(b).

Given a bi-infinite sequence $\{\varepsilon_k\}_{k\in\mathbb{Z}} \in \{-1;0;1\}^{\mathbb{Z}}$, a quick analysis of the transition graph $\mathcal{G}(\varphi) = \mathcal{G}(h\Pi h^{-1}|X)$ shows that the directed bi-infinite path $\mathcal{P} = \mathcal{P}(\{j_k\}_{k\in\mathbb{Z}})$ with $j_k = 3\sigma_k + 1 + \varepsilon_k$ (where—as defined in Theorem 1.1—$\sigma_0 = 0$ and $\sigma_{k+1} = \sigma_k + \varepsilon_k$ for $k \in \mathbb{Z}$) is a subgraph of $\mathcal{G}(\varphi)$. Trajectories satisfying (1.2) correspond to the directed bi-infinite path $\mathcal{P} = \mathcal{P}(\{j_k \mid k \in \mathbb{Z}\})$ and vice versa.

This provides all the necessary points for proving Theorem 1.1. Apply Corollary 2.2, and then we are done. ∎

The derivation of Theorem 1.1 follows the main argument in the Mischaikow–Mrozek framework for computer-assisted proofs. (Note that the invertibility of Π was not exploited, but it will be needed for the backward invariance of the set Λ in Corollary 2.3 below.) For the geometric background and details on the role of the computer, see section 3.

It is not hard to reformulate Theorem 1.1 in the language of symbolic dynamics [49], [40]. In fact, recall that $Q_j = h(K_j)$, and let $\Theta \subset X$ be the closure of all periodic points of φ that shadow the directed circles of $\mathcal{G}(\varphi)$. The set Θ is backward and forward invariant under φ. For $x \in \Theta$, the formula

$$(c(x))_k = j_k \quad \text{whenever} \quad \varphi^k(x) \in Q_{j_k} \quad \text{and} \quad k \in \mathbb{Z}$$

defines a continuous itinerary mapping $c : \Theta \to \mathbb{Z}^{\mathbf{Z}}$. The inverse of a homeomorphism h lifts everything to the Poincaré plane. Clearly $\Lambda = h^{-1}(\Theta)$ is backward and forward invariant under the Poincaré mapping Π, and, for $\lambda = (x, \dot{x}) \in \Lambda$ with $d(\lambda) = c(h(\lambda))$,

$$(d(\lambda))_k = j_k \quad \text{whenever} \quad \Pi^k(\lambda) \in K_{j_k}, \quad k \in \mathbb{Z}.$$

Letting S denote the shift operator on $\mathbb{Z}^{\mathbb{Z}}$, we may conclude that

$$c(\varphi(x)) = Sc(x) \quad \text{for each} \quad x \in \Theta \qquad \text{and} \qquad d(\Pi(\lambda)) = Sd(\mathbf{x}) \quad \text{for each} \quad \lambda \in \Lambda.$$

The entire construction is based on the horizontal 2π-translation symmetry of Π. The respective quotient maps are continuous and satisfy

$$\bar{d}(\bar{\Pi}(\bar{\lambda})) = \bar{S}\bar{d}(\bar{\lambda}) \qquad \text{for each} \quad \bar{\lambda} \in \bar{\Lambda}.$$

The quotient transition graph $\mathcal{G}(\bar{\varphi})$ is the complete directed graph on three vertices and thus the modulo 3 itinerary map $\bar{d} : \bar{\Lambda} \to \{0,1,2\}^{\mathbb{Z}}$ is onto. In particular, note that

$$(\bar{d}(\bar{\lambda}))_k = 1 + \varepsilon_k \quad \text{for} \quad \bar{\lambda} \in \bar{\Lambda} = \Lambda \cap \left\{(x, \dot{x}) \in \mathbf{R}^2 \;\middle|\; 0 < x < 2\pi\right\}, \quad k \in \mathbb{Z}.$$

The quotient results can be expressed in compact form as a corollary.

Corollary 2.3 (a continuation of Theorem 1.1). *The modulo 2π Poincaré mapping $\bar{\Pi}$ on $\bar{\Lambda}$ is semiconjugate to the shift operator $\bar{S}$ on Σ_3, the space of three symbols.*

In fact, as suggested by Hubbard [24], $\bar{d}$ is plausibly one-to-one, and thus $\bar{\Pi}|\bar{\Lambda}$ and $\bar{S}$ are conjugate. See Figure 3 again and compare it with Figures 4(a) and 4(b).

3. Chaos detection by computer. What the computer is used for in the Mischaikow–Mrozek framework of computer-assisted proofs for chaos is *to check certain subset relations* (like (2.5), (2.6), (2.7)) and, above all, *to find the subset relations to be checked*—in essence, to find a collection of "rectangular" subsets of the phase space like L_0, M_0, R_0 such that the associated transition graph has at least two different, but intersecting, circles. The hard part is to find the subset relations to be checked. If small perturbation arguments do not help, one cannot get by without a computer. The checking part is much easier and sometimes, in exceptional cases, like the equation $\ddot{x} + x = \sin(\sqrt{2}t) + 2^{-1}(|5x+1| - |5x-1|)$ [38], it can be done by hand. Still, the proof in [38] is computer-assisted. The successful collection of "rectangular" subsets is the result of trial-and-error computer experimentation with human overheads.

It is natural to ask to what extent the task of finding the successful subset relations can be left to the computer. The required subset relations determine a constrained satisfaction problem [10], and techniques of global optimization [39] apply. If we want to look for three quadrangles, the search domain of the optimization procedure is a subset of a 24-dimensional parameter space (eight dimensions for each quadrangle based on the coordinate pairs of the four vertices; the search for a successful collection of the "forbidden sets" $\mathcal{L}_0$, $\mathcal{L}_1$, and $\mathcal{E}_0$ requires the introduction of some additional parameters). And the smaller the search domain, the better. However, a "small" search domain corresponds to a "good" initial guess which can only be obtained from some a priori known theoretical or numerical results on the details of the dynamics. Typical candidates for members of a successful collection are quadrangles situated on the unstable manifold of a transversal homoclinic saddle.

In an interesting paper devoted to Hénon mapping with the classical parameters $a = 1.4$ and $b = 0.3$, Galias [19] describes the configuration of 29 polygons which leads to the rigorous entropy estimate $h(\mathcal{H}) > 0.430\ldots$, which is quite close to the generally accepted value of $h(\mathcal{H}) = 0.465\ldots$. All 29 polygons are narrow quadrangles—or quadrangles with some vertices

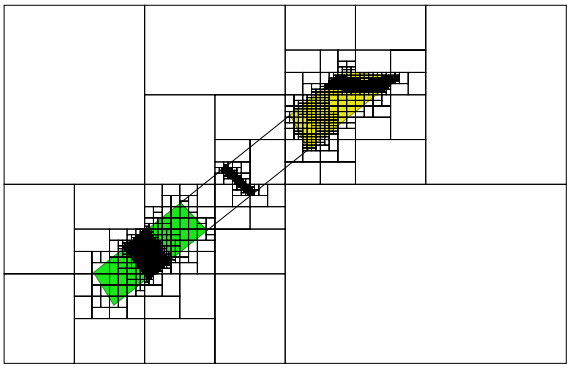
"chopped off"—situated along the unstable manifold of the homoclinic saddle. They were found by hand, based on an earlier search for periodic points of low periods. The well-known and highly automatized GAIO package [12], [13] is used to construct 247 GAIO polygons in a forthcoming paper by Day, Frongillo, and Trevino [11] proving the slightly better estimate of $h(\mathcal{H}) > 0.4318\ldots$. (If a global search is performed just on finding 29 or 247 segments of the unstable manifold, one needs 58 or 494 parameters, respectively. The second number is far too much for optimization methods currently available for this type of problem.) Nevertheless, it remains an open question whether a bootstrap application of global optimization procedures, keeping the number of parameters under 10, say, at each step of the gradual improvements along the consecutive local searches, can achieve a better estimate. We feel that it is not inappropriate here to call to the attention of the reader a forthcoming paper [5] of ours, where, within a 17-dimensional parameter space, the full power of the optimization method [4] is exploited. The main result is that $\mathcal{H}^k$, the kth iterate of Hénon's mapping with the classical parameters $a = 1.4$ and $b = 0.3$, has an embedded copy of the Σ_2 dynamics if and only if $k = 2$, $k = 4$, or $k \geq 6$. This is guaranteed by Smale's abstract theory of transversal homoclinic saddles only for $k \geq k_0$ sufficiently large. (Incidentally, all existence proofs (like [31], [14], [20]) for a transversal homoclinic saddle in the dynamics of $\mathcal{H}$ are, to the best of our knowledge, in some way or other, computer-assisted.)

In proving Theorem 1.1, the vertices of quadrangles L_0, M_0, R_0 (as well as of the "forbidden sets" $\mathcal{L}_0, \mathcal{L}_1, \mathcal{E}_0$) were chosen in the way shown in Hubbard [24]. Though the coordinates of the individual vertices were not explicitly given by him, it was straightforward to adjust them based on Figure 10 of his paper. This adjustment was made by hand. According to our estimates, our method [10] would have required several months of CPU time. Actually, what Hubbard works with are just three large quadrangles, the convex hulls of which we define as the sets $L_\ell \cup M_\ell \cup R_\ell$, $\ell = -1, 0, 1$, and the "forbidden sets" are not mentioned by him at all. At first sight, it is quite plausible that the twelve vertices $V_{ul}^{L_0}, \ldots, V_{lr}^{R_0}$ lie on the circumference of Hubbard's large quadrangle. However, we could not find such an arrangement. This indicates the differences between nonrigorous and rigorous computation. Just like the Hénon mapping $\mathcal{H}$, the Poincaré mapping Π of Hubbard's pendulum equation (1.1) also has a homoclinic saddle. This saddle point is

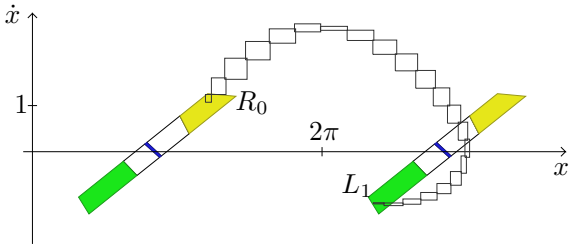
$$P = (2.634\ldots, 0.026\ldots) \quad \text{with eigenvalues} \quad \mu_1 = 321.836\ldots \quad \text{and} \quad \mu_2 = 0.001\ldots$$

(all our computations being rigorous). Of course P represents an unstable 2π-periodic solution which has bifurcated from the upward/top equilibrium position $x = \pi$, $\dot{x} = 0$ of the damped unforced pendulum. Equation (1.1) has a second, asymptotically stable, 2π-periodic solution which corresponds to the sink $Q = (4.236\ldots, 0.392\ldots)$ of the Poincaré mapping with eigenvalues $\mu_{1,2} = -0.725\cdots \pm i\,0.129\ldots$ and which has bifurcated from the bottom equilibrium position $x = 0$, $\dot{x} = 0$ of the damped unforced pendulum. Computer-assisted reasoning shows there are no further 2π-periodic solutions. Note that P is contained in M_0, and its unstable manifold intersects the carefully chosen quadrangles in the rather strange order of $R_{-1}, M_{-1}, L_{-1}, L_0, M_0, R_0, L_1, M_1, R_1$.

Unstable and stable manifolds of P intersect each other outside P. Apparently, this is a transversal intersection. But we did not verify transversality by rigorous computation. The reason is that transversality by itself, though guaranteeing the existence of a topological



(a) A typical detail in the computer-assisted part of proving Theorem 1.1



(b) A typical detail in the computer-assisted part of proving Theorem H

Figure 5. *Checking inclusions by interval computation.*

horseshoe, contains less information about the dynamics than a transition graph with carefully chosen "rectangular" subsets. The next logical step forward should be, rather, the verification of the Conley–Moser invariant cone field conditions [49] leading (if it is really the case) to transversality as well as to the conjugacy between $\bar{\Pi}|\bar{\Lambda}$ and $\bar{S}$. Unfortunately, *the verification of inclusions* (2.5), (2.6), (2.7) *takes almost an hour on a typical modern PC.* See Figure 5(a). Because of this, we think that there is little hope of checking the invariant cone field conditions in a reasonable amount of time. Nevertheless, the semiconjugacy of $\bar{\Pi}|\bar{\Lambda}$ to $\bar{S}$ established in Corollary 2.3 is not much worse than the conjugacy we expected. Semiconjugacy to $\bar{S}$ means that the dynamics is at least as complex as the full shift on the space of three symbols, while conjugacy would mean that the dynamics of $\bar{\Pi}|\bar{\Lambda}$ is just as complex as the one belonging to $\bar{S}$. What can be shown is that $m(\bar{\Lambda})$, the Lebesgue measure of $\bar{\Lambda}$, is zero. (This is clear because $\bar{\Pi}(\bar{\mathcal{C}}) \subset \bar{\mathcal{C}}$ for $\bar{\mathcal{C}} = \{(x, \dot{x}) \in \mathbb{R}^2 \mid 0 \le x < 2\pi,\ |\dot{x}| \le 12\}$, $\bar{\Lambda} \subset \cap_{k=0}^{\infty} \bar{\Pi}^k(\bar{\mathcal{C}})$, and $\bar{\Pi}$ contracts areas by a factor of $e^{-\pi/5}$, due to damping and the Liouville theorem [24].) Questions on additional chaos properties in Hubbard's pendulum equation (1.1)—like the Wada property experimentally observed by Hubbard [24] or fine ergodic properties like the existence of a unique Sinai–Ruelle–Bowen (SRB) measure (found for the Lorenz equation by Tucker [46]) and mixing (found for the Lorenz equation by Luzzato, Melbourne, and Paccaut [29])—remain open.

In conclusion, we note that the existence of a transition graph with two different but intersecting circles is implicit in a paper by Stoffer and Palmer [44] on shadowing. In essence, they prove that the existence of two hyperbolic periodic orbits which come sufficiently near each other without remaining too close for a long time (e.g., those whose minimal periods are highly nonresonant) implies the existence of an embedded horseshoe. This corresponds to the Levinson phenomenon which motivated Smale to construct the geometric horseshoe [43], [28]. For a comparison between the shadowing and the topological approach in computer-assisted proofs for chaos, see the recent paper by Coomes, Kocak, and Palmer [9].

4. Chaos in natural terms of the dynamics. The one-to-one correspondence between a set of the solutions to Hubbard's pendulum equation (1.1) and the set of all bi-infinite

sequences on three symbols manifests itself in natural terms of the dynamics.

Focusing on the pendulum, the quadrangles L_0, M_0, R_0 remain hidden, even to the most observant viewer. What can be easily seen are high speeds or low speeds, the number of consecutive clockwise or counterclockwise returns, changes in the direction of swing and/or rotation, and movements across the upper and/or lower vertical positions. When systematizing a range of dynamical behavior, the mind has a tendency to consider the consecutive occurrences of alternative, easily discernible events like a heads-or-tails sequence in coin-tossing.

Theorems H and 1.1 should be interpreted from this point of view. Any possible order of the mutually exclusive alternatives can occur. Both observations describe the same combinatorial aspect of Σ_3-chaos—the existence of "coin-tossing" (coins with three sides) label sequences [26] for itineraries. However, the alternatives in Theorem 1.1 are hard to observe whereas the alternatives in Theorem H are quite transparent. There exist uncountably many solutions of Hubbard's pendulum equation which can be distinguished from each other based on their combinatorially different qualitative behavior. This is what we might call combinatorial chaos in natural terms of the dynamics. Previous examples include symbolic dynamics in terms of consecutive return times in Alekseev's three-body system [2], [24]; consecutive maxima and minima in the Lorenz systems [22]; the number of sign changes in consecutive time intervals of equal length [7], [45]; and multibumps in bursting oscillations [41]. Their natural place is in the vicinity of bifurcating homoclinic/heteroclinic orbit connections.

Proof of Theorem H. In order to prove Theorem H, we need to examine what the solution map $(x(0), \dot{x}(0)) \to (x(t), \dot{x}(t))$ does between the Poincaré sections at $t_0 = 0$ and $t_1 = 2\pi$.

First, consider the collection of motions of the forced damped pendulum with initial position $(x(0), \dot{x}(0)) \in R_0$ and final position $(x(2\pi), \dot{x}(2\pi)) \in L_1 \cup M_1 \cup R_1$. It is not hard to check by rigorous/interval computation that $0 < x(t) < 4\pi$, whenever $0 \le t \le 2\pi$, and

$$\big\{(x(t), \dot{x}(t)) \in \mathbb{R}^2 \;\big|\; 0 \le t \le 2\pi\big\} \cap \big\{(x, \dot{x}) \in \mathbb{R}^2 \;\big|\; x = 2\pi \text{ and } \dot{x} \le 0\big\} = \emptyset.$$

Applying the intermediate value theorem, it follows that $x(t^*) = 2\pi$ for some $t^* \in (0, 2\pi)$, $x(t) \in (0, 2\pi)$ for $t \in [0, t^*)$, and $x(t) \in (2\pi, 4\pi)$ for $t \in (t^*, 2\pi]$. In other words, during the time interval $(0, 2\pi)$, the pendulum bob crosses the bottom position exactly once counterclockwise and does not point downward at the time instants $t_0 = 0$ and $t_1 = 2\pi$. This holds true for motions of the pendulum with initial position $(x(0), \dot{x}(0)) \in R_0$ and final position $(x(2\pi), \dot{x}(2\pi)) \in L_1 \cup M_1 \cup R_1$ (but not for all motions with initial position $(x(0), \dot{x}(0)) \in R_0$). This holds true especially for all $\sigma_0 = 0$, $\varepsilon_0 = 1$ (and, a fortiori, $\sigma_1 = 1$, $\varepsilon_1 \in \{-1, 0, 1\}$) motions of the pendulum described by Theorem 1.1. Parts of the necessary computations in subcase $\sigma_0 = 0$, $\varepsilon_0 = 1$, $\sigma_1 = 1$, $\varepsilon_1 = -1$ are shown in Figure 5(b).

The remaining cases $\sigma_0 = 0$, $\varepsilon_0 = 0$ and $\sigma_0 = 0$, $\varepsilon_0 = -1$ were settled in a similar way. *The total CPU time requested was under two minutes on a typical modern PC.* ∎

The connection between symbolic dynamics and oscillation patterns in (1.1) is worth further investigation. We would like to know whether symbolic dynamics appears regarding crossing the bottom *and the top* equilibrium positions.

5. Lemma 2.1 in a higher dimension. A simple proof. Let m, n be fixed nonnegative integers, and let $\mathbf{V} \subset \mathbb{Z}$ be a finite or countably infinite indexing set. Next, let the boundary and interior of a compact set S in a Euclidean space $\mathbb{R}^k$ be denoted by ∂S and $\mathrm{int}(S)$,

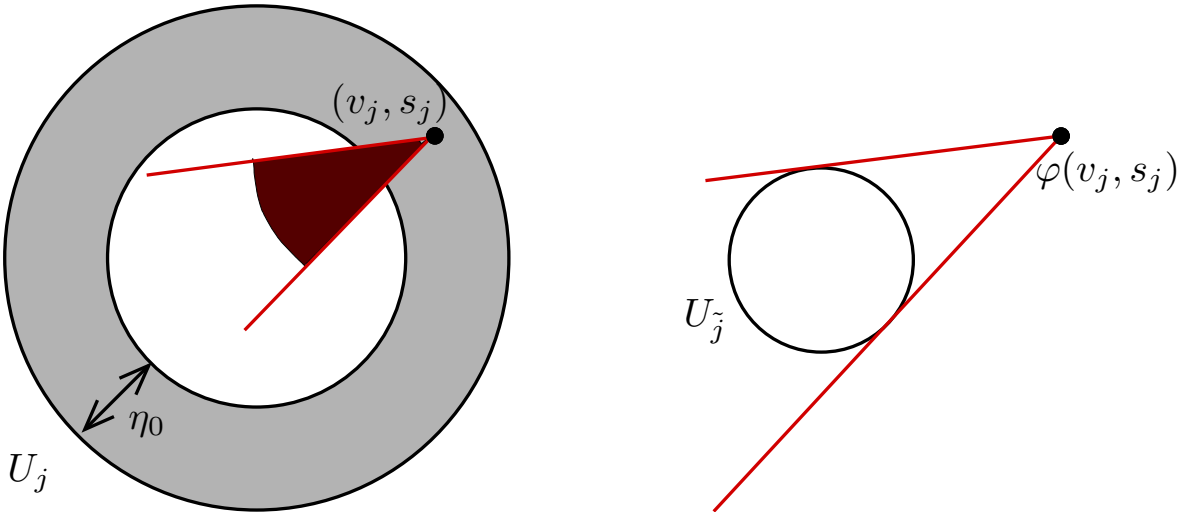


Figure 6. *Condition* (5.2) *for fixed* $(v_j, s_j) \in \{u_j \in U_j \mid d(u_j, \partial U_j) \le \eta_0\} \times S_j$.

respectively. The closed neighborhood of radius $R > 0$ of a point p and a set S in $\mathbb{R}^k$ will be denoted by $\mathcal{B}^k[p, R]$ and $\mathcal{B}^k[S, R]$, respectively. The norm and scalar product in $\mathbb{R}^k$ shall be denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$.

Now consider the collection of rectangular sets of the form

$$Q_j = \left\{x = (u, s) \in \mathbb{R}^m \times \mathbb{R}^n \;\middle|\; u \in U_j,\ s \in S_j\right\}, \quad j \in \mathbf{V},$$

where $\{U_j\}_{j\in\mathbf{V}}$ and $\{S_j\}_{j\in\mathbf{V}}$ are compact topological balls in $\mathbb{R}^m$ and in $\mathbb{R}^n$, respectively. Note that S_j is a retract of $\mathbb{R}^n$. Let $r_j : \mathbb{R}^n \to S_j$ be a retraction, where $j \in \mathbf{V}$.

Next, let $X = \cup_{j\in\mathbf{V}} Q_j \subset \mathbb{R}^m \times \mathbb{R}^n$ and consider a continuous mapping $\varphi : X \to \mathbb{R}^m \times \mathbb{R}^n$ with coordinate functions φ_u, φ_s. Afterward, suppose that $Q_j \cap Q_k = \emptyset$ for $j \ne k$ and that $\left\{j \in \mathbf{V} \;\middle|\; Q_j \cap \left\{(u, s) \in \mathbb{R}^m \times \mathbb{R}^n \;\middle|\; \|u\| + \|s\| < R\right\} \ne \emptyset\right\}$ is finite for any $R > 0$.

The *transition graph* $\mathcal{G}(\varphi)$ *of* φ is defined as a directed graph with vertex set $\mathbf{V}$. For $j, \tilde{j} \in \mathbf{V}$, the pair $(j, \tilde{j})$ belongs to the edge set $\mathbf{E}$ of $\mathcal{G}(\varphi)$ if

$$\varphi(Q_j) \subset \mathbb{R}^m \times \mathbb{R}^n \setminus U_{\tilde{j}} \times (\mathbb{R}^n \setminus S_{\tilde{j}}) \tag{5.1}$$

and, for some positive constants $\eta_0 = \eta_0(j, \tilde{j})$ and $\kappa_0 = \kappa_0(j, \tilde{j})$, one of the following two conditions holds true:

$$\begin{aligned} &v_j + \kappa(u_{\tilde{j}} - \varphi_u(v_j, s_j)) \in U_j \quad \text{whenever} \\ &v_j \in U_j,\ d(v_j, \partial U_j) \le \eta_0,\ s_j \in S_j,\ u_{\tilde{j}} \in U_{\tilde{j}},\ \text{and}\ 0 \le \kappa \le \kappa_0, \end{aligned} \tag{5.2}$$

or

$$\begin{aligned} &v_j - \kappa(u_{\tilde{j}} - \varphi_u(v_j, s_j)) \in U_j \quad \text{whenever} \\ &v_j \in U_j,\ d(v_j, \partial U_j) \le \eta_0,\ s_j \in S_j,\ u_{\tilde{j}} \in U_{\tilde{j}},\ \text{and}\ 0 \le \kappa \le \kappa_0. \end{aligned} \tag{5.3}$$

The definition of the transition graph in section 2 is more restrictive. If $m = n = 1$, then condition (5.1) is equivalent to $\varphi(Q_j) \subset \mathbb{R}^2 \setminus E_{\tilde{j}}$, a weakening of condition (2.1) discussed in Remark 2. Similarly, with $\eta_0 = 1 - \vartheta_0$ and κ_0 suitably chosen (it is enough to make both $\vartheta_0 > 0$ and $\kappa_0 = \kappa_0(\vartheta_0) > 0$ sufficiently small), conditions (5.2) and (5.3) are implied by conditions (2.2) and (2.3), respectively.

With the notion of the transition graph redefined in $\mathbb{R}^m \times \mathbb{R}^n$, $m, n \geq 1$, the wording of Lemma 2.1 in a higher dimension coincides with that of the original Lemma 2.1 verbatim. Now we turn to the proof of this generalization. Conditions (5.2) and (5.3) will be clarified and analyzed later.

Proof of Lemma 2.1 in $\mathbb{R}^m \times \mathbb{R}^n$. The strategy is to rewrite the system of equations

$$x_{k+1} = \varphi(x_k) \quad \text{and} \quad x_k \in Q_{j_k}, \quad k = 0, 1, \ldots, N,$$

as a fixed point equation $(x_0, x_1, \ldots, x_N) = \mathcal{F}(x_0, x_1, \ldots, x_N)$ in the product space $\prod_{k=0}^{N} Q_{j_k} \subset (\mathbb{R}^m \times \mathbb{R}^n)^{N+1}$ and to check that all conditions of Brouwer's fixed point theorem are satisfied.

Choose a positive constant

$$\kappa^* \leq \min_{k=0,1,\ldots,N} \kappa_0(j_k, j_{k+1}) \quad \text{such that} \quad \kappa^* C^* \leq \min_{k=0,1,\ldots,N} \eta_0(j_k, j_{k+1}),$$

where

$$C^* = \max_{k=0,1,\ldots,N} \max\bigl\{ \|u_{k+1} - \varphi_u(x_k)\| \;\big|\; u_{k+1} \in U_{j_{k+1}},\, x_k \in Q_{j_k} \bigr\}.$$

For $(x_0, x_1, \ldots, x_N) \in \prod_{k=0}^{N} Q_{j_k}$, coordinatewise we set

$$(\mathcal{F}(x_0, x_1, \ldots, x_N))_k = (u_k + \varepsilon_k \kappa^* (u_{k+1} - \varphi_u(x_k)), r_{j_k}(\varphi_s(x_{k-1}))) \in \mathbb{R}^m \times \mathbb{R}^n.$$

Here ε_k depends on the pair $(j, \tilde{j}) = (j_k, j_{k+1})$ taking $\varepsilon_k = 1$ if condition (5.2) applies and $\varepsilon_k = -1$ if condition (5.3) applies, where $k = 0, 1, \ldots, N$.

Since $x_{N+1} = x_0$, $x_{-1} = x_N$ by convention, we shift the index values in the $\mathbb{R}^n$-coordinate and see that the fixed point equation $(x_0, x_1, \ldots, x_N) = \mathcal{F}(x_0, x_1, \ldots, x_N)$ in $\prod_{k=0}^{N} Q_{j_k}$ is equivalent to the system of equations

$$u_{k+1} = \varphi_u(x_k) \quad \text{and} \quad s_{k+1} = r_{j_{k+1}}(\varphi_s(x_k)), \quad k = 0, 1, \ldots, N. \tag{5.4}$$

In view of condition (5.1), the first identity in (5.4) implies that $\varphi_s(x_k) \in S_{j_{k+1}}$. Hence $r_{j_{k+1}}(\varphi_s(x_k)) = \varphi_s(x_k)$, and system (5.4) simplifies to

$$u_{k+1} = \varphi_u(x_k) \quad \text{and} \quad s_{k+1} = \varphi_s(x_k), \quad \text{i.e.,} \quad x_{k+1} = \varphi(x_k), \quad k = 0, 1, \ldots, N.$$

It is clear that $\prod_{k=0}^{N} Q_{j_k}$ is a compact topological ball in $(\mathbb{R}^m \times \mathbb{R}^n)^{N+1}$ and $\mathcal{F} : \prod_{k=0}^{N} Q_{j_k} \to (\mathbb{R}^m \times \mathbb{R}^n)^{N+1}$ is a continuous function. Here all that remains is for us to prove that

$$(\mathcal{F}(x_0, x_1, \ldots, x_N))_k \in Q_{j_k} \quad \text{whenever} \quad (x_0, x_1, \ldots, x_N) \in \prod_{k=0}^{N} Q_{j_k},$$

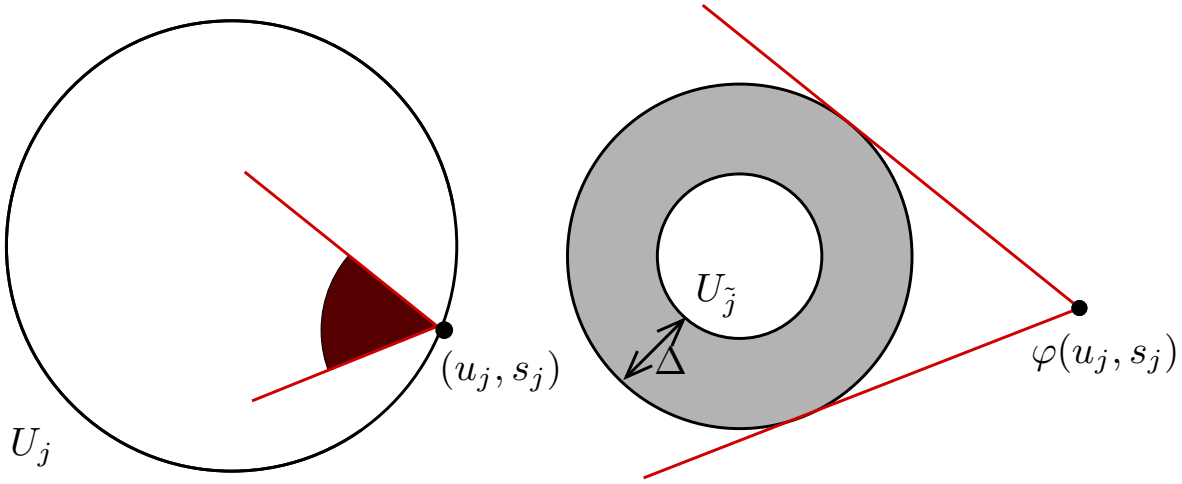


Figure 7. *Condition* (5.6) *for fixed* $(u_j, s_j) \in \partial U_j \times S_j$.

$k = 0, 1, \ldots, N$. Since $r_{j_k}(\varphi_s(x_{k-1})) \in S_{j_k}$, we can go to the $\mathbb{R}^m$-coordinate and just check that

$$u_k + \varepsilon_k \kappa^*(u_{k+1} - \varphi_u(x_k)) \in U_{j_k} \quad \text{if} \quad x_k = (u_k, s_k) \in Q_{j_k} \text{ and } u_{k+1} \in U_{j_{k+1}}. \tag{5.5}$$

If $u_k \in U_{j_k}$ with $d(u_k, \partial U_{j_k}) \le \eta_0(j_k, j_{k+1})$, then—depending on the value of ε_k—(5.5) reduces to (5.2) or (5.3) with $\kappa = \kappa^*$. On the other hand, if $u_k \in U_{j_k}$ with $d(u_k, \partial U_{j_k}) > \eta_0(j_k, j_{k+1})$, then (5.5) follows from the inequality $\kappa^* \|u_{k+1} - \varphi_u(x_k)\| \le \kappa^* C^* \le \eta_0(j_k, j_{k+1})$, $k = 0, 1, \ldots, N$. ■

From a geometric point of view, both condition (5.2) and the alternative condition (5.3) imply that $U_{\tilde{j}}$ is "surrounded by" $\varphi_u(\partial U_j \times S_j)$. In the special case, $U_j = U_{\tilde{j}} = \mathcal{B}^m[0, 1]$ and $S_j = \mathcal{B}^m[0, 1]$ (compact unit balls in the respective Euclidean spaces), so condition (5.2) is a consequence of the inequality

$$\langle \varphi_u(u, s) - \tilde{u},\, u \rangle > 0 \quad \text{whenever } u, \tilde{u} \in \mathbb{R}^m,\ s \in \mathbb{R}^n,\ \|u\| = 1,\ \|\tilde{u}\|, \|s\| \le 1,$$

which resembles certain geometric conditions in various versions of Brouwer's fixed point theorem [51].

The remaining part of this section will be devoted to a technical analysis of conditions (5.2) and (5.3). By a symmetry argument, this analysis reduces to investigating (5.2). Condition (5.2) will be replaced by the slightly stronger condition (5.6), which is stable with respect to small perturbations of φ_u, including numerical approximations with rounding errors. A second advantage of (5.6) over (5.2) is that condition (5.6) can be readily checked. To see this, compare Figure 7 with Figure 6. Overall, condition (5.6) is better suited to computer-assisted proofs than (5.2). The section ends with the somewhat more convenient and transparent condition (5.8), where uniformity with respect to λ is not required.

Proposition 5.1. *Condition* (5.2) *is a consequence of a simpler requirement. It is that there*

exist positive constants $\lambda_0 = \lambda_0(j,\tilde{j})$ *and* $\Delta = \Delta(j,\tilde{j})$ *such that*

$$u_j + \lambda(w_{\tilde{j}} - \varphi_u(u_j, s_j)) \in U_j \quad \text{whenever} \tag{5.6}$$
$$u_j \in \partial U_j,\ s_j \in S_j,\ w_{\tilde{j}} \in \mathcal{B}^m[U_{\tilde{j}}, \Delta],\ \text{and}\ 0 \le \lambda \le \lambda_0.$$

Proof. We omit indices $j, \tilde{j}$ in the following and write $U = U_j$, $S = S_j$, and $W = U_{\tilde{j}}$.

Now suppose that condition (5.6) is satisfied but (5.2) is not. Then there are sequences $\{v_\ell\} \subset U$, $\{s_\ell\} \subset S$, $\{w_\ell\} \subset W$, $\{\kappa_\ell\} \subset \mathbb{R}^+$, which have the following properties:

$$p_\ell = v_\ell + \kappa_\ell(w_\ell - \varphi_u(v_\ell, s_\ell)) \notin U \quad \text{for} \quad \ell = 1, 2, \dots \tag{5.7}$$

and both $v_\ell \to \partial U$ and $\kappa_\ell \to 0$ as $\ell \to \infty$.

Since $v_\ell \in U$ and $p_\ell \notin U$, there exists a $\kappa_\ell^* \in [0, \kappa_\ell)$ such that

$$z_\ell = v_\ell + \kappa_\ell^*(w_\ell - \varphi_u(v_\ell, s_\ell)) \in \partial U \quad \text{for} \quad \ell = 1, 2, \dots.$$

With the construction, $0 < \kappa_\ell - \kappa_\ell^* \le \lambda_0$, and (by using the uniform continuity of mapping φ_u on the compact set $U \times S$) $\|\varphi_u(z_\ell, s_\ell) - \varphi_u(v_\ell, s_\ell)\| \le \Delta$ for ℓ large enough. In view of condition (5.6), we conclude that

$$p_\ell = z_\ell + (\kappa_\ell - \kappa_\ell^*)[(w_\ell + \varphi_u(z_\ell, s_\ell) - \varphi_u(v_\ell, s_\ell)) - \varphi_u(z_\ell, s_\ell)] \in U$$

for large enough ℓ, which contradicts (5.7). ∎

Proposition 5.2. *Actually, condition* (5.6) *is a consequence of a simpler requirement. It is that there exists a positive constant* $\delta = \delta(j,\tilde{j})$ *such that*

$$u_j + \mu(w_{\tilde{j}} - \varphi_u(u_j, s_j)) \in \operatorname{int}(U_j) \quad \text{whenever} \tag{5.8}$$
$$u_j \in \partial U_j,\ s_j \in S_j,\ w_{\tilde{j}} \in \mathcal{B}^m[U_{\tilde{j}}, \delta],\ \text{and}\ 0 < \mu \le \mu_0\ \text{with some}\ \mu_0 = \mu_0(u_j, s_j, w_{\tilde{j}}).$$

Proof. As before, we write $U = U_j$, $S = S_j$, and $W = U_{\tilde{j}}$.

Fix $u^* \in \partial U$, $s^* \in S$, and $w^* \in W$. By compactness, it is sufficient to demonstrate the existence of two positive constants $\tau = \tau(u^*, s^*, w^*)$ and $\lambda^* = \lambda^*(u^*, s^*, w^*)$ such that, given $u \in \partial U$, $s \in S$, and $w \in \mathcal{B}^m[W, \delta]$ with $\|u - u^*\| \le \tau$, $\|s - s^*\| \le \tau$, $\|w - w^*\| \le \tau$, the following holds true:

$$u + \lambda(w - \varphi_u(u, s)) \in U \quad \text{whenever} \quad 0 \le \lambda \le \lambda^*.$$

By continuity, there is a $\sigma \in (0, \delta)$ such that, for arbitrary $w \in \mathcal{B}^m[w^*, \sigma]$ and $q \in U \cap \mathcal{B}^m[u^*, \sigma]$,

$$\tilde{w} - \varphi_u(q, s^*) = w - \varphi_u(u^*, s^*) \quad \text{for some} \quad \tilde{w} \in \mathcal{B}^m[w^*, \delta]. \tag{5.9}$$

In view of condition (5.8) with $(u^*, s^*, w^*) = (u_j, s_j, w_{\tilde{j}})$, we may assume that

$$u^* + \alpha_+^*(w^* - \varphi_u(u^*, s^*)) \in \operatorname{int}(U) \cap \partial\mathcal{B}^m[u^*, \sigma] \quad \text{for some} \quad \alpha_+^* > 0.$$

As a corollary, a simple geometric argument implies the existence of a constant $\eta \in (0, \sigma)$ such that, for arbitrary $p \in \mathcal{B}^m[u^*, \eta]$ and $w \in \mathcal{B}^m[w^*, \eta]$,

$$p + \alpha_+(w - \varphi_u(u^*, s^*)) \in \operatorname{int}(U) \cap \partial\mathcal{B}^m[u^*, \sigma] \quad \text{for some} \quad \alpha_+ = \alpha_+(p, w) > 0,$$

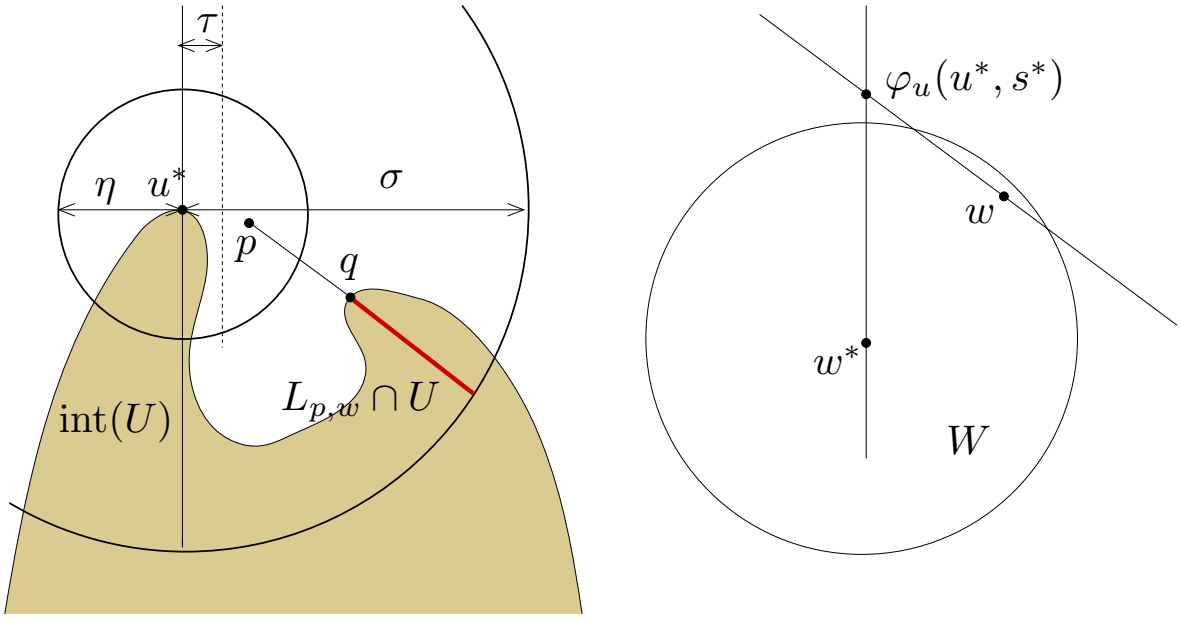


Figure 8. *Illustration of the proof of Proposition 5.2.*

where α_+ is unique, the function $(p, w) \to \alpha_+(p, w)$ is continuous, and $\alpha_+(u^*, w^*) = \alpha_+^*$. For future use, we note that

$$\alpha_* = \inf\{\alpha_+(p, w) \mid p \in \mathcal{B}^m[u^*, \eta],\ w \in \mathcal{B}^m[w^*, \eta]\} > 0$$

by compactness. Now consider the straight line segment

$$L_{p,w} = \{p + \lambda(w - \varphi_u(u^*, s^*)) \mid \lambda \geq 0\} \cap \mathcal{B}^m[u^*, \sigma],$$

and assume that $q = p + \gamma_0(w - \varphi_u(u^*, s^*)) \in \partial U \cap L_{p,w}$ for some $\gamma_0 < \alpha_+$. Applying property (5.9), condition (5.8) (with $(q, s^*, \tilde{w}) = (u_j, s_j, w_{\tilde{j}})$) implies that

$$p + \gamma(w - \varphi_u(u^*, s^*)) = q + (\gamma - \gamma_0)(\tilde{w} - \varphi_u(q, s^*)) \in \operatorname{int}(U)$$

for $\gamma > \gamma_0$, where $|\gamma - \gamma_0|$ is small. By an elementary connectedness argument in one dimension, we infer that $L_{p,w} \cap U$ is a compact interval with an endpoint on $\partial\mathcal{B}^m[u^*, \sigma]$. See Figure 8.

Similarly, observe that there exists a constant $\tau \in (0, \eta)$ such that, for arbitrary $w \in \mathcal{B}^m[w^*, \tau]$, $u \in U \cap \mathcal{B}^m[u^*, \tau]$, and $s \in S \cap \mathcal{B}^n[s^*, \tau]$,

$$\hat{w} - \varphi_u(u^*, s^*) = w - \varphi_u(u, s) \quad \text{for some} \quad \hat{w} \in \mathcal{B}^m[w^*, \eta]. \tag{5.10}$$

If, in particular, $u \in \partial U \cap \mathcal{B}^m[u^*, \tau]$, $s \in S \cap \mathcal{B}^n[s^*, \tau]$, and $w \in \mathcal{B}^m[w^*, \tau]$, then by property (5.10)

$$\{u + \lambda(w - \varphi_u(u, s)) \mid \lambda \geq 0\} \cap \mathcal{B}^m[u^*, \sigma] = L_{u,\hat{w}}$$

for some $\hat{w} \in \mathcal{B}^m[w^*, \eta]$. Since $u \in \partial U \subset U$ and $L_{u,\hat{w}} \cap U$ is a compact interval with an endpoint on $\partial\mathcal{B}^m[u^*, \sigma]$, we conclude that $L_{u,\hat{w}} \subset U$. Hence $u + \lambda(w - \varphi_u(u, s)) \in U$ for $0 \leq \lambda \leq \alpha_+(u, \hat{w})$, and thus $\lambda^* = \lambda^*(u^*, s^*, w^*)$ can be chosen for $\alpha_* > 0$. ∎

We do not know whether $\mathrm{int}(U_j)$ in (5.8) can actually be replaced by U_j. On the other hand, simple examples confirm that Proposition 5.1 does not hold true for $\Delta = 0$.

6. Lemma 2.1 and a recent four-dimensional example of Yang and Li [50]. As we mentioned earlier, conditions (5.2) and (5.3) can be readily checked for $m = 1$, but they are more complicated for $m > 1$. Regardless of the value of the positive integer n, condition (5.1) remains rather simple. It follows that for small, multidimensional perturbations of one-dimensional mappings which "contract" in the new directions, the $m = 1$, $n \geq 1$ case of Lemma 2.1 can be applied without difficulty. For example, Lemma 2.1 can be applied for the family of mappings investigated in [54] and simplifies the proofs therein.

As for the $m > 1$ case, it is reasonable to suppose that a twofold application of Lemma 2.1 leads to a rigorous proof of the existence of chaotic behavior in a recent four-dimensional neural network example of Yang and Li [50]. Our conjecture is supported by analyzing the figures in that paper.

We will now consider the autonomous system of ordinary differential equations [50]

$$\begin{aligned}
\dot{x}_1 &= -x_1 + 2.10f(x_1) + 2.50f(x_2),\\
\dot{x}_2 &= -x_2 - 2.60f(x_1) + 1.00f(x_2) + 3.00f(x_3),\\
\dot{x}_3 &= -x_3 - 2.80f(x_2) + 0.50f(x_3) - 1.10f(x_4),\\
\dot{x}_4 &= -100x_4 + 100f(x_3) + 160f(x_4),
\end{aligned} \tag{6.1}$$

which models a cellular neural network of Chua–Roska type [8] with $f(x_i) = 2^{-1}(|x_i + 1| - |x_i - 1|)$, $x_i \in \mathbb{R}$, $i = 1, 2, 3, 4$. A horseshoe in an appropriate Poincaré mapping Π was found by Yang and Li [50] numerically, via nonrigorous computation. Their paper does not say how the 14 coefficients/weights on the right-hand side of the above system of ordinary differential equations were chosen. The nice Figure 4 in [50] suggests that the successful Poincaré section was chosen by a trial-and-error process with human overheads.

Now we would like to show that, with the underlying sets properly chosen, the transition graph $\mathcal{G}(\Pi)$ is the complete directed graph on two vertices. The argument will be based on the case $m = 2$, $n = 1$ of the higher-dimensional generalization of Lemma 2.1 and, of course, on the geometric information contained in [50].

Simplified and schematic variants of Figures 7, 5, and 8 of [50] are presented here as Figures 9, 10, and 11, respectively. The two vertical prisms with quadrilateral bases in [50] correspond to our cylinders $\mathcal{C}_U = U \times S$ and $\mathcal{C}_W = W \times S$, while the vertical edges of the prisms correspond to the points A_ℓ and B_ℓ, $\ell = 1, 2, 3, 4$, respectively. The prisms are strongly contracted in the vertical direction. As for the two horizontal directions within, mapping Π is a modest expansion. Applied to our situation, the crucial observation in [50] is that vertical segments on the jacket of $\mathcal{C}_U$ and of $\mathcal{C}_W$ (i.e., segments of the form $\{A\} \times S$ and $\{B\} \times S$ with $A \in \partial U$ and $B \in \partial W$) are mapped onto "almost vertical" curves on $\partial\Pi(\mathcal{C}_U)$ and on $\partial\Pi(\mathcal{C}_W)$, respectively. This explains why $\Pi(\mathcal{C}_U)$ and $\Pi(\mathcal{C}_W)$ can be regarded as cylinders and implies that condition (5.6) or its alternative counterpart (i.e., there exist positive constants λ_0 and Δ such that

$$\begin{aligned}
&u_j - \lambda(w_{\tilde{j}} - \varphi_u(u_j, s_j)) \in U_j \quad \text{whenever}\\
&u_j \in \partial U_j,\ s_j \in S_j,\ w_{\tilde{j}} \in \mathcal{B}^m[U_{\tilde{j}}, \Delta],\ \text{and}\ 0 \leq \lambda \leq \lambda_0)
\end{aligned} \tag{6.2}$$

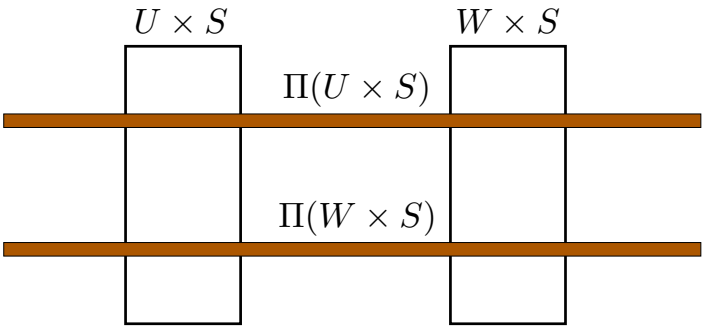


Figure 9. *The front view of the four cylinders* $\mathcal{C}_U = U \times S$, $\mathcal{C}_W = W \times S$, $\Pi(\mathcal{C}_U)$, *and* $\Pi(\mathcal{C}_W)$.

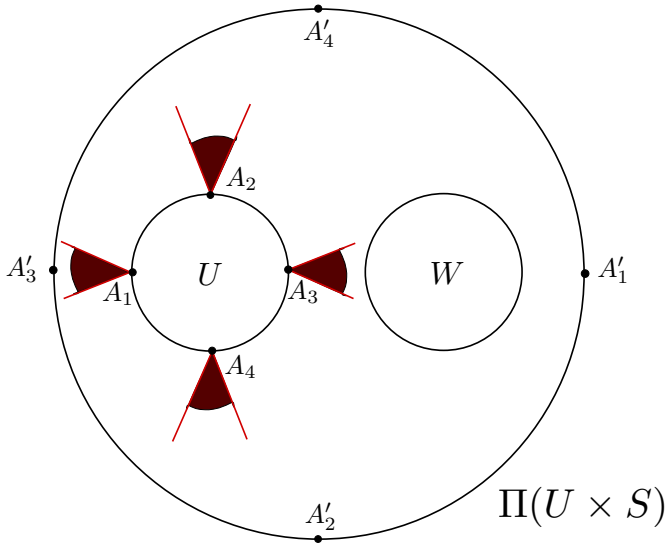


Figure 10. *The upper view of the cylinders* $\mathcal{C}_U = U \times S$, $\Pi(\mathcal{C}_U)$, *and* $\mathcal{C}_W = W \times S$.

has to be checked only for a single $s = s^* \in S$. There is no loss of generality in assuming that $S = [-1, 1] \subset \mathbb{R}$ and $s^* = 0$. For brevity, we write $A'_\ell = \Pi_u(A_\ell, 0)$ and $B'_\ell = \Pi_u(B_\ell, 0)$, $\ell = 1, 2, 3, 4$. The relative position of the four cylinders and the 2×8 special points in Figures 9, 10, and 11 are exactly like the computer pictures in [50].

In what follows we will show that Lemma 2.1 applies in this situation. We do this by examining if and how, with $U_j = U, W$ and $U_{\tilde{j}} = U, W$, the alternative pair of conditions (5.6) and (6.2) is satisfied. The final result will be that, with vertex set $\mathbf{V}(\mathcal{G}) = \{C_U, C_W\}$, the edge set of the transition graph $\mathcal{G}(\Pi)$ is $\mathbf{E}(\mathcal{G}) = \{(C_U, C_U), (C_U, C_W), (C_W, C_U), (C_W, C_W)\}$.

For a fixed $\ell \in \{1, 2, 3, 4\}$, the angular sector at A_ℓ in Figure 10 describes the two cones

$$\{A_\ell + \lambda(u - A'_\ell) \in \mathbb{R}^2 \mid u \in \mathcal{B}^2[U, \Delta] \text{ and } 0 \le \lambda \le \lambda_0\}$$

and

$$\{A_\ell + \lambda(w - A'_\ell) \in \mathbb{R}^2 \mid w \in \mathcal{B}^2[W, \Delta] \text{ and } 0 \le \lambda \le \lambda_0\}.$$

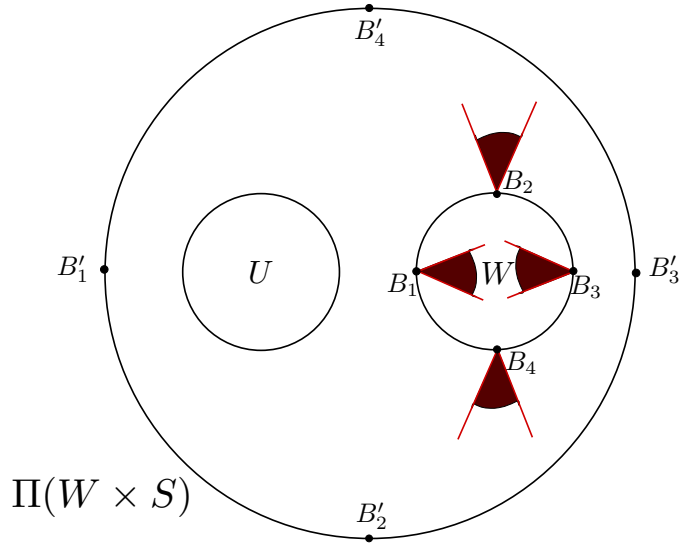


Figure 11. *The upper view of the cylinders $\mathcal{C}_W = W \times S$, $\Pi(\mathcal{C}_W)$, and $\mathcal{C}_U = U \times S$.*

(For the sake of simplicity, these two cones have been drawn in Figure 10 as a single angular sector with vertex A_ℓ for every ℓ. And, like all of figures in this section, the small positive constants Δ and λ_0 remain unspecified.) Based on the direction of these angular sectors, it seems plausible that condition (6.2) is satisfied for $\varphi = \Pi$, $S_j = [-1, 1]$, $U_j = U$, and $U_{\tilde{j}} = U, W$. Regarding the proof of Lemma 2.1, Proposition 5.1 implies that condition (5.5) is satisfied for $\varphi = \Pi$, $S_{j_k} = [-1, 1]$, $U_{j_k} = U$, $\varepsilon_k = -1$, and $U_{j_{k+1}} = U, W$.

Similarly, for $\ell = 1, 2, 3, 4$, the angular sector at B_ℓ in Figure 11 describes the two cones (or the union of the two cones)

$$\{B_\ell + \lambda(u - B'_\ell) \in \mathbb{R}^2 \mid u \in \mathcal{B}^2[U, \Delta] \text{ and } 0 \le \lambda \le \lambda_0\}$$

and

$$\{B_\ell + \lambda(w - B'_\ell) \in \mathbb{R}^2 \mid w \in \mathcal{B}^2[W, \Delta] \text{ and } 0 \le \lambda \le \lambda_0\}.$$

Unfortunately, neither condition (5.6) nor (6.2) is satisfied, and, if left unchanged, the proof of Lemma 2.1 breaks down in the present situation. However, the direction of the four angular sectors in Figure 11 suggests a simple way out.

Together with a combinatorial modification, the proof of Lemma 2.1 still holds true. The $\mathbb{R}^m = \mathbb{R}^2$ coordinate

$$u_k + \varepsilon_k \kappa^*(u_{k+1} - \varphi_u(x_k)) = u_k + \varepsilon_k \kappa^*(u_{k+1} - \Pi_u(x_k)) \in \mathbb{R}^2$$

of $(\mathcal{F}(x_0, x_1, \dots, x_N))_k$, $k = 0, 1, \dots, N$, is to be replaced by

$$(u^1_k + \varepsilon^1_k \kappa^*(u^1_{k+1} - \Pi^1_u(x_k)),\ u^2_k + \varepsilon^2_k \kappa^*(u^2_{k+1} - \Pi^2_u(x_k))) \in \mathbb{R} \times \mathbb{R},$$

where superscript 1 (resp., 2) stands for the first (= horizontal) (resp., second (= vertical)) coordinate in Figures 10 and 11, and

$$\varepsilon^1_k = \varepsilon^2_k = -1 \qquad \text{if} \quad U_{j_k} = U \quad \text{and} \quad U_{j_{k+1}} = U, W,$$

and

$$\varepsilon_k^1 = 1,\ \varepsilon_k^2 = -1 \qquad \text{if} \quad U_{j_k} = W \quad \text{and} \quad U_{j_{k+1}} = U, W.$$

No other changes are needed for the proof. After this refinement of the choice of parameter ε_k, condition (5.5) will be satisfied again.

We feel justified in concluding that, eventually, the argument outlined above leads to a rigorous proof for the existence of embedded Σ_2 dynamics in (6.1). At present, several details are missing. It is not enough to check the alternative conditions for two times four points in a simplified and schematic situation. The relation between the original dynamics creating Figures 7, 5, and 8 of [50] and its simplified representation in Figures 9, 10, and 11 has to be analyzed rigorously. This task is parallel to the one we performed in section 2 for Hubbard's forced damped pendulum equation (1.1).

Acknowledgment. The authors are grateful for the suggestions and comments of the referees that helped improve the paper.

REFERENCES

[1] G. Alefeld and G. Mayer, *Interval analysis: Theory and applications*, J. Comput. Appl. Math., 121 (2000), pp. 421–464.
[2] V. M. Alekseev, *On the capture orbits for the three-body problem for negative energy constant*, Uspekhi Mat. Nauk, 24 (1969), pp. 185–186 (in Russian).
[3] K. T. Alligood, T. D. Sauer, and J. A. Yorke, *Chaos. An Introduction to Dynamical Systems*, Springer-Verlag, Berlin, 1997.
[4] B. Bánhelyi, T. Csendes, and B. M. Garay, *Optimization and the Miranda approach in detecting horseshoe-type chaos by computer*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 735–747.
[5] B. Bánhelyi, T. Csendes, and B. M. Garay, *Σ_2-chaos for iterates of the classical Hénon mapping*, in preparation, 2008.
[6] E. Bosetto and E. Serra, *A variational approach to chaotic dynamics in periodically forced nonlinear oscillators*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 673–709.
[7] A. Capietto, W. Dambrosio, and D. Papini, *Superlinear indefinite equations on the real line and chaotic dynamics*, J. Differential Equations, 181 (2002), pp. 419–438.
[8] L. O. Chua and T. Roska, *Cellular Neural Networks and Visual Computing*, Cambridge University Press, Cambridge, UK, 2002.
[9] B. A. Coomes, H. Kocak, and K. J. Palmer, *Homoclinic shadowing*, J. Dynam. Differential Equations, 17 (2005), pp. 175–215.
[10] T. Csendes, B. M. Garay, and B. Bánhelyi, *A verified optimization technique to locate chaotic regions of a Hénon system*, J. Global Optim., 35 (2006), pp. 145–160.
[11] S. Day, R. Frongillo, and R. Trevino, *Algorithms for rigorous entropy bounds and symbolic dynamics*, submitted.
[12] M. Dellnitz, G. Froyland, and O. Junge, *The algorithms behind GAIO-set oriented numerical methods for dynamical systems*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, B. Fiedler, ed., Springer-Verlag, Berlin, 2001, pp. 145–174.
[13] M. Dellnitz and O. Junge, *Set oriented numerical methods for dynamical systems*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 221–264.
[14] V. Franceschini and L. Russo, *Stable and unstable manifolds of the Hénon mapping*, J. Statist. Phys., 25 (1981), pp. 757–769.
[15] A. Frommer and B. Lang, *Existence tests for solutions of nonlinear equations using Borsuk's theorem*, SIAM J. Numer. Anal., 43 (2005), pp. 1348–1361.
[16] A. Frommer, B. Lang, and M. Schnurr, *A comparison of the Moore and Miranda existence tests*, Computing, 72 (2004), pp. 349–354.

[17] M. Furi, M. Martelli, M. O'Neil, and C. Staples, *Chaotic orbits of a pendulum with variable length*, Electron. J. Differential Equations, (2004), no. 36.
[18] Z. Galias, *Positive topological entropy of Chua's circuit: A computer-assisted proof*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 7 (1997), pp. 331–349.
[19] Z. Galias, *Obtaining rigorous bounds for topological entropy for discrete time dynamical systems*, in Proceedings of the International Symposium on Nonlinear Theory and Its Applications (Xi'an, China), 2002, pp. 619–622.
[20] Z. Galias and P. Zgliczynski, *Abundance of homoclinic and heteroclinic orbits and rigorous bounds for the topological entropy for the Hénon map*, Nonlinearity, 14 (2001), pp. 909–932.
[21] M. Gidea and P. Zgliczynski, *Covering relations for multidimensional dynamical systems*, J. Differential Equations, 202 (2004), pp. 32–58.
[22] B. Hassard, B. Zhang, S. P. Hastings, and W. C. Troy, *A computer proof that the Lorenz equations have "chaotic" solutions*, Appl. Math. Lett., 7 (1994), pp. 79–83.
[23] S. P. Hastings and J. B. McLeod, *Chaotic motion of a pendulum with oscillatory forcing*, Amer. Math. Monthly, 100 (1993), pp. 563–572.
[24] J. H. Hubbard, *The forced damped pendulum: Chaos, complication and control*, Amer. Math. Monthly, 106 (1999), pp. 741–758.
[25] J. A. Kennedy, S. Kocak, and J. A. Yorke, *A chaos lemma*, Amer. Math. Monthly, 108 (2001), pp. 411–423.
[26] U. Kirchgraber and D. Stoffer, *On the definition of chaos*, Z. Angew. Math. Mech., 69 (1989), pp. 175–185.
[27] O. Knüppel, *PROFIL/BIAS—a fast interval library*, Computing, 53 (1994), pp. 277–287.
[28] M. Levi, *Qualitative analysis of the periodically forced relaxation oscillations*, Mem. Amer. Math. Soc., 32 (244) (1981).
[29] S. Luzzatto, I. Melbourne, and F. Paccaut, *The Lorenz attractor is mixing*, Comm. Math. Phys., 260 (2005), pp. 393–401.
[30] K. Mischaikow and M. Mrozek, *Chaos in the Lorenz equations: A computer-assisted proof*, Bull. Amer. Math. Soc., 32 (1995), pp. 66–72.
[31] M. Misiurewicz and B. Szewc, *Existence of a homoclinic point for the Hénon mapping*, Comm. Math. Phys., 75 (1980), pp. 285–291.
[32] N. S. Nedialkov, *VNODE—A Validated Solver for Initial Value Problems for Ordinary Differential Equations*, http://www.cas.mcmaster.ca/~nedialk/Software/VNODE/VNODE.shtml (2001).
[33] N. S. Nedialkov, K. R. Jackson, and G. F. Corliss, *Validated solutions of initial value problems for ordinary differential equations*, Appl. Math. Comput., 105 (1999), pp. 21–68.
[34] D. Papini and F. Zanolin, *Fixed points, periodic points, and coin-tossing sequences for mappings defined on two-dimensional cells*, Fixed Point Theory Appl., 2 (2004), pp. 113–134.
[35] D. Papini and F. Zanolin, *Some results on periodic points and chaotic dynamics arising from the study of the nonlinear Hill equations*, Rend. Semin. Mat. Univ. Politec. Torino, 65 (2007), pp. 115–157.
[36] L. C. Piccinini, G. Stampaccia, and G. Vidossich, *Ordinary Differential Equations in* $\mathbb{R}^n$, Springer-Verlag, Berlin, 1984.
[37] M. Pireddu and F. Zanolin, *Fixed points for dissipative-repulsive systems and topological dynamics of mappings defined on N-dimensional cells*, Adv. Nonlinear Stud., 5 (2005), pp. 411–440.
[38] A. Pokrovskii, O. Rasskazov, and D. Visetti, *Homoclinic trajectories and chaotic behaviour in a piecewise linear oscillator*, Discrete Contin. Dyn. Syst. Ser. B, 8 (2007), pp. 943–970.
[39] H. Ratschek and J. Rokne, *New Computer Methods for Global Optimization*, Ellis Horwood, Chichester, UK, 1988.
[40] C. Robinson, *Dynamical Systems. Stability, Symbolic Dynamics, and Chaos*, 2nd ed., CRC Press, Boca Raton, FL, 1999.
[41] J. Rubin and D. Terman, *Geometric singular perturbation analysis of neuronal dynamics*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 93–146.
[42] M. Schnurr, *On the proofs of some statements concerning the theorems of Kantorovich, Moore and Miranda*, Reliab. Comput., 11 (2005), pp. 77–85.
[43] S. Smale, *The Mathematics of Time*, Springer-Verlag, Berlin, 1980.
[44] D. Stoffer and K. J. Palmer, *Rigorous verification of chaotic behaviour of maps using validated shadowing*, Nonlinearity, 12 (1999), pp. 1683–1689.

[45] S. Terracini and G. Verzini, *Solutions of prescribed number of zeroes to a class of superlinear ODE's systems*, NoDEA Nonlinear Differential Equations Appl., 8 (2001), pp. 323–341.
[46] W. Tucker, *A rigorous ODE solver and Smale's* 14*th problem*, Found. Comput. Math., 2 (2002), pp. 53–117.
[47] X. S. Yang and Q. Li, *A computer-assisted proof of chaos in Josephson junctions*, Chaos Solitons Fractals, 27 (2006), pp. 25–30.
[48] S. Wiggins, *On the detection and dynamical consequences of orbits homoclinic to hyperbolic periodic orbits and normally hyperbolic invariant tori in a class of ordinary differential equations*, SIAM J. Appl. Math., 48 (1988), pp. 262–285.
[49] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, Berlin, 2003.
[50] X. S. Yang and Q. Li, *A horseshoe in a cellular neural network of four–dimensional autonomous ordinary differential equations*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 3211–3218.
[51] E. Zeidler, *Nonlinear Functional Analysis and Its Applications.* I. *Fixed-Point Theorems*, Springer-Verlag, Berlin, 1986.
[52] P. Zgliczynski, *Fixed point index for iterations of maps, topological horseshoe and chaos*, Topol. Methods Nonlinear Anal., 8 (1996), pp. 169–177.
[53] P. Zgliczynski, *Computer assisted proof of chaos in the Rössler equations and in the Hénon map*, Nonlinearity, 10 (1997), pp. 243–252.
[54] P. Zgliczynski, *Multidimensional perturbations of one-dimensional maps and stability of Sharkovski ordering*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 9 (1999), pp. 1867–1876.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 868–894
© 2008 Society for Industrial and Applied Mathematics

A Method for Determining Most Probable Errors in Nonlinear Lightwave Systems*

Elaine T. Spiller† and William L. Kath‡

Abstract. We present a method for computing bit-error ratios in soliton-based lightwave systems. The method uses soliton perturbation theory and calculus of variations to find approximate versions of the most probable paths through sample space leading to errors, combined with importance-sampled Monte Carlo simulations of the full set of equations around these approximate paths to compute the actual error rates. As a specific example, the method is applied to a differential phase-shift-keyed lightwave system. For this example the method not only computes the bit-error ratio but also predicts the set of failure modes leading to large pulse distortions, thus illuminating the specific manner in which errors occur.

Key words. solitons, perturbation theory, calculus of variations, rare events, importance sampling, lightwave systems, phase-shift keying, bit-error ratio

AMS subject classifications. 35Q51, 37K40, 65C05, 65C30, 65C50, 78A60

DOI. 10.1137/070708123

1. Introduction.

In-line optical amplification and nonlinear propagation are two of the primary features of modern long-distance lightwave communication systems. Optical amplification adds spontaneous emission noise to the signal [1], and nonlinearities present in the system can distort the combined signal plus noise during propagation [2], causing the output statistics to differ significantly from Gaussian random variables. Because the ultimate performance of such communication systems is often limited by noise, it is desirable to be able to predict how often errors occur. This is a challenging task, both because of the complicated interaction between signal and noise and because system designers typically require that errors be *rare events*, e.g., a bit-error ratio of 1 error per 10^9 bits or more.

Monte Carlo methods are one standard way to compute error probabilities, but doing so for such rare events is, of course, beyond the capability of standard methods. Recently, however, the application of both importance-sampled Monte Carlo [3, 4, 5, 6] and multicanonical Monte Carlo methods [7, 8, 9, 10, 11, 12] has demonstrated that it is possible to successfully overcome the limitations of standard Monte Carlo simulations for studying rare events in lightwave communication systems. Other methods are possible, of course [13, 14], and the development of new techniques is a subject of current active research. A common theme of

*Received by the editors November 11, 2007; accepted for publication (in revised form) by B. Sandstede April 1, 2008; published electronically July 25, 2008. This work was supported by the National Science Foundation (DMS-0406513 and DMS-0709070) and by the Air Force Office of Scientific Research (FA9550-04-1-0289).
http://www.siam.org/journals/siads/7-3/70812.html
†Department of Engineering Sciences and Applied Mathematics, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3125. Current address: Statistical and Applied Mathematical Sciences Institute, 19 T. W. Alexander Drive, P. O. Box 14006, Research Triangle Park, NC 27709-4006 (espiller@samsi.info).
‡Department of Engineering Sciences and Applied Mathematics, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3125, and Northwestern Institute on Complex Systems, 600 Foster Street, Evanston, IL 60208-4057 (kath@northwestern.edu).

such methods is the biasing of the noise used in the simulations so that samples from the region(s) of state space where the sought-after rare events (i.e., errors) occur arise much more frequently than they would normally. Once one accounts for the biasing, one can efficiently compute probability distribution functions (pdfs) that are accurate far down into the tails. Multicanonical Monte Carlo employs an iterative numerical procedure to determine the biasing toward the desired regions of state space, while importance-sampled Monte Carlo relies on physical or mathematical insight to accomplish this task.

A challenge associated with implementing importance sampling is in choosing a biasing distribution that is close to optimal, i.e., a biasing that *efficiently* samples from the regions of state space where the events of interest are most likely to occur [15]. Previously we have presented numerical results demonstrating that this can be done to simulate noise-induced transmission impairments in phase-modulated communication systems [16]. The purpose of this paper is to describe in detail the methods used to produce such numerical simulations, specifically, how the structure of the nonlinear Schrödinger (NLS) equation and results from approximate methods (in this case, soliton perturbation theory) can be used to guide importance-sampled Monte Carlo simulations of rare events in phase-modulated lightwave communications systems. Essentially, the search for an efficient biasing distribution breaks down into two optimization problems: At an individual amplifier, what is the most likely noise realization that will lead to a specified parameter change? And, over the entire transmission length, what combination of parameter changes at each amplifier is most likely to lead to large signal distortions and, ultimately, errors? The resolution of these questions not only allows the error rate to be computed but also provides insight into how errors arise.

This paper is organized as follows: in section 2 we present the NLS equation and review soliton perturbation theory. We also give an overview of some modulation formats, in particular, the manner in which information is encoded onto a uniform sequence of optical pulses. In section 3 we discuss some background concerning importance-sampled Monte Carlo simulations and present the solution to the optimal biasing problem at one amplifier. In section 3.2 we describe the full optimal biasing problem for a specific modulation format known as differential phase-shift keying (DPSK). The biasing problem is formulated here using a continuous approximation of the governing evolution equations and is thus an extension of previous versions which employed discrete approximations [3, 4]. This new version of the biasing problem allows more complicated problems to be treated, since they can be solved via calculus of variations. The resulting boundary value problem yields multiple solutions which we describe in detail and which we interpret as possible error modes. In section 5 we discuss the implementation of the importance-sampled Monte Carlo method for the differential phase-shift-keyed problem and present the results of simulations.

2. Formulation. A basic schematic of the differential phase-shift-keyed transmission system to be analyzed and simulated is included in Figure 1. In general terms, the system consists of a long length of fiber with a transmitter at the beginning and a receiver at the end [17]. The components of the transmitter are a laser acting as a source of pulses and a Mach–Zehnder modulator to encode data upon them. This data stream is then launched into the optical fiber. Because the signal experiences loss as it propagates, a gain element is usually inserted every few tens of kilometers to compensate for the loss [17]. Thus, the combination of propagation with loss followed by gain typically must be repeated a number of times to

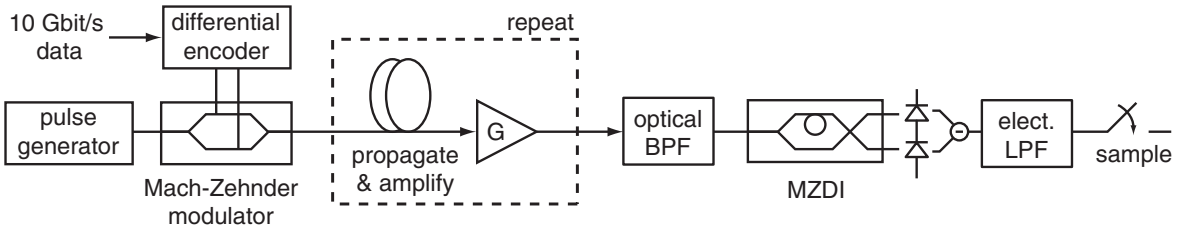


Figure 1. *Basic schematic of a differential phase-shift-keyed transmission system. For a specific description of the elements, please see the text.*

achieve the desired total optical transmission distance. A final set of components filters out noise and recovers the signal; here the elements of the receiver are an optical band-pass filter, a Mach–Zehnder delay interferometer, an electrical low-pass filter, and an electrical sampler.

2.1. Model of a soliton-based lightwave system. The longest part of the system is the optical fiber with periodic amplification. We will model this propagation using the NLS equation with periodically added noise [2]:

$$\frac{\partial u}{\partial z} = \frac{i}{2}\frac{\partial^2 u}{\partial t^2} + i|u|^2 u + \sum_{k=1}^{N_{\text{amp}}} n_k(t)\delta(z - kz_{\text{a}}). \tag{2.1}$$

Here u represents the optical signal's electromagnetic field envelope, z and t are dimensionless distance and retarded time, N_{amp} is the total number of amplifiers, and z_{a} is the dimensionless distance between them. As is customary in the optics literature, z is considered to be the propagation variable, and the "initial condition" (i.e., input signal) is the time-varying electric field envelope profile defined at the beginning of the fiber span, $z = 0$. The model also includes white Gaussian noise added by each amplifier,

$$\begin{aligned} \langle n_i(t)\rangle &= 0, \\ \langle n_i(t) n_j^*(t')\rangle &= \sigma^2 \delta_{ij}\delta(t - t'). \end{aligned} \tag{2.2}$$

This accounts for the effect of photons which are spontaneously emitted at each amplifier and which subsequently experience gain along with the signal (hence the name amplified spontaneous emission (ASE) noise [1]). Here * represents the complex conjugate, and σ^2 is the noise variance; physically, $\sigma^2 = (G-1)^2 \eta_{\text{sp}} T_{\text{w}} \gamma / D\, G \ln G$, where G is the amplifier (power) gain, η_{sp} is the amplifier's spontaneous emission factor, T_{w} is the pulse width, and γ and D are the fiber's nonlinear and dispersion coefficients [2]. The assumption of delta-correlated noise is not physically realistic because as written it contains infinite power, so this should be interpreted merely as a shorthand for the case where the noise spectrum has a bandwidth much larger than that of the signal [1]. Note that in (2.1) we have already averaged out the deterministic power fluctuations caused by loss and amplification [18] in order to more easily focus on the detrimental effects of ASE noise.

In the absence of noise, (2.1) has the well-known soliton solution given by [19]

$$u(t, z) = A \operatorname{sech}(A[t - T]) \exp(i\Omega[t - T] + i\varphi). \tag{2.3}$$

Here A, T, Ω, and φ are, respectively, the amplitude, position, frequency, and phase of the pulse. In the absence of noise A and Ω are constant and T and φ evolve according to $dT/dz = \Omega$ and $d\varphi/dz = (A^2 + \Omega^2)/2$. Note that the parameters in this solution arise from the invariances of the NLS equation [20], and so if a perturbation such as noise alters one or more of these parameters, the solution with changed parameters is a perfectly valid solution of the NLS equation. Thus, it is precisely because of these invariances that noise-induced perturbations can build into large deviations over long propagation distances.

2.2. Differential phase-shift keying (DPSK). A modulation format is the method by which information is encoded on an otherwise periodic optical signal. On-off keying (OOK), i.e., sending a single pulse of light to represent a logical "1" and none for a logical "0," is one traditional encoding format used in lightwave communication systems. An alternative method is phase-shift keying, where information is encoded on the optical signal's phase while the amplitude remains constant [21, 22, 23]. Differential phase-shift keying is a slight variant of this, in which information is encoded on the *phase difference* between two adjacent pulses. We will consider a binary differential phase-shift-keyed system where the transmitter consists of a source that produces an identical sequence of pulses followed by a modulator that encodes no phase difference between pulses to represent a logical "1" and a π phase difference to represent a logical "0." (The reverse could also be used, of course.)

At the receiver in a differential phase-shift-keyed system, the signal is detected, and adjacent bit slots[1] are compared to determine if the received signal represents a "1" or a "0." This step must also be modeled, because we are ultimately interested in calculating the probability of an error, i.e., the probability that a "1" was sent and a "0" was detected, or vice-versa. From an engineering perspective, it is difficult to directly measure the absolute phase of a pulse; it is substantially easier to interfere two adjacent pulses and calculate the resulting optical intensity as a voltage. In practice one does this with a device comprised of a Mach–Zehnder delay interferometer followed by a balanced detector, as shown in Figure 1.

A Mach–Zehnder delay interferometer splits the signal into two copies and delays one relative to the other by one bit period, T_b. The two optical signals are then both constructively and destructively interfered. This process is repeated for each pair of received optical pulses. The balanced detector produces a voltage by subtracting the intensity of the destructively interfered signal from the intensity of the constructively interfered signal. Therefore, if at the end of transmission $u(t)$ is the optical field envelope and $u(t - T_\mathrm{b})$ is the version delayed by T_b, then the voltage after interference and balanced detection (but before electrical filtering) is given by

$$V(t) = |u(t) + u(t - T_\mathrm{b})|^2 - |u(t) - u(t - T_\mathrm{b})|^2. \tag{2.4}$$

With this encoding format, the threshold voltage is at $V = 0$ and a positive voltage at the output of the detector in the center of a bit slot represents a "1," while a negative voltage represents a "0." Note that the zero voltage threshold corresponds to a phase difference of $\pi/2$ between the two pulses, independent of their amplitudes.

Figure 2 is a schematic outline of the detection process for a back-to-back signal, i.e., when detection immediately follows encoding without transmission. The top two plots are

[1] A bit slot is one of the finite-width temporal windows into which an optical signal is divided.

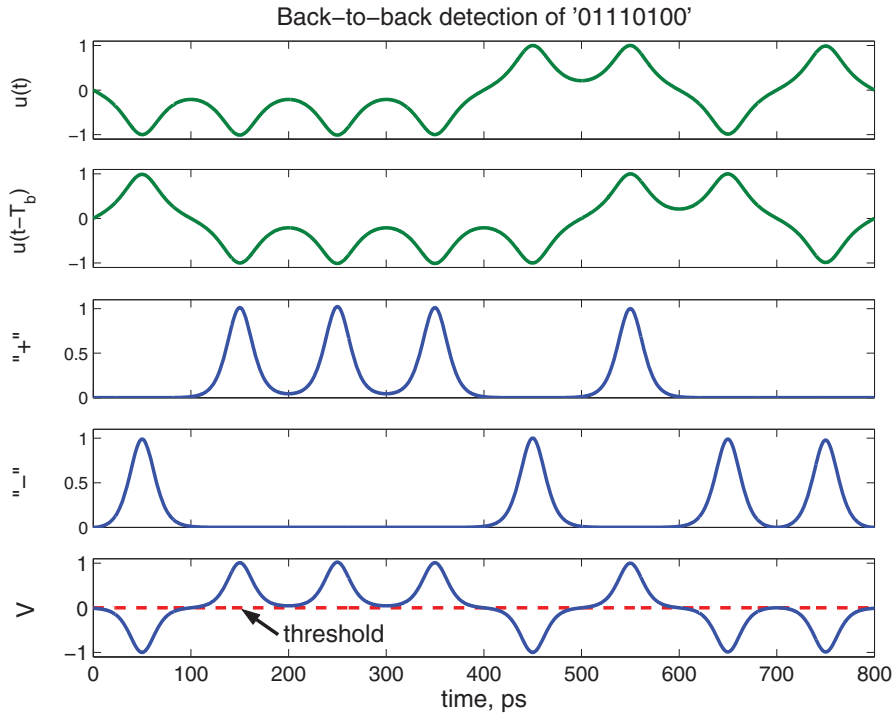


Figure 2. *In order from top to bottom: (The modulus of) the original and delayed optical signals (green), the constructively and destructively interfered electrical signals, and the output voltage V (blue), all in arbitrary units. The detection process is described in the text. On this plot, $V = 0$, and the signal is assumed to be periodic; the detection threshold (dashed red) is also included.*

the optical signal and the optical signal delayed one bit slot within the Mach–Zehnder delay interferometer. The next two plots are the output from the Mach–Zehnder delay interferometer, i.e., the electrical power of the two constructively and destructively interfered signals, respectively. The bottom plot is their difference and the detector's output voltage. The voltage threshold is also included on this plot; within each bit slot, a detected voltage above threshold is considered to be a "1," while each voltage below threshold is considered a "0." It is assumed that the detector samples the output voltage at the center of a given bit slot. This process is the same when transmission over a significant distance occurs between encoding and detection, although the received optical signal is then distorted by propagation and ASE noise. In practice (and in the simulations described in section 5) the signal is additionally filtered optically before the Mach–Zehnder delay interferometer and electrically just after; this is done to remove as much of the noise as possible. In what follows, the goal is to compute the pdf of the output voltage V and the resulting bit-error rate.

2.3. Soliton perturbation theory and noise-induced parameter changes. Consider a soliton solution, u_s, of the form given in (2.3), and a perturbation to it, δu. If we linearize the NLS equation about u_s, then the eigenmodes and generalized eigenmodes of the linear operator, $\mathcal{L}$, relate perturbations in the solution to changes in the soliton parameters [2, 24, 25]. Here we consider such a decomposition at the kth amplifier where the perturbation in the

underlying soliton solution, $u_s(t, kz_a)$, is comprised of noise, i.e., $\delta u(t, kz_a) = n_k(t)$. We then have [2, 24, 25]

$$n_k(t) = \delta A_k f_A + \delta\Omega_k f_\Omega + \delta T_k f_T + \delta\varphi_k f_\varphi + R(t, kz_a), \tag{2.5}$$

where f_K is the mode (or generalized mode) of $\mathcal{L}$ corresponding to the parameter $K \in \{A, \Omega, T, \varphi\}$, δK_k is the change in that parameter, and $R(t, kz_a)$ is the remainder of the perturbation, which manifests itself as dispersive radiation [2, 19, 24, 25]. In order to isolate the noise-induced change to a single parameter at the kth amplifier, one takes the inner product of (2.5) with the adjoint mode associated with the parameter of interest.[2] Thus, a noise-induced parameter change at the kth amplifier has the form

$$\delta K_k = \mathrm{Re} \int n_k(t) \underline{f}_K^* \, dt, \tag{2.6}$$

where * denotes the complex conjugate, and $\underline{f}_K$ is the adjoint or generalized adjoint mode of the linearized operator associated with parameter $K \in \{A, \Omega, T, \varphi\}$ [2, 3, 4, 24, 25].

We are interested in the instances when noise-induced parameter changes during propagation lead to significant distortion of the received signal. Although noise-induced parameter changes at each amplifier are small, several of these changes can combine to produce large deviations at the output. Combining the relation in (2.6) for the noise-induced parameter changes at each amplifier with how they vary between amplifiers in the absence of perturbations, we can write equations describing their evolution along the transmission line [2],

$$A_k = A_{k-1} + \delta A_k, \tag{2.7a}$$

$$T_k = T_{k-1} + z_a\Omega_{k-1} + \delta T_k, \tag{2.7b}$$

$$\Omega_k = \Omega_{k-1} + \delta\Omega_k, \tag{2.7c}$$

$$\varphi_k = \varphi_{k-1} + \frac{z_a}{2}(A_{k-1}^2 + \Omega_{k-1}^2) + \delta\varphi_k, \tag{2.7d}$$

where again z_a represents the distance between amplifiers. From the soliton solution given in (2.3), one can see that the additional changes to the phase and position arise because the rates at which they advance in the absence of perturbations depend upon the amplitude and frequency. In this notation φ is the phase of a pulse at the pulse center and $\delta\varphi_k$ represents direct phase perturbations at the kth amplifier. Note that A_k is the amplitude just after the kth amplifier and δA_k is the noise-induced amplitude change, which will depend on the soliton parameters from the previous amplifier (the $(k-1)$st, or the initial parameters if $k = 1$).

It is also useful to write the output voltage given by (2.4) in terms of the soliton parameters at the output. At the center of the bit slot, $t = 0$, and in the absence of position shifts ($T = 0$) and added noise, this voltage is

$$V = A_1^\circ A_2^\circ \cos(\varphi_1^\circ - \varphi_2^\circ), \tag{2.8}$$

where A_1°, A_2°, φ_1°, and φ_2° are the amplitudes and phases of two adjacent pulses of the output signal, i.e., evaluated at $z = z_{\mathrm{final}}$.

[2]Here we use the inner product $\langle v, w\rangle = \mathrm{Re} \int v^* w \, dt$.

3. Importance sampling. Recall that our objective is to find the probability of rare events in a nonlinear differential phase-shift-keyed lightwave system by direct simulation. To calculate the probability of an outcome, y, in a desired region, $\mathcal{R}$, one needs to compute the integral

$$P = Pr(y \in \mathcal{R}) = \int I(y(\boldsymbol{x}))p(\boldsymbol{x})\, d\boldsymbol{x}, \tag{3.1}$$

where I is an indicator function that is 1 when y is in $\mathcal{R}$ and 0 otherwise, and p is the pdf of the random variables $\boldsymbol{X}$ upon which the system's output depends. One can approximate this integral with a Monte Carlo quadrature given by

$$\hat{P} = \frac{1}{N}\sum_{k=1}^{N} I(y(\boldsymbol{X}_k)), \tag{3.2}$$

where $\boldsymbol{X}_k$ are random samples drawn from the distribution $p(\boldsymbol{x})$. Unfortunately, if P is very small (say, $P < 10^{-7}$), then standard Monte Carlo simulations require an unreasonably large number of samples to capture even a single event in this region, let alone a sufficient quantity to accurately approximate the integral.

To overcome this shortfall of Monte Carlo estimation we will utilize a variance reduction technique known as *importance sampling* [5, 26]. The main idea of importance sampling is to perform Monte Carlo quadrature by sampling from an alternative probability distribution, $p^*(\boldsymbol{x})$, that concentrates samples in the region of interest, $\mathcal{R}$, much more efficiently than the original distribution $p(\boldsymbol{x})$. However, if we draw random variables from $p^*(\boldsymbol{x})$, then we need to correct our statistics using the *likelihood ratio*, $L(\boldsymbol{x}) = p(\boldsymbol{x})/p^*(\boldsymbol{x})$, in order to obtain an unbiased result. Thus, (3.1) and (3.2) become

$$P = Pr(y \in \mathcal{R}) = \int I(y(\boldsymbol{x}))L(\boldsymbol{x})p^*(\boldsymbol{x})\, d\boldsymbol{x} \tag{3.3}$$

and

$$\hat{P}^* = \frac{1}{N}\sum_{k=1}^{N} I(y(\boldsymbol{X}_k^*))L(\boldsymbol{X}_k^*), \tag{3.4}$$

respectively, where the samples $\boldsymbol{X}_k^*$ are drawn from the distribution $p^*(\boldsymbol{x})$. Equations (3.1) and (3.3) are equivalent, but (3.4) will better estimate the probability P if the distribution $p^*(\boldsymbol{x})$ is chosen wisely. The challenge of effectively implementing importance-sampled Monte Carlo is to find a useful biasing distribution $p^*(\boldsymbol{x})$, i.e., one that can efficiently generate samples in $\mathcal{R}$ [15].

We wish to bias the noise in Monte Carlo simulations of this nonlinear lightwave system so that desired rare events occur more frequently than would normally be the case. If one corrects biased simulations with the likelihood ratio, the results are as if unbiased simulations had been run. In this manner we can accurately simulate rare events and compute pdfs of some quantity of interest far down into the tails. The difficulty, however, is that the number of random variables needed to simulate such systems is necessarily very large—in the tens or even hundreds of thousands. Because the dimensionality of the random state space is quite large,

it can be quite difficult to locate the specific regions that are most likely to produce errors. To bypass this problem, we use low-dimensional approximations of the system dynamics to determine the most probable manner in which perturbations can produce large deviations, and then we use this information to bias the noise and guide full importance-sampled Monte Carlo simulations of the full system. We will accomplish this by dividing the problem into two subproblems: the optimal biasing at a single amplifier and the optimal biasing across all amplifiers.

3.1. Optimal biasing at a single amplifier. First, we briefly summarize the method of biasing the ASE noise by changing the mean of individual Gaussian random variables associated with the noise added at the amplifiers as suggested by Moore, Biondini, and Kath [3], and as discussed in detail in [4]. The first step is to determine how to optimally bias a pulse at a single amplifier in order to achieve a desired parameter change.

The semianalytical approach is based upon some mathematical and physical insight. Studies exploring the growth of noise-induced phase jitter indicate that the variance in a pulse's phase is driven both directly and also by amplitude jitter that couples through self-phase modulation [27, 28]. (Determining the mean and variance alone, of course, is not sufficient information to calculate error probabilities when the distributions are non-Gaussian.) Both types of noise-induced fluctuations arise because solutions of the NLS equation cannot resist changes in the directions associated with the soliton parameters, as they arise from invariances of the equation [20]: any solution with different parameters is itself a perfectly valid solution. Thus, large phase and amplitude variations can build up from smaller ones, and these lead to significant deviations in the detected output voltage.

In what follows, we will neglect perturbations to the pulse's position (or timing) and frequency because we will consider the case of a system with small dispersion. In this limit, amplitude and phase fluctuations dominate. The first step in the analysis, then, is to determine the optimal biasing that produces the sought-after phase and amplitude changes at a single amplifier. When we perform the simulations we will use a discretized version of the NLS equation and the noise driving it, but to simplify the notation here we will describe the procedure using continuous functions of time. Following (2.1), at an amplifier a perturbation $\Delta u = n_k(t)$ is produced in the propagating signal by the noise. In the simulations, however, instead of using the unbiased noise $n_k(t) \equiv X(t)$, we will use a random variable of the form $X^*(t) = X(t) + b(t)$, where $b(t)$ is the biasing. The goal, again, is to choose the biasing $b(t)$ to make errors occur much more frequently than would be the case otherwise, and then to correct the estimate of these events' probabilities using the likelihood ratio.

The problem of finding the optimal biasing $b(t)$ has the same form for each of the four soliton parameters, so we will present it for the general parameter K, where $K \in \{A, T, \Omega, \varphi\}$. Recall that a parameter change, ΔK, due to a perturbation is the projection of the perturbation onto the adjoint or generalized adjoint mode of the linearized NLS equation associated with that parameter. In addition, maximizing the probability of achieving a desired outcome in the case of Gaussian noise is equivalent to maximizing the log-likelihood, i.e., minimizing the negative of the exponent. Thus, we want to minimize

$$\int_{-\infty}^{\infty} \langle |X(t) + b_K(t)|^2 \rangle \, dt = \sigma^2 + \int_{-\infty}^{\infty} |b_K(t)|^2 \, dt, \tag{3.5}$$

subject to the constraint of achieving (on average) a desired parameter change ΔK, i.e.,

$$\Delta K = \mathrm{Re}\int_{-\infty}^{\infty} \langle \underline{u}_K^*(t)(X(t)+b_K(t))\rangle\, dt = \mathrm{Re}\int_{-\infty}^{\infty} \underline{u}_K^*(t) b_K(t)\, dt = \text{constant}. \tag{3.6}$$

To solve this optimization problem, we reformulate it as a Lagrange multiplier problem:

$$F[b_K(t), b_K^*(t)] = \int_{-\infty}^{\infty} |b_K(t)|^2\, dt + \lambda\left[\Delta K - \mathrm{Re}\int_{-\infty}^{\infty} \underline{u}_K^* b_K(t)\, dt\right]. \tag{3.7}$$

To minimize F we take its functional derivative with respect to b_K and find the stationary point, $\delta F/\delta b_K = 0$. We see that F is minimized when

$$b_K(t) = \lambda \underline{u}_K(t), \tag{3.8}$$

which gives

$$\lambda = \frac{\Delta K}{\displaystyle\int_{-\infty}^{\infty} |\underline{u}_K(t)|^2\, dt}. \tag{3.9}$$

Thus, combining (3.8) and (3.9), we have the optimal functions to bias the soliton amplitude and phase,

$$b_A(t) = \frac{\Delta A}{\displaystyle\int_{-\infty}^{\infty} |\underline{u}_A|^2\, dt}\, \underline{u}_A(t), \tag{3.10a}$$

$$b_\varphi(t) = \frac{\Delta \varphi}{\displaystyle\int_{-\infty}^{\infty} |\underline{u}_\varphi|^2\, dt}\, \underline{u}_\varphi(t). \tag{3.10b}$$

Note that the latter equation is valid if $\Omega = 0$.

3.2. Optimal biasing across all amplifiers. The next question to address is how the changes across all of the amplifiers along the full transmission line should be arranged in order to best achieve a targeted output voltage. Is it more probable for large perturbations to occur at just a few amplifiers or for them to occur in a more distributed manner?

To answer this question, we will assume that the biasing function at each amplifier is a linear combination of the optimal biasing functions determined in the previous section and let the coefficients vary from amplifier to amplifier. Note that these adjoint eigenmodes are functions of time, but they also depend explicitly upon the parameters (A, T, Ω, φ) of the underlying soliton at each amplifier. The desired optimal biasing over all amplifiers then has the form

$$B_n = \alpha_n \underline{u}_A + \beta_n \underline{u}_\varphi, \tag{3.11}$$

where $1 \le n \le N_{\text{amp}}$. The goal then reduces to determining the optimal values of α_n and β_n at each amplifier along the transmission line.

Recall that the output voltage depends on the amplitudes and phases of two adjacent pulses. If we label two such adjacent pulses (1) and (2), we have, at the center of the bit slot,

$$V = A^{(1)}A^{(2)}\cos(\varphi^{(1)} - \varphi^{(2)}). \tag{3.12}$$

As a result, we must expand the biasing to include the adjoint modes associated with the amplitudes and phases of both adjacent pulses, i.e.,

$$B_n = \alpha_n^{(1)}\underline{u}_A^{(1)} + \beta_n^{(1)}\underline{u}_\varphi^{(1)} + \alpha_n^{(2)}\underline{u}_A^{(2)} + \beta_n^{(2)}\underline{u}_\varphi^{(2)}. \tag{3.13}$$

To simplify the optimization problem, we will assume that, during propagation, terms labeled (1) do not overlap in time with terms labeled (2), which means that all inner products of $\underline{u}_K^{(1)}$ and $\underline{u}_K^{(2)}$ are zero; i.e., $\langle \underline{u}_K^{(1)}, \underline{u}_K^{(2)} \rangle = \text{Re}\int_{-\infty}^{\infty} \underline{u}_K^{*(1)}\underline{u}_K^{(2)}\,dt = 0$ for $K \in \{A, \varphi\}$.

The goal, then, is the same as before: to choose the appropriate combination of biasing functions such that the probability of a desired rare event at the end of transmission is maximized, subject to the constraint that the biasing will achieve a desired output voltage. Again, because the noise is Gaussian, this is equivalent to minimizing the overall norm of the biasing functions; i.e., we wish to minimize

$$\sum_{n=1}^{N_{\text{amp}}} \|B_n\|^2 = \sum_{n=1}^{N_{\text{amp}}} (\alpha_n^{(1)})^2\|\underline{u}_A^{(1)}\|^2 + (\beta_n^{(1)})^2\|\underline{u}_\varphi^{(1)}\|^2 + (\alpha_n^{(2)})^2\|\underline{u}_A^{(2)}\|^2 + (\beta_n^{(2)})^2\|\underline{u}_\varphi^{(2)}\|^2 \tag{3.14}$$

subject to

$$A_{N_{\text{amp}}}^{(1)}A_{N_{\text{amp}}}^{(2)}\cos(\varphi_{N_{\text{amp}}}^{(1)} - \varphi_{N_{\text{amp}}}^{(2)}) = V^{\circ}. \tag{3.15}$$

In reality, (3.15) will not exactly represent the voltage at the output due to pulse distortions that occur during propagation and because of filtering, which has not been included in the analysis. It will be highly correlated with the output voltage, however, and thus sufficient to guide the biasing of the full simulations.

In this problem we have an additional constraint that the soliton parameters, A and φ, vary from amplifier to amplifier as dictated by soliton perturbation theory. For example, the amplitude at a particular amplifier is the value at the previous amplifier plus the noise-induced amplitude change. The phase at an amplifier, however, is the value at the previous amplifier plus the noise-induced phase change *plus* the amplitude-induced phase change that has accumulated between amplifiers. In addition, the adjoint eigenmodes at each amplifier depend on the soliton parameters, since the norms of the amplitude and phase adjoint eigenmodes are given by

$$\|\underline{u}_A\|^2 = \int_{-\infty}^{\infty} |\underline{u}_A|^2\,dt = 2A, \tag{3.16a}$$

$$\|\underline{u}_\varphi\|^2 = \int_{-\infty}^{\infty} |\underline{u}_\varphi|^2\,dt = \frac{\pi^2 + 12}{18A}. \tag{3.16b}$$

Thus, the goal is to minimize

$$\sum_{n=1}^{N_{\text{amp}}} 2A_n^{(1)}(\alpha_n^{(1)})^2 + \frac{\pi^2+12}{18A_n^{(1)}}(\beta_n^{(1)})^2 + 2A_n^{(2)}(\alpha_n^{(2)})^2 + \frac{\pi^2+12}{18A_n^{(2)}}(\beta_n^{(2)})^2 \tag{3.17}$$

subject to

$$V^\circ = A_1^\circ A_2^\circ \cos(\varphi_1^\circ - \varphi_2^\circ), \tag{3.18a}$$

$$A_n^{(k)} = A_{n-1}^{(k)} + \langle \underline{u}_{A_{n-1}}^{(k)}, B_n \rangle, \tag{3.18b}$$

$$\varphi_n^{(k)} = \varphi_{n-1}^{(k)} + \frac{1}{2}(A_{n-1}^{(k)})^2 z_{\text{a}} + \langle \underline{u}_{\varphi_{n-1}}^{(k)}, B_n \rangle, \quad k = 1, 2. \tag{3.18c}$$

Here, A_n and φ_n represent the values of the amplitude and phase, respectively, just after the nth amplifier. The $^\circ$ symbol denotes evaluation at $z = z_L$, i.e., $A_k^\circ = A_{N_{\text{amp}}}^{(k)}$. The last terms on the right-hand sides of (3.18b) and (3.18c) represent the biased noise B_n added at the nth amplifier projected onto the amplitude and phase modes; recall that they depend on the parameter values before the amplifier, i.e., those from just after the $(n-1)$st amplifier (or the initial parameter values if at the first amplifier).

It is not clear how to solve the discrete constrained minimization problem given by (3.17) and (3.18) even approximately, as was done previously in less complicated situations [3, 4]. We therefore follow an alternative approach in which the discrete problem is approximated by a continuous version. Equations (3.18b) and (3.18c) can be rewritten, for example, as

$$\frac{A_n^{(k)} - A_{n-1}^{(k)}}{z_{\text{a}}} = \frac{2\alpha_n^{(k)}}{z_{\text{a}}} A_{n-1}^{(k)}, \tag{3.19a}$$

$$\frac{\varphi_n^{(k)} - \varphi_{n-1}^{(k)}}{z_{\text{a}}} = \frac{1}{2}(A_{n-1}^{(k)})^2 + \frac{\beta_n^{(k)}}{z_{\text{a}}} \frac{\pi^2+12}{18A_{n-1}^{(k)}}, \tag{3.19b}$$

which shows that their continuous approximations should be

$$\frac{dA_k}{dz} = \frac{2\alpha_k}{z_{\text{a}}} A_k, \tag{3.20a}$$

$$\frac{d\varphi_k}{dz} = \frac{1}{2}A_k^2 + \frac{\beta_k}{z_{\text{a}}} \frac{\pi^2+12}{18A_k}, \quad k = 1, 2. \tag{3.20b}$$

The continuous version of the problem therefore reduces to minimizing a functional with differential equation constraints, which can be solved using calculus of variations [29]. In Lagrange multiplier form the problem is

$$\begin{aligned} F = \int_0^{z_L} \sum_{j=1}^{2} \Bigg[& 2A_j\alpha_j^2 + \frac{\pi^2+12}{18A_j}\beta_j^2 + \lambda_j^A(z)\left(\frac{dA_j}{dz} - \frac{2\alpha_j}{z_a}A_j\right) \\ & + \lambda_j^\varphi(z)\left(\frac{d\varphi_j}{dz} - \frac{1}{2}A_j^2 - \frac{\pi^2+12}{18z_aA_j}\beta_j\right)\Bigg] dz + \lambda_{V^\circ}(A_1^\circ A_2^\circ \cos(\varphi_1^\circ - \varphi_2^\circ) - V^\circ). \end{aligned} \tag{3.21}$$

Each λ above is a Lagrange multiplier and is labeled to refer to its corresponding constraint (all four evolution equations—two equations for each pulse—from soliton perturbation theory and the targeted output voltage). Where z_a remains in this approximation, it is a finite parameter.

To find the most likely soliton parameter paths that lead to large voltage distortions, we thus obtain a boundary value problem given by the Euler–Lagrange equations associated with the functional F,

$$\frac{d^2A_1}{dz^2} - \frac{1}{2A_1}\left(\frac{dA_1}{dz}\right)^2 - \frac{\pi^2+12}{18}\frac{\beta^2}{A_1} + 2g\beta A_1^2 = 0, \tag{3.22a}$$

$$\frac{d^2A_2}{dz^2} - \frac{1}{2A_2}\left(\frac{dA_2}{dz}\right)^2 - \frac{\pi^2+12}{18}\frac{\beta^2}{A_2} - 2g\beta A_2^2 = 0, \tag{3.22b}$$

$$A_1(0) = 1, \tag{3.22c}$$

$$A_2(0) = 1, \tag{3.22d}$$

$$\sin(\varphi)\frac{dA_1}{dz}(1) + 2\beta\cos(\varphi) = 0, \tag{3.22e}$$

$$\sin(\varphi)\frac{dA_2}{dz}(1) + 2\beta\cos(\varphi) = 0, \tag{3.22f}$$

$$\frac{g}{2}\int_0^1 \left(A_1^2 - A_2^2\right) dz + \frac{\pi^2+12}{18z_a}\beta\int_0^1 \left(\frac{1}{A_1} + \frac{1}{A_2}\right) dz = \varphi, \tag{3.22g}$$

$$A_1(1)A_2(1)\cos\varphi = V^\circ. \tag{3.22h}$$

Here, we have eliminated the Lagrange multipliers from the equations, as well as α_1 and α_2. Because the functional depends upon the phases only through the phase difference $\varphi \equiv \varphi_1^\circ - \varphi_2^\circ$, the resulting Euler equations also involve only this difference. In turn, this leads to $\beta_1 = -\beta_2 \equiv \beta$. In the above, we have rescaled $z \to z_L\, z$ for numerical convenience so that the distance varies from $z = 0$ to $z = 1$. We have also rescaled $A \to A_0\, A$, where A_0 is the initial amplitude of the pulse, so that in the rescaled equations the initial amplitude is unity. It is then natural to rescale $\beta \to A_0\beta/z_L$ and introduce $g = A_0^2 z_L$ as an effective nonlinear parameter. In the remainder of this paper, when we discuss the phase we will refer to the relative phase, i.e., the phase difference between adjacent pulses, unless explicitly mentioned otherwise. Note that (3.22g) is merely the integrated form of the evolution equation for this phase difference. In the above we have assumed that the initial phase difference between the pulses is zero; i.e., we have assumed that a "1" was initially encoded. For simplicity, we will focus on this particular case in what follows. The case of a "0," i.e., two pulses with an initial π phase difference, is almost identical.

Thus, the continuous optimization problem has been reduced to a coupled boundary value problem describing the "optimal" evolution of the amplitude and phase parameters, i.e., a problem whose solution gives the most probable paths leading to a specified output voltage. It should be reemphasized that the paths obtained from the solution of this problem will not be directly used to approximate the probability of the sought-after output voltage. Instead, they will be used to guide simulations of the full NLS equation; i.e., they will be used to bias the noise in such a way to induce parameter fluctuations lying close to these paths. The

biasing, of course, is accounted for by updating the likelihood ratio. The approximations made in calculating these optimal paths are thus not detrimental because the randomness associated with the importance-sampled Monte Carlo simulations allows for a search of state space around these paths. Another view of this is that the low-dimensional description of the problem given by soliton perturbation theory is merely used to determine the most significant regions of state space. Then, full importance-sampled Monte Carlo simulations of the NLS equation are used to sample the rare events of interest and directly calculate the probability of occurrence of the sought-after rare events.

Before continuing, it is useful to consider some general features of (3.22). If $\beta = 0$, one can see that evolution equations for the two amplitudes are the same. Thus, the output phase is zero, $\varphi = 0$, and the boundary conditions are satisfied. In this case, one can solve the equations exactly and obtain

$$A_1 = A_2 = A(z) = (1 - cz)^2. \tag{3.23}$$

The constant c is related to the targeted output voltage by $c = 1 - \sqrt[4]{V}$. This solution identifies *one* most probable manner in which large voltage changes can be achieved. In this way, however, errors are not obtained since a change of sign of the voltage must occur for this to happen. This suggests that *another* optimal pulse distortion mode exists—one that can achieve negative output voltages.

Predictions based upon the growth of the phase variance indicate that large phase fluctuations at the output do occur. If these phase fluctuations are large enough, the cosine term in the formula for the output voltage will change sign, leading to an error. There are additional mathematical indications that another error mode should exist. If one uses perturbation theory to solve (3.22) for small β, one obtains a consistency result of the form $\varphi \approx m \sin(\varphi)$, where m is a constant depending upon g and $V°$. If $m > 1$, this problem has a bifurcation to solutions with either positive or negative output phase.

4. Pulse distortion modes. We therefore look numerically for solutions where φ is nonzero and specifically for solutions with an output phase that is at least $\pi/2$ so that a change of sign of output voltage occurs. To find them, we let β parameterize the family of solutions and use the numerical continuation and bifurcation package AUTO [30] incorporated into XPP [31] to solve (3.22).

The first solution to be obtained numerically is the one described by (3.23) that produces *correlated* amplitude fluctuations. For this mode, large changes in the output voltage are most likely to occur when, on average, the amplitudes of two adjacent pulses decrease together. Again, there is no phase difference between the two pulses at output produced by this mode. Figure 3 shows typical amplitude and phase solution paths for this correlated pulse distortion mode.

In addition, the numerical solution of the equations with AUTO also reveals a bifurcation to *anticorrelated* amplitude and phase changes; this bifurcation occurs as the output voltage is reduced. In this case, on average, one pulse's amplitude increases, which advances its phase through self-phase modulation, while the adjacent pulse's amplitude decreases, retarding its phase. Consequently, this distortion mode can result in a nonzero output phase difference. This anticorrelated pulse distortion mode can lead to negative targeted output voltages and

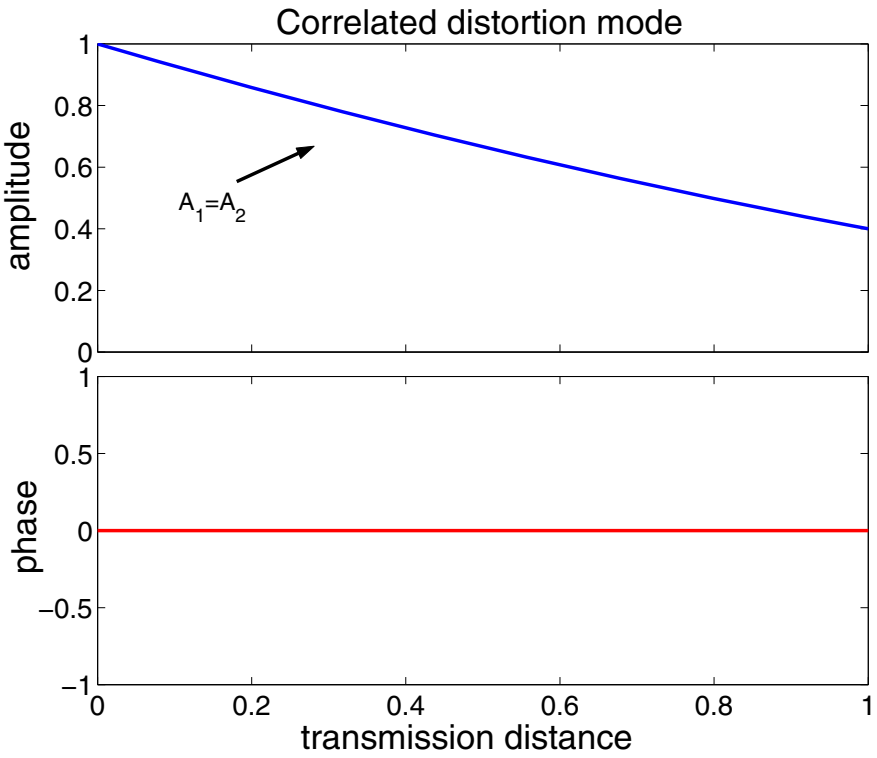


Figure 3. *Typical mean path that amplitudes of adjacent pulse and their phase difference are most likely to take on average to reach a desired output voltage when pulse variations are correlated. For this mode it is most likely that the amplitudes of two adjacent pulses decrease (or increase) together resulting in no phase change. The horizontal axis represents the fraction of total transmission distance.*

thus produce errors. Figure 4 shows typical solution paths for the this particular distortion mode; the other solution is symmetric under the interchange of A_1 and A_2 and a change of sign of φ.

In the descriptions of the pulse distortion modes above we have not yet addressed the issue of which of the two modes is more probable. We will see shortly that correlated fluctuations are most probable when the voltage is large, i.e., near its initial value. Anticorrelated fluctuations, on the other hand, become more probable as the voltage decreases, and especially when the voltage is near zero, i.e., near the threshold at which an error occurs. There is also a transition region of voltage values where these two distortion modes are of approximately equal importance, and the location and width of this region varies, to a certain extent, from problem to problem depending on the physical system parameters. Specifically, when changes to the physical parameters alter the effective nonlinear coefficient, g, then the location of the bifurcation point on the solution curve associated with (3.23) from which the second solution appears will vary.

To measure the importance of a pulse distortion mode for a given output voltage we will calculate a quantity that we refer to as the *biasing strength*. The biasing strength, denoted as $\mathcal{S}$, is merely a measure of the size of the minimum functional which was used to obtain the optimal biasing across all amplifiers. Recall that this functional is essentially the sum over all

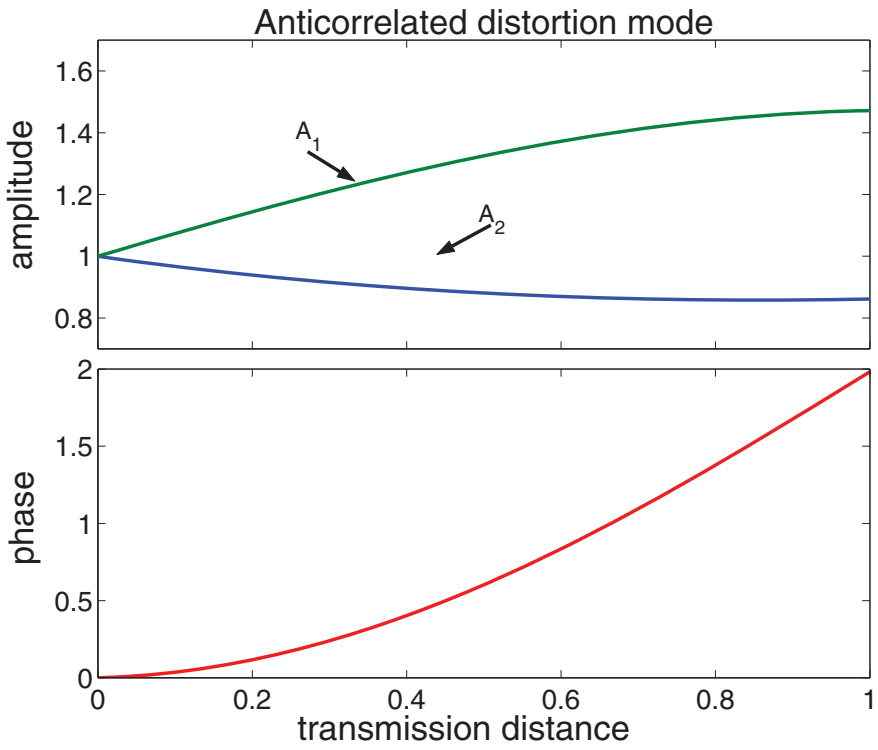


Figure 4. *Typical mean path that amplitudes of adjacent pulses and their phase difference are most likely to take, on average, to reach a desired output voltage in the case of anticorrelated noise-induced perturbations. In this case, we can see that one pulse's amplitude tends to increase while the neighboring pulse's amplitude tends to decrease. The opposite phase changes produced by these amplitude changes, plus equal and opposite direct noise-induced phase changes, lead to a large output phase difference. The horizontal axis represents fraction of total transmission distance.*

amplifiers of the norms of the biasing applied to the mean ASE noise,

$$\mathcal{S} = \sum_{n=1}^{N_{\text{amp}}} \|B_n\|^2, \tag{4.1}$$

so that it is a measure of the total amount of noise. Recall that the ASE noise is modeled as white zero-mean Gaussian noise. Thus, when we bias the mean of the ASE noise at each amplifier using B_n, the probability of a resulting outcome is proportional to the negative exponential of the biasing strength, $\mathcal{S}$; i.e., $\mathcal{S}$ is essentially the negative of the log-likelihood associated with the Gaussian noise. Thus, smaller biasing strengths correspond to more likely events.

In Figure 5 we have plotted a bifurcation diagram summarizing the optimal biasing solutions of (3.22), with $\mathcal{S}$ included. The precise details of these curves are expected to depend somewhat upon the nonlinearity coefficient, g. Here they are plotted for $g = 2.076$, which is the effective nonlinearity of the physical system discussed later in this paper. The curve labeled "correlated" corresponds to the correlated pulse distortion mode and represents solutions similar to Figure 3. The curve labeled "anticorrelated" corresponds to the anticorrelated pulse

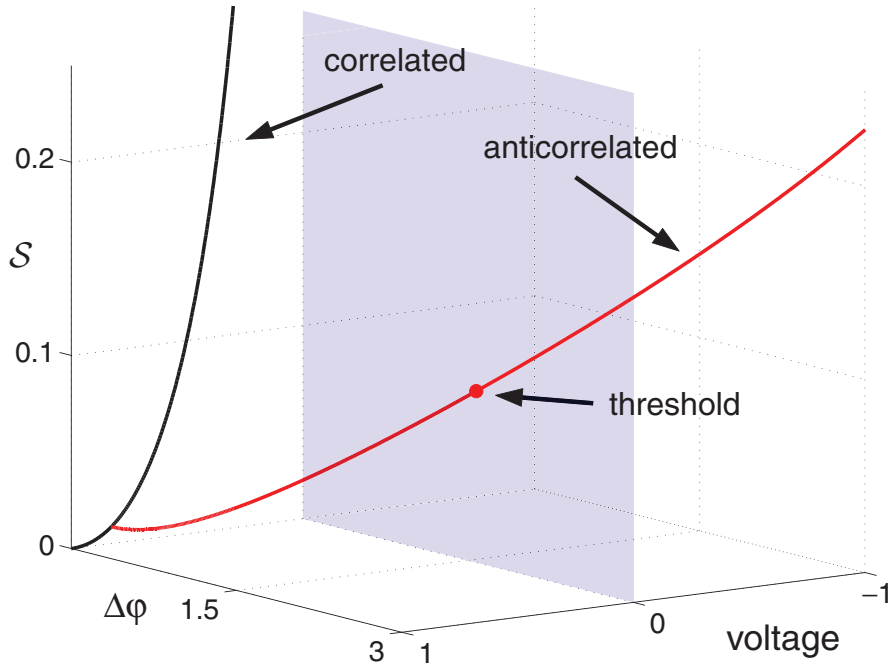


Figure 5. *General representation of the solution curves of the optimal biasing equations for $g = 2.076$. The axes represent output phase, output voltage, and biasing strength, $\mathcal{S}$. The error threshold ($V = 0$) is also highlighted for reference. As anticipated, the curve of correlated solutions has positive output voltage and cannot reach threshold. The anticorrelated solution curve passes through threshold, so anticorrelated amplitude changes are the mechanism most likely to produce errors in this differential phase-shift-keyed lightwave system. Near the bifurcation point (in the lower left-hand corner), the distortions due to each mode have approximately equal importance.*

distortion mode and represents solutions similar to those in Figure 4. These solution curves predict that only the anticorrelated distortion mode can achieve voltages beyond threshold, i.e., output voltages that correspond to errors.

5. Biasing for the soliton-based differential phase-shift-keyed system. Now that we know how to optimally bias the noise at each amplifier, we will use this biasing to guide importance-sampled Monte Carlo simulations. As previously stated, the goal of these simulations is to accurately determine the probability of noise-induced output voltage distortions and errors when errors are extremely rare events. The reason we do full simulations is that the optimal biasing functions and optimal parameter paths that we have calculated are only approximate. By doing full nonlinear simulations, any errors due to linear approximations are avoided. The approximations need not be perfect, of course, because they are used only to guide the full simulations, and thus the full importance-sampled Monte Carlo simulations can still correctly determine the probabilities associated with the rare events.

The outline of the process used to perform one trial of an importance-sampled Monte Carlo simulation of the differential phase-shift-keyed lightwave system is as follows (please note that computational details of these steps are carefully explained in section 6.2):

1. Generate the initial differential phase-shift-keyed signal.

2. Propagate the signal numerically to the next amplifier.
3. Extract the soliton parameters for one pair of adjacent pulses from the noisy signal.
4. Compute the adjoint eigenmodes of the NLS equation linearized about the solitons determined in the previous step using the appropriate analytic formulas [2].
5. Generate zero-mean Gaussian white noise with variance prescribed by (2.2).
6. Bias the mean of the random variables with an optimal linear combination of adjoint eigenmodes determined by the solution of (3.22).
7. Update the likelihood ratio. (Each amplifier's noise is independent, and hence the overall likelihood ratio is the product of the individual likelihood ratios.)
8. Repeat steps 2–7 until the end of transmission line is reached.
9. Apply any desired optical filter, detect the signal, convert it to a voltage with a balanced Mach–Zehnder delay interferometer, and apply the electrical filter.
10. Update the desired statistics of the output voltage signal.

A typical simulation has several biasing directions, i.e., targeted output voltages, per biasing mode (correlated or anticorrelated targeted amplitude changes). Several thousand trials are performed for each direction and the results are combined using multiple importance sampling [5, 26, 32]. In this manner the pdf of the output voltage can be computed well down into the tails. To validate results from the importance-sampled Monte Carlo simulations, we can compare them, at least for not-too-small probabilities, to results from standard Monte Carlo simulations. We also keep track of the coefficient of variation, the sample mean divided by the sample standard deviation, for each bin [5]. Good agreement between the biased and unbiased simulations, and a coefficient of variation that decreases smoothly like $1/\sqrt{N}$ with the number of trials N, are indications that the results from the importance-sampled Monte Carlo simulations are accurate.

For importance-sampled Monte Carlo simulations to be implemented properly, *all* important biasing directions should be included. For example, if one were to bias to target only anticorrelated variations, the simulation would not converge properly for larger probabilities—the pdf would not "visually" converge (would not be smooth and/or the tails might drop off rapidly), the pdf might not agree properly with the unbiased simulations (in the region where unbiased simulations are valid), and/or the coefficient of variation would contain large fluctuations. These large fluctuations would occur in voltage regions where correlated variations are the more likely events, because they would still occur occasionally even though the less likely anticorrelated events had been targeted, and when they did occur they would dominate the probability distribution.

Important biasing directions must not be overlooked. For example, for the anticorrelated biasing directions, one must consider *both* the amplitude of pulse (1) increasing with the amplitude of pulse (2) decreasing *and* the amplitude of pulse (2) increasing with the amplitude of pulse (1) decreasing as distinct biasing directions even if they target the same output voltage. Both cases, of course, will produce the same output phase difference on average. If only one of the cases were considered for each targeted voltage, however, the simulation would miss *half* of the most probable events.

As an example of how the random sampling around the biasing paths works, Figure 6 shows random paths of amplitude and phase from approximately 30 samples where the noise has been biased to follow an anticorrelated path to produce a low voltage. This figure demonstrates

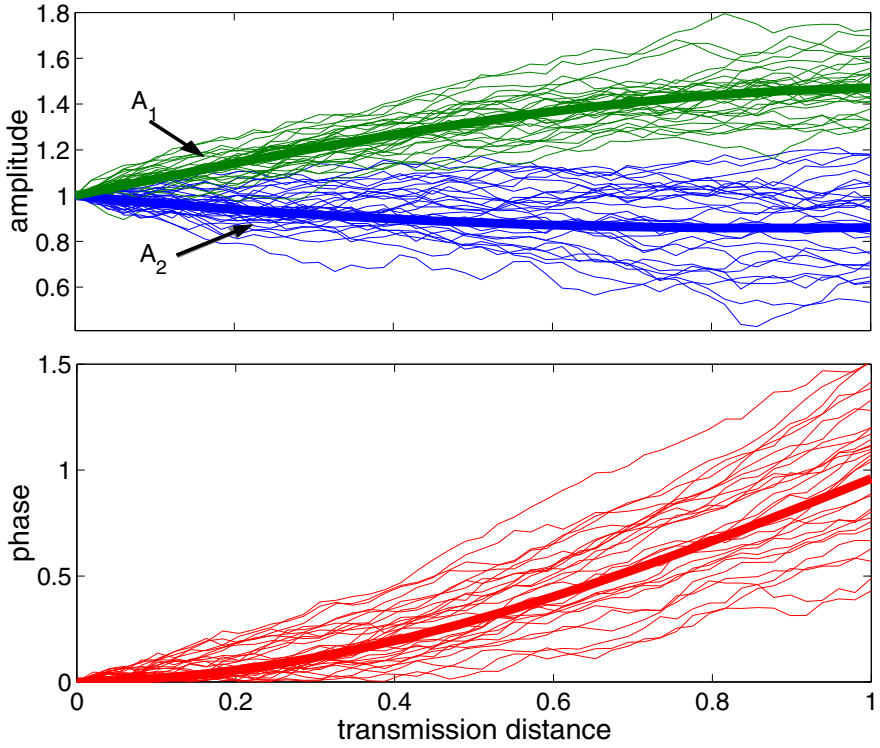


Figure 6. *Simulation showing the random paths (jagged, thin lines) from approximately* 30 *biased samples near a typical anticorrelated optimal biasing direction (smooth, bold curves). The amplitude parameters are plotted in the top figure (green for* A_1*, blue for* A_2*), while the phase difference is plotted in the bottom figure. The horizontal axis represents the normalized transmission distance.*

that the importance-sampled Monte Carlo scheme outlined above successfully biases noise to induce larger-than-normal amplitude and phase fluctuations in the manner predicted by soliton perturbation theory. It also demonstrates how the randomness of biased Monte Carlo simulations enables a search of state space near the predicted optimal path. In this manner, even if the predicted optimal path is not exact, the random nature of the simulation enables the most probable path through state space to a desired rare event to be properly sampled.

6. Simulation of an example differential phase-shift-keyed system.

6.1. System parameters. As a test of the method, we simulate a 10 Gbit/s soliton-based differential phase-shift-keyed transmission system, which means that each pulse lies within a 100 ps bit slot. This bit rate has been of interest for communication applications [22, 23]. Here we assume a fiber loss of 0.21 dB/km, a nonlinear coefficient of 1.7 $(\text{W-km})^{-1}$, an average dispersion of 0.15 ps^2/km, and 30 ps full-width half-max (FWHM) pulses. The FWHM is the time interval over which the instantaneous power is larger than 1/2 of its maximum. A related parameter, the pulse width, T_{w}, is used to nondimensionalize time. This scaling factor varies from pulse shape to pulse shape, but for hyperbolic secant pulses, $T_{\text{w}} = T_{\text{FWHM}}/1.76$.

We assume further that the amplifiers are spaced every 80 km, and we use a spontaneous emission factor $\eta_{\text{sp}} = 1.25$. The total transmission distance is 4,000 km, and the average power

is 0.1 mW (the pulse peak power is 0.3 mW). This choice of parameters results in a system with an optical signal to noise ratio of 14.5 dB (using a .2 nm filter). It is worth reiterating at this point that we have averaged over deterministic power fluctuations as described in section 2.1 after (2.2). At the end of transmission the signal is filtered using a 50 GHz optical bandpass filter, detected using a balanced Mach–Zehnder delay interferometer, and then filtered electrically by a Bessel filter with a bandwidth that is 80% of the bit rate.

6.2. Numerical parameters and implementation. In the split-step numerical simulations the pseudorandom bit pattern "01110100" (or a cyclic permutation) is used, encoded using a π phase change between adjacent pulses to represent a "0" and no phase change for a "1." This 8-bit pattern contains all possible 3-bit combinations. Spatial evolution steps are taken to be 1/5th of the amplifier spacing. Time is discretized by dividing each bit slot by 64. Thus, 64 Fourier modes describe each pulse, and $N = 512$ Fourier modes describe the time-periodic 8-bit sequence.

To simulate the noise, $n_k(t)$, we generate independently and identically distributed unit Gaussian random variables, X_i, $i = 1, \ldots, 2N$, and then scale them by an appropriate standard deviation. Random variables must be added to both the real and imaginary parts of the field at each time point (or to each frequency component), resulting in the factor of two above. The variance of the numerical noise is set to be equal to the variance of the ASE noise determined by physical parameters, σ^2 (see (2.2) and [4]). Because the noise $n_k(t)$ is delta-correlated, we must have

$$\int \langle n_k(t) n_k^*(t') \rangle \, dt = \sigma^2. \tag{6.1}$$

We wish to find a scaling factor, a, which when applied to standard Gaussians makes them satisfy the equivalent relationship in the discrete case, namely,

$$\sum_i \langle aX_i \, aX_j \rangle \Delta t \approx a^2 \langle X_j^2 \rangle \Delta t = \sigma^2/2. \tag{6.2}$$

(The factor of 1/2 here is because $n_k(t)$ contains both real and imaginary parts, each contributing half of the total.) Thus, we see that the numerical standard deviation a should be $\sigma/\sqrt{2\Delta t}$.

The optical bandpass filter is implemented by taking the product of the output signal and a Gaussian (again, 50 GHz) in the frequency domain and then transforming the filtered optical signal back to the time domain. For the electrical filter, we use a 5th order Bessel filter (which is a causal approximation of a perfect time delay) and apply the filter in the frequency domain of the optical intensity. The mean delay associated with the filter is removed before detection. Note that these filters are incorporated into the simulation because such filters are typically used in practice to eliminate as much noise as possible from the output signal.

To bias the noise at each amplifier, one must "extract" the underlying soliton from a noisy pulse. This underlying pulse is the soliton about which the NLS equation is linearized at each amplifier. Recall that a soliton solution of the NLS equation is determined by four parameters: amplitude, timing, frequency, and phase. In order to find the underlying soliton, one must determine at least approximate values for these parameters. Given a pulse, one

can approximate the soliton parameters by using moment integrals [4], which arise from the invariances of the NLS equation's Lagrangian. For example, given part of a signal $u(t, z_k)$ that represents one bit slot, one can calculate an approximate soliton amplitude with the equation

$$A(z_k) = \frac{1}{2} \int |u|^2 \, dt. \tag{6.3}$$

Similar moment integrals can be used to find the other parameters (for the phase, no exact local definition is available, but ad hoc versions can be created). For a very noisy pulse, however, it is possible that the moment integral given by (6.3) will overestimate the amplitude. One can improve parameter estimates by filtering the high frequency components of the pulse to remove some of the noise before calculating the moment integrals [4].

Recall that the parameters are used to construct the underlying soliton and associated linearized modes. These modes are used to bias the noise added to the signal, and more precise biasing leads to faster statistical convergence of the importance-sampled Monte Carlo simulations. To this end iterative improvement to the soliton parameter estimates can also be obtained. Once initial approximations for the soliton parameters have been found, one can compute the residual $\delta u = u(t, z) - u_s(t, z)$. Next, one can project the residual δu onto the adjoint eigenmodes of the NLS equation linearized about the approximate soliton. If these projections are not close to zero (as compared to some tolerance), then the underlying soliton has not been estimated properly. To remedy this, one can update the parameter estimates by the projected amounts and repeat the process until sufficient accuracy is achieved.

For the simulations presented here, 340,000 biased Monte Carlo trials were performed. More precisely, we targeted 17 different output voltages so that the entire voltage range of interest is covered; for each of these 17 "biasing directions" 20,000 trials were used, and the results were combined with multiple importance sampling [5, 26, 32]. The simulation cycles through the eight possible pulse pairs from the bit pattern given above. That is, for each importance-sampled Monte Carlo trial, an individual pair of pulses was targeted (i.e., the noise was biased for that pair and not for the other six pulses). The resulting voltage (associated with each particular pulse pair biased for a given trial) is sampled at the center of the corresponding bit slot, and the pdf of the output voltage is computed by dividing the output voltage range into 80 bins. Two pdfs are computed per simulation—one conditioned on a "0" having been sent initially and another conditioned on a "1" having been sent initially.

6.3. Simulation results. In Figure 7 we plot the probability density functions computed from the importance-sampled Monte Carlo simulation of the soliton differential phase-shift-keyed lightwave system with the physical parameters described previously. Also plotted are the results from an unbiased Monte Carlo simulation that used 10^7 trials. Two overall pdfs are plotted—one conditioned on a "1" having been sent initially and another conditioned on a "0" having been sent. Note the importance-sampled results lie on top of the unbiased simulations in the region where the latter give results, but go down more than 10 orders of magnitude farther in probability. The simulation clearly shows that importance sampling outperforms standard Monte Carlo simulations by many orders of magnitude. In the figure we also plot the individual contributions to the combined pdf from the trials that targeted the correlated distortion mode and the trials that targeted the anticorrelated distortion mode.

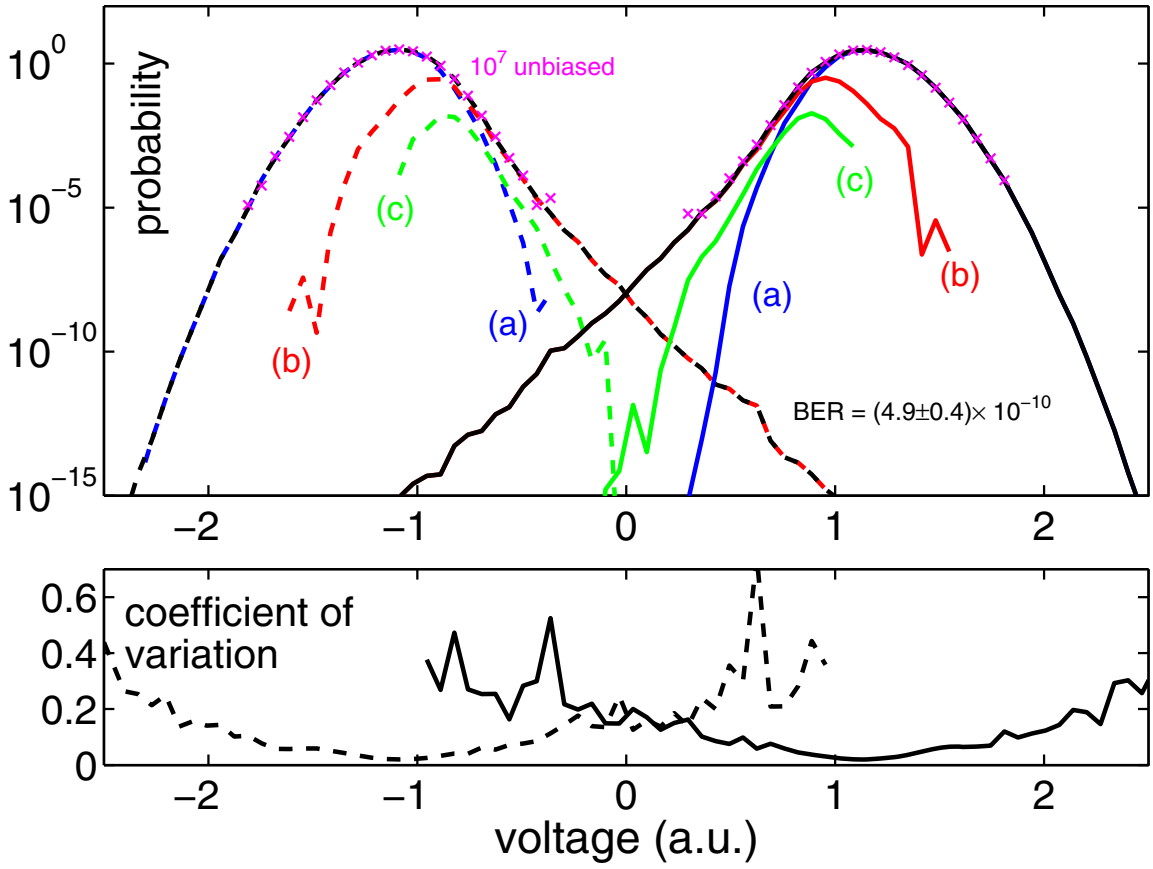


Figure 7. *Top: Black curves represent probability distribution functions (pdfs) for output voltage, conditioned on either a "1" (solid) or a "0" (dashed) being sent initially. The lower curves indicate individual contributions to the total probability from each of the two predicted distortion modes discussed previously and the one-pulse distortion mode—*(a) *correlated amplitude variations;* (b) *anticorrelated phase and amplitude variations;* (c) *single pulse variations. The contributions from each of these modes are combined to compute the overall pdfs (black curves). Bottom: The coefficient of variation for the importance sampling simulation. (The C.V., or relative variance, is the ratio of the sample standard deviation divided to the mean; the smaller the relative variance, the more accurate the estimate is likely to be.)*

In some regions the overall pdfs come primarily from one of the two modes, and so one of the individual curves overlaps the overall pdf in these regions. In particular, the pdfs at large voltages, $|V| > 1.5$, are comprised only of the contributions from correlated amplitude variations, i.e., the curves labeled (a). Similarly, the pdfs near and past the threshold at zero voltage are comprised solely of the contribution from anticorrelated amplitude and phase variations, i.e., curves labeled (b); thus, for these system parameters this distortion mode solely determines the bit-error ratio.

For this particular simulation we also biased the noise in an additional direction not predicted by the biasing theory for a pair of pulses; the contribution from these trials is represented by curve (c). The impetus for including this additional biasing direction, which targets changes in the output voltage due to variations in a single pulse, rather than a pair of

pulses, was to speed up convergence of the simulation and reduce the size of the coefficient of variation. Including this additional biasing direction did not significantly change the resulting pdf but did lower the total number of trials needed to achieve a desired (small) coefficient of variation at moderately small probabilities. In this case, the optimal biasing problem is equivalent to (3.14) and (3.15), except that the amplitude of one of the pulses is constant. As a result, the boundary value problem that one must solve to optimize the biasing problem is equivalent to (3.22) when one eliminates the equations governing the evolution of A_2 and its boundary conditions and replaces the A_2 in the other instances with $A_2 = 1$. This single pulse biasing mode exhibits solutions similar to the anticorrelated distortion mode, in that the voltage distortion is most likely to occur if the pulse's amplitude decreases, causing a phase retardation. There is no equivalent to the correlated distortion mode in the singe-pulse case, however. The result of the single-pulse biasing shows a contribution to the overall pdf that is smaller than that from the other two modes. This explains why this solution did not show up when solving the pair-pulse biasing problem (3.22); it is not a local extremum of the function given by (3.21). Nevertheless, the contribution to the overall pdf from this mode is a noticeable fraction of the total, which is why including it speeds up the convergence of the simulations.

To further illustrate the character of the large variation modes that were included in the simulation, Figure 8 shows specific simulation trials for each. These amplitude, phase, and voltage profiles were recorded when a particular voltage threshold was crossed (a different threshold was used for each type of distortion). The amplitude and phase curves are shown after optical filtering, while the voltage curve is shown after electrical filtering. Modes (a), (b), and (c) are examples of the result of correlated, anticorrelated, and single-pulse amplitude and phase variations, respectively, corresponding to Figure 7. In each case it is clear that the actual pulse fluctuations show deviations from the optimal biasing paths; e.g., the amplitude profiles shown in mode (a) are neither identical nor smooth; it is only the *mean* biasing that is described properly using soliton perturbation theory. It is also worth noting that although one particular voltage bit slot has been targeted, the fluctuations also produce significant deviations in nearby bit slots. Thus, each fluctuation produces a particular pattern of correlated deviations in the voltage signal.

For these system parameters, the individual pdfs for the one and zero voltage rails cross in Figure 7 at a value just below 10^{-8}. The probability of an error is the average of the probability that a "0" was sent and a "1" was detected and the probability that a "1" was sent and a "0" was detected. Integrating the area under the parts of the pdfs that lie beyond the zero voltage point (the decision threshold) and multiplying by 1/2 (the probability of a "1" or a "0" being sent) gives a total bit-error ratio of $(4.9 \pm .4) \times 10^{-10}$ for these parameters. To calculate the variance in this multiple importance-sampled simulation, we used the method of Biondini, Kath, and Menyuk [5].

As a comparison, we have performed another set of Monte Carlo simulations where a noiseless signal is propagated and a comparable amount of white Gaussian noise is then added at the end of the transmission line. This is a standard approximation valid when the noise does not interact with the signal during propagation, but only at the detector [33]. In this simulation, the amount of white Gaussian noise added at the end of the transmission line was adjusted so that the optical signal to noise ratio at the end was the same as for the

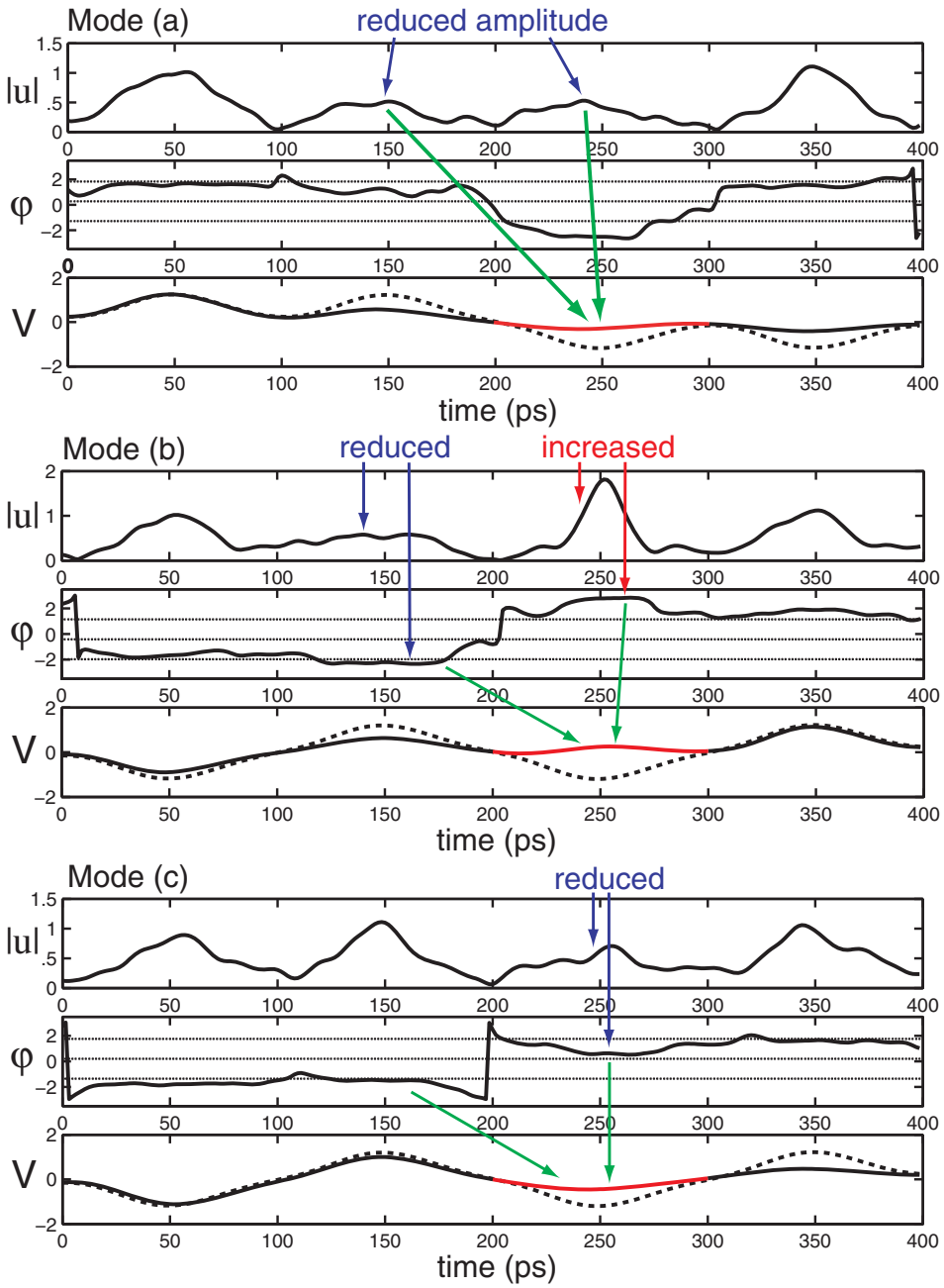


Figure 8. *Single-trial examples of each of the error modes. Modes* (a), (b), *and* (c) *correspond to correlated, anticorrelated, and single-pulse amplitude and phase variations, respectively. The dashed lines correspond to the voltage of the undistorted signal. Note that a different sequence of pulses was targeted in case* (a) *than in* (b) *and* (c). *In each case, reductions or increases in amplitude and phase are indicated. The optical pulses that combine to produce the specific targeted voltage bit slot are also indicated.*

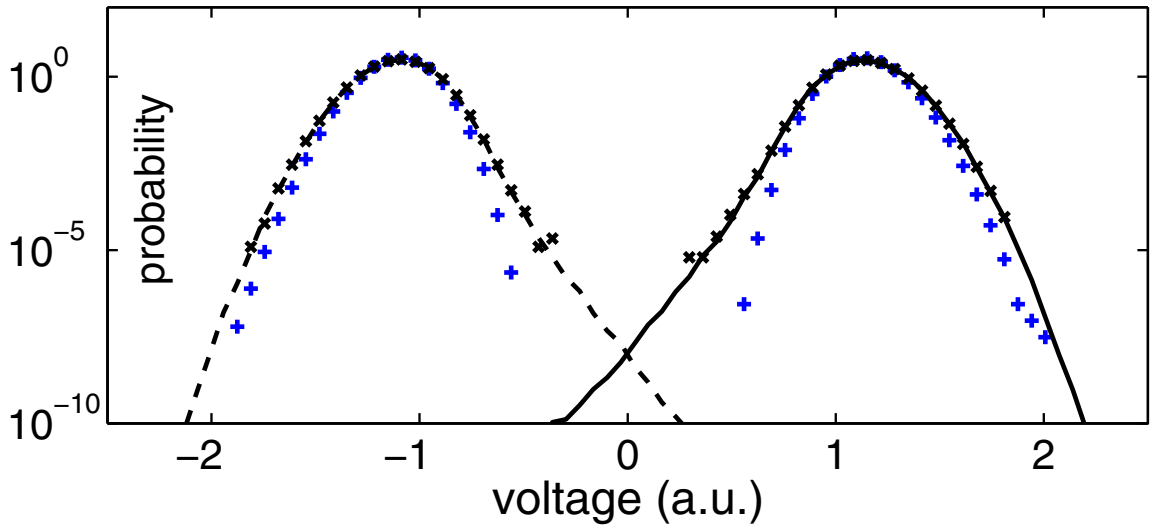


Figure 9. *Comparison between full importance-sampled Monte Carlo simulations (black curves, as in Figure 7) and another set of unbiased Monte Carlo simulations where a noiseless signal is propagated and then a comparable amount of noise is added at the end of the transmission line just before detection (+ signs; see text for more details). Black ×'s are pdfs from unbiased Monte Carlo simulations of the full nonlinear system, as in Figure 7. The results show that neglecting the interaction between signal and noise during propagation can significantly underestimate the bit-error ratio.*

full simulations. The result is shown by the + signs in Figure 9; the main simulation results from Figure 7 are also included for comparison purposes. For this comparison simulation, 10^9 unbiased Monte Carlo trials were used (more trials are possible in this case; since only one nonlinear propagation is necessary, it is a much simpler simulation). Although this number of trials is insufficient to actually estimate a bit-error ratio, it is clear that the pdfs conditioned on either a "1" or a "0" being sent will cross at a probability that falls significantly below the value seen in the full importance-sampled simulations. This demonstrates that estimates which neglect the interaction between signal and noise during propagation (including those based upon assuming Gaussian statistics for the received "1"s and "0"s [34]) can significantly underestimate the actual bit-error ratio.

7. Discussion. We have presented a method for computing errors in soliton-based lightwave systems. The method uses soliton perturbation theory and calculus of variations to find approximate versions of the most probable paths through sample space leading to errors, followed by importance-sampled Monte Carlo simulations of the full set of equations around these approximate paths to compute the actual error rates. A specific example differential phase-shift-keyed lightwave system was simulated, and the system's bit-error ratio was computed. The method predicts the most probable ways in which large pulse distortions occur, thus providing an explanation as to the specific manner in which errors occur.

One of the main difficulties associated with simulating such systems is the large dimensionality of the state space. Recall that 512 Fourier modes were used for the simulations of the example system in section 6, meaning that 1,024 independent Gaussian random variables are added at each amplifier. With a total distance of 4,000 km, and an amplifier spacing of

80 km, this gives 50 amplifiers, and thus a total of 51,200 Gaussian random variables. This is just an example, of course, and it is possible that even more variables might be used in other cases. Ultimately we are searching for regions in a very large-dimensional state space that produce large deviations in a *single* output quantity, the voltage, and furthermore, we want to determine which of these regions are the most probable.

The large dimensionality of the state space associated with such problems can make finding the regions most likely to contribute to errors difficult. The only sure way of doing so, of course, is by an exhaustive search of state space, something that is clearly beyond the capability of existing computational methods and hardware. In the present case, we have employed low-dimensional approximations of the system dynamics and analytical methods to guide the simulations to the locations of such regions. In this we regard the mathematical structure imposed by the equations describing the system as constraints which limit the regions of state space in which errors may occur.

By contrast, when iterative methods such as the multicanonical Monte Carlo method are used, a performance measure is monitored (such as the voltage at the center of the bit slot), and each set of simulations is analyzed in order to determine directions in state space for further exploration. To do this no analytical approximations are needed, of course, and so in one sense such iterative methods are more straightforward. For such methods to work, however, the regions of state space that produce one set of voltage values should be contiguous with regions that produce values with slightly smaller probabilities, something that is difficult to verify a priori in any specific application. If this is not the case, of course, then iterative sampling may fail to locate these newer values. In the present case, although it requires a nontrivial effort to construct the analytical approximation used to guide the biasing, the insight provided by the analytic approach gives one much more confidence that nothing has been missed when one preforms the Monte Carlo sampling.

A detailed comparison between the two approaches is beyond the scope of the present work, of course. We hope to make such a comparison in the future, however, and thus further clarify the pros and cons associated with the application of each type of method of simulating lightwave systems.

Acknowledgments. The authors would like to thank Gino Biondini, Curtis Menyuk, Colin McKinstrie, Richard Moore, and John Zweck for many helpful discussions.

REFERENCES

[1] E. Desurvire, *Erbium-Doped Fiber Amplifiers: Theory and Applications*, Wiley, New York, 1994.
[2] E. Iannone, F. Matera, A. Mecozzi, and M. Settembre, *Nonlinear Optical Communication Networks*, Wiley, New York, 1998.
[3] R. O. Moore, G. Biondini, and W. L. Kath, *Importance sampling for noise-induced amplitude and timing jitter in soliton transmission systems*, in Nonlinear Physics: Theory and Experiment II, World Scientific, River Edge, NJ, 2003, pp. 383–390.
[4] R. O. Moore, G. Biondini, and W. L. Kath, *A method to compute statistics of large, noise-induced perturbations of nonlinear Schrödinger solitons*, SIAM J. Appl. Math., 67 (2007), pp. 1418–1439.
[5] G. Biondini, W. L. Kath, and C. R. Menyuk, *Importance sampling for polarization-mode dispersion: Techniques and applications*, J. Lightwave Technology, 22 (2004), pp. 1201–1215.

[6] A. Tonello, S. Wabnitz, I. Gabitov, and R. Indik, *Importance sampling of Gordon-Mollenauer soliton phase noise in optical fibers*, IEEE Phot. Tech. Lett., 18 (2006), pp. 886–888.
[7] D. Yevick, *Multicanonical communication system modeling-application to PMD statistics*, IEEE Phot. Tech. Lett., 14 (2002), pp. 1512–1514.
[8] T. Kamalakis, D. Varoutas, and T. Sphicopoulos, *Statistical study of in-band crosstalk noise using the multicanonical Monte Carlo method*, IEEE Phot. Tech. Lett., 16 (2004), pp. 2242–2244.
[9] Y. Yadin, M. Shtaif, and M. Orenstein, *Bit-error rate of optical DPSK in fiber systems by multicanonical Monte Carlo simulations*, IEEE Phot. Tech. Lett., 17 (2005), pp. 1355–1357.
[10] W. Pellegrini, J. Zweck, C. R. Menyuk, and R. Holzlohner, *Computation of bit error ratios for a dense WDM system using the noise covariance matrix and multicanonical Monte Carlo methods*, IEEE Phot. Tech. Lett., 17 (2005), pp. 1644–1646.
[11] R. Holzlöhner and C. R. Menyuk, *Use of multicanonical Monte Carlo simulations to obtain accurate bit error rates in optical communications systems*, Optics Letters, 28 (2003), pp. 1894–1896.
[12] I. Neokosmidis, T. Kamalakis, A. Chipouras, and T. Sphicopoulos, *Estimation of the four-wave mixing noise probability-density function by the multicanonical Monte Carlo method*, Optics Letters, 30 (2005), pp. 11–13.
[13] P. Del Moral and J. Garnier, *Genealogical particle analysis of rare events*, Ann. Appl. Probab., 15 (2005), pp. 2496–2534.
[14] R. Y. Rubinstein, *A stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation*, Methodol. Comput. Appl. Probab., 7 (2005), pp. 5–50.
[15] J. S. Sadowsky and J. A. Bucklew, *On large deviations theory and asymptotically efficient Monte-Carlo estimation*, IEEE Trans. Inform. Theory, 36 (1990), pp. 579–588.
[16] E. T. Spiller, W. L. Kath, R. O. Moore, and C. J. McKinstrie, *Computing large signal distortions and bit-error ratios in DPSK transmission systems*, IEEE Phot. Tech. Lett., 17 (2005), pp. 1022–1024.
[17] G. P. Agrawal, *Fiber-Optic Communication Systems*, John Wiley & Sons, New York, 1992.
[18] A. Hasegawa and Y. Kodama, *Solitons in Optical Communications*, Oxford Series in Optical and Imaging Sciences 7, Clarendon Press, Oxford, UK, 1995.
[19] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM Stud. Appl. Math. 4, SIAM, Philadelphia, 1981.
[20] W. L. Kath, *A modified conservation law for the phase of the nonlinear Schrödinger soliton*, Methods Appl. Anal., 4 (1997), pp. 141–155.
[21] A. H. Gnauck and P. J. Winzer, *Optical phase-shift-keyed transmission*, J. Lightwave Technology, 23 (2005), pp. 115–130.
[22] J.-X. Cai, G. Foursa, Y. Cai, G. Domagala, H. Li, L. Liu, W. W. Patterson, A. N. Pilipetskii, M. Nissov, and N. S. Bergano, *A DWDM demonstration of* 3.73 *Tb/s over* 11,000 *km using* 373 *RZ-DPSK channels at* 10 *Gb/s*, in OFC 2003 Technical Digest, Optical Society of America, Washington, DC, 2003, p. PD22.
[23] K. Ishida, T. Kobayashi, J. Abe, K. Kinjo, S. Kuroda, and T. Mizuouchi, *A comparative study of* 10 *Gb/s RZ-DPSK and RZ-ASK WDM transmission over transoceanic distances*, in OFC 2003 Technical Digest, Optical Society of America, Washington, DC, 2003, pp. 451–453.
[24] D. Kaup, *Perturbation theory for solitons in optical fibers*, Phys. Rev. A (3), 42 (1990), pp. 5689–5694.
[25] H. Haus and Y. Lai, *Quantum theory of soliton squeezing: A linearized approach*, J. Opt. Soc. Amer. B Opt. Phys., 7 (1990), pp. 386–392.
[26] A. Owen and Y. Zhou, *Safe and effective importance sampling*, J. Amer. Statist. Assoc., 95 (2000), pp. 135–143.
[27] J. P. Gordon and L. F. Mollenauer, *Phase noise in photonic communications systems using linear amplifiers*, Optics Letters, 15 (1990), pp. 1351–1353.
[28] C. J. McKinstrie and C. Xie, *Phase jitter in single-channel soliton systems with constant dispersion*, IEEE J. Sel. Top. Quant. Elect., 8 (2002), pp. 616–625.
[29] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
[30] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. J. Wang, *AUTO97: Continuation and Bifurcation Software for Ordinary Differential Equations*, available from http://cmvl.cs.concordia.ca, 1997.
[31] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.

[32] E. Veach, *Robust Monte Carlo Methods for Light Transport Simulation*, Ph.D. thesis, Stanford University, Stanford, CA, 1997.
[33] D. Marcuse, *Derivation of analytical expressions for the bit-error probability in lightwave systems with optical amplifiers*, J. Lightwave Technology, 8 (1990), pp. 1816–1823.
[34] G. Bosco and P. Poggiolini, *On the Q factor inaccuracy in the performance analysis of optical direct-detection DPSK systems*, IEEE Phot. Tech. Lett., 16 (2004), pp. 665–667.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 895–932
© 2008 Society for Industrial and Applied Mathematics

C^1 Approximation of Vector Fields Based on the Renormalization Group Method*

Hayato Chiba†

Abstract. The renormalization group (RG) method for differential equations is one of the perturbation methods for obtaining solutions which approximate exact solutions for a long time interval. This article shows that, for a differential equation associated with a given vector field on a manifold, a family of approximate solutions obtained by the RG method defines a vector field which is close to the original vector field in the C^1 topology under appropriate assumptions. Furthermore, some topological properties of the original vector field, such as the existence of a normally hyperbolic invariant manifold and its stability, are shown to be inherited from those of the RG equation. This fact is applied to the bifurcation theory.

Key words. renormalization group method, singular perturbation method, bifurcation theory

AMS subject classifications. 34C15, 34C20, 37D10, 37G10

DOI. 10.1137/070694892

1. Introduction.

The renormalization group (RG) method for differential equations is one of the perturbation methods for obtaining solutions which approximate exact solutions for a long time interval. In their papers [1, 2], Chen, Goldenfeld, and Oono established the RG method for ODEs of the form

$$\dot{x} = \frac{dx}{dt} = f(t,x) + \varepsilon g(t,x), \quad x \in \mathbf{R}^n, \tag{1.1}$$

where $\varepsilon > 0$ is a small parameter. For this equation, the method for deriving approximate solutions of the form

$$x(t) = x_0(t) + \varepsilon x_1(t) + \varepsilon^2 x_2(t) + \cdots \tag{1.2}$$

is called the *naive expansion* or the *regular perturbation method*, where $x_i(t)$'s are governed by inhomogeneous linear ODEs obtained by putting (1.2) into (1.1) and equating the coefficients of ε^i of both sides of (1.1). It is well known that approximate solutions constructed by the naive expansion are valid only in a time interval of $O(1)$ in general, since secular terms diverge as $t \to \infty$. Many techniques for obtaining approximate solutions which are valid in a long time interval have been developed until now; these are collectively called singular perturbation methods.

The RG method proposed by Chen, Goldenfeld, and Oono is one of the singular perturbation methods that look like the variation-of-constant method, in which the secular terms

*Received by the editors June 20, 2007; accepted for publication (in revised form) by T. Kaper April 22, 2008; published electronically July 25, 2008.
http://www.siam.org/journals/siads/7-3/69489.html
†Department of Applied Mathematics and Physics, Kyoto University, Kyoto, 606-8501, Japan (chiba@amp.i.kyoto-u.ac.jp).

included in $x_1(t), x_2(t), \dots$ of (1.2) are renormalized into the integral constant of $x_0(t)$. The ODE to be satisfied by the renormalized integral constant is called the *RG equation*. Chen, Goldenfeld, and Oono showed that the RG method unifies the conventional singular perturbation methods, such as the multiscale method, the boundary layer technique, WKB analysis, and the reductive perturbation method, by giving explicit examples. Though the multiscale method requires occasionally fractional power laws or logarithmic functions of ε in the expansion of $x(t)$, the RG method needs only a power-series expansion of $x(t)$ in ε, and it starts with the naive expansion of $x(t)$ to reach the same result the multiscale method does.

Kunihiro [3, 4] interpreted the RG method as a theory of envelopes for approximate solutions constructed by the naive expansion. His insight revealed why the RG method works well. Nozaki and Oono [5] and Goto, Masutomi, and Nozaki [6] proposed a proto-RG equation or translational Lie group method to renormalize secular terms up to arbitrary order and to obtain higher order approximate solutions. Ei, Fujii, and Kunihiro [7] apply the RG method to obtain approximate center manifolds and slow manifolds. Ziane [8] and DeVille et al. [9] proved that an orbit constructed on the RG method approximates an exact solution for a long time interval. Further, DeVille et al. [9] showed that if the unperturbed part of a given ODE is linear and diagonalizable, the RG equation for the ODE is equivalent to the normal form of the vector field.

Despite the active interest in the RG method, little attention has been paid to date to the question of whether a family of approximate solutions to exact solutions of the original ODE (vector field), which is obtained by varying initial values, forms a well-defined vector field or not. Put another way, a question to be asked is whether approximate solutions intersect with one other or not. Further, the RG method has been applied to differential equations only on the Euclidean space but has not yet been extended to a method applicable to differential equations on manifolds.

In the present paper, it is shown that for a given vector field of the form $f(t,x)+\varepsilon g(t,x)$ on an arbitrary manifold, approximate solutions obtained by the RG method define a vector field which is close to the original vector field in the C^1 topology on appropriate assumptions of boundedness for the flow of $f(t,x)$ and for other functions. This implies that the approximate vector field works well in investigating properties of the original vector field that are persistent under C^1 perturbation. In particular, if the approximate vector field has a normally hyperbolic invariant manifold, then the original vector field is expected also to have an invariant manifold because the Fenichel theory ensures that normally hyperbolic invariant manifolds are persistent under C^1 perturbation. In fact, it is shown that the existence of an invariant manifold and its stability are inherited from those of the RG equation since the flow of the RG equation is proved to be conjugate to that of the approximate vector field. In view of this, it is desirable that the RG equation be easier to solve than the original equation. In fact, it will be proved that the RG equation has larger symmetry than the original equation. This method will be applied in the bifurcation theory to show that a periodic orbit emerges far away from a fixed point, which is an example of the global bifurcation other than the ordinary Hopf bifurcation.

In particular, the RG method is applied to a time-dependent linear equation of the form

$$\dot{x} = F(t)x + \varepsilon G(t)x, \quad x \in \mathbf{R}^n, \tag{1.3}$$

where $F(t)$ and $G(t)$ are $n \times n$ matrix functions. On appropriate assumptions, the stability of the trivial solution $x = 0$ of (1.3) is shown to coincide with that of the RG equation for (1.3), which is a time-*independent* linear equation. By using this result, synchronous solution of coupled oscillators is shown to be stable.

This paper is organized as follows: Section 2 presents basic facts and definitions in dynamical systems. Section 3 contains a simple example of the RG method. In section 4, a main theorem on approximate vector fields is proved. Section 5 gives a few properties of the RG equation in term of symmetries. In section 6, an invariant manifold of a given equation is shown to be inherited from its RG equation. In section 7, the RG method is applied to time-dependent linear equations (1.3). In Appendix A, we discuss the higher order RG equation to prove Theorem 6.1.

2. Notation. Let f be a time-independent C^r vector field on a C^r manifold M and $\varphi : \mathbf{R} \times M \to M$ its flow. We denote by $\varphi_t(x_0) \equiv x(t)$, $t \in \mathbf{R}$, a solution to the ODE $\dot{x} = f(x)$ through $x_0 \in M$, which satisfies $\varphi_t \circ \varphi_s = \varphi_{t+s}$, $\varphi_0 = \mathrm{id}_M$, where id_M denotes the identity map of M. For a fixed $t \in \mathbf{R}$, $\varphi_t : M \to M$ defines a diffeomorphism of M. We assume φ_t is defined for all $t \in \mathbf{R}$.

For a time-dependent vector field, let $x(t, \tau, \xi)$ denote a solution to an ODE $\dot{x}(t) = f(t, x)$ through ξ at $t = \tau$, which defines a flow $\varphi : \mathbf{R} \times \mathbf{R} \times M \to M$ by $\varphi_{t,\tau}(\xi) = x(t, \tau, \xi)$. For fixed $t, \tau \in \mathbf{R}$, $\varphi_{t,\tau} : M \to M$ is a diffeomorphism of M satisfying

$$\varphi_{t,t'} \circ \varphi_{t',\tau} = \varphi_{t,\tau}, \quad \varphi_{t,t} = \mathrm{id}_M. \tag{2.1}$$

Conversely, a family of diffeomorphisms $\varphi_{t,\tau}$ of M, which are C^1 with respect to t and τ, satisfying the above equality for any $t, \tau \in \mathbf{R}$ defines a time-dependent vector field on M through

$$f(t, x) = \frac{d}{d\tau}\Big|_{\tau=t} \varphi_{\tau,t}(x). \tag{2.2}$$

3. A brief review of the renormalization group method. Before describing a general theory of the RG method in the next section, we review the RG method for obtaining approximate solutions of an ODE with a simple example.

Let us consider an ODE

$$\ddot{x} + x + \varepsilon x^3 = 0, \quad x \in \mathbf{R}, \ |\varepsilon| << 1. \tag{3.1}$$

Assume that the ODE admits a solution of the form $x(t) = x_0(t) + \varepsilon x_1(t) + O(\varepsilon^2)$. Then the substitution provides

$$\ddot{x}_0 + \varepsilon \ddot{x}_1 + x_0 + \varepsilon x_1 + \varepsilon (x_0 + \varepsilon x_1)^3 + O(\varepsilon^2) = 0.$$

Expanding this into a power series in ε and equating the coefficients of $\varepsilon^0, \varepsilon^1$ to zero, respectively, we get

$$\ddot{x}_0 + x_0 = 0, \tag{3.2}$$

$$\ddot{x}_1 + x_1 = -x_0^3. \tag{3.3}$$

We denote a general solution of the former whose initial time is $t = 0$ by

$$x_0(t, 0, A) = Ae^{it} + \overline{A}e^{-it}, \quad A \in \mathbf{C}. \tag{3.4}$$

Then (3.3) and (3.4) are put together to give

$$\ddot{x}_1 + x_1 = -(A^3 e^{3it} + 3|A|^2 Ae^{it} + 3|A|^2 \overline{A}e^{-it} + \overline{A}^3 e^{-3it}).$$

A special solution of this equation, whose initial time is $t = \tau$, is written as

$$x_1(t, \tau; A) = \frac{A^3}{8}e^{3it} + \frac{3i}{2}|A|^2 A(t - \tau)e^{it} + \text{c.c.}, \tag{3.5}$$

where c.c. is the complex conjugate of the first two terms of the right-hand side. Note that a secular term arises which diverges to infinity as $t \to \infty$. The reason for taking the initial time $t = \tau$ is that we want to construct a family of curves parameterized by τ since approximate solutions obtained by the RG method are given as envelopes of the family (see Kunihiro [3, 4]).

Now let us define $\hat{x}$ as

$$\hat{x}(t, \tau; A) = x_0(t, 0, A) + \varepsilon x_1(t, \tau; A).$$

Then $\hat{x}$ is an approximate solution to (3.1) on short time intervals. Indeed, $\hat{x}$ satisfies the equation

$$\ddot{\hat{x}} + \hat{x} + \varepsilon\hat{x}^3 = 3\varepsilon^2(Ae^{it} + \overline{A}e^{-it})^2 \left(\frac{A^3}{8}e^{3it} + \frac{3i}{2}|A|^2 A(t - \tau)e^{it} + \text{c.c.}\right) + O(\varepsilon^3), \tag{3.6}$$

which implies that if A is bounded and t is sufficiently close to τ, then $\hat{x}$ approximates to an exact solution of (3.1) well. This procedure for obtaining a local approximate solution is called *naive expansion*.

The RG method employs two additional steps to obtain solutions approximating to exact solutions on a long time intervals. At first, we regard the constant A as a differentiable function of τ and differentiate $\hat{x}$ with respect to τ at t:

$$\begin{aligned}\frac{d\hat{x}}{d\tau}\Big|_{\tau=t}(t, \tau, A(\tau)) &= \frac{\partial x_0}{\partial A}\frac{dA}{d\tau}\Big|_{\tau=t} + \varepsilon\frac{\partial x_1}{\partial \tau}\Big|_{\tau=t} + \varepsilon\frac{\partial x_1}{\partial A}\frac{dA}{d\tau}\Big|_{\tau=t} \\ &= A'e^{it} + \overline{A}'e^{-it} + \varepsilon\left(-\frac{3i}{2}|A|^2 Ae^{it} + \frac{3A^2}{8}A'e^{3it} + \text{c.c.}\right).\end{aligned}$$

We impose the condition on $A(t)$ that $d\overline{x}/d\tau|_{\tau=t} = 0$, which is called the *RG condition*. Then we obtain the following ODE for $A(t)$:

$$\frac{dA}{dt} = \varepsilon\frac{3i}{2}|A|^2 A + O(\varepsilon^2).$$

Truncating the higher order term $O(\varepsilon^2)$, we obtain the *RG equation*

$$\frac{dA}{dt} = \varepsilon\frac{3i}{2}|A|^2 A, \tag{3.7}$$

which is solved by

$$A(t) := A(t,a,\theta) = \frac{1}{2}a\exp i\left(\frac{3\varepsilon}{8}a^2t+\theta\right), \tag{3.8}$$

where a,θ are arbitrary constants. With this $A(t)$, we define $X(t,a,\theta)$ by

$$X(t,a,\theta) := \hat{x}(t,t,A(t,a,\theta)). \tag{3.9}$$

Then this $X(t)$ gives a solution which approximates an exact solution of (3.1) for a long time interval. The condition $d\overline{x}/d\tau|_{\tau=t}=0$ means that the curve $X(t,a,\theta)=\hat{x}(t,t;A(t,a,\theta))$ is an envelope for the family of curves $\{\hat{x}(t,\tau;A(\tau,a,\theta))\}_{\tau\in\mathbf{R}}$ (see Kunihiro [3, 4]). Our general definition of the RG equation is shown in the next section.

4. Main theorem. In this section, under appropriate assumptions, we prove that a family of orbits constructed by the RG method defines a vector field which approximates the original vector field in the C^1 topology. Though we show Theorem 4.4 for vector fields on Euclidean space, it can be easily extended to vector fields on an arbitrary manifold. See Remark 4.5.

Let $f(t,x)$ and $g(t,x)$ be C^4 and C^3 time-dependent vector fields on $\mathbf{R}^n$, respectively, and consider an ODE

$$\dot{x}(t) = f(t,x)+\varepsilon g(t,x) \tag{4.1}$$

and its unperturbed system

$$\dot{x}_0(t) = f(t,x_0). \tag{4.2}$$

We denote a general solution to the latter by

$$x_0(t) := x_0(t,0,A) = \varphi^0_{t,0}(A), \tag{4.3}$$

whose initial value is $x_0(0)=A\in\mathbf{R}^n$ at $t=0$, and where φ^0 is its flow. With this x_0, we further consider an ODE

$$\dot{x}_1(t) = \frac{\partial f}{\partial x}(t,x_0)x_1+g(t,x_0). \tag{4.4}$$

A general solution to this equation is written as

$$x_1 = (D\varphi^0_{t,0})_A\circ(D\varphi^0_{\tau,0})_A^{-1}h(\tau,A)+(D\varphi^0_{t,0})_A\int_\tau^t(D\varphi_{s,0})_A^{-1}g(s,\varphi^0_{s,0}(A))ds, \tag{4.5}$$

where τ is an initial time, $h(\tau,A)$ is an initial value, and $(D\varphi^0_{t,0})_A$ is the derivative of $\varphi^0_{t,0}$ at A. In what follows, we denote by $\mathbf{R}_{\geq T}$ the set of the real numbers which are larger than or equal to $T\in\mathbf{R}$: $\mathbf{R}_{\geq T}=\{t\in\mathbf{R}\mid t\geq T\}$. Set $\mathbf{R}_{\geq T}=\mathbf{R}$ if $T=-\infty$.

Definition 4.1. *A function $p(t)$ is said to be Krylov–Bogolyubov–Mitropolskii (KBM) on $\mathbf{R}_{\geq T}$ if the number*

$$\lim_{t\to\infty}\frac{1}{t-t_0}\int_{t_0}^t p(s)ds \tag{4.6}$$

converges for all $t_0 \geq T$.

The notation of KBM vector fields was introduced in [14] and used in DeVille et al. [9] to define the RG equation. Note that periodic functions and almost periodic functions are KBM on $\mathbf{R}$ (see Fink [13]).

The next definition is proposed by DeVille et al. [9].

Definition 4.2. *Suppose that* $(D\varphi^0_{t,0})^{-1}_A g(t, \varphi^0_{t,0}(A))$ *is KBM on* $\mathbf{R}_{\geq T}$ *for each* $A \in \mathbf{R}^n$. *Then a* C^3 *function* $R : \mathbf{R}^n \to \mathbf{R}^n$ *defined by*

$$R(A) = \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (D\varphi^0_{s,0})^{-1}_A g(s, \varphi^0_{s,0}(A)) ds \tag{4.7}$$

is called the resonance *or* secular *part for the solution* x_1 *defined by* (4.5).

By using (4.7), equation (4.5) is rewritten as

$$x_1 = (D\varphi^0_{t,0})_A \circ (D\varphi^0_{\tau,0})^{-1}_A h(\tau, A) + (D\varphi^0_{t,0})_A \int_\tau^t \left((D\varphi^0_{s,0})^{-1}_A g(s, \varphi^0_{s,0}(A)) - R(A)\right) ds + (D\varphi^0_{t,0})_A R(A)(t-\tau).$$

Define the initial value $h(\tau, A)$ to be

$$h(\tau, A) := (D\varphi^0_{\tau,0})_A \int^\tau \left((D\varphi^0_{s,0})^{-1}_A g(s, \varphi^0_{s,0}(A)) - R(A)\right) ds, \tag{4.8}$$

where $\int^\tau$ is the indefinite integral, whose integral constant is fixed arbitrarily. Then, x_1 is expressed as

$$\begin{aligned} x_1 := x_1(t, \tau; A) &= (D\varphi^0_{t,0})_A \int^t \left((D\varphi^0_{s,0})^{-1}_A g(s, \varphi^0_{s,0}(A)) - R(A)\right) ds + (D\varphi^0_{t,0})_A R(A)(t-\tau) \\ &= h(t, A) + (D\varphi^0_{t,0})_A R(A)(t-\tau). \end{aligned} \tag{4.9}$$

In perturbation theory, the second term of the right-hand side is called the *secular term*. The reason for defining the initial value $h(\tau, A)$ as (4.8) is that we want to divide x_1 into two terms: one is the secular term which diverges as $t \to \infty$, and the other is the bounded term $h(t, A)$ (see also the norm conditions (N) below). With this $x_1(t, \tau; A)$, we associate a curve defined by

$$\hat{x}(t) := \hat{x}(t, \tau; A) = x_0(t, 0, A) + \varepsilon x_1(t, \tau; A), \tag{4.10}$$

which provides a locally approximate solution of (4.1). Now we define the RG equation.

Definition 4.3. *Suppose that* $(D\varphi^0_{t,0})^{-1}_A g(t, \varphi^0_{t,0}(A))$ *is KBM on* $\mathbf{R}_{\geq T}$ *for each* $A \in \mathbf{R}^n$. *Then, the equation defined by*

$$\frac{dA}{dt} = \varepsilon R(A), \quad A \in \mathbf{R}^n, \tag{4.11}$$

is called the RG equation *for* $f + \varepsilon g$, *and the vector field* $\varepsilon R(A)$ *on* $\mathbf{R}^n$ *is called the* RG vector field *for* $f + \varepsilon g$. *We denote by* φ^{RG}_t *the flow generated by the RG vector field.*

In the literature, the RG equation is defined so that its solution $A := A(t)$ may satisfy $d\hat{x}/d\tau|_{\tau=t}(t,\tau;A(\tau)) = 0$. According to our definition of the RG vector field, $d\hat{x}/d\tau|_{\tau=t}$ is calculated as

$$\frac{d\hat{x}}{d\tau}\Big|_{\tau=t}(t,\tau;A(\tau)) = \varepsilon^2 \frac{\partial x_1}{\partial A}(t,t;A(t))R(A(t)). \tag{4.12}$$

The higher order term $O(\varepsilon^2)$ is truncated; (4.12) then implies that solutions to (4.11) satisfy $d\hat{x}/d\tau|_{\tau=t}(t,\tau;A(\tau)) = 0$.

To state our main theorem, we assume the following norm conditions (N) for the functions $f(t,x)$, $g(t,x)$, $x_0(t,0,A)$, and $h(t,A) = x_1(t,t;A)$ on $\mathbf{R}_{\geq T} \times \mathbf{R}^n$. These conditions will be used to prove that the vector field F_ε defined in (4.16) is sufficiently close to the original vector field $f + \varepsilon g$ in the C^1 topology (see (4.18), (4.19)).

Norm conditions (N). Let $K \subset \mathbf{R}^n$ be an arbitrary compact subset. We assume that there exists T such that $(D\varphi^0_{t,0})^{-1}_A g(t, \varphi^0_{t,0}(A))$ is KBM on $\mathbf{R}_{\geq T}$ for each $A \in K$ and the following functions are bounded uniformly on $\mathbf{R}_{\geq T} \times K$:
(N1) $h(t,A)$,
(N2) $\partial^2 f/\partial x^2$, $\partial f/\partial x$, $\partial g/\partial x$, $x_0(-t,0,A)$, $(\partial x_0/\partial A)^{-1}$, $\partial^2 x_0/\partial A^2$, $\partial h/\partial A$, $\partial h^2/\partial A^2$,
(N3) f, $\partial^2 f/\partial x \partial t$, $\partial^3 f/\partial x^3$, $\partial^3 f/\partial x^2 \partial t$, g, $\partial^2 g/\partial x^2$, $\partial^2 g/\partial x \partial t$, $\partial^3 x_0/\partial A^3$, $\partial^3 h/\partial A^3$.

In section 6 and Appendix A, we consider a system of the form

$$\dot{x} = Fx + \varepsilon g(t,x), \quad x \in \mathbf{R}^n, \tag{4.13}$$

where F is a diagonalizable $n \times n$ constant matrix, all of whose eigenvalues lie on the imaginary axis. In this case, the following are sufficient conditions for this system to satisfy the norm conditions (N1) to (N3):

(i) $g(t,x)$ is polynomial in x and periodic in t.
(ii) $g(t,x)$ is polynomial in x and almost periodic in t, the set of whose Fourier exponents has no accumulation points.

See Appendix A for the proof. The case where F has eigenvalues on the left half plane will be treated in a forthcoming paper. In Example 4.7, we show another example satisfying norm conditions (N) whose unperturbed part is nonlinear.

In what follows, we fix an open subset $U \subset \mathbf{R}^n$ such that $\overline{U}$ is compact. Define $\alpha_t : U \to \mathbf{R}^n$ to be

$$\alpha_t(A) = x_0(t,0,A) + \varepsilon h(t,A) \tag{4.14}$$

for all $t \in \mathbf{R}_{\geq T}$. The set U is defined so that α_t is a diffeomorphism on U (see the proof of Theorem 4.4(i) below). Note that the smaller $|\varepsilon|$ is, the larger set U we can take.

Our main theorem is stated as follows.

Theorem 4.4. *Let f, g, $x_0(t,0,A)$, $x_1(t,\tau;A)$ be vector fields and solutions to differential equations defined in (4.1) to (4.4) and (4.9), respectively. Let $\varepsilon R(A)$ be the RG vector field for $f + \varepsilon g$, and denote its integral curves, whose initial time is t_0 and initial value is $\xi \in U$, by $A(t) := A(t,t_0,\xi) = \varphi^{RG}_{t-t_0}(\xi)$. Then, there exist $\varepsilon_0 > 0$ such that the following hold for all $|\varepsilon| < \varepsilon_0$:*

(i) *Suppose that the norm condition* (N1) *is satisfied. Then,*

$$\Phi_{t,t_0} := \alpha_t \circ \varphi^{RG}_{t-t_0} \circ \alpha_{t_0}^{-1} : \alpha_{t_0}(U) \to \mathbf{R}^n \tag{4.15}$$

defines a flow on $U_\varepsilon := \{(t,x) \mid t \in \mathbf{R}_{\geq T},\ x \in \alpha_t(U)\}$ *associated with a time-dependent vector field*

$$F_\varepsilon(t,x) := \frac{d}{da}\Big|_{a=t} \Phi_{a,t}(x). \tag{4.16}$$

The integral curves of F_ε *are put in the form*

$$X(t,t_0;\xi) := \hat{x}(t,t;A(t,t_0,\xi)), \tag{4.17}$$

where $\hat{x}$ *is defined by* (4.10).

(ii) *Suppose that the norm conditions* (N1)–(N2) *are satisfied. Then, there exists a non-negative constant* L_1 *such that the vector field* F_ε *defined by* (4.16) *satisfies an inequality*

$$\sup_{U_\varepsilon} ||f + \varepsilon g - F_\varepsilon|| < \varepsilon^2 L_1. \tag{4.18}$$

(iii) *Suppose that the norm conditions* (N1)–(N3) *are satisfied. Then, there exists a non-negative constant* L_2 *such that the vector field* F_ε *satisfies an inequality*

$$\sup_{U_\varepsilon} ||D_{t,x} f + \varepsilon D_{t,x} g - D_{t,x} F_\varepsilon|| < \varepsilon^2 L_2, \tag{4.19}$$

where $D_{t,x} f = (\partial f/\partial t, \partial f/\partial x)$ *and* $||D_{t,x} f|| = ||\partial f/\partial x|| + ||\partial f/\partial t||$. *In particular,* F_ε *is sufficiently close to* $f + \varepsilon g$ *in the* C^1 *topology if* $|\varepsilon|$ *is sufficiently small.*

Proof of (i). Since $h(t,x)$ is bounded on $\mathbf{R}_{\geq T} \times U$ by the norm condition (N1), $\varepsilon h(t,x)$ can be sufficiently close to a null function as a C^3 function of x for sufficiently small ε. Since the flow $\varphi^0_{t,0}$ is a C^4 diffeomorphism and since the set of diffeomorphisms is open in the space of C^1 maps in the C^1 topology, it follows that for a sufficiently small ε, the map α_t given by (4.14) is a diffeomorphism from U into $\mathbf{R}^n$ for each $t \in \mathbf{R}_{\geq T}$. Therefore, the map $\Phi_{t,t_0} : \alpha_{t_0}(U) \to \mathbf{R}^n$ defined by (4.15) is a diffeomorphism from $\alpha_{t_0}(U)$ into $\mathbf{R}^n$ as well and satisfies $\Phi_{t,t'} \circ \Phi_{t',t_0} = \Phi_{t,t_0}$, $\Phi_{t,t} = id_{\alpha_t(U)}$. This shows that Φ_{t,t_0} is a flow associated with a vector field F_ε defined by (4.16). Then, it turns out that

$$\Phi_{t,t_0}(\alpha_{t_0}(\xi)) = \alpha_t \circ \varphi^{RG}_{t-t_0}(\xi) = \alpha_t(A(t,t_0,\xi)) = \hat{x}(t,t;A(t,t_0,\xi)) = X(t,t_0;\xi),$$

which implies that $X(t,t_0;\xi)$ gives an integral curve of F_ε, namely,

$$\frac{dX}{dt}(t,t_0;\xi) = F_\varepsilon(t, X(t,t_0;\xi)). \tag{4.20}$$

This ends the proof. ■

Proof of (ii), (iii). Denote $h(t,A)$ as $h_t(A)$. The vector field $F_\varepsilon(t,x)$ is calculated as

$$\begin{aligned}
F_\varepsilon(t,x) &= \frac{d}{da}\Big|_{a=t}\left((\varphi^0_{a,0}+\varepsilon h_a)\circ\varphi^{RG}_{a-t}\circ\alpha_t^{-1}(x)\right)\\
&= \frac{d}{da}\Big|_{a=t}(\varphi^0_{a,0}+\varepsilon h_a)\circ\alpha_t^{-1}(x)+(D\varphi^0_{t,0}+\varepsilon Dh_t)_{\alpha_t^{-1}(x)}\frac{d}{da}\Big|_{a=t}\varphi^{RG}_{a-t}\circ\alpha_t^{-1}(x)\\
&= f(t,x_0(t,0,\alpha_t^{-1}(x)))+\varepsilon\frac{\partial f}{\partial x}(t,x_0(t,0,\alpha_t^{-1}(x)))x_1(t,t;\alpha_t^{-1}(x))+\varepsilon g(t,x_0(t,0,\alpha_t^{-1}(x)))\\
&\quad+\varepsilon\frac{d}{da}\Big|_{a=t}x_1(t,a,\alpha_t^{-1}(x))+\varepsilon(D\varphi^0_{t,0}+\varepsilon Dh_t)_{\alpha_t^{-1}(x)}R(\alpha_t^{-1}(x))\\
&= f(t,x_0(t,0,\alpha_t^{-1}(x)))+\varepsilon g(t,x_0(t,0,\alpha_t^{-1}(x)))\\
&\quad+\varepsilon\frac{\partial f}{\partial x}(t,x_0(t,0,\alpha_t^{-1}(x)))h_t(\alpha_t^{-1}(x))+\varepsilon^2(Dh_t)_{\alpha_t^{-1}(x)}R(\alpha_t^{-1}(x)).
\end{aligned}$$

On account of $\alpha_t(x)=x_0(t,0,x)+\varepsilon h_t(x)$, the above equation is expanded as

$$\begin{aligned}
F_\varepsilon(t,x) &= f(t,x)+\varepsilon\frac{df}{d\varepsilon}\Big|_{\varepsilon=0}(t,x_0(t,0,\alpha_t^{-1}(x)))+\frac{\varepsilon^2}{2}\frac{d^2f}{d\varepsilon^2}\Big|_{\varepsilon=\theta_1\varepsilon}(t,x_0(t,0,\alpha_t^{-1}(x)))+\varepsilon g(t,x)\\
&\quad+\varepsilon^2\frac{dg}{d\varepsilon}\Big|_{\varepsilon=\theta_2\varepsilon}(t,x_0(t,0,\alpha_t^{-1}(x)))+\varepsilon\frac{\partial f}{\partial x}(t,x)h_t((\varphi^0_{t,0})^{-1}(x))\\
&\quad+\varepsilon^2\frac{\partial f}{\partial x}(t,x)\frac{dh_t}{d\varepsilon}\Big|_{\varepsilon=\theta_3\varepsilon}(\alpha_t^{-1}(x))\\
&\quad+\varepsilon^2\frac{d}{d\varepsilon}\Big|_{\varepsilon=\theta_4\varepsilon}\left(\frac{\partial f}{\partial x}(t,x_0(t,0,\alpha_t^{-1}(x)))\right)h_t(\alpha_t^{-1}(x))+\varepsilon^2(Dh_t)_{\alpha_t^{-1}(x)}R(\alpha_t^{-1}(x)),
\end{aligned}$$

where $0<\theta_1,\theta_2,\theta_3,\theta_4<1$ are constants in the Taylor formula. The second term of the right-hand side of the above is calculated as

$$\begin{aligned}
\frac{df}{d\varepsilon}\Big|_{\varepsilon=0}(t,x_0(t,0,\alpha_t^{-1}(x))) &= \frac{\partial f}{\partial x}(t,x)\frac{\partial x_0}{\partial A}(t,0,(\varphi^0_{t,0})^{-1}(x))\frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\alpha_t^{-1}(x)\\
&= -\frac{\partial f}{\partial x}(t,x)h_t((\varphi^0_{t,0})^{-1}(x)).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
F_\varepsilon(t,x)-f(t,x)-\varepsilon g(t,x) &= \frac{\varepsilon^2}{2}\frac{d^2f}{d\varepsilon^2}\Big|_{\varepsilon=\theta_1\varepsilon}(t,x_0(t,0,\alpha_t^{-1}(x)))+\varepsilon^2\frac{dg}{d\varepsilon}\Big|_{\varepsilon=\theta_2\varepsilon}(t,x_0(t,0,\alpha_t^{-1}(x)))\\
&\quad+\varepsilon^2\frac{\partial f}{\partial x}(t,x)\frac{dh_t}{d\varepsilon}\Big|_{\varepsilon=\theta_3\varepsilon}(\alpha_t^{-1}(x))\\
&\quad+\varepsilon^2\frac{d}{d\varepsilon}\Big|_{\varepsilon=\theta_4\varepsilon}\left(\frac{\partial f}{\partial x}(t,x_0(t,0,\alpha_t^{-1}(x)))\right)h_t(\alpha_t^{-1}(x))\\
&\quad+\varepsilon^2(Dh_t)_{\alpha_t^{-1}(x)}R(\alpha_t^{-1}(x)).
\end{aligned}\tag{4.21}$$

We have to estimate the norm of the right-hand side of the above equation. At first, $df/d\varepsilon$ is given by

$$\frac{df}{d\varepsilon}(t, x_0(t, 0, \alpha_t^{-1}(x)))$$

$$= -\frac{\partial f}{\partial x}(t, x_0(t, 0, \alpha_t^{-1}(x)))\frac{\partial x_0}{\partial A}(t, 0, \alpha_t^{-1}(x))\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1} h_t(\alpha_t^{-1}(x)). \tag{4.22}$$

Note that equations

$$\alpha_t^{-1}(x) = (\varphi_{t,0}^0 + \varepsilon h_t)^{-1}(x) = (id + \varepsilon(\varphi_{t,0}^0)^{-1} \circ h_t)^{-1} \circ (\varphi_{t,0}^0)^{-1}(x), \tag{4.23}$$

$$x_0(t, 0, \alpha_t^{-1}(x)) = \varphi_{t,0}^0 \circ \alpha_t^{-1}(x) = (id - \varepsilon h_t \circ \alpha_t^{-1})(x), \tag{4.24}$$

$$\frac{\partial x_0}{\partial A}(t, 0, \alpha_t^{-1}(x))\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1}$$

$$= id - \varepsilon\left(\frac{\partial h_t}{\partial A}\right)_{\alpha_t^{-1}(x)} \sum_{k=0}^{\infty}\left(-\varepsilon\left(\frac{\partial x_0}{\partial A}\right)_{\alpha_t^{-1}(x)}^{-1} \circ \left(\frac{\partial h_t}{\partial A}\right)_{\alpha_t^{-1}(x)}\right)^k \circ \left(\frac{\partial x_0}{\partial A}\right)_{\alpha_t^{-1}(x)}^{-1} \tag{4.25}$$

hold and the left-hand sides of the above three equations are bounded by the norm conditions (N1)–(N2). Therefore, the right-hand side of (4.22) is bounded uniformly in $\mathbf{R}_{\geq T}$. To show the boundedness of the first term of the right-hand side of (4.21), it is sufficient to show that the derivative of each factor of the right-hand side of (4.22) is bounded. They are calculated as

$$\frac{d}{d\varepsilon}\frac{\partial f}{\partial x}(t, x_0(t, 0, \alpha_t^{-1}(x)))$$

$$= -\frac{\partial^2 f}{\partial x^2}(t, x_0(t, 0, \alpha_t^{-1}(x)))\frac{\partial x_0}{\partial A}(t, 0, \alpha_t^{-1}(x))\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1} h_t(\alpha_t^{-1}(x)), \tag{4.26}$$

$$\frac{d}{d\varepsilon}\frac{\partial x_0}{\partial A}(t, 0, \alpha_t^{-1}(x))$$

$$= -\frac{\partial^2 x_0}{\partial A^2}(t, 0, \alpha_t^{-1}(x))\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1} h_t(\alpha_t^{-1}(x)), \tag{4.27}$$

$$\frac{d}{d\varepsilon}\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1} \tag{4.28}$$

$$= -\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1} \frac{d}{d\varepsilon}\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1},$$

$$\frac{d}{d\varepsilon}\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)$$

$$= \left(\frac{\partial h_t}{\partial A}\right)_{\alpha_t^{-1}(x)} - \left(\frac{\partial^2}{\partial A^2}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1} h_t(\alpha_t^{-1}(x)), \tag{4.29}$$

$$\frac{d}{d\varepsilon}h_t(\alpha_t^{-1}(x))$$

$$= -\left(\frac{\partial h_t}{\partial A}\right)_{\alpha_t^{-1}(x)}\left(\frac{\partial}{\partial A}(\varphi_{t,0}^0 + \varepsilon h_t)_{\alpha_t^{-1}(x)}\right)^{-1} h_t(\alpha_t^{-1}(x)). \tag{4.30}$$

By the norm conditions and (4.23), (4.24), and (4.25), these are bounded uniformly in $\mathbf{R}_{\geq T}$. Therefore, the first term of the right-hand side of (4.21) is bounded.

The boundedness of the second term of the right-hand side of (4.21) is verified from (4.22) by using g instead of f, and the boundedness of the other terms of the right-hand side of (4.21) is verified from (4.26), (4.30), and the norm conditions (N1)–(N2). This proves Theorem 4.4(ii). Theorem 4.4(iii) is verified by differentiating both sides of (4.21) with respect to x, t and estimating the norm as above. This calculation is elementary and is omitted here. ■

Remark 4.5. Though we have treated the vector field F_ε on an open set of $\mathbf{R}^n$, the vector field F_ε may be defined in the case of an arbitrary manifold M. Let $\{U_i\}_{i\in\Lambda}$ be an open covering of M such that each $\overline{U}_i$ is compact. We identify U_i with an open subset on $\mathbf{R}^n$. Suppose that $U_i \cap U_j \neq \emptyset$ and let $\psi_{ij} : U_i \cap U_j \to U_i \cap U_j$ be a coordinate transformation function from U_i to U_j. Let $\varepsilon R^i(A)$ and $\varepsilon R^j(A)$ be the RG vector fields constructed on U_i and U_j, respectively, and let $\varphi_t^{RG(i)}$, $\varphi_t^{RG(j)}$ be respective flows. By (4.7), it is easy to verify that $R^i(A) = (D\psi_{ij})^{-1}R^j(\psi_{ij}(A))$ and $\varphi_t^{RG(i)} = \psi_{ij}^{-1} \circ \varphi_t^{RG(j)} \circ \psi_{ij}$. Let F_ε^i, F_ε^j be approximate vector fields constructed on U_i, U_j defined by (4.16), respectively. Then F_ε^i is transformed by the coordinate transformation as follows:

$$\begin{aligned} D\psi_{ij}F_\varepsilon^i(t,x) &= D\psi_{ij}\frac{d}{da}\Big|_{a=t}\Phi_{a,t}(x) \\ &= \frac{d}{da}\Big|_{a=t}\psi_{ij}\circ\alpha_t\circ\varphi_{t-t_0}^{RG(i)}\circ\alpha_{t_0}^{-1}(x) \\ &= \frac{d}{da}\Big|_{a=t}\psi_{ij}\circ(x_0+\varepsilon h)\circ\psi_{ij}^{-1}\circ(\psi_{ij}\circ\varphi_{t-t_0}^{RG(i)}\circ\psi_{ij}^{-1}) \\ &\qquad\qquad \circ(\psi_{ij}\circ(x_0+\varepsilon h)\circ\psi_{ij}^{-1})^{-1}(\psi_{ij}(x)), \end{aligned}$$

where $\psi_{ij}\circ x_0(t,0,\psi_{ij}^{-1}(x))$ and $\psi_{ij}\circ h(t,\psi_{ij}^{-1}(x)) = \psi_{ij}\circ x_1(t,t,\psi_{ij}^{-1}(x))$ are coordinate representations on U_j of $x_0(t,0,x)$ and of $x_1(t,t,x)$, respectively, which are represented in the coordinates on U_i. This means that

$$D\psi_{ij}F_\varepsilon^i(t,x) = F_\varepsilon^j(t,\psi_{ij}(x)), \quad x\in U_i. \tag{4.31}$$

Let $\{\rho_i\}_{i\in\Lambda}$ be a partition of unity subordinate to the cover $\{U_i\}_{i\in\Lambda}$ and define $F_\varepsilon(t,x) := \sum_{i\in\Lambda}\rho_i(x)F_\varepsilon^i(t,x)$; then F_ε is a well-defined vector field on M which approximates to $f+\varepsilon g$.

Remark 4.6. Now that we have the approximate vector field $F_\varepsilon(t,x) = f(t,x)+\varepsilon g(t,x)+O(\varepsilon^2)$, the Gronwall inequality immediately proves the error estimate for approximate solutions.

Let $x(t,t_0)$ be a solution of (4.1) satisfying the norm conditions (N) whose initial time is t_0. Let $X(t,t_0;\xi)$ be a curve defined by (4.17). Suppose that $x(t_0,t_0) = X(t_0,t_0;\xi) \in \alpha_t(U)$. Then, there exist positive constants ε_0, T, C such that the inequality

$$||x(t,t_0) - X(t,t_0;\xi)|| < C\varepsilon, \quad 0 < t < T/\varepsilon, \tag{4.32}$$

holds for $0 < \varepsilon < \varepsilon_0$.

This fact was essentially proved in Ziane [8] and DeVille et al. [9]. Note that DeVille et al. also treated the case that the norm conditions (N) are not satisfied—for example,

$g(t,x) = x/\sqrt{t}$. The above fact is also followed by putting $m = 1$ and replacing e^{Ft} by $(D\varphi^0_{t,0})_A$ in the proof of Theorem A.8, in which the error estimate for a higher order case by using the higher order RG equation is proved.

In the next example, the RG method is applied to a vector field whose unperturbed part is nonlinear. Application to vector fields with linear unperturbed parts will be treated in section 6.

Example 4.7. Consider a system on $\{(x,y) \mid x > 0,\ y \in \mathbf{R}\} \subset \mathbf{R}^2$,

$$\begin{cases} \dot{x} = xy + \varepsilon xy^2, \\ \dot{y} = -\log x + \varepsilon y, \end{cases} \tag{4.33}$$

where $\varepsilon \in \mathbf{R}$ is a small constant. Note that unperturbed part is nonlinear. In order to obtain approximate solutions to (4.33), we apply the RG method. The unperturbed system of (x_0, y_0) is written as $\dot{x}_0 = x_0 y_0$, $\dot{y}_0 = -\log x_0$. Its general solution, whose initial value is $(x_0(0), y_0(0)) = (A, B)$, is given by

$$x_0(t) = e^{B\sin t + (\log A)\cos t}, \quad y_0(t) = B\cos t - (\log A)\sin t. \tag{4.34}$$

The RG equation defined by (4.11) is calculated as

$$\frac{d}{dt}\begin{pmatrix} A \\ B \end{pmatrix} = \frac{\varepsilon}{2}\begin{pmatrix} A\log A \\ B \end{pmatrix}, \tag{4.35}$$

which is solved as

$$A(t) = \exp\left(pe^{\varepsilon t/2}\right), \quad B(t) = qe^{\varepsilon t/2}, \tag{4.36}$$

where $p, q \in \mathbf{R}$ are arbitrary constants. On the other hand, $h(t, A, B)$ defined by (4.8) is given by $h(t, A, B) = (D\varphi^0_{t,0})_{(A,B)} M(t)$, where

$$(D\varphi^0_{t,0})_{(A,B)} = \begin{pmatrix} \cos t \cdot e^{B\sin t + (\log A)\cos t}/A & \sin t \cdot e^{B\sin t + (\log A)\cos t} \\ -\sin t/A & \cos t \end{pmatrix}, \tag{4.37}$$

(4.38)

$$M(t) = \begin{pmatrix} \dfrac{A(\log A)^2 - AB^2}{3}\sin^3 t + \dfrac{2AB\log A}{3}\cos^3 t - \dfrac{AB}{2}\sin^2 t + AB^2\sin t - \dfrac{A\log A}{4}\sin^2 t \\ \dfrac{(\log A)^2 - B^2}{3}\cos^3 t - \dfrac{2B\log A}{3}\sin^3 t - \dfrac{\log A}{2}\sin^2 t - (\log A)^2\cos t + \dfrac{B}{4}\sin 2t \end{pmatrix}.$$

It is easy to verify that the norm conditions (N) are satisfied. According to (4.17) with the present $A(t), B(t)$, an approximate solution to (4.33) is given by

$$\begin{pmatrix} X(t) \\ Y(t) \end{pmatrix} = \begin{pmatrix} e^{B(t)\sin t - (\log A(t))\cos t} \\ B(t)\cos t - (\log A(t))\cos t \end{pmatrix} + \varepsilon h(t, A(t), B(t)). \tag{4.39}$$

Note that the RG vector field $\frac{\varepsilon}{2}(x\log x, y)$ commutes with the vector field $(xy, -\log x)$, which is the unperturbed part of (4.33) with respect to the Lie bracket product. This fact is proved generally in the next section.

5. RG vector fields with symmetry. In this section, we consider an autonomous equation on a manifold M:

$$\dot{x} = f(x) + \varepsilon g(x), \quad x \in M. \tag{5.1}$$

For this equation, we suppose that $(D\varphi_s^0)_A^{-1} g(\varphi_s^0(A))$ is KBM on $\mathbf{R}_{\geq T}$ and the RG equation for $f + \varepsilon g$

$$\frac{dA}{dt} = \varepsilon R(A) = \varepsilon \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (D\varphi_s^0)_A^{-1} g(\varphi_s^0(A)) ds \tag{5.2}$$

is defined, where φ^0 is a flow of $f(x)$ satisfying $\varphi_{t+t'}^0 = \varphi_t^0 \circ \varphi_{t'}^0$.

Assume that a Lie group G acts on the manifold M. If a vector field f on M satisfies

$$(Da)_x f(x) = f(ax) \quad \forall a \in G, \forall x \in M, \tag{5.3}$$

then f is called *invariant* under the action of G, where $(Da)_x$ is the derivative at x of the map determined by $a : M \to M$ at x.

Proposition 5.1. *If vector fields f and g are invariant under the action of a Lie group G, then so is the RG vector field for $f + \varepsilon g$.*

Proof. For all $a \in G$, $R(aA)$ is calculated as

$$\begin{aligned} R(aA) &= \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (D\varphi_s^0)_{aA}^{-1} g(\varphi_s^0(aA)) ds \\ &= \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (Da)_A (D\varphi_s^0)_A^{-1} (Da)_A^{-1} g(a\varphi_s^0(A)) ds \\ &= (Da)_A \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (D\varphi_s^0)_A^{-1} (Da)_A^{-1} (Da)_A g(\varphi_s^0(A)) ds = (Da)_A R(A). \end{aligned}$$

This proves the proposition. ∎

The next proposition was proved by Ziane [8] for the case that $f(t,x)$ is a linear vector field.

Proposition 5.2. *The RG vector field $\varepsilon R(A)$ for $f + \varepsilon g$ commutes with f with respect to the Lie bracket product. Equivalently, $R(A)$ satisfies*

$$(D\varphi_t^0)_A R(A) = R(\varphi_t^0(A)) \tag{5.4}$$

for all $t \in \mathbf{R}$ and all $A \in M$.

Proof. For all $s' \in \mathbf{R}$ and for all $A \in M$, $R(\varphi_{s'}^0(A))$ is calculated as

$$\begin{aligned} R(\varphi_{s'}^0(A)) &= \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (D\varphi_s^0)_{\varphi_{s'}^0(A)}^{-1} g(\varphi_s^0 \circ \varphi_{s'}^0(A)) ds \\ &= \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (D\varphi_{s'}^0)_A \circ (D\varphi_s^0)_A^{-1} \circ (D\varphi_{s'}^0)_A^{-1} g(\varphi_{s+s'}^0(A)) ds \\ &= (D\varphi_{s'}^0)_A \lim_{t\to\infty} \frac{1}{t-T} \int_T^t (D\varphi_{s+s'}^0)_A^{-1} g(\varphi_{s+s'}^0(A)) ds. \end{aligned}$$

Putting $s+s'=s''$ provides

$$\begin{aligned}R(\varphi^0_{s'}(A)) =&(D\varphi^0_{s'})_A \lim_{t\to\infty}\frac{1}{t-T}\int_{T+s'}^{t+s'}(D\varphi^0_{s''})_A^{-1}g(\varphi^0_{s''}(A))ds''\\ =&(D\varphi^0_{s'})_A R(A)+(D\varphi^0_{s'})_A \lim_{t\to\infty}\frac{1}{t-T}\int_{t}^{t+s'}(D\varphi^0_{s''})_A^{-1}g(\varphi^0_{s''}(A))ds''\\ &-(D\varphi^0_{s'})_A \lim_{t\to\infty}\frac{1}{t-T}\int_{T}^{T+s'}(D\varphi^0_{s''})_A^{-1}g(\varphi^0_{s''}(A))ds''\\ =&(D\varphi^0_{s'})_A R(A).\end{aligned}$$

This proves the proposition. ■

Propositions 5.1 and 5.2 show that if vector fields f and g are invariant under the action of a Lie group G, then the RG vector field $\varepsilon R(A)$ is invariant under the action of G and the one-parameter group $\{\varphi^0_t\}_{t\in\mathbf{R}}$. In this sense, the RG vector field has a simpler structure than the original vector field $f+\varepsilon g$.

6. Invariant manifolds.

In this section, we consider an equation of the form

$$\dot{x}=Fx+\varepsilon g(x),\quad x\in\mathbf{R}^n, \tag{6.1}$$

where F is a diagonalizable $n\times n$ constant matrix, all of whose eigenvalues lie on the imaginary axis, and where g is a polynomial vector field on $\mathbf{R}^n$. Note that in this situation, the norm conditions (N) are satisfied.

Theorem 6.1. *If the RG vector field $\varepsilon R(x)$ for (6.1) has a boundaryless compact normally hyperbolic invariant manifold N, then (6.1) also has a normally hyperbolic invariant manifold N_ε for sufficiently small $\varepsilon>0$. This invariant manifold N_ε is diffeomorphic to N, and its stability coincides with that of N.*

We will prove this theorem in Appendix A, while we give a brief sketch of the proof below.

Suppose that the RG vector field has a normally hyperbolic invariant manifold N. Then, the approximate vector field $F_\varepsilon(t,x)$ defined by (4.16) has a normally hyperbolic invariant manifold $\tilde{N}$ which is diffeomorphic to $\mathbf{R}\times N$ in the (t,x) space since the flow of the approximate vector field is related to the flow of the RG vector field through (4.15). Now we need Fenichel's theorem.

Theorem (Fenichel [10]). *Let M be a C^r manifold ($r\geq 1$) and $\mathcal{X}^r(M)$ the set of C^r vector fields on M with the C^1 topology. Let f be a C^r vector field on M, and suppose that $N\subset M$ is a boundaryless compact connected normally hyperbolic f-invariant manifold. Then, the following hold:*

(i) *There is a neighborhood $\mathcal{U}\subset\mathcal{X}^r(M)$ of f such that there exists a normally hyperbolic g-invariant C^r manifold $N_g\subset M$ for all $g\in\mathcal{U}$.*

(ii) *N_g is diffeomorphic to N and the diffeomorphism $h:N_g\to N$ is close to the identity $\mathrm{id}:N\to N$ in the C^1 topology.*

See [10, 11, 12] for the proof of the theorem and the definition of normal hyperbolicity. Since the approximate vector field $F_\varepsilon(t,x)$ is C^1 close to the original vector field $Fx+\varepsilon g(x)$, we expect that Fenichel's theorem concludes that the original vector field $Fx+\varepsilon g(x)$ has

an invariant manifold which is diffeomorphic to $\mathbf{R} \times N$ in the (t, x) space. Since (6.1) is an autonomous equation, $Fx + \varepsilon g(x)$ has an invariant manifold which is diffeomorphic to N in the x space.

The above argument needs to be modified because the approximate vector field is a time-dependent vector field even if the original vector is independent of t, while Fenichel's theorem holds for time-independent vector fields. In Appendix A, we define the higher order RG equation to refine the error estimate of the approximate vector field to prove Theorem 6.1.

Note that for the case of compact normally hyperbolic invariant manifolds with boundary, Fenichel's theorem is modified as follows: If a vector field f has a compact connected normally hyperbolic invariant manifold N with a boundary, then a vector field g, which is C^1 close to f, has a *locally* invariant manifold N_g which is diffeomorphic to N. In this case, an orbit of the flow of g through a point on N_g may go out from N_g through its boundary. According to this theorem, Theorem 6.1 has to be modified so that N_ε is locally invariant if N has a boundary.

Example 6.2. Consider the system on $\mathbf{R}^2$:

$$\begin{cases} \dot{x} = y - x^3 + \varepsilon x, \\ \dot{y} = -x. \end{cases} \tag{6.2}$$

The unperturbed system $\dot{x} = y - x^3$, $\dot{y} = -x$ has the origin as a fixed point which is *not* hyperbolic. By using Theorem 6.1, we show the occurrence of the Hopf bifurcation at $\varepsilon = 0$, and a stable periodic orbit appears for $\varepsilon > 0$.

Changing the coordinate by $(x, y) = (\varepsilon X, \varepsilon Y)$, we obtain

$$\begin{cases} \dot{X} = Y + \varepsilon(X - \varepsilon X^3), \\ \dot{Y} = -X. \end{cases} \tag{6.3}$$

We want to regard the term $\varepsilon^2 X^3$ as a *first* order term with respect to ε since, at this time, we define only the *first* order RG equation while the higher order RG equation will be defined in Appendix A. To do so, define the function $\varepsilon_0(t)$ by $\varepsilon_0(t) \equiv \varepsilon$, and rewrite (6.3) as

$$\begin{cases} \dot{X} = Y + \varepsilon(X - \varepsilon_0 X^3), \\ \dot{Y} = -X, \\ \dot{\varepsilon}_0 = 0. \end{cases} \tag{6.4}$$

Then this system takes the form (6.1). The RG method is applicable to (6.4). Substitute $X = X_0 + \varepsilon X_1$, $Y = Y_0 + \varepsilon Y_1$ into (6.4) and equate the coefficients of $\varepsilon^0, \varepsilon^1$ to zero, respectively. Then we get

$$\begin{cases} \dot{X}_0 = Y_0, \\ \dot{Y}_0 = -X_0, \end{cases} \quad \begin{cases} \dot{X}_1 = Y_1 + X_0 - \varepsilon_0 X_0^3, \\ \dot{Y}_1 = -X_1. \end{cases} \tag{6.5}$$

We denote a solution to the former by

$$X_0(t) = Ae^{it} + \overline{A}e^{-it}, \quad A \in \mathbf{C}. \tag{6.6}$$

With this $X_0(t)$, a special solution to the latter defined by (4.9), whose initial time is $t = \tau$, is written as

$$X_1(t) = \frac{1}{2}(A - 3\varepsilon_0 A|A|^2)(t-\tau)e^{it} + \frac{3i}{8}A^3 e^{3it} + \text{c.c.}, \tag{6.7}$$

where c.c. is the complex conjugate of the first two terms of the right-hand side. Therefore, the RG equation for (6.3) is given by

$$\frac{dA}{dt} = \frac{1}{2}\varepsilon(A - 3\varepsilon_0 A|A|^2). \tag{6.8}$$

Substituting $A = re^{i\theta}$ into the above equation provides

$$\begin{cases} \dot{r} = \dfrac{\varepsilon}{2}(r - 3\varepsilon_0 r^3), \\ \dot{\theta} = 0. \end{cases} \tag{6.9}$$

Fixed points of this system are $r = 0$ and $r = \sqrt{1/3\varepsilon_0} := r_0$, when $\varepsilon_0 > 0$. Further, we obtain

$$\frac{d}{dr}\Big|_{r=r_0} \frac{\varepsilon}{2}(r - 3\varepsilon_0 r^3) = \frac{\varepsilon}{2}\left(1 - 9\varepsilon_0 \cdot \frac{1}{3\varepsilon_0}\right) = -\varepsilon < 0.$$

This means that the RG equation (6.9) has a circle $\{r = r_0\}$ as a stable normally hyperbolic invariant manifold (the set of fixed points) if $\varepsilon > 0$. By Theorem 6.1, the system (6.2) also has a stable periodic orbit if $\varepsilon > 0$ is sufficiently small. This proves that the Hopf bifurcation occurs for (6.2). Note that the radius of the invariant circle for the RG equation is of order $O(1/\sqrt{\varepsilon})$. In the original coordinate (x, y), the radius of the periodic orbit for the system (6.2) is of order $O(\sqrt{\varepsilon})$. Indeed, the periodic solution is approximately given by $x(t) = 2\sqrt{\varepsilon/3}\cos t$ in the (x, y) coordinate.

We can show that the second order RG equation defined in Definition A.5 for (6.3) is given as $\dot{r} = \varepsilon(r - 3\varepsilon r^3)/2$, $\dot{\theta} = -\varepsilon^2/8$. Thus we can obtain the same result as above without introducing ε_0 by using the second order RG equation, although it provides a modification to the motion in the θ direction.

We have just seen in Example 6.2 that the RG method can be used on problems in which there is an ordinary Hopf bifurcation. In the next example, we show that the RG method can also be used for systems in which a limit cycle is created far away from a fixed point, namely, with $O(1)$ radius.

Example 6.3. Consider the system on $\mathbf{R}^2$

$$\begin{cases} \dot{x} = y + \varepsilon(x - x^3), \\ \dot{y} = -x. \end{cases} \tag{6.10}$$

Substituting $x = x_0 + \varepsilon x_1$, $y = y_0 + \varepsilon y_1$ into (6.10) and equating the coefficients of $\varepsilon^0, \varepsilon^1$ to zero, respectively, we get

$$\begin{cases} \dot{x}_0 = y_0, \\ \dot{y}_0 = -x_0, \end{cases} \qquad \begin{cases} \dot{x}_1 = y_1 + x_0 - x_0^3, \\ \dot{y}_1 = -x_1. \end{cases} \tag{6.11}$$

We denote a solution to the former by

$$x_0(t) = Ae^{it} + \overline{A}e^{-it}, \quad A \in \mathbf{C}. \tag{6.12}$$

With this $x_0(t)$, a special solution to the latter defined by (4.9), whose initial time is $t = \tau$, is written as

$$x_1(t) = \frac{1}{2}(A - 3A|A|^2)(t-\tau)e^{it} + \frac{3i}{8}A^3 e^{3it} + \text{c.c.}, \tag{6.13}$$

where c.c. is the complex conjugate of the first two terms of the right-hand side. Therefore, the RG equation for (6.10) is given by

$$\frac{dA}{dt} = \frac{1}{2}\varepsilon(A - 3A|A|^2). \tag{6.14}$$

Substituting $A = re^{i\theta}$ into the above equation provides

$$\begin{cases} \dot{r} = \dfrac{\varepsilon}{2}(r - 3r^3), \\ \dot{\theta} = 0. \end{cases} \tag{6.15}$$

Fixed points of this system are $r = 0$ and $r = \sqrt{1/3} := r_0$, when $\varepsilon > 0$. It is easy to verify that $r = r_0$ is the stable fixed point. Therefore, the system (6.10) has a stable periodic orbit if $\varepsilon > 0$ is sufficiently small. Note that since the radius of the invariant circle for the RG equation is of $O(1)$, the radius of the periodic orbit of the system (6.10) is also of $O(1)$. This can be verified numerically. For each ε, points $y_0 > 0$ at which the periodic orbit for the system (6.10) crosses the y axis are calculated numerically to provide Figure 1. The radius y_0 is almost independent of ε when $\varepsilon > 0$ is sufficiently small.

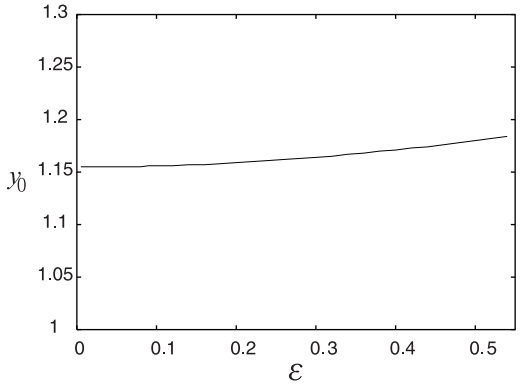


Figure 1. *The radius y_0 of the periodic orbit of the system (6.10) for each ε.*

7. Linear equations.

We apply the RG method to a time-dependent linear equation

$$\dot{x} = F(t)x + \varepsilon G(t)x, \quad x \in \mathbf{R}^n, \tag{7.1}$$

where $F(t)$ and $G(t)$ are $n \times n$ matrix functions which are of C^1 class with respect to t. A solution to the equation $\dot{x}_0 = F(t)x_0$ is denoted by $x_0(t, 0, v) = X(t)v$, where $X(t)$ is the

fundamental matrix and $v \in \mathbf{R}^n$ is an initial value. We assume that $X(t)^{-1}G(t)X(t)$ is KBM on $t \geq 0$, and we define a constant matrix

$$R := \lim_{t\to\infty} \frac{1}{t}\int_0^t X^{-1}(s)G(s)X(s)ds. \tag{7.2}$$

We call it a *secular matrix* for (7.1). Then, a special solution to an equation $\dot{x}_1 = F(t)x_1 + G(t)x_0(t,0,v)$ defined by (4.9) is given by

$$x_1(t,\tau;v) = X(t)\widetilde{G}(t)v + X(t)(t-\tau)Rv,\ \ \widetilde{G}(t) = \int^t (X(s)^{-1}G(s)X(s) - R)ds, \tag{7.3}$$

and the RG equation for (7.1) is given by a linear equation

$$\dot{v} = \varepsilon Rv,\ \ v \in \mathbf{R}^n. \tag{7.4}$$

If $X(t)$ and $\widetilde{G}(t)$ are bounded in $t \geq 0$, then Theorem 4.4(i) holds and the flow Φ_{t,t_0} defined by (4.15) is put in the form

$$\Phi_{t,t_0} = X(t)(I + \varepsilon\widetilde{G}(t))e^{\varepsilon R(t-t_0)}(I + \varepsilon\widetilde{G}(t_0))^{-1}X(t_0)^{-1}, \tag{7.5}$$

where I is the $n \times n$ identity matrix. Accordingly, the approximate vector field F_ε defined by (4.16) is expressed as

$$F_\varepsilon(t,x) = F(t)x+\varepsilon G(t)X(t)(I+\varepsilon\widetilde{G}(t))^{-1}X(t)^{-1}x+\varepsilon^2 X(t)\widetilde{G}(t)R(I+\varepsilon\widetilde{G}(t))^{-1}X(t)^{-1}x. \tag{7.6}$$

The following proposition means that the stability of $X(t)^{-1}x(t)$ is inherited from that of the RG equation if $\varepsilon > 0$ is sufficiently small. In fact, the proposition shows that if real parts of all eigenvalues of R are negative, then $||X(t)^{-1}x(t)|| \to 0$ as $t \to \infty$ for an arbitrary solution $x(t)$ of (7.1), and that if there exists an eigenvalue of R whose real part is positive, then there exists a solution $x(t)$ of (7.1) such that $||X(t)^{-1}x(t)|| \to \infty$ as $t \to \infty$.

Proposition 7.1. *Suppose that $X(t)$ and $\widetilde{G}(t)$ defined in (7.3) are bounded in $t \geq 0$. Let R be a secular matrix for (7.1) and $\lambda_1, \dots, \lambda_n$ its eigenvalues. Then, for each integer k with $1 \leq k \leq n$, there exist positive constants D_1, D_2, t_0, a positive valued function $\phi(\varepsilon)$ with $\phi(\varepsilon) \to 0$ as $\varepsilon \to 0$, and a solution $x(t)$ of (7.1) such that the inequality*

$$D_2 e^{\varepsilon \mathrm{Re}(\lambda_k)t - 2\varepsilon\phi(\varepsilon)t} \leq ||X(t)^{-1}x(t)|| \leq D_1 e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t} \tag{7.7}$$

holds for $t \geq t_0$.

Proof. Since $\widetilde{G}(t) = \int_0^t (X(s)^{-1}G(s)X(s) - R)ds$ is bounded, $(I + \varepsilon\widetilde{G}(t))^{-1}$ is expanded into the Neumann series as $(I + \varepsilon\widetilde{G}(t))^{-1} = \sum_{n=0}^\infty (-\varepsilon)^n \widetilde{G}(t)^n$. With this expansion inserted into (7.6), $F_\varepsilon(t,x)$ is rewritten as

$$F_\varepsilon(t,x) = F(t)x + \varepsilon G(t)x + \varepsilon^2 H(t,\varepsilon)x, \tag{7.8}$$

$$H(t,\varepsilon) := \sum_{n=0}^\infty (-\varepsilon)^n \left(X(t)\widetilde{G}(t)R\widetilde{G}(t)^n X(t)^{-1} - G(t)X(t)\widetilde{G}(t)^{n+1}X(t)^{-1}\right). \tag{7.9}$$

Let us rewrite (7.1) as

$$\dot{x} = F_\varepsilon(t,x) - \varepsilon^2 H(t,\varepsilon)x. \tag{7.10}$$

Introducing a new function $y(t)$ by $x(t) = X(t)y(t)$, we verify that y satisfies the differential equation

$$\dot{y} = \widetilde{F}_\varepsilon(t)y - \varepsilon^2 \widetilde{H}(t,\varepsilon)y, \tag{7.11}$$

where

$$\widetilde{F}_\varepsilon(t) := \varepsilon X(t)^{-1}G(t)X(t) + \varepsilon^2 X(t)^{-1}H(t,\varepsilon)X(t), \tag{7.12}$$

$$\widetilde{H}(t,\varepsilon) := \sum_{n=0}^{\infty}(-\varepsilon)^n \left(\widetilde{G}(t)R\widetilde{G}(t)^n - X(t)^{-1}G(t)X(t)\widetilde{G}(t)^{n+1}\right), \tag{7.13}$$

and further that the flow of the linear vector field $\widetilde{F}_\varepsilon(t)y$ is given by

$$\widetilde{\Phi}_{t,t_0} = (I + \varepsilon\widetilde{G}(t))e^{\varepsilon R(t-t_0)}(I + \varepsilon\widetilde{G}(t_0))^{-1}. \tag{7.14}$$

To prove the proposition, we can suppose that the secular matrix R is put in the Jordan form. In fact, if we change the variable x in (7.1) by $x \mapsto Px$, where P is an arbitrary nonsingular constant matrix, then $F(t)$, $G(t)$, and $X(t)^{-1}G(t)X(t)$ are brought into $P^{-1}F(t)P$, $P^{-1}G(t)P$, and $P^{-1}X(t)^{-1}G(t)X(t)P$, respectively. This means that R turns into $P^{-1}RP$. In what follows, we assume that R is of the Jordan form

$$R = \begin{pmatrix} \lambda_1 & p_1 & & & \\ & \lambda_2 & p_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & p_{n-1} \\ & & & & \lambda_n \end{pmatrix}, \tag{7.15}$$

where λ_i $(i = 1,\dots,n)$ are the eigenvalues of R such that $\mathrm{Re}(\lambda_1) \le \cdots \le \mathrm{Re}(\lambda_n)$ and where p_i $(i = 1,\dots,n-1)$ are either 0 or 1.

Now let us fix an integer $k < n$ such that $\mathrm{Re}(\lambda_{k+1}) - \mathrm{Re}(\lambda_k) > 0$. The case that $n = k$ and the case that there are no such $k < n$ are treated later. Define matrices $Q_1(t), Q_2(t)$ to be upper triangle matrices

$$Q_1(t) = \left(\begin{array}{ccc|ccc} e^{\varepsilon\lambda_1 t} & & * & & & \\ & \ddots & & & 0 & \\ & & e^{\varepsilon\lambda_k t} & & & \\ \hline & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{array}\right), \quad Q_2(t) = \left(\begin{array}{ccc|ccc} 0 & & & & & \\ & \ddots & & & 0 & \\ & & 0 & & & \\ \hline & & & e^{\varepsilon\lambda_{k+1}t} & & * \\ & & & & \ddots & \\ & & & & & e^{\varepsilon\lambda_n t} \end{array}\right)$$

such that $Q_1(t) + Q_2(t) = e^{\varepsilon Rt}$. Then, a solution $y(t)$ to (7.11) satisfies an integral equation

$$\begin{aligned} y(t) = \widetilde{\Phi}_{t,0}\boldsymbol{e}_k &- \varepsilon^2 \int_0^t (I + \varepsilon\widetilde{G}(t))Q_1(t-s)(I + \varepsilon\widetilde{G}(s))^{-1} \circ \widetilde{H}(s,\varepsilon)y(s)ds \\ &+ \varepsilon^2 \int_t^\infty (I + \varepsilon\widetilde{G}(t))Q_2(t-s)(I + \varepsilon\widetilde{G}(s))^{-1} \circ \widetilde{H}(s,\varepsilon)y(s)ds, \end{aligned} \tag{7.16}$$

where $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ are the canonical bases of $\mathbf{R}^n$. The first term of the right-hand side of the above is written as $\widetilde{\Phi}_{t,0}\boldsymbol{e}_k = (I+\varepsilon\widetilde{G}(t))(q_1(t)e^{\varepsilon\lambda_k t}, \dots, q_{k-1}(t)e^{\varepsilon\lambda_k t}, e^{\varepsilon\lambda_k t}, 0, \dots, 0)^t$, where $q_i(t)$ $(i = 1, \dots, k-1)$ are monomials of t whose degrees are at most $k-1$. The fact that $\widetilde{G}(t) = \int_0^t (X(s)^{-1}G(s)X(s) - R)ds$ is bounded uniformly in t implies that $(I+\varepsilon\widetilde{G}(t))^{\pm 1}$ and $X(t)^{-1}G(t)X(t)$ are also bounded uniformly in t, and thereby so is $\widetilde{H}(t,\varepsilon)$. Consequently, there exist positive constants C_0, C_1 such that

$$||\widetilde{H}(t,\varepsilon)|| \le C_0, \quad ||(I+\varepsilon\widetilde{G}(t))^{\pm 1}|| \le C_1. \tag{7.17}$$

Further, there exist positive constants C_2, C_3 and a positive valued function $\phi(\varepsilon)$ satisfying $\phi(\varepsilon) \to 0$ as $\varepsilon \to 0$ such that

$$\begin{aligned} ||Q_1(t)|| &\le \frac{C_2}{\phi(\varepsilon)^n} e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t} \quad \text{for } t \ge 0, \\ ||\widetilde{\Phi}_{t,0}\boldsymbol{e}_k|| &\le \frac{C_1C_2}{\phi(\varepsilon)^n} e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t} \quad \text{for } t \ge 0, \\ ||Q_2(t)|| &\le \frac{C_3}{\phi(\varepsilon)^n} e^{\varepsilon \mathrm{Re}(\lambda_{k+1})t - \varepsilon\phi(\varepsilon)t} \quad \text{for } t \le 0. \end{aligned} \tag{7.18}$$

Indeed, if $\varepsilon t \ge 1$, there exists a constant C such that $||Q_1(t)|| \le C\varepsilon^n t^n e^{\varepsilon \mathrm{Re}(\lambda_k)t}$. Suppose that there exists a function $q(\varepsilon)$ such that

$$||Q_1(t)|| \le C\varepsilon^n t^n e^{\varepsilon \mathrm{Re}(\lambda_k)t} \le Cq(\varepsilon) e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t}.$$

This inequality is equivalent to the inequality $\varepsilon t \le q(\varepsilon)^{1/n} e^{\varepsilon\phi(\varepsilon)t/n}$, and it is easy to verify that this inequality holds when $q(\varepsilon) = (n/(\phi(\varepsilon)e))^n$. Putting $C_2 = C(n/e)^n$, we obtain $||Q_1(t)|| \le \frac{C_2}{\phi(\varepsilon)^n} e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t}$ for $\varepsilon t \ge 1$. This inequality also holds when $0 \le \varepsilon t < 1$ because $||Q_1(t)|| \le Ce^{\varepsilon \mathrm{Re}(\lambda_k)t}$ holds if $0 \le \varepsilon t < 1$. The inequalities for $||\widetilde{\Phi}_{t,0}\boldsymbol{e}_k||$ and $||Q_2(t)||$ above are verified in a similar way.

We define a sequence of functions $\{y_m(t)\}_{m\ge 0}$ by

$$\begin{aligned} y_0(t) &= \widetilde{\Phi}_{t,0}\boldsymbol{e}_k, \\ y_{m+1}(t) &= y_0(t) - \varepsilon^2 \int_0^t (I+\varepsilon\widetilde{G}(t))Q_1(t-s)(I+\varepsilon\widetilde{G}(s))^{-1} \circ \widetilde{H}(s,\varepsilon)y_m(s)ds \\ &\quad + \varepsilon^2 \int_t^\infty (I+\varepsilon\widetilde{G}(t))Q_2(t-s)(I+\varepsilon\widetilde{G}(s))^{-1} \circ \widetilde{H}(s,\varepsilon)y_m(s)ds. \end{aligned}$$

We need two lemmas to prove the proposition.

Lemma 7.2. *Let $\phi(\varepsilon) = \varepsilon^{1/(2n+2)}$ and fix $\varepsilon > 0$ small so that $\mathrm{Re}(\lambda_{k+1}) - \mathrm{Re}(\lambda_k) - 3\phi(\varepsilon) > 0$. Then there exists a constant $0 < p < 1$ such that*

$$||y_m(t) - y_{m-1}(t)|| \le p^m e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t}, \quad m = 1, 2, \dots, \tag{7.19}$$

for $t \ge 0$.

Proof. We prove (7.19) by induction. For $m = 1$, the quantity $||y_1(t) - y_0(t)||$ is estimated as follows:

$$\begin{aligned}
||y_1 - y_0|| &\le \varepsilon^2 \int_0^t ||I + \varepsilon\widetilde{G}(t)|| \cdot ||Q_1(t-s)|| \cdot ||(I + \varepsilon\widetilde{G}(s))^{-1}|| \cdot ||\widetilde{H}(s,\varepsilon)|| \cdot ||y_0(s)||ds \\
&\quad + \varepsilon^2 \int_t^\infty ||I + \varepsilon\widetilde{G}(t)|| \cdot ||Q_2(t-s)|| \cdot ||(I + \varepsilon\widetilde{G}(s))^{-1}|| \cdot ||\widetilde{H}(s,\varepsilon)|| \cdot ||y_0(s)||ds \\
&\le \frac{\varepsilon^2 C_0 C_1^3 C_2^2}{\phi(\varepsilon)^{2n}} \int_0^t e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t} e^{\varepsilon\phi(\varepsilon)s} ds \\
&\quad + \frac{\varepsilon^2 C_0 C_1^3 C_2 C_3}{\phi(\varepsilon)^{2n}} \int_t^\infty e^{\varepsilon \mathrm{Re}(\lambda_{k+1})t - \varepsilon\phi(\varepsilon)t} e^{-\varepsilon \mathrm{Re}(\lambda_{k+1})s + \varepsilon \mathrm{Re}(\lambda_k)s + 3\varepsilon\phi(\varepsilon)s} ds \\
&\le \frac{\varepsilon C_0 C_1^3 C_2}{\phi(\varepsilon)^{2n+1}} \left(C_2 + \frac{C_3\phi(\varepsilon)}{\mathrm{Re}(\lambda_{k+1}) - \mathrm{Re}(\lambda_k) - 3\phi(\varepsilon)} \right) e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t} \\
&= \varepsilon^{1/(2n+2)} C_0 C_1^3 C_2 \left(C_2 + \frac{C_3\phi(\varepsilon)}{\mathrm{Re}(\lambda_{k+1}) - \mathrm{Re}(\lambda_k) - 3\phi(\varepsilon)} \right) e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t}.
\end{aligned}$$

Define

$$p = \varepsilon^{1/(2n+2)} C_0 C_1^3 C_2 \left(C_2 + \frac{C_3\phi(\varepsilon)}{\mathrm{Re}(\lambda_{k+1}) - \mathrm{Re}(\lambda_k) - 3\phi(\varepsilon)} \right); \tag{7.20}$$

then (7.19) holds for $m = 1$. Further, if ε is sufficiently small, the inequality $0 < p < 1$ holds. With this p, if we suppose that (7.19) holds, then we can verify that $||y_{m+1} - y_m|| \le p^{m+1} e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t}$ by the same calculation as above. ∎

This lemma implies that the sequence $\{y_m(t)\}_{m\ge 0}$ converges to a solution of (7.11).

Lemma 7.3. *Under the same conditions as Lemma 7.2, there exist positive constants D_1 and t_0 such that*

$$||y_m(t)|| \le D_1 e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t}, \quad m = 0, 1, \ldots, \tag{7.21}$$

for $t \ge t_0$.

Proof. We prove the lemma by induction. When $m = 0$, the above inequality is clear if $D_1 \ge C_1 C_2/\phi(\varepsilon)^n$. Suppose that the above inequality holds for m; then

$$\begin{aligned}
||y_{m+1}|| &\le ||y_0|| + \varepsilon^2 \int_0^t ||I + \varepsilon\widetilde{G}(t)|| \cdot ||Q_1(t-s)|| \cdot ||(I + \varepsilon\widetilde{G}(s))^{-1}|| \cdot ||\widetilde{H}(s,\varepsilon)|| \cdot ||y_m(s)||ds \\
&\quad + \varepsilon^2 \int_t^\infty ||I + \varepsilon\widetilde{G}(t)|| \cdot ||Q_2(t-s)|| \cdot ||(I + \varepsilon\widetilde{G}(s))^{-1}|| \cdot ||\widetilde{H}(s,\varepsilon)|| \cdot ||y_m(s)||ds \\
&= D_1 e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t} + \frac{\varepsilon^2 C_0 C_1^2 C_2 D_1}{\phi(\varepsilon)^n} \int_0^t e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t} e^{\varepsilon\phi(\varepsilon)s} ds \\
&\quad + \frac{\varepsilon^2 C_0 C_1^2 C_3 D_1}{\phi(\varepsilon)^n} \int_t^\infty e^{\varepsilon \mathrm{Re}(\lambda_{k+1})t - \varepsilon\phi(\varepsilon)t} e^{-\varepsilon \mathrm{Re}(\lambda_{k+1})s + \varepsilon \mathrm{Re}(\lambda_k)s + 3\varepsilon\phi(\varepsilon)s} \\
&\le D_1 e^{\varepsilon \mathrm{Re}(\lambda_k)t + \varepsilon\phi(\varepsilon)t} + \frac{\varepsilon C_0 C_1^2 C_2 D_1}{\phi(\varepsilon)^{n+1}} e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t} \\
&\qquad + \frac{\varepsilon C_0 C_1^2 C_3 D_1}{\phi(\varepsilon)^n (\mathrm{Re}(\lambda_{k+1}) - \mathrm{Re}(\lambda_k) - 3\phi(\varepsilon))} e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t}
\end{aligned}$$

$$\leq D_1 e^{\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t} \left(e^{-\varepsilon\phi(\varepsilon)t} + \frac{\varepsilon^{n/(2n+2)}}{C_1 C_2} p \right),$$

where p is defined by (7.20). Since $0 < p < 1$, we can take sufficiently large t_0 and sufficiently small ε such that

$$0 < e^{-\varepsilon\phi(\varepsilon)t} + \frac{\varepsilon^{n/(2n+2)}}{C_1 C_2} p < 1$$

for $t \geq t_0$. This proves the lemma. ■

We return to the proof of Proposition 7.1. By taking the limit $m \to \infty$ in (7.21), we obtain a solution $y(t) = y^{(k)}(t)$ of (7.11) satisfying the right part of the inequality (7.7) when $k \neq n$. If there exist eigenvalues $\lambda_{k'}$ of R satisfying $\mathrm{Re}(\lambda_{k'}) = \mathrm{Re}(\lambda_k)$, we repeat the above discussion with $\boldsymbol{e}_{k'}$ instead of $\boldsymbol{e}_k$ included in (7.16). Then we obtain a solution $y^{(k')}(t)$ of (7.11), which is linearly independent of $y^{(k)}(t)$, satisfying the right part of the inequality (7.7). To prove the same inequality for $k = n$, instead of (7.16), we use the integral equation

$$y(t) = \widetilde{\Phi}_{t,0}\boldsymbol{e}_n - \varepsilon^2 \int_0^t (I + \varepsilon\widetilde{G}(t)) e^{\varepsilon R(t-s)} (I + \varepsilon\widetilde{G}(s))^{-1} \circ \widetilde{H}(s,\varepsilon) y(s) ds. \tag{7.22}$$

The same procedure as above applied to this equation yields the right part of inequality (7.7) for $k = n$.

We proceed to prove the left part of inequality (7.7). Let $y^{(k)}(t)$ be a solution of (7.11) which satisfies the right part of inequality (7.7), and denote the fundamental matrix to (7.11) by Y, whose column vectors are $y^{(1)}(t), \ldots, y^{(n)}(t)$. Define a matrix Z by $Z = Y^{-t}$, where Y^{-t} is the abbreviation of $(Y^{-1})^t$; this notation will be used in what follows. Each column vector $z^{(1)}(t), \ldots, z^{(n)}(t)$ of Z satisfies an adjoint equation of (7.11):

$$\dot{z} = -(\widetilde{F}_\varepsilon(t) - \varepsilon^2 \widetilde{H}(t,\varepsilon))^t z. \tag{7.23}$$

Since the flow of the linear vector field $-(\widetilde{F}_\varepsilon)^t z$ is given by $\widetilde{\Phi}_{t,t_0}^{-t} = (I + \varepsilon\widetilde{G}(t))^{-t} e^{-\varepsilon R^t (t-t_0)} (I + \varepsilon\widetilde{G}(t_0))^t$, we can prove that there exists a solution $z(t) = u^{(k)}(t)$ of (7.23) such that

$$||u^{(k)}(t)|| \leq D_1' e^{-\varepsilon \mathrm{Re}(\lambda_k)t + 2\varepsilon\phi(\varepsilon)t}, \quad k = 1, \ldots, n, \quad t \geq t_0, \tag{7.24}$$

by the same procedure as that for the proof of the inequality (7.7), where D_1' is some positive constant. Let U be the fundamental matrix for (7.23) whose column vectors are $u^{(1)}(t), \ldots, u^{(n)}(t)$. By the uniqueness of solutions of (7.23), there exists a constant matrix K such that $U = ZK$. Let k_{ij} be the (i,j) component of K. Since $Y^t U = K$, the inequality $|k_{ii}| = |(y^{(i)}(t), u^{(i)}(t))| \leq ||y^{(i)}(t)|| \cdot ||u^{(i)}(t)||$ holds, which then proves the left part of inequality (7.7) if $k_{ii} \neq 0$, where $(\ ,\ )$ denotes the standard inner product of the $\mathbf{R}^n$. If $k_{ii} = 0$, we define $\widetilde{u}^{(k)}(t)$ by $\widetilde{u}^{(k)}(t) = u^{(k)}(t) + \sum_{i=k+1}^n \alpha_i u^{(i)}(t)$, with $\alpha_i \in \mathbf{R}$. And we define a matrix $\widetilde{U}$ whose column vectors are $\widetilde{u}^{(1)}(t), \ldots, \widetilde{u}^{(n)}(t)$. Each $\widetilde{u}^{(k)}(t)$ satisfies the inequality (7.24) for some constant D_1'. Then there exists a constant matrix $\widetilde{K}$ such that $Y^t \widetilde{U} = \widetilde{K}$, and we can assume that its diagonal component $k_{ii} \neq 0$ by defining $\alpha_i \in \mathbf{R}$ appropriately. This ends the proof of Proposition 7.1. ■

Corollary 7.4. *Consider an equation* $\dot{x} = Fx + \varepsilon G(t)x$ *with* $x \in \mathbf{R}^n$, *where* F *is an* $n \times n$ *constant matrix and* $G(t)$ *is an* $n \times n$ *matrix which is of* C^1 *class with respect to* t. *Suppose that all eigenvalues of* F *lie on the imaginary axis, and suppose that* $G(t)$ *and* $\widetilde{G}(t)$ *defined by* (7.3) *are bounded in* $t \in \mathbf{R}$. *If* $\varepsilon > 0$ *is sufficiently small, then the stability of a trivial solution* $x(t) \equiv 0$ *coincides with that of a trivial solution of the RG equation* $\dot{v} = \varepsilon R v$, *where* R *is a secular matrix for* $Fx + \varepsilon G(t)x$.

In the above corollary, the boundedness of $G(t)$ and $\widetilde{G}(t)$ are satisfied if $G(t)$ is a periodic or almost periodic function in t whose Fourier exponents do not have accumulation points in $\mathbf{R}$.

Example 7.5. Let us consider the Mathieu equation:

$$\ddot{y} = -(a + 2\varepsilon \cos t)y, \tag{7.25}$$

where a and ε are positive parameters. It is well known that there exists an area in the (a, ε) plane such that the origin is an unstable fixed point for (7.25) if (a, ε) is in this area. We calculate the area approximately by the RG method.

Let $a = a_0 + \varepsilon a_1$ and $y = y_0 + \varepsilon y_1$. Substituting them into (7.25) and comparing the coefficients of ε^0 and ε^1 in both sides of (7.25) provides

$$\ddot{y}_0 = -b^2 y_0, \tag{7.26}$$

$$\ddot{y}_1 = -b^2 y_1 - a_1 y_0 - 2\cos t \cdot y_0, \tag{7.27}$$

where $a_0 = b^2$. A general solution to the former is given by

$$y_0(t) = Ae^{ibt} + \overline{A}e^{-ibt}, \quad A \in \mathbf{C}. \tag{7.28}$$

With this y_0, (7.27) is rewritten as

$$\ddot{y}_1 = -b^2 y_1 - \left(a_1 A e^{ibt} + A e^{i(1+b)t} + \overline{A} e^{i(1-b)t} + \text{c.c.}\right). \tag{7.29}$$

If $b = 1/2$ (i.e., $a_0 = 1/4$), the secular term appears for all a_1. In fact, the equation

$$\ddot{y}_1 = -\frac{1}{4} y_1 - \left(a_1 A e^{it/2} + A e^{3it/2} + \overline{A} e^{it/2} + \text{c.c.}\right) \tag{7.30}$$

admits a special solution given by

$$y_1(t, \tau; A) = i(a_1 A + \overline{A})(t - \tau)e^{it/2} + \frac{A}{2} e^{3it/2} + \text{c.c.}, \tag{7.31}$$

where the initial time has been chosen to be $t = \tau$. Then, the RG equation for (7.25) is given by

$$\dot{A} = i\varepsilon(a_1 A + \overline{A}). \tag{7.32}$$

Putting $A = B + iC$, $B, C \in \mathbf{R}$, we break up (7.32) into

$$\begin{cases} \dot{B} = \varepsilon(1 - a_1)C, \\ \dot{C} = \varepsilon(1 + a_1)B. \end{cases} \tag{7.33}$$

A general solution to this equation is given by

$$B(t) = \begin{cases} pe^{\varepsilon\sqrt{1-a_1^2}t} + qe^{-\varepsilon\sqrt{1-a_1^2}t} & (|a_1| \le 1), \\ pe^{i\varepsilon\sqrt{a_1^2-1}t} + qe^{-i\varepsilon\sqrt{a_1^2-1}t} & (|a_1| > 1), \end{cases} \tag{7.34}$$

where $p, q \in \mathbf{R}$ are arbitrary constants. This shows that the origin is an unstable fixed point for the RG equation (7.33) if $|a_1| < 1$. This proves the instability of the fixed point of the Mathieu equation (7.25) if $a = 1/4 + \varepsilon a_1 + O(\varepsilon^2)$, $|a_1| < 1$.

Example 7.6. Consider the coupled Mathieu equations

$$\begin{cases} \ddot{x} = -(a + 2\varepsilon\cos t)x - \varepsilon p(x-y) - \varepsilon q(\dot{x}-\dot{y}), \\ \ddot{y} = -(a + 2\varepsilon\cos t)y - \varepsilon p(y-x) - \varepsilon q(\dot{y}-\dot{x}), \end{cases} \tag{7.35}$$

where $\varepsilon > 0$ and $a, p, q \in \mathbf{R}$ are constants. Put $u = x + y$; then the equation for $u(t)$ is the Mathieu equation (7.25). In Example 7.5, we proved that if $a = 1/4$, the trivial solution $u = 0$ of the Mathieu equation (7.25) is unstable. In what follows, we assume that $a = 1/4$. Put $z = x - y$. Then z satisfies the equation

$$\ddot{z} = -\frac{1}{4}z + \varepsilon(-2q\dot{z} - 2pz - 2\cos t \cdot z). \tag{7.36}$$

Put further $z = z_0 + \varepsilon z_1$, where z_0 is subjected to the unperturbed equation $\ddot{z}_0 = -\frac{1}{4}z_0$ and has a general solution of the form $z_0(t) = Ae^{it/2} + \overline{A}e^{-it/2}$. With this $z_0(t)$, the equation for z_1 proves to be given by

$$\ddot{z}_1 = -\frac{1}{4}z_1 - iqAe^{it/2} - 2pAe^{it/2} - Ae^{3it/2} - \overline{A}e^{it/2} + \text{c.c.}, \tag{7.37}$$

where c.c. denotes the complex conjugate of the last four terms of the right-hand side. A special solution of this equation, whose initial time is $t = \tau$, is given by

$$z_1(t) = i(iqA + 2pA + \overline{A})(t-\tau)e^{it/2} + \frac{A}{2}e^{3it/2} + \text{c.c.} \tag{7.38}$$

Therefore, the RG equation for (7.36) is put in the form

$$\dot{A} = i\varepsilon(iqA + 2pA + \overline{A}), \quad A \in \mathbf{C}. \tag{7.39}$$

Put $A = \alpha + i\beta$, $\alpha, \beta \in \mathbf{R}$. Then the above equation is rewritten as

$$\frac{d}{dt}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \varepsilon \begin{pmatrix} -q & -2p+1 \\ 2p+1 & -q \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{7.40}$$

Eigenvalues of the matrix in the right-hand side of the above equation are $\lambda_\pm = -q \pm \sqrt{1-4p^2}$. Therefore, the stability of the trivial solution $(\alpha, \beta) = (0, 0)$ of the RG equation is as given in Figure 2.

Corollary 7.4 shows that the stability of the trivial solution $z(t) = 0$ of (7.36) coincides with that of the stability of $(\alpha, \beta) = (0, 0)$. This proves that if $\mathrm{Re}(\lambda_\pm) < 0$, then $|x(t) - y(t)| \to 0$ as $t \to \infty$, although each $|x(t)|, |y(t)|$ diverges as $t \to \infty$.

A numerical solution to (7.35) for $\varepsilon = 0.01$, $x(0) = 0.5$, $y(0) = 0.1$ is presented in Figure 3.

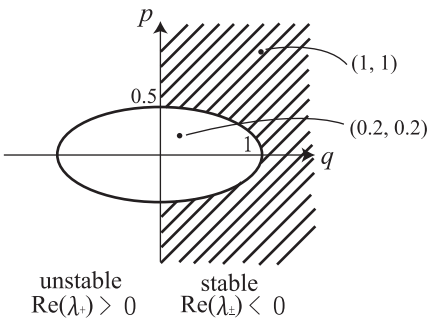


Figure 2. *The trivial solution* $(\alpha, \beta) = (0, 0)$ *is stable on the shaded area.*

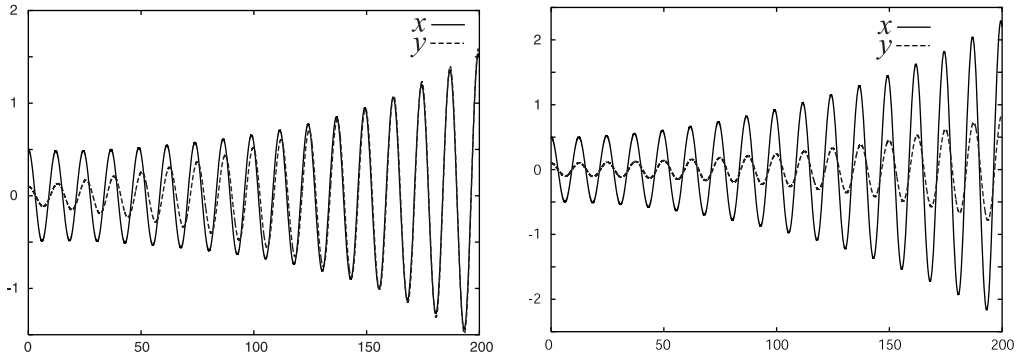


Figure 3. *Numerical results for* (7.35). *The synchronous solution* $x(t) = y(t)$ *is* (a) *stable if* $p = q = 1$, (b) *unstable if* $p = q = 0.2$.

Appendix A. Higher order RG equation. In this appendix, we define the higher order RG equation for constructing an approximate vector field which is $O(\varepsilon^{m+1})$ close to a given original vector field. The result is used in proving Theorem 6.1.

Let F be a diagonalizable $n \times n$ matrix, all of whose eigenvalues lie on the imaginary axis, and let $g_1(t, x), \ldots, g_m(t, x)$ be C^∞ vector fields on $\mathbf{R} \times \mathbf{R}^n$ which are polynomial in x and periodic in t. Consider an ODE

$$\dot{x} = Fx + \varepsilon g_1(t, x) + \varepsilon^2 g_2(t, x) + \cdots + \varepsilon^m g_m(t, x), \quad x \in \mathbf{R}^n, \tag{A.1}$$

where $\varepsilon \in \mathbf{R}$ is a small parameter. Put $x = x_0 + \varepsilon x_1 + \cdots + \varepsilon^m x_m$. Then the above equation is rewritten as

$$\dot{x}_0 + \varepsilon \dot{x}_1 + \cdots + \varepsilon^m \dot{x}_m = F(x_0 + \varepsilon x_1 + \cdots + \varepsilon^m x_m) + \sum_{i=1}^{m} \varepsilon^i g_i(t, x_0 + \varepsilon x_1 + \cdots + \varepsilon^m x_m). \tag{A.2}$$

Expanding the right-hand side of the above equation with respect to ε and equating the coefficients of each ε^i of both sides of the above, we obtain ODEs for $x_0, x_1, \ldots, x_m$,

$$\dot{x}_0 = Fx_0, \tag{A.3}$$

$$\dot{x}_1 = Fx_1 + G_1(t, x_0), \tag{A.4}$$

$$\vdots$$

$$\dot{x}_i = Fx_i + G_i(t, x_0, x_1, \dots, x_{i-1}), \tag{A.5}$$

$$\vdots$$

$$\dot{x}_m = Fx_m + G_m(t, x_0, x_1, \dots, x_{m-1}), \tag{A.6}$$

where G_i is some smooth function of $t, x_0, x_1, \dots, x_{i-1}$ which is periodic in t. For example, G_1, G_2, and G_3 are given by

$$G_1(t, x_0) = g_1(t, x_0), \tag{A.7}$$

$$G_2(t, x_0, x_1) = \frac{\partial g_1}{\partial x}(t, x_0)x_1 + g_2(t, x_0), \tag{A.8}$$

$$G_3(t, x_0, x_1, x_2) = \frac{1}{2}\frac{\partial^2 g_1}{\partial x^2}(t, x_0)x_1^2 + \frac{\partial g_1}{\partial x}(t, x_0)x_2 + \frac{\partial g_2}{\partial x}(t, x_0)x_1 + g_3(t, x_0), \tag{A.9}$$

respectively. We have to solve the above equations. At first, we denote by $x_0(t, 0, A) = X(t)A$ a solution of the unperturbed part $\dot{x}_0 = Fx_0$, where $X(t) = e^{Ft}$ is the fundamental matrix and $A \in \mathbf{R}^n$ is an initial value. With this x_0, by a discussion similar to that in section 4, a solution of (A.4) is given by

$$x_1(t, \tau; A) = h_t^{(1)}(A) + X(t)R_1(A)(t - \tau), \tag{A.10}$$

where $h_t^{(1)}(A)$ and $R_1(A)$ are defined by

$$R_1(A) = \lim_{t\to\infty} \frac{1}{t}\int^t X(s)^{-1}G_1(s, X(s)A)ds, \tag{A.11}$$

$$h_t^{(1)}(A) = X(t)\int^t \left(X(s)^{-1}G_1(s, X(s)A) - R_1(A)\right)ds, \tag{A.12}$$

respectively. The integral constants of the indefinite integrals in (A.11), (A.12) and (A.13), (A.14) below are fixed arbitrarily. By choosing these integral constants appropriately, we can reduce the RG equation. This will be done in a forthcoming paper. Note that since $X(t)$ and $G_1(t, x)$ are almost periodic in t, $X(t)^{-1}G_1(t, X(t)A)$ is bounded uniformly in $t \in \mathbf{R}$ and $R_1(A)$ is well defined. With this x_0 and x_1, we solve the equation for x_2, as will be shown in Proposition A.1. This process is performed step by step until a solution x_m to (A.6) is obtained.

Proposition A.1. *Define functions $R_i(A)$ and $h_t^{(i)}(A)$, $i = 2, \dots, m$, by*

$$R_i(A) := \lim_{t\to\infty} \frac{1}{t}\int^t \Bigg(X(s)^{-1}G_i(s, X(s)A, h_s^{(1)}(A), \dots, h_s^{(i-1)}(A)) - X(s)^{-1}\sum_{k=1}^{i-1}(Dh_s^{(k)})_A R_{i-k}(A)\Bigg)ds, \tag{A.13}$$

$$h_t^{(i)}(A) := X(t)\int^t \Big(X(s)^{-1} G_i(s, X(s)A, h_s^{(1)}(A), \dots, h_s^{(i-1)}(A)) - X(s)^{-1}\sum_{k=1}^{i-1} (Dh_s^{(k)})_A R_{i-k}(A) - R_i(A)\Big) ds. \tag{A.14}$$

Then, the curve defined by

$$x_i := x_i(t,\tau;A) = h_t^{(i)}(A) + y_1^{(i)}(t,A)(t-\tau) + y_2^{(i)}(t,A)(t-\tau)^2 + \cdots + y_i^{(i)}(t,A)(t-\tau)^i \tag{A.15}$$

gives a solution to (A.5) *for* $i = 1, 2, \dots, m$, *where* $y_1^{(i)}, \dots, y_i^{(i)}$ *are defined by*

$$y_1^{(i)}(t,A) = X(t)R_i(A) + \sum_{k=1}^{i-1} (Dh_t^{(k)})_A R_{i-k}(A), \tag{A.16}$$

$$y_j^{(i)}(t,A) = \frac{1}{j}\sum_{k=1}^{i-1} \frac{\partial y_{j-1}^{(k)}}{\partial A}(t,A) R_{i-k}(A), \quad j = 2, 3, \dots, i-1, \tag{A.17}$$

$$y_i^{(i)}(t,A) = \frac{1}{i}\sum_{k=1}^{i-1} \frac{\partial y_{i-1}^{(k)}}{\partial A}(t,A) R_{i-k}(A) = \frac{1}{i}\frac{\partial y_{i-1}^{(i-1)}}{\partial A}(t,A) R_1(A), \tag{A.18}$$

$$y_j^{(i)}(t,A) = 0, \quad j > i. \tag{A.19}$$

Proof. We prove Proposition A.1 by induction. Assume that $x_1, \dots, x_{i-1}$ defined by (A.15) are solutions of (A.5) for $i = 1, 2, \dots, i-1$. In order to prove that x_i defined by (A.15) is a solution of (A.5), we substitute (A.15) into (A.5) to obtain

$$\begin{aligned} &Fh_t^{(i)}(A) + G_i(t, X(t)A, h_t^{(1)}(A), \dots, h_t^{(i-1)}(A)) - \sum_{k=1}^{i-1} (Dh_t^{(k)})_A R_{i-k}(A) - X(t)R_i(t) \\ &+ \sum_{k=1}^{i} \dot{y}_k^{(i)}(t,A)(t-\tau)^k + \sum_{k=1}^{i} y_k^{(i)}(t,A) k (t-\tau)^{k-1} \\ &\quad = Fh_t^{(i)}(A) + F\sum_{k=1}^{i} y_k^{(i)}(t,A)(t-\tau)^k + G_i(t, x_0, x_1, \dots, x_{i-1}). \end{aligned} \tag{A.20}$$

It is easy to verify that $G_i(t, x_0, x_1, \dots, x_{i-1})$ with $x_0, x_1, \dots, x_{i-1}$ defined by (A.15) is a polynomial in $t-\tau$ whose degree is at most $i-1$. We denote it by

$$G_i(t, x_0, \dots, x_{i-1}) = \sum_{k=0}^{i-1} \widetilde{G}_i^{(k)}(t, x_0, \dots, x_{i-1})(t-\tau)^k. \tag{A.21}$$

Note that $\widetilde{G}_i^{(0)}(t, x_0, \dots, x_{i-1}) = G_i(t, X(t)A, h_t^{(1)}(A), \dots, h_t^{(i-1)}(A))$. Equating the coefficients of $(t-\tau)^k$ of both sides of (A.20) with (A.21), we obtain

$$y_1^{(i)}(t,A) = \sum_{k=1}^{i-1} (Dh_t^{(k)})_A R_{i-k}(A) + X(t)R_i(A), \tag{A.22}$$

$$\dot{y}_k^{(i)}(t,A) + (k+1)y_{k+1}^{(i)}(t,A) = Fy_k^{(i)}(t,A) + \widetilde{G}_i^{(k)}(t,x_0,\dots,x_{i-1}),\ \ k=1,\dots,i-1, \tag{A.23}$$

$$\dot{y}_i^{(i)}(t,A) = Fy_i^{(i)}(t,A). \tag{A.24}$$

These equations can determine $y_1^{(i)}, y_2^{(i)}, \dots, y_i^{(i)}$. Equation (A.22) gives (A.16). From (A.23) for $k=1$, we obtain

$$\begin{aligned}
2y_2^{(i)}(t,A) &= Fy_1^{(i)}(t,A) - \dot{y}_1^{(i)}(t,A) + \widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1}) \\
&= \sum_{k=1}^{i-1} F(Dh_t^{(k)})_A R_{i-k}(A) + FX(t)R_i(A) \\
&\quad - \sum_{k=1}^{i-1} \frac{\partial}{\partial t}(Dh_t^{(k)})_A R_{i-k}(A) - \frac{\partial}{\partial t}X(t)R_i(A) + \widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1}) \\
&= \sum_{k=1}^{i-1} F(Dh_t^{(k)})_A R_{i-k}(A) - \sum_{k=1}^{i-1} \frac{\partial}{\partial A}\Bigg(Fh_t^{(k)}(A) + G_k(t,X(t)A,h_t^{(1)}(A),\dots,h_t^{(k-1)}(A)) \\
&\quad - \sum_{j=1}^{k-1}(Dh_t^{(j)})_A R_{k-j}(A) - X(t)R_k(A)\Bigg) R_{i-k}(A) + \widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1}) \\
&= \sum_{k=1}^{i-1} \frac{\partial}{\partial A}\Bigg(\sum_{j=1}^{k-1}(Dh_t^{(j)})_A R_{k-j}(A) + X(t)R_k(A)\Bigg) R_{i-k}(A) \\
&\quad + \widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1}) - \sum_{k=1}^{i-1}\frac{\partial}{\partial A}G_k(t,X(t)A,h_t^{(1)}(A),\dots,h_t^{(k-1)}(A))R_{i-k}(A) \\
&= \sum_{k=1}^{i-1} \frac{\partial}{\partial A} y_1^{(k)}(t,A)R_{i-k}(A) \\
&\quad + \widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1}) - \sum_{k=1}^{i-1}\frac{\partial}{\partial A}G_k(t,X(t)A,h_t^{(1)}(A),\dots,h_t^{(k-1)}(A))R_{i-k}(A).
\end{aligned} \tag{A.25}$$

If the equality

$$\widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1}) = \sum_{k=1}^{i-1}\frac{\partial}{\partial A}G_k(t,X(t)A,h_t^{(1)}(A),\dots,h_t^{(k-1)}(A))R_{i-k}(A) \tag{A.26}$$

holds, then (A.17) for $j=2$ is obtained. The left-hand side of the above is calculated as

$$\begin{aligned}
\widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1}) &= -\frac{\partial}{\partial \tau}\Big|_{\tau=t} G_i(t,x_0,\dots,x_{i-1}) \\
&= -\sum_{j=1}^{i-1}\lim_{\tau\to t}\frac{\partial G_i}{\partial x_j}(t,x_0,\dots,x_{i-1})\frac{\partial}{\partial \tau}\Big|_{\tau=t} x_j(t,\tau;A) \\
&= \sum_{j=1}^{i-1}\lim_{\tau\to t}\frac{\partial G_i}{\partial x_j}(t,x_0,\dots,x_{i-1})\Bigg(X(t)R_j(A) + \sum_{k=1}^{j-1}(Dh_t^{(k)})_A R_{j-k}(A)\Bigg).
\end{aligned} \tag{A.27}$$

The right-hand side of (A.26) is calculated as

$$\begin{aligned}
&\sum_{k=1}^{i-1}\frac{\partial}{\partial A}G_k(t,X(t)A,h_t^{(1)}(A),\dots,h_t^{(k-1)}(A))R_{i-k}(A)\\
&=\sum_{k=1}^{i-1}\sum_{j=1}^{k-1}\lim_{\tau\to t}\frac{\partial G_k}{\partial x_j}(t,x_0,\dots,x_{i-1})(Dh_t^{(j)})_A R_{i-k}(A)\\
&\quad+\sum_{k=1}^{i-1}\lim_{\tau\to t}\frac{\partial G_k}{\partial x_0}(t,x_0,\dots,x_{i-1})X(t)R_{i-k}(A). \qquad \text{(A.28)}
\end{aligned}$$

Now we need a simple lemma.

Lemma A.2. *For integers i,j with $i>j$, the equality*

$$\frac{\partial G_i}{\partial x_j}=\frac{\partial G_{i-1}}{\partial x_{j-1}}=\cdots=\frac{\partial G_{i-j}}{\partial x_0} \qquad \text{(A.29)}$$

holds.

We will prove this lemma after the proof of Proposition A.1 is completed. According to Lemma A.2, (A.27) and (A.28) are brought into

(A.30)

$$\widetilde{G}_i^{(1)}(t,x_0,\dots,x_{i-1})=\sum_{j=1}^{i-1}\lim_{\tau\to t}\frac{\partial G_{i-j}}{\partial x_0}(t,x_0,\dots,x_{i-1})\left(X(t)R_j(A)+\sum_{k=1}^{j-1}(Dh_t^{(k)})_A R_{j-k}(A)\right),$$

and

$$\begin{aligned}
&\sum_{k=1}^{i-1}\frac{\partial}{\partial A}G_k(t,X(t)A,h_t^{(1)}(A),\dots,h_t^{(k-1)}(A))R_{i-k}(A)\\
&=\sum_{k=1}^{i-1}\sum_{j=1}^{k-1}\lim_{\tau\to t}\frac{\partial G_{k-j}}{\partial x_0}(t,x_0,\dots,x_{i-1})(Dh_t^{(j)})_A R_{i-k}(A)\\
&\quad+\sum_{k=1}^{i-1}\lim_{\tau\to t}\frac{\partial G_k}{\partial x_0}(t,x_0,\dots,x_{i-1})X(t)R_{i-k}(A), \qquad \text{(A.31)}
\end{aligned}$$

respectively. This proves (A.26), and (A.17) for $j=2$ is verified.

By using (A.23), $y_3^{(i)},\dots,y_{i-1}^{(i)},y_i^{(i)}$ are calculated in the same way as above, and (A.17) and (A.18) are proved, but we omit the detailed calculation here. Next, we have to show that $y_i^{(i)}$ given by (A.18) satisfies (A.24). To show this, according to $y_1^{(1)}(t,A)=X(t)R_1(t)$, we rewrite (A.18) as

$$\begin{aligned}
y_i^{(i)}(t,A)&=\frac{1}{i}\frac{\partial y_{i-1}^{(i-1)}}{\partial A}R_1(A)\\
&=\frac{1}{i(i-1)}\frac{\partial}{\partial A}\left(\frac{\partial y_{i-2}^{(i-2)}}{\partial A}R_1(A)\right)R_1(A)
\end{aligned}$$

$$
\begin{aligned}
&= \ \vdots \\
&= \frac{1}{i!}\frac{\partial}{\partial A}\left(\frac{\partial}{\partial A}\left(\cdots\frac{\partial}{\partial A}\left(\frac{\partial y_1^{(1)}}{\partial A}R_1(A)\right)\cdots\right)R_1(A)\right)R_1(A)\\
&= X(t)\frac{1}{i!}\frac{\partial}{\partial A}\left(\frac{\partial}{\partial A}\left(\cdots\frac{\partial}{\partial A}\left(\frac{\partial R_1}{\partial A}R_1(A)\right)\cdots\right)R_1(A)\right)R_1(A).
\end{aligned}
$$

Since $X(t)$ is the fundamental matrix of the equation $\dot{y} = Fy$, $y_i^{(i)}$ satisfies (A.24). Therefore, x_i defined by (A.15) satisfies (A.5). This ends the proof of Proposition A.1. ■

Proof of Lemma A.2. By definition, $G_i(t, x_0, \ldots, x_{i-1})$ is written as

$$
G_i(t, x_0, \ldots, x_{i-1}) = \sum_{k=1}^{i-1} \frac{1}{k!}\frac{d^k}{d\varepsilon^k}\Big|_{\varepsilon=0} g_{i-k}(t, \textstyle\sum_{l=0}^m \varepsilon^l x_l) + g_i(t, x_0).
$$

On the other hand, $G_{i-1}(t, x_0, \ldots, x_{i-2})$ is rewritten as

$$
\begin{aligned}
G_{i-1}(t, x_0, \ldots, x_{i-2}) &= \sum_{k=0}^{i-2} \frac{1}{k!}\frac{d^k}{d\varepsilon^k}\Big|_{\varepsilon=0} g_{i-k-1}(t, \textstyle\sum_{l=0}^m \varepsilon^l x_l)\\
&= \sum_{k=1}^{i-1} \frac{1}{(k-1)!}\frac{d^{k-1}}{d\varepsilon^{k-1}}\Big|_{\varepsilon=0} g_{i-k}(t, \textstyle\sum_{l=0}^m \varepsilon^l x_l).
\end{aligned}
$$

To show the equality $\partial G_i/\partial x_j = \partial G_{i-1}/\partial x_{j-1}$, it is sufficient to prove the equality

$$
\frac{\partial}{\partial x_j}\frac{1}{k!}\frac{d^k}{d\varepsilon^k}\Big|_{\varepsilon=0} g_{i-k}(t, \textstyle\sum_{l=0}^m \varepsilon^l x_l) = \frac{\partial}{\partial x_{j-1}}\frac{1}{(k-1)!}\frac{d^{k-1}}{d\varepsilon^{k-1}}\Big|_{\varepsilon=0} g_{i-k}(t, \textstyle\sum_{l=0}^m \varepsilon^l x_l) \tag{A.32}
$$

for $k = 1, 2, \ldots, i-1$. For simplicity, we denote $g_{i-k}(t, x)$ by $g(x)$. Consider the trivial equality

$$
\frac{\partial}{\partial x_j} g(\textstyle\sum_{l=0}^m \varepsilon^l x_l) = \varepsilon \frac{\partial}{\partial x_{j-1}} g(\textstyle\sum_{l=0}^m \varepsilon^l x_l), \quad j = 1, \ldots, m. \tag{A.33}
$$

Expanding both sides of the above equation with respect to ε, we obtain

$$
\begin{aligned}
&\frac{\partial}{\partial x_j}\left(\sum_{p=0}^{k} \frac{\varepsilon^p}{p!}\frac{d^p}{d\varepsilon^p}\Big|_{\varepsilon=0} g(\textstyle\sum_{l=0}^m \varepsilon^l x_l) + \widetilde{R}(\varepsilon, x_0, \ldots, x_m)\right)\\
&= \varepsilon\frac{\partial}{\partial x_{j-1}}\left(\sum_{p=0}^{k} \frac{\varepsilon^p}{p!}\frac{d^p}{d\varepsilon^p}\Big|_{\varepsilon=0} g(\textstyle\sum_{l=0}^m \varepsilon^l x_l) + \widetilde{R}(\varepsilon, x_0, \ldots, x_m)\right),
\end{aligned}
$$

where $\widetilde{R}$ is some function satisfying $\widetilde{R} \sim o(|\varepsilon|^{k+1})$. Equating the coefficients of ε^k of both sides of the above, we obtain (A.32). This ends the proof of Lemma A.2. ■

Remark A.3. Proposition A.1 also holds for a time-dependent matrix $F(t)$ as long as the fundamental matrix $X(t)$ of $F(t)$ is periodic in t. Further, for Proposition A.1, we do not

need to assume that functions g_i in (A.1) are polynomial in x. These assumptions are used to prove statements below.

Lemma A.4. *For* (A.1)*, functions* $h_t^{(i)}(A)$ *with* $i = 1, 2, \dots, m$ *defined by* (A.12) *and* (A.14) *are bounded uniformly in* t*.*

To prove this lemma, we need a theory of almost periodic functions. Indeed, we can show that functions $h_t^{(i)}(A)$ are almost periodic functions. This fact also holds even if the $g_i(t,x)$'s in (A.1) are not periodic in t but almost periodic in t as long as the set of Fourier exponents of the $g_i(t,x)$'s does not have accumulation points in $\mathbf{R}$. See Fink [13] for the definitions and basic facts of almost periodic functions.

Proof of Lemma A.4. We prove the proposition by induction. At first, note that $G_1(t,x_0)$ defined by (A.7) is almost periodic uniformly in x_0 because it is periodic in t and polynomial in x_0. Therefore, a function $X(t)^{-1}G_1(t,X(t)A)$ included in (A.12) is almost periodic uniformly in A (see Theorem 2.11 of Fink [13]). Each component of the vector-valued function $X(t)^{-1}G_1(t,X(t)A)$ is of the form $\sum_{k=1}^{p} b_k(t)e^{i\xi_k t}$, where $b_k(t)$ are some periodic functions and $\xi_k \in \mathbf{R}$ are some constants. Since each $b_k(t)$ can be expanded as a Fourier series in the ordinary sense, the set of Fourier exponents of $\sum_{k=1}^{p} b_k(t)e^{i\xi_k t}$ does not have accumulation points on $\mathbf{R}$. Since the Fourier coefficient corresponding to the zero Fourier exponent, if it exists, is $R_1(A)$ defined by (A.11), $X(t)^{-1}G_1(t,X(t)A) - R_1(A)$ does not have a zero as a Fourier exponent. Therefore, $\int^t (X(s)^{-1}G_1(s,X(s)A) - R_1(A))ds$ is almost periodic (we use Theorem 4.12 of Fink [13]), and this proves Lemma A.4 for $h_t^{(1)}(A)$.

Suppose that Lemma A.4 holds for $h_t^{(1)}(A), \dots, h_t^{(i-1)}(A)$. Like the above, the integrand in (A.14) is almost periodic uniformly in A because $G_1(t,x_0,\dots,x_{i-1})$ is periodic in t and polynomial in $x_0,\dots,x_{i-1}$. Since $X(s)A, h_s^{(1)}(A), \dots, h_s^{(i-1)}(A)$, the set of whose Fourier exponents has no accumulation points by the assumption of induction, are almost periodic the set of Fourier exponents of the function

$$p(s,A) := X(s)^{-1}G_i(s,X(s)A,h_s^{(1)}(A),\dots,h_s^{(i-1)}(A)) - X(s)^{-1}\sum_{k=1}^{i-1}(Dh_s^{(k)})_A R_{i-k}(A)$$

included in (A.14) does not have accumulation points. Since $R_i(A)$ defined by (A.13) gives the Fourier coefficient corresponding to the zero Fourier exponent of $p(s,A)$, if it exists, there exists $M > 0$ such that all Fourier exponents λ of the integrand in (A.14) satisfy $|\lambda| \geq M$. Then Theorem 4.12 of Fink [13] proves that $h_t^{(i)}(A)$ is almost periodic. ■

Definition A.5. *Along with* $R_1(A), \dots, R_m(A)$ *defined by* (A.11) *and* (A.13)*, we define the* mth order RG equation *for* (A.1) *by*

$$\dot{A} = \varepsilon R_1(A) + \varepsilon^2 R_2(A) + \cdots + \varepsilon^m R_m(A), \quad A \in \mathbf{R}^n, \tag{A.34}$$

and we call $\varepsilon R_1(A) + \cdots + \varepsilon^m R_m(A)$ *the* mth order RG vector field *for* (A.1)*. We denote by* $\varphi_t^{(m)}$ *the flow generated by the* m*th order RG vector field.*

Fix an open set $U \subset \mathbf{R}^n$ such that $\overline{U}$ is compact. Define the map α_t to be

$$\alpha_t(A) := X(t)A + \varepsilon h_t^{(1)}(A) + \varepsilon^2 h_t^{(2)}(A) + \cdots + \varepsilon^m h_t^{(m)}(A) \tag{A.35}$$

for all $t \in \mathbf{R}$. Now we are in a position to restate Theorem 4.4 in the present situation.

Theorem A.6. *Let $\varphi_t^{(m)}$ be the flow of the mth order RG equation for (A.1) and α_t the map defined by (A.35). Then, there exists $\varepsilon_0 > 0$ such that the following holds for all $|\varepsilon| < \varepsilon_0$: A map*

$$\Phi_{t,t_0} := \alpha_t \circ \varphi_{t-t_0}^{(m)} \circ \alpha_{t_0}^{-1} : \alpha_{t_0}(U) \to \mathbf{R}^n \tag{A.36}$$

defines a flow on $U_\varepsilon := \{(t,x) \mid t \in \mathbf{R},\ x \in \alpha_t(U)\}$ associated with a time-dependent vector field

$$F_\varepsilon(t,x) := \frac{d}{da}\Big|_{a=t} \Phi_{a,t}(x). \tag{A.37}$$

Further, there exists a vector field $\widetilde{F}_\varepsilon(t,x)$, which is bounded in t and bounded as $\varepsilon \to 0$, satisfying

$$F_\varepsilon(t,x) = Fx + \varepsilon g_1(t,x) + \cdots + \varepsilon^m g_m(t,x) + \varepsilon^{m+1}\widetilde{F}_\varepsilon(t,x). \tag{A.38}$$

Proof. The proof of the fact that the map Φ_{t,t_0} defines a flow is the same as that of Theorem 4.4(i). We prove (A.38). The vector field defined by (A.37) is calculated as

$$\begin{aligned}
F_\varepsilon(t,x) &= \frac{d}{da}\Big|_{a=t} \alpha_a \circ \varphi_{a-t}^{(m)} \circ \alpha_t^{-1}(x) \\
&= \frac{d}{da}\Big|_{a=t} \left(x_0(a,0,\alpha_t^{-1}(x)) + \varepsilon x_1(a,a;\alpha_t^{-1}(x)) + \cdots + \varepsilon^m x_m(a,a;\alpha_t^{-1}(x))\right) \\
&\quad + \left(X(t) + \varepsilon (Dh_t^{(1)})_{\alpha_t^{-1}(x)} + \cdots + \varepsilon^m (Dh_t^{(m)})_{\alpha_t^{-1}(x)}\right) \\
&\quad \circ \left(\varepsilon R_1(\alpha_t^{-1}(x)) + \cdots + \varepsilon^m R_m(\alpha_t^{-1}(x))\right) \\
&= \frac{d}{da}\Big|_{a=t} \left(x_0(a,0,\alpha_t^{-1}(x)) + \varepsilon x_1(a,t;\alpha_t^{-1}(x)) + \cdots + \varepsilon^m x_m(a,t;\alpha_t^{-1}(x))\right) \\
&\quad + \frac{d}{da}\Big|_{a=t} \left(\varepsilon x_1(t,a;\alpha_t^{-1}(x)) + \cdots + \varepsilon^m x_m(t,a;\alpha_t^{-1}(x))\right) \\
&\quad + \left(X(t) + \varepsilon (Dh_t^{(1)})_{\alpha_t^{-1}(x)} + \cdots + \varepsilon^m (Dh_t^{(m)})_{\alpha_t^{-1}(x)}\right) \\
&\quad \circ \left(\varepsilon R_1(\alpha_t^{-1}(x)) + \cdots + \varepsilon^m R_m(\alpha_t^{-1}(x))\right).
\end{aligned} \tag{A.39}$$

Since $x_i(a,t;\alpha_t^{-1}(x))$ is a solution of (A.5), it satisfies

$$\begin{aligned}
&\frac{d}{da}\Big|_{a=t} x_i(a,t;\alpha_t^{-1}(x)) \\
&= Fx_i(t,t;\alpha_t^{-1}(x)) + G_i(t,x_0(t,0,\alpha_t^{-1}(x)),\dots,x_{i-1}(t,t;\alpha_t^{-1}(x))) \\
&= Fh_t^{(i)}(\alpha_t^{-1}(x)) + G_i(t,x_0,h_t^{(1)}(\alpha_t^{-1}(x)),\dots,h_t^{(i-1)}(\alpha_t^{-1}(x))).
\end{aligned} \tag{A.40}$$

And, according to (A.15) and (A.16), the equality

$$\frac{d}{da}\Big|_{a=t} x_i(t,a;\alpha_t^{-1}(x)) = -y_1^{(i)}(t,\alpha_t^{-1}(x)) = -X(t)R_i(\alpha_t^{-1}(x)) - \sum_{k=1}^{i-1} (Dh_t^{(k)})_{\alpha_t^{-1}(x)} R_{i-k}(\alpha_t^{-1}(x)) \tag{A.41}$$

holds. Substituting (A.40) and (A.41) into (A.39), we obtain

(A.42)

$$\begin{aligned} &F_\varepsilon(t,x)\\ &= Fx_0(t,0,\alpha_t^{-1}(x)) + \sum_{k=1}^{m} \varepsilon^k \left(Fh_t^{(k)}(\alpha_t^{-1}(x)) + G_k(t,x_0,h_t^{(1)}(\alpha_t^{-1}(x)),\dots,h_t^{(k-1)}(\alpha_t^{-1}(x))) \right)\\ &\qquad + O(\varepsilon^{m+1})\\ &= Fx + \varepsilon g_1(t,x) + \cdots + \varepsilon^m g_m(t,x) + O(\varepsilon^{m+1}). \end{aligned}$$

It is hard to write out the term $O(\varepsilon^{m+1})$ explicitly. However, it is easy to prove that the term $O(\varepsilon^{m+1})$ is bounded uniformly in t, because it consists of the almost periodic functions $X(t)$, $g_i(t,x)$, $h_t^{(i)}$, α_t^{-1}. This ends the proof of Theorem A.6. ■

Theorem 6.1 follows immediately as a corollary of the next theorem.

Theorem A.7. *Consider an autonomous equation*

$$\dot{x} = Fx + \varepsilon g_1(x) + \cdots + \varepsilon^m g_m(x), \quad x \in \mathbf{R}^n, \tag{A.43}$$

where F is a diagonalizable $n\times n$ constant matrix, all of whose eigenvalues lie on the imaginary axis, and where $g_1,\dots,g_m$ are polynomial vector fields on $\mathbf{R}^n$. Suppose that its mth order RG vector field satisfies

$$R_1(A) = \cdots = R_{k-1}(A) = 0, \quad R_k(A) \neq 0, \quad k \leq 2m. \tag{A.44}$$

If the vector field $R_k(A)$ has a compact normally hyperbolic invariant manifold N, then (A.43) also has a normally hyperbolic invariant manifold N_ε for sufficiently small $\varepsilon > 0$. The N_ε is diffeomorphic to N and its stability coincides with that of N.

Proof. Before proving the theorem, we point out that the condition $k \leq 2m$ is not essential because we can take $m \in \mathbf{N}$ sufficiently large. Let us denote by $F_\varepsilon(t,x)$ the approximate vector field for (A.43) defined by (A.37). From Theorem A.6, we can rewrite (A.43) as

$$\dot{x} = F_\varepsilon(t,x) - \varepsilon^{m+1}\widetilde{F}_\varepsilon(t,x). \tag{A.45}$$

On account of (A.36), the RG vector field $\varepsilon^k R_k(x) + \cdots + \varepsilon^m R_m(x)$ satisfies the equation

$$\begin{aligned} \varepsilon^k R_k(x) + \cdots + \varepsilon^m R_m(x) &= \frac{d}{da}\Big|_{a=t} \alpha_a^{-1} \circ \Phi_{a,t} \circ \alpha_t(x)\\ &= \frac{d\alpha_a^{-1}}{da}\Big|_{a=t}(\alpha_t(x)) + (D\alpha_t^{-1})_{\alpha_t(x)} \frac{d}{da}\Big|_{a=t} \Phi_{a,t} \circ \alpha_t(x)\\ &= -(D\alpha_t)_x^{-1} \frac{d\alpha_t}{dt}(x) + (D\alpha_t)_x^{-1} F_\varepsilon(t,\alpha_t(x)). \end{aligned} \tag{A.46}$$

Introducing a new function $y(t)$ by $x(t) = \alpha_t \circ y(t)$ and substituting it into (A.45), we obtain

$$\frac{d\alpha_t}{dt}(y(t)) + (D\alpha_t)_{y(t)}\dot{y}(t) = F_\varepsilon(t,\alpha_t(y(t))) - \varepsilon^{m+1}\widetilde{F}_\varepsilon(t,\alpha_t(y(t))).$$

This equation is put together with (A.46) to yield

$$\dot{y} = \varepsilon^k R_k(y) + \cdots + \varepsilon^m R_m(y) - \varepsilon^{m+1}(D\alpha_t)_y^{-1} \circ \widetilde{F}_\varepsilon(t, \alpha_t(y)). \tag{A.47}$$

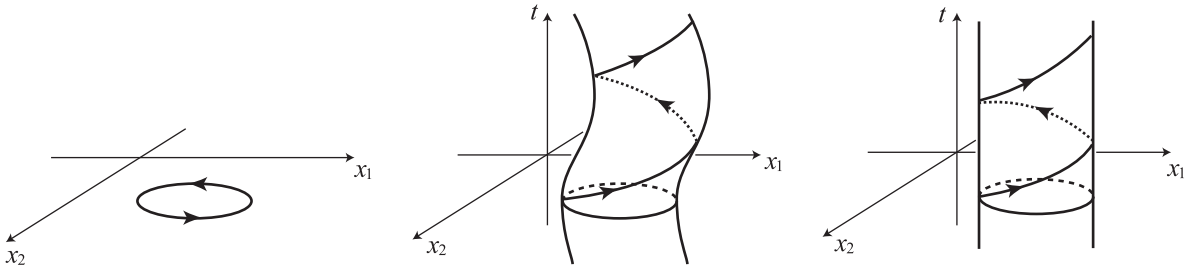
We introduce a newly scaled time s by $t = s/\varepsilon^k$. Then the above equation is rewritten as

$$\frac{dy}{ds} = R_k(y) + \varepsilon R_{k+1}(y) + \cdots + \varepsilon^{m-k} R_m(y) - \varepsilon^{m-k+1}(D\alpha_{s/\varepsilon^k})_y^{-1} \circ \widetilde{F}_{\varepsilon^k}(s/\varepsilon^k, \alpha_{s/\varepsilon^k}(y)). \tag{A.48}$$

Since α_t, $(D\alpha_t)_y$, and $\widetilde{F}_\varepsilon(t, y)$ are bounded uniformly in $t \in \mathbf{R}$, $(D\alpha_{s/\varepsilon^k})_y^{-1} \circ \widetilde{F}_{\varepsilon^k}(s/\varepsilon^k, \alpha_{s/\varepsilon^k}(y))$ is also bounded as $s \to \pm\infty$ and $\varepsilon \to 0$. Therefore, the time-dependent vector field $H(s, y)$ defined by the right-hand side of the above equation is sufficiently close to the vector field $R_k(y)$ in the C^1 topology if $\varepsilon > 0$ is sufficiently small.

Now we use Fenichel's theorem. We regard the vector field $R_k(y)$ on $\mathbf{R}^n$ as a vector field on $\mathbf{R} \times \mathbf{R}^n$ by putting $R_k(t, y) := R_k(y)$. If $R_k(y)$ has a normally hyperbolic invariant manifold N, then $R_k(t, y)$ has a normally hyperbolic invariant manifold $\mathbf{R} \times N$ in (t, y) space. Since $H(s, y)$ is sufficiently close to $R_k(t, y)$ as a vector field on $\mathbf{R} \times \mathbf{R}^n$ in the C^1 topology, $H(s, y)$ also has a normally hyperbolic invariant manifold $\widetilde{N}_\varepsilon$ which is diffeomorphic to $\mathbf{R} \times N$. Since $x(t) = \alpha_t \circ y(t)$ and since $D\alpha_t$ is bounded, (A.43) for $x(t)$ has a normally hyperbolic invariant manifold $\hat{N}_\varepsilon$ which is diffeomorphic to $\mathbf{R} \times N$ in (t, x) space.

Since (A.43) is autonomous, the manifold $\hat{N}_\varepsilon$ must be straight along the time axis (see Figure 4). Consequently, (A.43) has a normally hyperbolic invariant manifold on $\mathbf{R}^n$ which is diffeomorphic to N. ■



flow of $R_k(x)$ flow of (A.48) flow of (A.43)

Figure 4. *The case that $R_k(x)$ has an invariant circle. In this case, the flows of (A.48) and (A.43) have invariant cylinders in the (t, x) space.*

Let $A(t)$ be a solution of the mth order RG equation (A.34) for (A.1), and define the curve $\tilde{x}(t)$ to be

$$\tilde{x}(t) := \alpha_t(A(t)) = X(t)A(t) + \varepsilon h_t^{(1)}(A(t)) + \cdots + \varepsilon^m h_t^{(m)}(A(t)). \tag{A.49}$$

Then, $\tilde{x}(t)$ is an integral curve of the approximate vector field $F_\varepsilon(t, x)$ defined by (A.37), and it gives an approximate solution for (A.1).

Theorem A.8. *There exist positive constants ε_0, C, T and a compact subset $V = V(\varepsilon) \subset \mathbf{R}^n$ including the origin such that for all $|\varepsilon| < \varepsilon_0$, every solution $x(t)$ of (A.1) and $\widetilde{x}(t)$ defined*

by (A.49) *with* $x(0) = \widetilde{x}(0) \in V$ *satisfies the inequality*

$$||x(t) - \widetilde{x}(t)|| < C\varepsilon^m \quad \textit{for } 0 \le t \le T/\varepsilon. \tag{A.50}$$

Proof of Theorem A.8. Suppose that $||x(0)|| < K$. At first, we show that there exists $T > 0$ such that $||x(t)|| < 2K$ for $0 \le t \le T/\varepsilon$. We rewrite (A.1) as the integral equation

$$x(t) = e^{Ft}x(0) + e^{Ft}\int_0^t e^{-Fs}\varepsilon g(s, x(s), \varepsilon)ds, \tag{A.51}$$

where $g(t, x, \varepsilon) := g_1(t, x) + \varepsilon g_2(t, x) + \cdots + \varepsilon^{m-1}g_m(t, x)$. Choose $t \ge 0$ so that $||x(s)|| < 2K$ if $0 \le s \le t$. Then, there exists a positive constant $K' > 0$ such that $||g(s, x(s), \varepsilon)|| < K'$ and the inequality

$$\begin{aligned} ||x(t)|| &\le ||x(0)|| + \int_0^t \varepsilon ||g(s, x(s), \varepsilon)||ds \\ &\le K + \int_0^t \varepsilon K' ds = K\left(1 + \frac{K'}{K}\varepsilon t\right) \end{aligned}$$

holds. When $0 \le t \le K/(K'\varepsilon)$, we have $||x(t)|| < 2K$ so that we put $T := K/K'$ for the existence of T.

By Theorem A.6, an approximate solution $\widetilde{x}(t)$ satisfies an ODE

$$\dot{\widetilde{x}}(t) = F_\varepsilon(t, \widetilde{x}) = F\widetilde{x} + \varepsilon g_1(t, \widetilde{x}) + \cdots + \varepsilon^m g_m(t, \widetilde{x}) + \varepsilon^{m+1}\widetilde{F}_\varepsilon(t, \widetilde{x}). \tag{A.52}$$

Fix a positive number K such that the closed ball B_{2K} of radius $2K$ centered at the origin is included in the open set $\alpha_t(U)$, where U is an open set on which α_t is a diffeomorphism. Then, we can verify that $||\widetilde{x}(t)|| < 2K$ if $||\widetilde{x}(0)|| < K$ and if $0 \le t \le T/\varepsilon$ in the same way as above.

For $x(t)$ and $\widetilde{x}(t)$ such that $x(0) = \widetilde{x}(0), ||x(0)|| < K$, we put $\xi(t) = \alpha_t^{-1} \circ x(t)$, $\eta(t) = \alpha_t^{-1} \circ \widetilde{x}(t)$. They satisfy respective ODEs

$$\dot{\xi}(t) = \varepsilon R_1(\xi) + \varepsilon^2 R_2(\xi) + \cdots + \varepsilon^m R_m(\xi) + \varepsilon^{m+1}\widetilde{G}_\varepsilon(t, \xi), \tag{A.53}$$

$$\dot{\eta}(t) = \varepsilon R_1(\eta) + \varepsilon^2 R_2(\eta) + \cdots + \varepsilon^m R_m(\eta), \tag{A.54}$$

where $\widetilde{G}_\varepsilon$ is a smooth function which is bounded uniformly in $t \in \mathbf{R}$ and bounded as $\varepsilon \to 0$ for each $\xi \in \mathbf{R}^n$. Let W be the image of the closed ball B_{2k} under the map α_t^{-1}. Then $\xi(t)$ and $\eta(t)$ are sitting in the compact set W if $0 \le t \le T/\varepsilon$. Let $L_1 > 0$ be a Lipschitz constant for $R_1(\xi) + \varepsilon R_2(\xi) + \cdots + \varepsilon^{m-1}R_m(\xi)$ on W and suppose that $\sup_{t\in\mathbf{R}, \xi\in W} ||\widetilde{G}_\varepsilon(t, \xi)|| < L_2$. Then, for $0 \le t \le T/\varepsilon$, the inequality

$$||\xi(t) - \eta(t)|| \le \varepsilon L_1 \int_0^t ||\xi(s) - \eta(s)||ds + \varepsilon^{m+1}L_2 t \tag{A.55}$$

holds. Then, the Gronwall inequality implies that

$$||\xi(t) - \eta(t)|| \le \frac{L_2}{L_1}\varepsilon^m(e^{\varepsilon L_1 t} - 1) \le \frac{L_2}{L_1}\varepsilon^m(e^{L_1 T} - 1), \ \ 0 \le t \le T/\varepsilon. \tag{A.56}$$

This shows that there exists a positive constant C such that $||x(t) - \widetilde{x}(t)|| = ||\alpha_t \circ \xi(t) - \alpha_t \circ \eta(t)|| \leq C\varepsilon^m$ holds if $0 \leq t \leq T/\varepsilon$. ■

The next theorem is a simple extension of Propositions 5.1 and 5.2.

Theorem A.9. *Consider an autonomous equation* (A.43).

(i) *If vector fields Fx and $g_1(x), g_2(x), \ldots$ are invariant under the action of a Lie group G, then the mth order RG equation is also invariant under the action of G.*

(ii) *The mth order RG equation commutes with the linear vector field Fx with respect to the Lie bracket product. Equivalently, each $R_i(A)$, $i = 1, 2, \ldots$, satisfies*

$$X(t)R_i(A) = R_i(X(t)A), \quad A \in \mathbf{R}^n. \tag{A.57}$$

Proof of Theorem A.9. Recall that G_i in (A.5) is independent of t since (A.43) is autonomous.

(i) We prove by induction that $R_i(A)$ and $h_t^{(i)}(A)$, $i = 1, 2, \ldots$, are invariant under the action of a Lie group G. Since $aX(t)A = X(t)aA$ and $ag_1(x) = g_1(ax)$ hold for all $a \in G$, $R_1(aA)$ is brought into the form

$$\begin{aligned} R_1(aA) &= \lim_{t\to\infty} \frac{1}{t} \int^t X(s)^{-1} G_1(X(s)aA) ds \\ &= a \lim_{t\to\infty} \frac{1}{t} \int^t X(s)^{-1} G_1(X(s)A) ds = aR_1(A). \end{aligned}$$

And the invariance of $h_t^{(1)}$, $h_t^{(1)}(aA) = ah_t^{(1)}(A)$, is verified in a similar way. Suppose that $R_k(aA) = aR_k(A)$ and $h_t^{(k)}(aA) = ah_t^{(k)}(A)$ hold for $k = 1, 2, \ldots, i-1$. Then, it is easy to verify that

$$(Dh_t^{(k)})_{aA} = a(Dh_t^{(k)})_A a^{-1}, \tag{A.58}$$

$$G_k(X(t)aA, h_t^{(1)}(aA), \ldots, h_t^{(k-1)}(aA)) = aG_k(X(t)A, h_t^{(1)}(A), \ldots, h_t^{(k-1)}(A)) \tag{A.59}$$

for $k = 1, 2, \ldots, i-1$. This and (A.13), (A.14) imply that $R_i(aA) = aR_i(A)$ and $h_t^{(i)}(aA) = ah_t^{(i)}(A)$.

(ii) We prove by induction that $R_i(X(t)A) = X(t)R_i(A)$ and $h_t^{(i)}(X(t')A) = h_{t+t'}^{(i)}(A)$ hold for $i = 1, 2, \ldots$. For all $s' \in \mathbf{R}$, $R_1(X(s')A)$ takes the form

$$\begin{aligned} R_1(X(s')A) &= \lim_{t\to\infty} \frac{1}{t} \int^t X(s)^{-1} G_1(X(s)X(s')A) ds \\ &= X(s') \lim_{t\to\infty} \frac{1}{t} \int^t X(s+s')^{-1} G_1(X(s+s')A) ds. \end{aligned}$$

Putting $s + s' = s''$, we verify that

$$\begin{aligned} R_1(X(s')A) &= X(s') \lim_{t\to\infty} \frac{1}{t} \int^{t+s'} X(s'')^{-1} G_1(X(s'')A) ds'' \\ &= X(s')R_1(A) + X(s') \lim_{t\to\infty} \frac{1}{t} \int_t^{t+s'} X(s'')^{-1} G_1(X(s'')A) ds'' \\ &= X(s')R_1(A). \end{aligned}$$

Next, $h_t^{(1)}(X(s')A)$ is calculated as

$$\begin{aligned} h_t^{(1)}(X(s')A) &= X(t)\int^t \left(X(s)^{-1}G_1(X(s)X(s')A) - R_1(X(s')A)\right) ds \\ &= X(t)X(s')\int^t \left(X(s')^{-1}X(s)^{-1}G_1(X(s)X(s')A) - R_1(A)\right) ds \\ &= X(t+s')\int^t \left(X(s+s')^{-1}G_1(X(s+s')A) - R_1(A)\right) ds. \end{aligned}$$

Putting $s+s'=s''$ provides

$$h_t^{(1)}(X(s')A) = X(t+s')\int^{t+s'} \left(X(s'')^{-1}G_1(X(s'')A) - R_1(A)\right) ds'' = h_{t+s'}^{(1)}(A). \tag{A.60}$$

Suppose that $R_k(X(t)A) = X(t)R_i(A)$ and $h_t^{(k)}(X(t')A) = h_{t+t'}^{(k)}(A)$ hold for $k = 1, 2, \ldots, i-1$. Then, $R_i(X(s')A)$ is calculated as

$$\begin{aligned} R_i(X(s')A) &= \lim_{t\to\infty}\frac{1}{t}\int^t \Bigg(X(s)^{-1}G_i(X(s)X(s')A, h_s^{(1)}(X(s')A), \ldots, h_s^{(i-1)}(X(s')A)) \\ &\qquad - X(s)^{-1}\sum_{k=1}^{i-1}(Dh_s^{(k)})_{X(s')A}R_{i-k}(X(s')A)\Bigg) ds \\ &= X(s')\lim_{t\to\infty}\frac{1}{t}\int^t \Bigg(X(s+s')^{-1}G_i(X(s+s')A, h_{s+s'}^{(1)}(A), \ldots, h_{s+s'}^{(i-1)}(A)) \\ &\qquad - X(s+s')^{-1}\sum_{k=1}^{i-1}(Dh_{s+s'}^{(k)})_A R_{i-k}(A)\Bigg) ds. \end{aligned}$$

Putting $s+s'=s''$ provides

$$\begin{aligned} R_i(X(s')A) &= X(s')\lim_{t\to\infty}\frac{1}{t}\int^{t+s'} \Bigg(X(s'')^{-1}G_i(X(s'')A, h_{s''}^{(1)}(A), \ldots, h_{s''}^{(i-1)}(A)) \\ &\qquad - X(s'')^{-1}\sum_{k=1}^{i-1}(Dh_{s''}^{(k)})_A R_{i-k}(A)\Bigg) ds'' \\ &= X(s')R_i(A) + X(s')\lim_{t\to\infty}\frac{1}{t}\int_t^{t+s'} \Bigg(X(s'')^{-1}G_i(X(s'')A, h_{s''}^{(1)}(A), \ldots, h_{s''}^{(i-1)}(A)) \\ &\qquad - X(s'')^{-1}\sum_{k=1}^{i-1}(Dh_{s''}^{(k)})_A R_{i-k}(A)\Bigg) ds'' \\ &= X(s')R_i(A). \end{aligned}$$

We can show that $h_t^{(i)}(X(t')A) = h_{t+t'}^{(i)}(A)$ in a similar way. ■

Acknowledgments. The author would like to thank Professor Toshihiro Iwai and Professor Hiroshi Kokubu for critical reading of the manuscript and for useful comments. The author is also grateful to Assistant Professor Yoshiyuki Y. Yamaguchi for bringing the RG method to his attention.

REFERENCES

[1] L. Y. Chen, N. Goldenfeld, and Y. Oono, *Renormalization group theory for global asymptotic analysis*, Phys. Rev. Lett., 73 (1994), pp. 1311–1315.
[2] L. Y. Chen, N .Goldenfeld, and Y. Oono, *Renormalization group and singular perturbations: Multiple scales, boundary layers, and reductive perturbation theory*, Phys. Rev. E (3), 54 (1996), pp. 376–394.
[3] T. Kunihiro, *A geometrical formulation of the renormalization group method for global analysis*, Progr. Theoret. Phys., 94 (1995), pp. 503–514.
[4] T. Kunihiro, *The renormalization-group method applied to asymptotic analysis of vector fields*, Progr. Theoret. Phys., 97 (1997), pp. 179–200.
[5] K. Nozaki and Y. Oono, *Renormalization-group theoretical reduction*, Phys. Rev. E (3), 63 (2001), 046101.
[6] S. Goto, Y. Masutomi, and K. Nozaki, *Lie-group approach to perturbative renormalization group method*, Progr. Theoret. Phys., 102 (1999), pp. 471–497.
[7] S. Ei, K. Fujii, and T. Kunihiro, *Renormalization-group method for reduction of evolution equations; invariant manifolds and envelopes*, Ann. Physics, 280 (2000), pp. 236–298.
[8] M. Ziane, *On a certain renormalization group method*, J. Math. Phys., 41 (2000), pp. 3290–3299.
[9] R. E. Lee Deville, A. Harkin, M. Holzer, K. Josić, and T. Kaper, *Analysis of a renormalization group method and normal form theory for perturbed ordinary differential equations*, Phys. D, 237 (2008), pp. 1029–1052.
[10] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.
[11] M. W. Hirsch, C. C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.
[12] S. Wiggins, *Normally Hyperbolic Invariant Manifolds in Dynamical Systems*, Springer-Verlag, New York, 1994.
[13] A. M. Fink, *Almost Periodic Differential Equations*, Lecture Notes in Math. 377, Springer-Verlag, New York, 1974.
[14] N. N. Bogoliubov and Y. A. Mitropolski, *Asymptotic Methods in the Theory of Non-Linear Oscillations*, Gordon and Breach, New York, 1961.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 933–961
© 2008 Society for Industrial and Applied Mathematics

Stability Analysis of Two-Dimensional Pool-Boiling Systems*

M. Speetjens†, A. Reusken‡, S. Maier-Paape§, and W. Marquardt¶

Abstract. In this paper we consider a model for pool-boiling systems known from the literature. This model involves only the temperature distribution within the heater and models the heat exchange with the boiling medium via a nonlinear boundary condition imposed on the fluid-heater interface. The model allows multiple homogeneous (i.e., spatially constant) and multiple heterogeneous steady-state solutions. The structure of this family of steady-state solutions has been studied by means of a bifurcation analysis in two recent papers by Speetjens, Reusken, and Marquardt [*Commun. Nonlinear Sci. Numer. Simul.*, 13 (2008), pp. 1475–1494; *Commun. Nonlinear Sci. Numer. Simul.*, 13 (2008), pp. 1518–1537]. The present study concentrates on stability properties of these steady-state solutions. To this end, a generic linear and a case-specific nonlinear stability analysis are performed which show that only the homogeneous steady-state solutions of complete nucleate or complete film boiling are linearly stable. All heterogeneous steady-state solutions appear linearly unstable. These stability results are consistent with laboratory observations.

Key words. pool boiling, stability, bifurcation analysis, numerical simulation

AMS subject classifications. 35K05, 47J35, 35B35, 35B41, 37M05

DOI. 10.1137/070706823

1. Introduction.

Pool boiling refers to boiling processes that lean on natural convection as a means for heat transfer through the boiling medium and is the key mode of thermal transport in many practical applications. Local heat-transfer phenomena near heating walls in industrial boiling equipment (e.g., evaporators and kettle reboilers) are essentially pool-boiling processes [3]. Furthermore, pool boiling is emerging as a novel cooling technique for electronics components [4]. Despite its importance, many aspects of (pool) boiling remain largely unexplored to date, mainly due to the immense complexity of the process induced by the intricate interplay between hydro- and thermodynamics. Studies on boiling known in the literature are mainly experimental and empirical. Theoretical investigations of fundamental phenomena in pool boiling, on the other hand, are scarce. This is the primary motivation for our recent studies, reported in [1] and [2], as well as for the follow-up study presented in this paper.

*Received by the editors October 30, 2007; accepted for publication (in revised form) by C. Wayne April 11, 2008; published electronically August 6, 2008.
http://www.siam.org/journals/siads/7-3/70682.html

†Energy Technology, Eindhoven University of Technology, Eindhoven, The Netherlands (m.f.m.speetjens@tue.nl). This author's research was partly funded by the German Research Foundation through the Research Training Group "Hierarchy and Symmetry in Mathematical Models."

‡Numerical Analysis, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany (reusken@igpm.rwth-aachen.de).

§Mathematics, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany (maier@instmath.rwth-aachen.de).

¶Process Systems Engineering, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany (Wolfgang.Marquardt@avt.rwth-aachen.de).

The central topic of the present study is the stability behavior of pool-boiling systems. Laboratory experiments indicate that, without active control, pool-boiling systems allow only two stable steady-state solutions, namely nucleate boiling and film boiling [5, 6, 7, 8]. Other states belong to the transition-boiling regime and are inherently unstable. Nucleate boiling is, as opposed to film boiling, an efficient and safe mode of heat transfer and the sought-after boiling mode in most practical applications [9]. However, for typical operating conditions, the system admits both nucleate boiling and film boiling as steady states [1], and, consequently, the stable state eventually attained by the system is a priori unclear. Whether a given unstable state in the transition-boiling regime evolves towards either the nucleate-boiling or the film-boiling state is of major practical importance, though. This is intimately related to the stability properties of boiling states. Stability analyses of pool-boiling systems are hitherto restricted to highly idealized models such as, for instance, heated wires [10, 11, 12], heated foils [13], heated cylinders with homogeneous boiling conditions [14, 15], and rectangular "thick" heaters with artificial heterogeneous boiling conditions [13]. Similar studies for more sophisticated models including both realistic heater geometries and realistic heterogeneous boiling conditions are not known in the literature. This is the impetus for the study presented in this paper.

The stability analysis in this paper concerns the stability properties of the multiple heterogeneous boiling states that have been found in [1] for a spatially two-dimensional (2D) heater. Key to the modelling approach is the phenomenological connection between the local state of aggregation of the boiling medium and the local temperature at the fluid-heater interface at mesoscopic length and time scales:[1] "lower" and "higher" temperatures correspond to the liquid and vapor phases, respectively. This allows a description of the (qualitative) behavior of the pool-boiling problem entirely in terms of the temperature field within the heater. Thus the pool-boiling problem is reduced to a heat-transfer problem for the heater with a nonlinear heat-flux relation at the interface between the heater and the boiling medium. This heater-only model is based on the approach used in [13, 14]. Section 2 provides a concise description. Further details can be found in [1].

The nonlinear heat-transfer model resembles nonlinear evolution equations of parabolic type (e.g., reaction-diffusion and pattern-formation equations) known from mathematical physics [17, 18]. The dynamics of such systems are typically dominated by a global attractor, consisting of the steady-state solutions and their heteroclinic connections, to which initial conditions converge if time evolves [18]. Said resemblance suggests that the pool-boiling model may exhibit similar dynamical behavior. However, in the pool-boiling model the nonlinearity of the problem is due to the nonlinear heat-flux condition at the heater-fluid interface and is not due to a nonlinearity in the partial differential equation itself as in "conventional" nonlinear parabolic evolution equations. Thus the concepts known for the latter problem class cannot be applied directly to the pool-boiling problem. We are unaware of rigorous mathematical studies on, e.g., existence, smoothness, and asymptotic stability of solutions of problems involving nonlinear boundary conditions such as our pool-boiling problem. A rigorous mathematical analysis is beyond the present scope, however. Instead, preliminary

[1]Here mesoscopic means locally averaged in space and time over intervals larger than bubble dimensions and bubble lifetimes in order to smooth out microscopic short-term fluctuations [16].

results are given that indicate the existence of a global attractor consisting of steady-state solutions and their heteroclinic connections (section 3.1). This strongly suggests dynamical behavior akin to that of "conventional" nonlinear evolution equations and justifies a study on the existence of (multiple) steady-state solutions and their stability properties.

The set of steady-state solutions of the 2D boiling problem has been studied extensively in [1]; an extension to the three-dimensional (3D) case is given in [2]. The present study investigates the stability properties of these steady states by a linear stability analysis. This analysis hinges on linearization of the nonlinear problem at a given steady-state solution. Treatment of the resulting linearized model with a separation-of-variables technique results in a linear eigenvalue problem that governs the eigenmodes and corresponding eigenvalues (i.e., temporal growth rates) of steady-state solutions (section 3.2.1). Analysis of this eigenvalue problem yields generic stability properties (section 3.2.2). The eigenmodes and eigenvalues of a given steady state can be computed (approximately) using a Fourier-collocation discretization method (section 3.2.3). The generic linear stability analysis and the eigenmode decomposition are demonstrated for a representative set of steady-state solutions (section 4.1). A brief recapitulation of steady-state solutions determined in [1] is given in section 3.3. Numerical simulation of the *non*linear evolution of linearly unstable steady states is performed by a spectral algorithm. These simulations validate the linear analysis and yield first insight into the *non*linear (in)stability behavior (section 4.2). Conclusions are drawn in section 5.

2. Model problem. The stability of pool-boiling systems is investigated in terms of a model problem considered in [1]. An extensive discussion and motivation of this model is provided in that paper. Here we restrict ourselves to a concise description of the nondimensional formulation of this model.

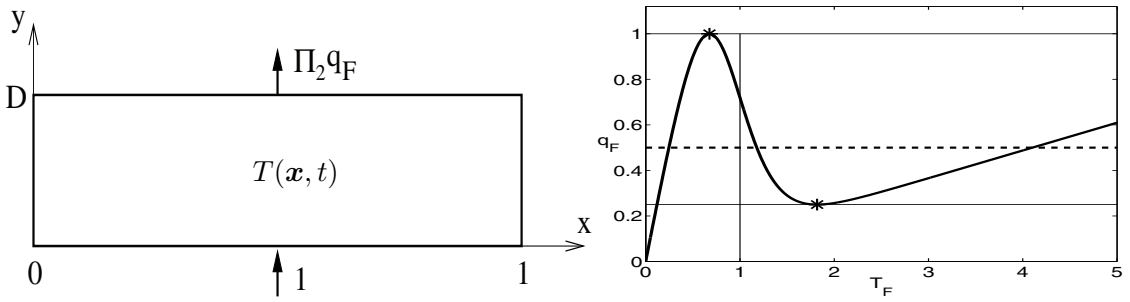
We consider the heat transfer within the 2D rectangular heater $\mathcal{D} := [0,1] \times [0,D]$ (Figure 1(a)). Its boundary is given by $\Gamma = \partial\mathcal{D} = \Gamma_H \cup \Gamma_A \cup \Gamma_F$ and comprises the boundary segments Γ_H: $y = 0$ (constant heat supply), Γ_A: $x = 0, 1$ (adiabatic sidewalls), and Γ_F: $y = D$ (nonconstant heat extraction by boiling process). The heat transfer within $\mathcal{D}$ is modeled by

$$\begin{aligned} \frac{\partial T}{\partial t} &= \kappa \nabla^2 T \qquad && \text{in } \ \mathcal{D} \times [0, t_{\text{end}}], \\ T(\boldsymbol{x}, 0) &= T_0(\boldsymbol{x}) \qquad && \text{for } \ \boldsymbol{x} := (x, y) \in \mathcal{D}, \\ \frac{\partial T}{\partial \nu} &= g(\boldsymbol{x}, T) \qquad && \text{on } \ \Gamma, \end{aligned} \tag{2.1}$$

with ν the outward normal of Γ. The boundary condition on Γ introduces a nonlinearity due to the dependence of g on $T(x, D, t) =: T_F(x, t)$. Note that T_F denotes the value of T at the fluid-heater interface Γ_F. The function g is given by

$$g(\boldsymbol{x}, Z) = \begin{cases} 0 & \text{for } x = 0 \text{ or } x = 1, \\ 1/\Lambda & \text{for } y = 0, \\ -\Pi_1 \, q_F(Z; \Pi_2, \Pi_3)/\Lambda & \text{for } y = D. \end{cases} \tag{2.2}$$

The function $T = T(\boldsymbol{x}, t)$ is the nondimensional temperature excess (i.e., the temperature relative to the boiling point of the medium). System parameters are D (aspect ratio of the heater), Λ (nondimensional thermal conductivity), κ (nondimensional thermal diffusivity),



(a) *Heater configuration.* (b) *Heat-flux function.*

Figure 1. *Nondimensional model problem:* (a) *heater configuration and* (b) *heat-flux function* q_F. *The dashed line represents the normalized heat supply* Π_1^{-1}.

and Π_1 (nondimensional critical heat flux of boiling process), all of which are positive. The nonlinear heat-flux function $q_F(\cdot;\Pi_2,\Pi_3)$ accounts for the heat exchange between the heater and the boiling medium. This function is specified in the appendix and introduces two further control parameters, viz. Π_2 and Π_3, resulting in a total of six parameters. However, physical considerations suggest $\Lambda D/\kappa = |1-\Pi_1|$, and thus the model contains five independent control parameters.

Physical considerations further suggest that the heat-flux function q_F, which describes the *local* heat exchange between fluid and heater, should be qualitatively similar to the so-called boiling curve. The latter describes the *global* heat exchange between fluid and heater obtained via averaging over the fluid-heater interface Γ_F. Therefore, as before [1] we use a heat-flux function schematically shown in Figure 1(b), which has the typical shape of a boiling curve; see the appendix for an explicit expression. The heat-flux function is parameterized by Π_2 (ratio of extremal heat fluxes) and Π_3 (ratio of extremal temperatures) and consists of three distinct regimes that correspond to one of the local boiling modes and associated mesoscopic states: nucleate boiling (left of local maximum; fluid-rich state); transition boiling (in between both extrema; transitional state); film boiling (right of local minimum; vapor-rich state).[2]

Important to note is that, despite being of the same shape, imposition of the heat-flux function q_F on the fluid-heater interface Γ_F is not equivalent to imposition of the boiling curve on Γ_F. Heat-flux function q_F describes the *local* heat flux as a function of the *local* interface temperature and holds for both homogeneous and heterogeneous interface states. The boiling curve, on the other hand, describes the *mean* heat flux as a function of the *mean* interface temperature and strictly holds only for homogeneous interface states. (Homogeneous states satisfying q_F automatically satisfy the boiling curve and form a subset of the total set of solutions to the proposed model.) The local nature of q_F is essential for the present model to admit the (multiple) heterogeneous boiling states that characterize physical boiling systems.

3. Unsteady pool-boiling problem: Generic analysis. The nonlinear heat-transfer model (2.1) bears resemblance to nonlinear parabolic problems known from mathematical physics [18]. However, a fundamental difference is that in the heat-transfer model (2.1) the nonlinearity resides in the boundary condition rather than in the partial differential equation itself.

[2]Further physical background for this heater-only model is given by [1].

This is a direct consequence of the finite thickness ($D > 0$) of the present heater configuration and distinguishes our problem from "thin" ($D \to 0$) configurations. For vanishing D the model reduces to a partial differential equation with a nonlinear source term (see, e.g., [13]), and generic concepts for the analysis of parabolic systems—in particular of the reaction-diffusion type [17]—can be applied. Examples of such approaches to pool-boiling systems are in [10, 11, 12, 13].

Rather than providing a rigorous mathematical analysis, which is beyond the present scope, in this section we indicate that concepts similar to those introduced for "conventional" parabolic evolution equations [18] can be applied to the pool-boiling problem, too. We assume that solutions of the pool-boiling problem fulfill all the regularity conditions that admit application of such concepts. We note that the results of the generic analysis and those of the case study in section 4 are in good agreement.

3.1. Generic dynamical behavior. The dynamical behavior of an evolutionary (parabolic) system is commonly investigated based on its corresponding weak formulation [18]. We use the Sobolev space $H^1(\mathcal{D})$ and the notation

$$(u,v)_{L^2(\mathcal{D})} := \int_{\mathcal{D}} uv \, dx \, dy, \quad B(T,\varphi) := \int_{\mathcal{D}} \nabla T \cdot \nabla \varphi \, dx \, dy.$$

For the heat-transfer problem (2.1) the weak formulation is as follows: find $u = u(t) = T(\cdot,t) \in H^1(\mathcal{D})$ with $\frac{du(t)}{dt} \in L^2(\mathcal{D})$ such that

$$\left(\frac{du(t)}{dt}, \varphi\right)_{L^2(\mathcal{D})} = \int_{\partial\mathcal{D}} g(\boldsymbol{x},u)\,\varphi \, d\sigma - \kappa B(u,\varphi) \quad \text{for all } \varphi \in H^1(\mathcal{D}). \tag{3.1}$$

If a solution u of (3.1) is sufficiently smooth, it clearly provides an L_2-solution of (2.1). Conversely, regular solutions to (2.1) naturally are solutions to (3.1).

The weak formulation describes the evolution of the system from its initial state $u(0)$ to its current state $u(t)$; i.e., $u(t) = \Phi_t(u(0))$ defines a flow. Thus the weak formulation (3.1) defines a dynamical system for the weak solution $u(t)$. Existence and uniqueness of a solution of (3.1) depend on smoothness and growth properties of g. This topic is not studied in the present paper. We *assume* that g is such that the following property holds.

Property 1. *System* (3.1) *results in a global semiflow:* $\Phi_t : H^1(\mathcal{D}) \to H^1(\mathcal{D})$, $t \in [0,\infty)$.

The dynamical system (3.1) has a gradient structure; i.e., the functional $E : H^1(\mathcal{D}) \to \mathbb{R}$,

$$E(u) := \frac{\kappa}{2} B(u,u) - \int_{\partial\mathcal{D}} G(\boldsymbol{x},u)\, d\sigma, \tag{3.2}$$

with $G(\boldsymbol{x},u)$ such that $D_u G(\boldsymbol{x},u) = g(\boldsymbol{x},u)$, defines an energy or Lyapunov function for the solutions u of (3.1). Under certain smoothness and growth assumptions on g we have $DE(u) \in \mathcal{L}\big(H^1(\mathcal{D}),\mathbb{R}\big)$ and

$$DE(u)[\varphi] = \kappa B(u,\varphi) - \int_{\partial\mathcal{D}} g(\boldsymbol{x},u)\,\varphi \, d\sigma \quad \text{for all } u,\varphi \in H^1(\mathcal{D}). \tag{3.3}$$

Along a solution curve $u(t)$ of (3.1) the energy E decays monotonically in time:

$$\frac{d}{dt}\left[E(u(t))\right] = DE(u)\left[\frac{du}{dt}\right] = -\left(\frac{du}{dt}, \frac{du}{dt}\right)_{L^2(\mathcal{D})} \leq 0, \tag{3.4}$$

with equality (for a range of t-values) only if $u(t) = T(\cdot, t) = T_\infty(\cdot)$ is a steady-state solution. This property implies a loss of energy of solutions $u(t) = T(\cdot, t)$ with progressing time. This already provides important information on the long-term dynamical behavior, namely that for any initial condition u_0 the corresponding solution converges to a steady-state solution.

Our case study for the pool-boiling problem further indicates the existence of a global attractor $\mathcal{A}$ for the semiflow $\Phi_\cdot$, i.e., a strictly positive invariant subset of $H^1(\mathcal{D})$, which is compact and attracts all bounded subsets of $H^1(\mathcal{D})$ (cf. [18], Definition I. 1.3). We *assume* the following property to guarantee this existence.

Property 2. *Assume that the global semiflow $\Phi_\cdot$ generated by (3.1) has some compactness property, e.g., in the sense of* [18, *Theorem* I. 1.12 *or* I. 1.13]*, and further assume the existence of a bounded set $M \subset H^1(\mathcal{D})$, which attracts all bounded sets in $H^1(\mathcal{D})$.*

In order to elucidate the consequences of the above properties we introduce the following notation. Let $\mathcal{E}$ be the set of equilibria:

$$\mathcal{E} = \left\{u \in H^1(\mathcal{D}) \colon\ \Phi_t(u) = u \quad \text{for all } t \geq 0\right\}.$$

For $u_0, v_0 \in \mathcal{E}$ let $\mathcal{C}(u_0, v_0)$ be the set of heteroclinic connections between u_0 and v_0, i.e., all full orbits that approach u_0 as $t \to -\infty$ and v_0 as $t \to \infty$. Using these notions we can formulate the following result.

Theorem 3.1. *We assume Properties 1 and 2. Then the semiflow $\Phi_\cdot$ possesses a global compact attractor $\mathcal{A}$. This global attractor equals the unstable set of $\mathcal{E}$; i.e., it consists of full orbits which approach the set of steady-state solutions for $t \to -\infty$. If, furthermore, $\mathcal{E}$ is discrete, then the global attractor consists of steady-state solutions $u(t) = T(\cdot, t) = T_\infty(\cdot) \in H^1(\mathcal{D})$ of (3.1), i.e.,*

$$\kappa B(T_\infty, \varphi) = \int_{\partial\mathcal{D}} g(\boldsymbol{x}, T_\infty)\, \varphi \, d\sigma \quad \textit{for all } \varphi \in H^1(\mathcal{D}), \tag{3.5}$$

and their heteroclinic connections

$$\mathcal{A} = \left(\bigcup_{u_0 \in \mathcal{E}} \{u_0\}\right) \cup \bigcup_{u_0, v_0 \in \mathcal{E}} \mathcal{C}(u_0, v_0). \tag{3.6}$$

Proof. [18, Theorems I.1.1 and VII.4.1]. ∎

In general, unstable steady-state solutions u_0 have a stable manifold $\mathcal{M}_s(u_0) = \{w \in H^1(\mathcal{D}) : \Phi_t(w) \to u_0 \text{ for } t \to \infty\}$ with a finite nonzero codimension. Therefore it is most likely for a generic initial condition u_0 to lie in the stable manifold of a stable equilibrium $g \in \mathcal{E}$. Hence in realistic systems due to physical imperfections as well as numerical simulations due to rounding errors the evolution process basically always converges towards a stable steady state.

Isolation of the global attractor of the boiling problem requires identification of its steady-state solutions and determination of the corresponding stability properties. The steady-state solutions have been studied extensively in [1]; a brief recapitulation is given in section 3.3. The present paper concerns the corresponding stability properties. Section 3.2 gives a generic linear stability analysis; section 4 demonstrates and validates this linear analysis by way of numerical simulation of the nonlinear evolution of unstable steady-state solutions. This also offers first insight into the nonlinear stability behavior of the system. Moreover, it may enable a more detailed investigation of the structure of the attractor by the approach proposed in [19] along the lines of the analysis of the Cahn–Hilliard equation in [20] and [21].

3.2. Linear stability analysis of steady-state solutions.

3.2.1. Linearized heat-transfer model. The stability analysis of steady-state solutions that we present is based on the linear theory of stability; cf. [22, section I.6]. To determine stability of a steady-state solution $T_\infty(\boldsymbol{x})$, one subjects this steady-state solution to a small initial perturbation $v_0(\boldsymbol{x}) = v(\boldsymbol{x}, 0)$. We assume that for the nonlinear problem (2.1) the *principle of linearized stability* holds (cf. [23]); i.e., solutions of the nonlinear problem in a neighborhood of a steady-state solution T_∞ and of the linearized problem (linearization at T_∞) have the same qualitative behavior. This assumption justifies the analysis of stability properties of the nonlinear problem by means of a stability analysis of the linearized problem. In this paper, unless stated otherwise explicitly, the notions stable and unstable are always meant in the sense of this linear theory of stability.

For the stability analysis we introduce the linearization of (2.1). Let $T_\infty(\boldsymbol{x})$ be a regular steady-state solution of (2.1); i.e., T_∞ satisfies the following Laplace equation with a nonlinear Neumann boundary condition:

$$\nabla^2 T_\infty = 0 \quad \text{in } \mathcal{D}, \qquad \frac{\partial T_\infty}{\partial \nu} = g(\boldsymbol{x}, T_\infty) \quad \text{on } \Gamma = \partial\mathcal{D} \tag{3.7}$$

(g as in (2.2)). Properties of the nonlinear steady-state problem (3.7) are derived in [1] and summarized in section 3.3. The corresponding linearized problem at T_∞ for the perturbation $v(\boldsymbol{x}, t)$ induced by an initial perturbation $v_0(\boldsymbol{x})$ is given by

$$\begin{aligned} \frac{\partial v}{\partial t} &= \kappa \nabla^2 v && \text{in } \mathcal{D} \times [0, t_{\text{end}}], \\ v(\boldsymbol{x}, 0) &= v_0(\boldsymbol{x}) && \text{for } \boldsymbol{x} \in \mathcal{D}, \\ \frac{\partial v}{\partial \nu} &= f(\boldsymbol{x}) v && \text{on } \Gamma. \end{aligned} \tag{3.8}$$

The Neumann boundary condition is given by

$$f(\boldsymbol{x}) = \begin{cases} 0 & \text{on } \Gamma \setminus \Gamma_F, \\ -\gamma(x) & \text{on } \Gamma_F, \end{cases} \tag{3.9}$$

with

$$\gamma(x) = \frac{\Pi_1}{\Lambda} \frac{dq_F}{dZ} \left(T_{F,\infty}(x) \right), \qquad x \in [0, 1], \tag{3.10}$$

where $T_{F,\infty}(x) := T_\infty(x, D)$ is the steady-state temperature profile at the fluid-heater interface Γ_F. Thus the original nonlinear condition on Γ_F simplifies to a standard linear Neumann condition with an x-dependent coefficient determined by the interface temperature $T_{F,\infty}(x)$ of the steady-state solution.

As ansatz (based on separation of variables) we seek solutions of (3.8) of the form

$$v(\boldsymbol{x}, t) = \mathrm{e}^{-\kappa\mu t}\psi(\boldsymbol{x}). \tag{3.11}$$

Substitution of (3.11) into (3.8) leads to the following linear elliptic eigenvalue problem for ψ:

$$\nabla^2\psi + \mu\psi = 0 \quad \text{in} \quad \mathcal{D}, \tag{3.12}$$

$$\frac{\partial\psi}{\partial\nu} = 0 \quad \text{on} \quad \Gamma \setminus \Gamma_F, \tag{3.13}$$

$$\frac{\partial\psi}{\partial\nu} + \gamma(x)\psi = 0 \quad \text{on} \quad \Gamma_F. \tag{3.14}$$

The weak formulation of this eigenvalue problem is as follows: determine $\mu_n \in \mathbb{R}$, $\psi_n \in H^1(\mathcal{D})$ such that

$$-B(\psi_n, \varphi) - \int_{\Gamma_F} \gamma\,\psi_n\,\varphi\,d\sigma + \mu_n \int_{\mathcal{D}} \psi_n\,\varphi\,dx\,dy = 0 \quad \text{for all } \varphi \in H^1(\mathcal{D}). \tag{3.15}$$

The eigenpairs that solve this problem are denoted by (μ_n, ψ_n), $n = 1, 2, \ldots$. The eigenfunctions are scaled such that $\|\psi_n\|_{L^2(\mathcal{D})} = 1$. Whether, for generic γ (e.g., $\gamma \in L^\infty(\Gamma_F)$), the eigenfunctions $(\psi_n)_{n\geq 1}$ form a complete orthogonal basis of $L^2(\mathcal{D})$ is an open question. We do not study this topic in the present paper. Instead, we assume that these eigenfunctions span a space that is sufficiently large such that it makes sense to restrict the choice of the initial perturbations $v_0(\boldsymbol{x})$ to this space. We obtain the following representation for the unique solution v of (3.8) induced by an initial perturbation $v_0 \in \operatorname{span}\{\psi_n \mid n \geq 1\}$:

$$\text{Let} \quad v_0(\boldsymbol{x}) = \sum_{k=0}^{\infty} \eta_k\psi_k, \quad \text{with} \quad \eta_k = (v_0, \psi_k)_{L^2(\mathcal{D})}; \tag{3.16}$$

$$\text{then} \quad v(\boldsymbol{x}, t) = \sum_{k=1}^{\infty} \eta_k \mathrm{e}^{-\kappa\mu_k t}\psi_k(\boldsymbol{x}). \tag{3.17}$$

Note that the eigenvalues μ_k depend (via γ) on T_∞ yet *not* on the perturbation v_0.

Representation (3.17) implies that the steady-state solution T_∞ of (2.1) is linearly stable w.r.t. all perturbations of the form (3.16) if all μ_k are nonnegative. Conversely, T_∞ is an unstable steady-state solution if at least one μ_k is negative. Relation (3.15), in turn, implies that $\mu_k \geq 0$ for all k if $\gamma \geq 0$. Thus it follows that $\gamma \geq 0$ on the fluid-heater interface Γ_F is a sufficient condition for linear stability of the steady-state solution.

3.2.2. Analysis of the eigenvalue problem. The stability properties of steady-state solutions T_∞ of (2.1) are directly related to the eigenvalues of problem (3.12)–(3.14). In this section we study this eigenvalue problem.

We use an approach based on Fourier analysis. First, for a *fixed* $\mu \in \mathbb{R}$ we consider the problem (3.12)–(3.13) and apply separation of variables to construct bivariate Fourier modes that satisfy (3.12)–(3.13). Let $\phi(\boldsymbol{x}) = \alpha(x)\beta(y)$. An elementary computation shows that $c\,\alpha(x)\beta(y)$, $c \in \mathbb{R} \setminus \{0\}$, solves (3.12)–(3.13) if and only if

$$\alpha(x) = \cos(n\pi x), \qquad \beta(y) = \cosh\bigl(\sqrt{(n\pi)^2 - \mu}\, y\bigr), \quad n = 0, 1, 2, \ldots.$$

Note that for $z < 0$ we have $\cosh(\sqrt{z}) = \cosh(i\sqrt{|z|}) = \cos(\sqrt{|z|})$. We now make the ansatz that the whole solution space of (3.12)–(3.13) is obtained by superposition of these Fourier modes; i.e., all solutions of (3.12)–(3.13) lie in the space

$$S_\mu := \left\{ \psi(\boldsymbol{x}) = \sum_{n=0}^{\infty} A_n \cos(n\pi x) \cosh(\sqrt{\alpha_{n,\mu}}\, y), \quad A_n \in \mathbb{R},\ \alpha_{n,\mu} := (n\pi)^2 - \mu \right\}.$$

If we take the boundary condition (3.14) into account, then the solutions of (3.12)–(3.14) form a subspace of S_μ. A function $\phi(\boldsymbol{x}) = \sum_{n=0}^{\infty} A_n \cos(n\pi x) \cosh(\sqrt{\alpha_{n,\mu}}\, y) \in S_\mu$ solves (3.14) if and only if

$$\sum_{n=0}^{\infty} A_n \sqrt{\alpha_{n,\mu}} \sinh(\sqrt{\alpha_{n,\mu}}\, D) \cos(n\pi x) + \gamma(x)\phi(x, D) = 0 \quad \text{for all} \quad x \in [0, 1]. \tag{3.18}$$

Thus the problem of finding the eigenvalues of (3.12)–(3.14) is transformed to the problem of finding μ such that (3.18) has a nontrivial solution $\phi_F(x) := \phi(x, D)$. Note that the latter problem is spatially *one*-dimensional (1D). Furthermore, for $z < 0$ we have $\sqrt{z}\sinh(\sqrt{z}) = -\sqrt{|z|}\sin(\sqrt{|z|})$, and thus for all $\alpha_{n,\mu} \in \mathbb{R}$ (3.18) is real. We will solve this equation using a univariate Fourier analysis. To this end some notation is introduced. Let $\mathcal{F} : L^2([0, 1]) \to \ell^2$ be the Fourier transform:

$$\mathcal{F}\left(\sum_{n=0}^{\infty} c_n \cos(n\pi \cdot) \right) = (c_n)_{n \geq 0}.$$

For $\mathbf{c}, \mathbf{d} \in \ell^2$ the elementwise multiplication is denoted by $\mathbf{c} * \mathbf{d} = (c_n d_n)_{n \geq 0}$. Furthermore, we write for $\phi \in S_\mu$ restricted to fluid-heater interface Γ_F

$$\begin{aligned} \phi_F(x) := \phi(x, D) &= \sum_{n=0}^{\infty} A_n \cosh(\sqrt{\alpha_{n,\mu}}\, D) \cos(n\pi x) = \sum_{n=0}^{\infty} \widetilde{\phi}_n \cos(n\pi x), \\ \widetilde{\phi}_n &:= A_n \cosh(\sqrt{\alpha_{n,\mu}}\, D). \end{aligned} \tag{3.19}$$

Using this, (3.18) can be rewritten as follows: find $\phi_F(x) = \sum_{n=0}^{\infty} \widetilde{\phi}_n \cos(n\pi x)$ such that

$$\sum_{n=0}^{\infty} \sqrt{\alpha_{n,\mu}} \tanh(\sqrt{\alpha_{n,\mu}}\, D) \widetilde{\phi}_n \cos(n\pi x) + \gamma(x)\phi_F(x) = 0 \quad \text{for all} \quad x \in [0, 1]. \tag{3.20}$$

For $z < 0$ we have $\sqrt{z}\tanh(\sqrt{z}) = -\sqrt{|z|}\tan(\sqrt{|z|})$. Define

$$\mathbf{d}_\mu \in \ell^2, \quad (d_\mu)_n := \sqrt{\alpha_{n,\mu}} \tanh(\sqrt{\alpha_{n,\mu}}\, D), \quad n = 0, 1, \ldots.$$

Then problem (3.20) has the following compact formulation: find ϕ_F such that

$$J_\mu \phi_F := \mathcal{F}^{-1}\big(\mathbf{d}_\mu * \mathcal{F}\phi_F\big) + \gamma\phi_F = 0. \tag{3.21}$$

In this formulation it is implicitly assumed that a solution ϕ_F is sufficiently smooth such that $\mathbf{d}_\mu * \mathcal{F}\phi_F \in \text{range}(\mathcal{F})$. The linear operator J_μ is well defined on a dense subspace of $L^2([0,1])$. We are interested in values for μ for which (3.21) has a nontrivial solution ϕ_F. Furthermore, we are interested in the sign of these eigenvalues μ, as they determine the stability of corresponding steady-state solutions. In the analysis below we distinguish two cases: constant and nonconstant γ. For $\gamma(x) = \gamma =$ constant the relevant properties of μ can be determined analytically. For the general case of a smooth but not necessarily constant function $\gamma(x)$ certain properties can still be derived analytically. However, for full insight we must resort to discretization of (3.21) and study its properties via numerical computations. Note that, for brevity, ϕ hereafter refers to both the full and the boundary solutions; its meaning readily follows from the context.

Homogeneous temperature on fluid-heater interface: Constant γ. If $\gamma(x) = \gamma$ is constant, the following holds.

Theorem 3.2. *There exist sequences $(z_k^+)_{k\in\mathbb{N}}$ with $z_k^+ \in [(k-1)\pi, (k-\frac{1}{2})\pi)$, $k \geq 1$, and $(z_k^-)_{k\in\mathbb{N}}$ with $z_1^- < 0$, $z_{k+1}^- \in [(k-\frac{1}{2})\pi, k\pi)$, $k \geq 1$, such that the following holds. If (μ, ϕ), with $\phi \neq 0$, solves (3.21), then $\mu \in (\mu_{k,n})_{k,n\in\mathbb{N}}$ with $\mu_{k,n}$ defined by*

$$\mu_{k,n} = z_k^+ + (n\pi)^2 \quad \textit{if } \gamma \geq 0, \tag{3.22}$$

$$\mu_{k,n} = z_k^- + (n\pi)^2 \quad \textit{if } \gamma < 0. \tag{3.23}$$

For all k, n, the pair $\mu = \mu_{k,n}$, $\phi(x) = \cos(n\pi x)$ is a solution of (3.21).

Proof. Note that

$$\begin{aligned} & \mathcal{F}^{-1}\big(\mathbf{d}_\mu * \mathcal{F}\phi\big) + \gamma\phi = 0 \\ \Leftrightarrow\quad & \mathbf{d}_\mu * \mathcal{F}\phi + \gamma\mathcal{F}\phi = 0 \\ \Leftrightarrow\quad & \text{for all } n: \ (d_\mu)_n + \gamma = 0 \ \text{ or } \ \widetilde{\phi}_n = (\mathcal{F}\phi)_n = 0. \end{aligned}$$

Take an $n \in \mathbb{N}$ such that $\widetilde{\phi}_n \neq 0$. Then $(d_\mu)_n + \gamma = 0$ must hold. We consider the equation

$$\mu \to (d_\mu)_n + \gamma = \sqrt{\alpha_{n,\mu}}\tanh\big(\sqrt{\alpha_{n,\mu}}\,D\big) + \gamma = 0, \quad \alpha_{n,\mu} = (n\pi)^2 - \mu. \tag{3.24}$$

Thus we look for the roots of the function $g(z) := \sqrt{z}\tanh(\sqrt{z}\,D) + \gamma$. For $z > 0$ and $\gamma \geq 0$ the equation $g(z) = 0$ has no solution. For $z > 0$ and $\gamma < 0$ there is a unique root $z^* > 0$. Define $z_1^- := -z^*$. This induces a corresponding $\mu_1 := z_1^- + (n\pi)^2$ that solves (3.24). For $z \leq 0$ we have $g(z) = -\sqrt{-z}\tan(\sqrt{-z}\,D) + \gamma$. An elementary analysis shows that for $\gamma \geq 0$ the equation $g(z) = 0$ has negative roots z^* with $-z^* =: z_k^+ \in [(k-1)\pi, (k-\frac{1}{2})\pi)$ for $k = 1, 2, \ldots$. For $\gamma < 0$ the equation $g(z) = 0$ has negative roots z^* with $-z^* =: z_{k+1}^- \in [(k-\frac{1}{2})\pi, k\pi)$ for $k = 1, 2, \ldots$. Due to $-z = -\alpha_{n,\mu} = \mu - (n\pi)^2$ we obtain corresponding solutions of (3.24): $\mu_k = z_k^+ + (n\pi)^2$, $k = 1, 2, \ldots$, for the case $\gamma \geq 0$ and $\mu_k = z_k^- + (n\pi)^2$, $k = 2, 3, \ldots$, for the case $\gamma < 0$. Combining the results for the cases $z > 0$ and $z \leq 0$ reveals that all possible roots

of (3.24) are given by (3.22)–(3.23). If we take $\widetilde{\phi}_n = 1$, $\widetilde{\phi}_m = 0$ for all $m \neq n$, then $(\mu_{k,n}, \phi)$ with $\phi(x) = \cos(n\pi x)$ solves (3.21). ■

Remark 1. Relation (3.22) implies that for the case $\gamma \geq 0$ all eigenvalues $\mu = \mu_{k,n}$ of (3.12)–(3.14) are positive. Relation (3.23) yields that for the case $\gamma < 0$ there always exists some integer $n_0 > 0$ such that $\mu_{1,n} < 0$ for all $0 \leq n \leq n_0$. *Thus Theorem 3.2 describes the linear stability properties of the pool-boiling problem in case of constant γ. Steady-state solutions for which $\gamma \geq 0$ holds are linearly stable, whereas steady-state solutions for which $\gamma < 0$ holds are unstable.*

Remark 2. Using Theorem 3.2 and the representation (3.19) we obtain that the eigenfunction corresponding to $\mu = \mu_{k,n}$ is given by $\phi(x, y) = \cos(n\pi x)\cosh(\sqrt{(n\pi)^2 - \mu_{k,n}}\, y)$. Note that $(n\pi)^2 - \mu_{k,n} = -z_k^+$ for $\gamma \geq 0$ and $(n\pi)^2 - \mu_{k,n} = -z_k^-$ for $\gamma < 0$. Hence, for $\gamma \geq 0$ an eigenfunction corresponding to $\mu = \mu_{k,n}$ is given by $\phi(x, y) = \cos(n\pi x)\cos(\sqrt{z_k^+}\, y)$, and for $\gamma < 0$ we obtain $\phi(x, y) = \cos(n\pi x)\cosh(\sqrt{-z_1^-}\, y)$ if $k = 1$ and $\phi(x, y) = \cos(n\pi x)\cos(\sqrt{z_k^-}\, y)$ if $k \geq 2$.

Heterogeneous temperature on fluid-heater interface: Nonconstant γ. The linear operator J_μ in (3.21) is *symmetric* on its domain in $L^2([0, 1])$. This allows a general analysis using real eigenvalues and energy arguments. We introduce the notation $I := [0, 1]$ and $\gamma_{min} := \min_{x \in I} \gamma(x)$. We derive a similar stability condition as before.

Theorem 3.3. *Assume that $\gamma_{min} \geq 0$ holds. Then for all solutions (μ, ϕ) of (3.21), with $\phi \neq 0$, we have $\mu \geq 0$.*

Proof. If $\mu, \phi \neq 0$ satisfies (3.21), we attain

$$\big(\mathcal{F}^{-1}(\mathbf{d}_\mu * \mathcal{F}\phi), \phi\big)_{L^2(I)} + (\gamma\phi, \phi)_{L^2(I)} = 0. \tag{3.25}$$

Suppose $\mu < 0$. Then $\alpha_{n,\mu} = (n\pi)^2 - \mu > 0$ for all n, and thus $(d_\mu)_n > 0$ for all n. This implies, with $c_n := \|\cos(n\pi x)\|_{L^2(I)}$, that

$$\big(\mathcal{F}^{-1}(\mathbf{d}_\mu * \mathcal{F}\phi), \phi\big)_{L^2(I)} = \sum_{n=0}^{\infty} (d_\mu)_n c_n^2 \widetilde{\phi}_n^2 > 0.$$

Combined with $(\gamma\phi, \phi)_{L^2(I)} \geq \gamma_{min}(\phi, \phi)_{L^2(I)} \geq 0$, this results in a contradiction with (3.25). ■

Below we derive properties of eigenvalues $\mu < 0$, i.e., eigenvalues for which corresponding unstable stationary solutions exist. Due to Theorem 3.3 these exist only if $\gamma_{min} < 0$ holds.

Lemma 3.4. *Let $\mu < 0$ be such that $J_\mu\phi = 0$ for a $\phi \neq 0$; i.e., (3.21) holds. Then $\mu \in [\mu^*, 0)$ holds, where $\mu^* < 0$ is the unique solution of*

$$\lambda^*(\mu^*) = 0, \qquad \lambda^*(\mu) = \sqrt{-\mu}\tanh\big(\sqrt{-\mu}\, D\big) + \gamma_{min}. \tag{3.26}$$

Proof. Due to $\gamma_{min} < 0$ relation (3.26) has a unique solution. Through $J_\mu\phi = 0$ we arrive at

$$\big(\mathcal{F}^{-1}(\mathbf{d}_\mu * \mathcal{F}\phi), \phi\big)_{L^2(I)} + (\gamma\phi, \phi)_{L^2(I)} = 0. \tag{3.27}$$

For $\mu < \mu^*$ and arbitrary $\phi \neq 0$ we have

$$\begin{aligned}\big(\mathcal{F}^{-1}(\mathbf{d}_\mu * \mathcal{F}\phi), \phi\big)_{L^2(I)} &\geq (d_\mu)_0 \|\phi\|^2_{L^2(I)} = \sqrt{-\mu}\tanh\big(\sqrt{-\mu}\,D\big)\|\phi\|^2_{L^2(I)} \\ &> \sqrt{-\mu^*}\tanh\big(\sqrt{-\mu^*}\,D\big)\|\phi\|^2_{L^2(I)} \\ &= \gamma_{min}\|\phi\|^2_{L^2(I)} \geq -(\gamma\phi, \phi)_{L^2(I)}.\end{aligned}$$

Hence (3.27) cannot hold for $\mu < \mu^*$. ∎

Lemma 3.4 implies that negative eigenvalues must lie in the finite interval $\mu \in [\mu^*, 0)$. The result in the following theorem characterizes the number of negative eigenvalues in this interval.

Theorem 3.5. *Let $\mu^* \leq \mu_1 < \mu_2 < \cdots < \mu_s < 0$ be all μ_i such that $J_{\mu_i}\phi_i = 0$ for some $\phi_i \neq 0$. Assume that for all i the null space of J_{μ_i} is 1D. Let Σ_J be the number of strictly negative eigenvalues (the "index") of $J_0 =: J$ and assume that for each of these eigenvalues the corresponding eigenspace is 1D. Then $s = \Sigma_J$ holds.*

Proof. Let $\lambda(\mu), v(\mu)$ be an eigenpair of J_μ, i.e., $J_\mu v(\mu) = \lambda(\mu)v(\mu)$, with $\|v(\mu)\|_{L^2(I)} = 1$. This implies that

$$\lambda(\mu) = \big(\lambda(\mu)v(\mu), v(\mu)\big)_{L^2(I)} = \big(J_\mu v(\mu), v(\mu)\big)_{L^2(I)}.$$

Differentiation w.r.t. μ is denoted by a prime. Using the symmetry of J_μ we obtain

$$\begin{aligned}\lambda'(\mu) &= \big(J'_\mu v(\mu), v(\mu)\big)_{L^2(I)} + 2\big(J_\mu v(\mu), v'(\mu)\big)_{L^2(I)} \\ &= \big(J'_\mu v(\mu), v(\mu)\big)_{L^2(I)} + 2\lambda(\mu)\big(v(\mu), v'(\mu)\big)_{L^2(I)} \\ &= \big(J'_\mu v(\mu), v(\mu)\big)_{L^2(I)}.\end{aligned}$$

The last equality follows from differentiation of $\|v(\mu)\|^2_{L^2(I)} = 1$ w.r.t. μ. The linear operator J'_μ is given by $J'_\mu\phi = \mathcal{F}^{-1}\big(\hat{\mathbf{d}}_\mu * \mathcal{F}\phi\big)$, with

$$(\hat{d}_\mu)_n = \frac{d}{d\mu}(d_\mu)_n = \frac{d}{d\mu}\Big[\sqrt{(n\pi)^2 - \mu}\tanh(\sqrt{(n\pi)^2 - \mu}\,D)\Big].$$

An elementary computation yields that $-c_0 \leq (\hat{d}_\mu)_n < 0$ holds, where c_0 is a constant independent of n and μ. Thus $-c_0 \leq \big(J'_\mu v(\mu), v(\mu)\big)_{L^2(I)} = \lambda'(\mu) \leq 0$ holds for all $\mu \leq 0$. This means that $\lambda(\mu)$ is a decreasing function with a bounded derivative. For $J_{\mu_i}\phi_i = 0$, $\phi_i \neq 0$ (with a 1D null space), this implies a unique eigenvalue curve $\lambda(\mu)$ which passes through μ_i, i.e., $\lambda(\mu_i) = 0$. Due to the monotonicity of $\lambda(\mu)$ this curve must intersect the negative y-axis, resulting in a corresponding negative eigenvalue of J_0. This implies $s \leq \Sigma_J$. Conversely, let ξ be a negative eigenvalue of J_0. Then there is a unique eigenvalue curve $\lambda(\mu)$ with $\lambda(0) = \xi$. Due to Lemma 3.4 this curve must intersect the μ-axis in the interval $[\mu^*, 0)$. Hence, a unique corresponding $\mu_i \in [\mu^*, 0)$ exists such that $\lambda(\mu_i) = 0$. This implies $\Sigma_J \leq s$. ∎

From Theorem 3.5 it follows that steady-state solutions with $\gamma_{min} < 0$ and $\Sigma_J > 0$ are always unstable. The necessary condition for instability $\gamma_{min} < 0$ is satisfied only if the fluid-heater temperature falls (at least locally) within the transition regime; this implies a fundamental relation between transition boiling and instability.

The results of Theorem 3.5 induce a method for computing the number of negative eigenvalues $\mu \in [\mu^*, 0)$. This number equals the number of negative eigenvalues of the operator $J = J_0$. For the general case of a smooth but nonconstant γ the eigenvalues of J_μ cannot be determined analytically. Thus we apply a discretization method (section 3.2.3) for numerical approximation. Using a sufficiently high resolution, this allows us to determine the correct number of negative eigenvalues (i.e., the same number as for the continuous problem). Moreover, an accurate approximation of the negative continuous eigenvalues can be computed. The resolution needed is not very high due to the fact that the eigenfunctions corresponding to the negative eigenvalues are dominated by low-frequency modes. This is explained in the following remark.

Remark 3. Consider ϕ with $\|\phi\|_{L^2(I)} = 1$ and $\xi < 0$ such that $J\phi = \xi\phi$. Represent ϕ in the cosine basis as $\phi(x) = c_0 + \sum_{n=1}^{\infty} c_n \sqrt{2}\cos(n\pi x)$ (with $\|\phi\|^2_{L^2(I)} = \sum_{n=0}^{\infty} c_n^2 = 1$). Then

$$\left(\mathcal{F}^{-1}(\mathbf{d}_0 * \mathcal{F}\phi), \phi\right)_{L^2(I)} + (\gamma\phi, \phi)_{L^2(I)} = \xi(\phi, \phi)_{L^2(I)} < 0$$

holds, and

$$\begin{aligned}\sum_{n=0}^{\infty} c_n^2 n\pi \tanh(n\pi) &= \left(\mathcal{F}^{-1}(\mathbf{d}_0 * \mathcal{F}\phi), \phi\right)_{L^2(I)} \\ &= -(\gamma\phi, \phi)_{L^2(I)} \le \|\gamma\|_{L^\infty(I)} \|\phi\|^2_{L^2(I)} = \|\gamma\|_{L^\infty(I)}.\end{aligned}$$

Thus c_n must become "smaller" for "larger" n, meaning that, in this sense, the eigenfunction ϕ is dominated by low-frequency modes.

3.2.3. Discretization method. As explained above, the problem of finding eigenvalues of (3.12)–(3.14) has been transformed to the problem of finding μ such that (3.21) has a nontrivial solution. If for some μ an eigenvalue $\lambda(\mu)$ of J_μ equals zero, this μ is an eigenvalue of (3.12)–(3.14). For a general smooth function γ the eigenvalue curves $\lambda(\mu)$ of the operator J_μ cannot be determined analytically. We introduce a discretization method that is used to discretize J_μ and thus determine the eigenvalue curves approximately. We use a Fourier-collocation method [24]: determine $\phi_F(x) := \sum_{n=0}^{N} \widetilde{\phi}_n \cos(n\pi x)$ such that

$$\sum_{n=0}^{N} \{\sqrt{\alpha_{n,\mu}} \tanh(\sqrt{\alpha_{n,\mu}} D) + \gamma(x_k)\} \widetilde{\phi}_n \cos(n\pi x_k) = 0 \quad \text{for all } 0 \le k \le N \tag{3.28}$$

holds, with $x_k = k/N$, $k = 0, \ldots, N$, the collocation points. Note that this is a discrete version of the continuous problem in (3.20). The $N+1$ equations (3.28) for the $N+1$ unknowns $\widetilde{\phi}_n$ can be represented in a compact matrix-vector formulation. To this end we introduce some notation. Let $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_N)^T$ be the vector of nodal values $\phi_n := \phi_F(x_n)$. The latter relate to the truncated Fourier spectrum $\widetilde{\boldsymbol{\phi}} = (\widetilde{\phi}_0, \ldots, \widetilde{\phi}_N)^T$ via

$$\boldsymbol{\phi} = \boldsymbol{V}\,\widetilde{\boldsymbol{\phi}}, \qquad \widetilde{\boldsymbol{\phi}} = \boldsymbol{V}^{-1}\boldsymbol{\phi}, \tag{3.29}$$

with

$$\boldsymbol{V} := \begin{bmatrix} 1 & \cos(\pi x_0) & \ldots & \cos(N\pi x_0) \\ \vdots & \vdots & & \vdots \\ 1 & \cos(\pi x_N) & \ldots & \cos(N\pi x_N) \end{bmatrix} = \boldsymbol{V}^T. \tag{3.30}$$

The relation

$$\boldsymbol{V}^{-1} = \frac{2}{N}\,\boldsymbol{D}\,\boldsymbol{V}\,\boldsymbol{D}, \qquad \boldsymbol{D} = \operatorname{diag}\left(\frac{1}{2}, 1, \ldots, 1, \frac{1}{2}\right) \tag{3.31}$$

holds; i.e., the matrix $\sqrt{\frac{2}{N}}\boldsymbol{V}\boldsymbol{D}$ is orthogonal. Define

$$\boldsymbol{K}_\mu = \boldsymbol{V}\boldsymbol{K}_{S,\mu}\boldsymbol{V}^{-1}, \quad \boldsymbol{K}_{S,\mu} = \operatorname{diag}\big(\sqrt{\alpha_{n,\mu}}D\tanh(\sqrt{\alpha_{n,\mu}}D)\big)_{0\le n\le N}, \tag{3.32}$$

$$\boldsymbol{Q} = \operatorname{diag}\big(D\gamma(x_n)\big)_{0\le n\le N}. \tag{3.33}$$

Note that $\boldsymbol{K}_{S,\mu}$ and $\boldsymbol{Q}$ are both diagonal matrices. The discrete problem (3.28) can be formulated in matrix-vector form as

$$\boldsymbol{J}_\mu\boldsymbol{\phi} = \boldsymbol{0}, \qquad \boldsymbol{J}_\mu := \boldsymbol{K}_\mu + \boldsymbol{Q}. \tag{3.34}$$

In spectral form this becomes

$$\boldsymbol{J}_{S,\mu}\widetilde{\boldsymbol{\phi}} = \boldsymbol{0}, \quad \boldsymbol{J}_{S,\mu} := \boldsymbol{V}^{-1}\boldsymbol{J}_\mu\boldsymbol{V} = \boldsymbol{K}_{S,\mu} + \boldsymbol{Q}_S, \quad \boldsymbol{Q}_S := \boldsymbol{V}^{-1}\boldsymbol{Q}\boldsymbol{V}. \tag{3.35}$$

The eigenvalues μ and eigenfunctions ϕ_F are approximated by those $\mu \in \mathbb{R}$ and $\boldsymbol{\phi} \in \mathbb{R}^{N+1}$ for which $\boldsymbol{\phi}$ is a nontrivial null-vector of $\boldsymbol{J}_\mu$.

Numerical tests for the case study in section 4 revealed that in these (approximate) eigenfunctions, for sufficiently high N, the Fourier coefficients decay exponentially with increasing wave number n. This is consistent with Remark 3 in that the low-frequency modes are indeed dominant. In all our experiments we use a resolution with $N = 128$. This resolution allows a correct determination of the number of negative eigenvalues μ as well as an accurate approximation of their numerical value.

3.3. Steady-state solutions. A detailed analysis of the steady-state behavior of the pool-boiling problem is given in [1]. The approach is in essence similar to that adopted above. The nonlinear 2D steady-state problem (3.7) is reduced to a 1D boundary model via the method of separation of variables. This 1D model is solved (approximately) through numerical approximation with a Fourier-collocation discretization method. The issues relevant in the present context are summarized below.

Application of separation of variables to (3.7) yields a (formal) representation of the solution of the Laplace equation and the linear Neumann boundary conditions on $\Gamma \setminus \Gamma_F$. This results in

$$T_\infty(x,y) = \sum_{n=0}^{\infty} \widetilde{T}_n \frac{\cosh(n\pi y)}{\cosh(n\pi D)}\cos(n\pi x) + \frac{D-y}{\Lambda}, \tag{3.36}$$

which can easily be checked by substitution. The coefficients $\widetilde{T}_n$ form the spectrum of the Fourier cosine expansion

$$T_{F,\infty}(x) := T_\infty(x,D) = \sum_{n=0}^{\infty} \widetilde{T}_n \cos(n\pi x) \tag{3.37}$$

of the temperature profile at the fluid-heater interface Γ_F. These coefficients are determined by the nonlinear Neumann boundary condition on Γ_F. Substitution of (3.36) into the nonlinear boundary condition on Γ_F leads to

$$\sum_{n=0}^{\infty} n\pi \tanh(n\pi D)\widetilde{T}_n \cos(n\pi x) + \eta(T_{F,\infty}(x))T_{F,\infty}(x) - \frac{1}{\Lambda} = 0, \tag{3.38}$$

for all $x \in [0,1]$, where $\eta(T_F) = \frac{\Pi_1}{\Lambda}\frac{q_F(T_F)}{T_F}$ is the scaled heat-transfer coefficient. The nonlinear equation (3.38) is the characteristic equation that determines the particular properties of the steady-state solutions of (3.7). Note the resemblance to relation (3.20).

The reduced steady-state problem (3.38) admits trivial and nontrivial solutions. Trivial solutions are homogeneous interface temperatures, for which $T_{F,\infty}(x) = \widetilde{T}_0$ and $\widetilde{T}_n = 0$ for $n \geq 1$ holds. Then the nonlinear condition (3.38) simplifies to

$$q_F(\widetilde{T}_0) = \Pi_1^{-1}. \tag{3.39}$$

Thus homogeneous solutions coincide with intersection(s) between the heat-flux function q_F and the normalized heat supply given by Π_1^{-1} (Figure 1(b)). Nontrivial, i.e., heterogeneous, solutions $T_{F,\infty}(x)$ that satisfy (3.38) cannot be determined analytically. However, certain properties of such solutions (if they exist) can be derived. One important property, proved in [1], is that such solutions always occur *as conjugate pairs*

$$T_{F,\infty}(x) \quad \text{and} \quad T^*_{F,\infty}(x) := T_{F,\infty}(x + 1/k) \quad \text{for a} \quad k \in \mathbb{N}. \tag{3.40}$$

This means that if $T_{F,\infty}(x)$ is a solution, then (for a certain $k \in \mathbb{N}$) the shifted function $T_{F,\infty}(x + 1/k)$ is a solution, too. This implies nonuniqueness of heterogeneous solutions. Homogeneous steady-state solutions can easily be determined by a standard root-finding method applied to (3.39). Heterogeneous steady-state solutions are computed (approximately) by using a Fourier-collocation discretization method, as described in section 3.2.3, applied to (3.38). This results in a nonlinear system of equations of the form

$$\boldsymbol{\mathcal{G}}(\boldsymbol{T}) := \big(\boldsymbol{K} + \boldsymbol{M}(\boldsymbol{T})\big)\boldsymbol{T} - \boldsymbol{G} = \boldsymbol{0}, \tag{3.41}$$

$\boldsymbol{T} = (T_0, \ldots, T_N)^T$, $T_n := T_{F,\infty}(x_n)$, $\boldsymbol{K} = \boldsymbol{K}_0$, as in (3.32) and

$$\boldsymbol{M} = \operatorname{diag}\big(\eta_n\big)_{0\leq n\leq N}, \quad \eta_n := \eta(T_{F,\infty}(x_n)), \quad \boldsymbol{G} = (1/\Lambda, \ldots, 1/\Lambda)^T. \tag{3.42}$$

The discrete system (3.41) defines a nonlinear set of equations that is solved by a continuation procedure. To this end we introduce a parameterized heat-flux function

$$q_F(T_F; P) := C_D\big(F_1 - PF_2H(C_DT_F - 1)\big)T_F, \quad 0 \leq P \leq 1 \tag{3.43}$$

(cf. (A.1)). In this modified heat-flux function the degree of nonlinearity is controlled through the nonlinearity parameter P. For $P = 0$ function (3.43) reduces to a linear form; for $P = 1$ the physical heat-flux $q_F(T_F)$ is recovered. This P-dependence of q_F induces a P-dependence

of the matrix $\boldsymbol{M}$ via the function η. This is expressed by the notation $\boldsymbol{M}_P$. The continuation procedure is applied to the function

$$P \to \boldsymbol{\mathcal{G}}(\boldsymbol{T}, P) := \big(\boldsymbol{K} + \boldsymbol{M}_P(\boldsymbol{T})\big)\boldsymbol{T} - \boldsymbol{G} = \boldsymbol{0}. \tag{3.44}$$

For each $P \in [0,1]$ the set of *homogeneous* solutions of this system can be easily computed. Starting on a branch of homogeneous solutions, the continuation algorithm determines (pitchfork) bifurcations at which the conjugate solution pairs (3.40) branch off from the homogeneous branch. An extensive treatment of these bifurcation results is given in [1]. In section 4.1 a bifurcation diagram for a representative case study is given.

4. Unsteady boiling problem: An illustrative case study. In this section the concepts introduced above are demonstrated by means of a representative case study. Unless indicated otherwise, the fixed parameter set $\Lambda = 0.2$, $D = 0.2$, $\Pi_2 = 4$, $\Pi_1 = 2$, and $\Pi_3 = 0.37$ (corresponding with $W = 1$; see appendix) is used. The steady-state solutions and corresponding *linear* stability properties are treated in section 4.1. The *nonlinear* (long-term) evolution of perturbed unstable steady-state solutions is investigated in section 4.2 via numerical simulation of (2.1). These results give a numerical validation of the linear stability analysis (section 4.1) and provide insight into the nonlinear stability behavior (section 4.2).

4.1. Steady-state solutions and their linear stability properties.

4.1.1. Homogeneous steady-state solutions. Homogeneous solutions are determined through relation (3.39) and coincide with the intersections between the heat-flux function q_F (solid line) and the normalized heat supply given by Π_1^{-1} (dashed line) in Figure 1(b). Three nondegenerate situations can occur:

Regime $\Pi_1 > \Pi_2$. Relation (3.39) admits one steady-state solution $T_{F,\infty}$ in the nucleate-boiling regime (Figure 2(a)). The local positive slope of the boiling curve ($\dot{q}_F := dq_F/dT > 0$) implies $\gamma > 0$, and thus by Theorem 3.2 we have stability of $T_{F,\infty}$.

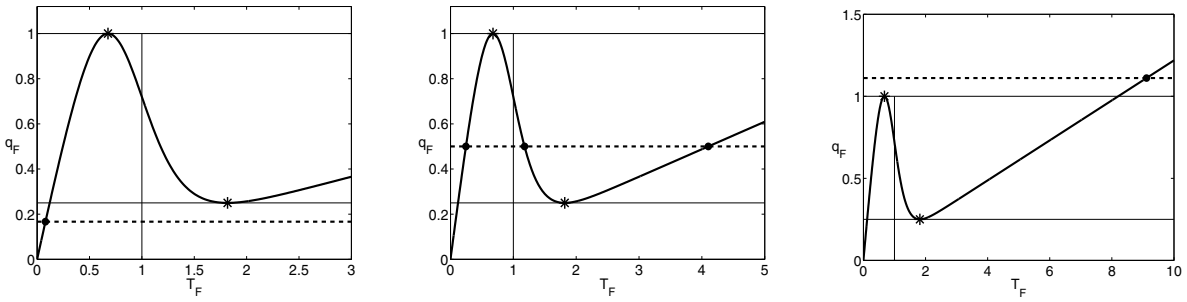
Regime $1 < \Pi_1 < \Pi_2$. Relation (3.39) yields three steady-state solutions (Figure 2(b)): $T_{F,\infty}^{(1)}$ (nucleate boiling); $T_{F,\infty}^{(2)}$ (transition boiling); $T_{F,\infty}^{(3)}$ (film boiling). For $T_{F,\infty}^{(1,3)}$ we have $\gamma > 0$ (due to $\dot{q}_F > 0$), and thus by Theorem 3.2 we have stability of these steady-state solutions. For $T_{F,\infty}^{(2)}$ we have $\dot{q}_F < 0$ and thus $\gamma < 0$. From Theorem 3.2 we conclude that $T_{F,\infty}^{(2)}$ is unstable.

Regime $\Pi_1 < 1$. Relation (3.39) has one steady-state solution $T_{F,\infty}$ in the film-boiling regime (Figure 2(c)). From $\gamma > 0$ and Theorem 3.2 we conclude that this solution is stable.

Cases $\Pi_1 = \Pi_2$ and $\Pi_1 = 1$ are the degenerate cases through which the system switches between one and three homogeneous solutions.

4.1.2. Heterogeneous steady-state solutions.

Steady-state behavior. The study of [1] strongly suggests that heterogeneous solution pairs (3.40) emerge only from pitchfork bifurcations that occur on branches of homogeneous solutions. For *homogeneous* solutions, $\boldsymbol{T} = T_F = \text{constant}$, the Jacobian w.r.t. T of $\boldsymbol{\mathcal{G}}(\boldsymbol{T}; P)$ is



(a) $\Pi_1 > \Pi_2$ (b) $1 < \Pi_1 < \Pi_2$ (c) $\Pi_1 < 1$

Figure 2. *Homogeneous steady-state solution(s) as a function of the system parameters* Π_1 *and* Π_2*. The solutions (dots) coincide with the intersections between the heat-flux function* q_F *(solid lines) and normalized heat supply* Π_1^{-1} *(dashed lines).*

given by

$$\frac{\partial \boldsymbol{\mathcal{G}}(\boldsymbol{T};P)}{\partial \boldsymbol{T}} =: \hat{\boldsymbol{J}}_P(\boldsymbol{T}) = \boldsymbol{V}\boldsymbol{\Lambda}_P\boldsymbol{V}^{-1}, \qquad \boldsymbol{\Lambda}_P = \operatorname{diag}\big(n\pi\tanh(n\pi D) + \gamma_P(T_F)\big)_{0\le n\le N}, \tag{4.1}$$

with $\gamma_P(Z) = \frac{\Pi_1}{\Lambda}\frac{\partial q_F(Z;P)}{\partial Z}$. Note that for $P = 1$ (i.e., $q_F(Z;1) = q_F(Z)$) *this Jacobian is equal to the matrix* $\boldsymbol{J}_0$ in (3.34): $\hat{\boldsymbol{J}}_1 = \boldsymbol{J}_0$. The origin of this identity lies in the fact that for $\mu = 0$ the *linear* eigenvalue problem (3.12)–(3.14) (which has a corresponding discrete boundary operator $\boldsymbol{J}_0$ as in (3.34)) is the same as the *linearization* of the stationary problem in (3.7) (which has a corresponding discrete boundary operator $\hat{\boldsymbol{J}}_1$ as in (4.1)).

The eigenvalues and corresponding eigenvectors of $\hat{\boldsymbol{J}}_P$ are given by

$$\lambda_n = n\pi\tanh(n\pi D) + \gamma_P(T_F), \quad \boldsymbol{v}_n = (\cos(n\pi x_0),\dots,\cos(n\pi x_N))^T, \quad 0 \le n \le N. \tag{4.2}$$

The eigenvector $\boldsymbol{v}_n$ coincides with the nth Fourier mode. The Jacobian is singular if one or more of its eigenvalues λ_n vanish. Because $n\pi\tanh(n\pi D) \ge 0$ for all $n \ge 0$, this can happen only if $\gamma(T_F) \le 0$. Thus a bifurcation on a homogeneous solution branch can occur only for those T_F for which the boiling curve has a negative slope. Only the intersection $T_F^{(2)}$ satisfies this criterion. This explains why bifurcations are restricted to the $T_F^{(2)}$-branch in the bifurcation diagram (Figure 3). This implies that bifurcations—and thus multiple (heterogeneous) solutions—can occur only for surface temperature values in the transition range of the boiling curve. Figure 3 shows the bifurcation diagram as a function of the nonlinearity parameter P [1]. The heavy curves are the solution branches corresponding to the homogeneous solutions $T_{F,\infty}^{(1,2,3)}$. The lower (nearly horizontal) branch coincides with the intersection $T_{F,\infty}^{(1)}$ that exists for all $0 \le P \le 1$; the upper branch, with turning point at P_B (here $P_B \approx 0.926$), coincides with the two intersections $T_{F,\infty}^{(2,3)}$ that exist only in the interval $P_B \le P \le 1$. The lower and upper legs of this upper branch (connected at the turning point) correspond to $T_{F,\infty}^{(2)}$ and $T_{F,\infty}^{(3)}$, respectively. The solid curves are the heterogeneous solution branches that originate from pitchfork bifurcations (dots) on the $T_{F,\infty}^{(2)}$-branch and from left to right correspond with the single vanishing eigenvalue $\lambda_n = 0$ for $n = 1, 2, 3$. The corresponding

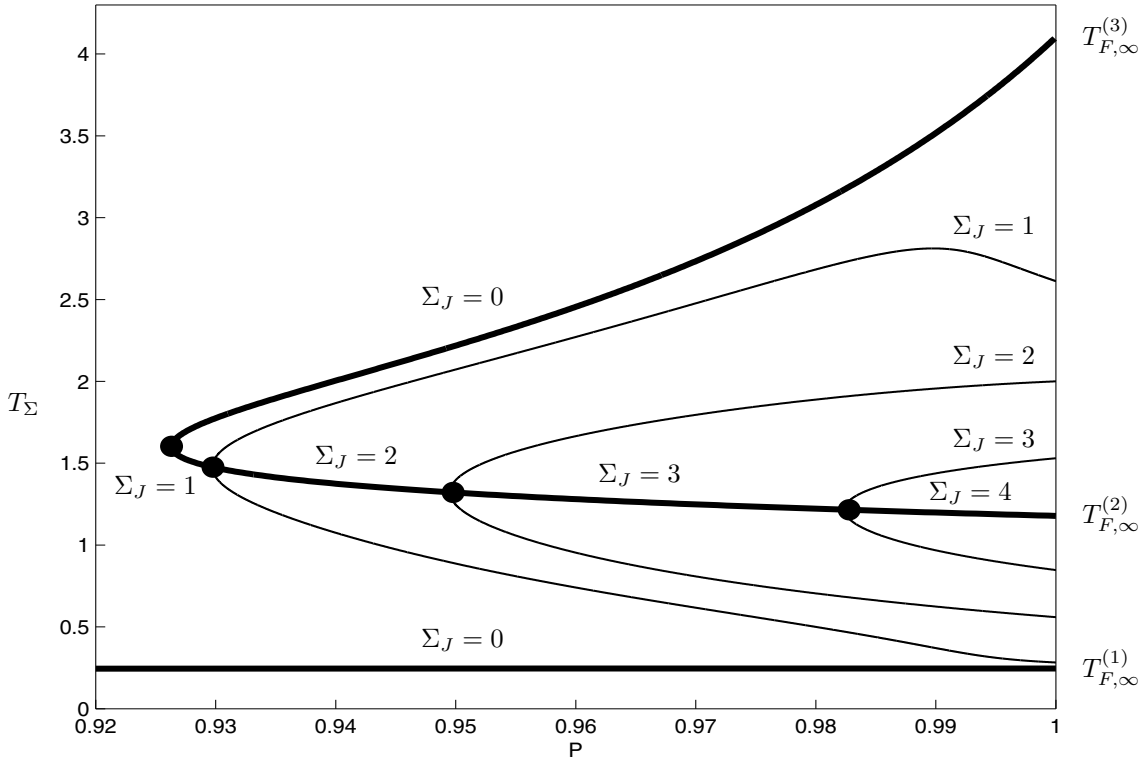
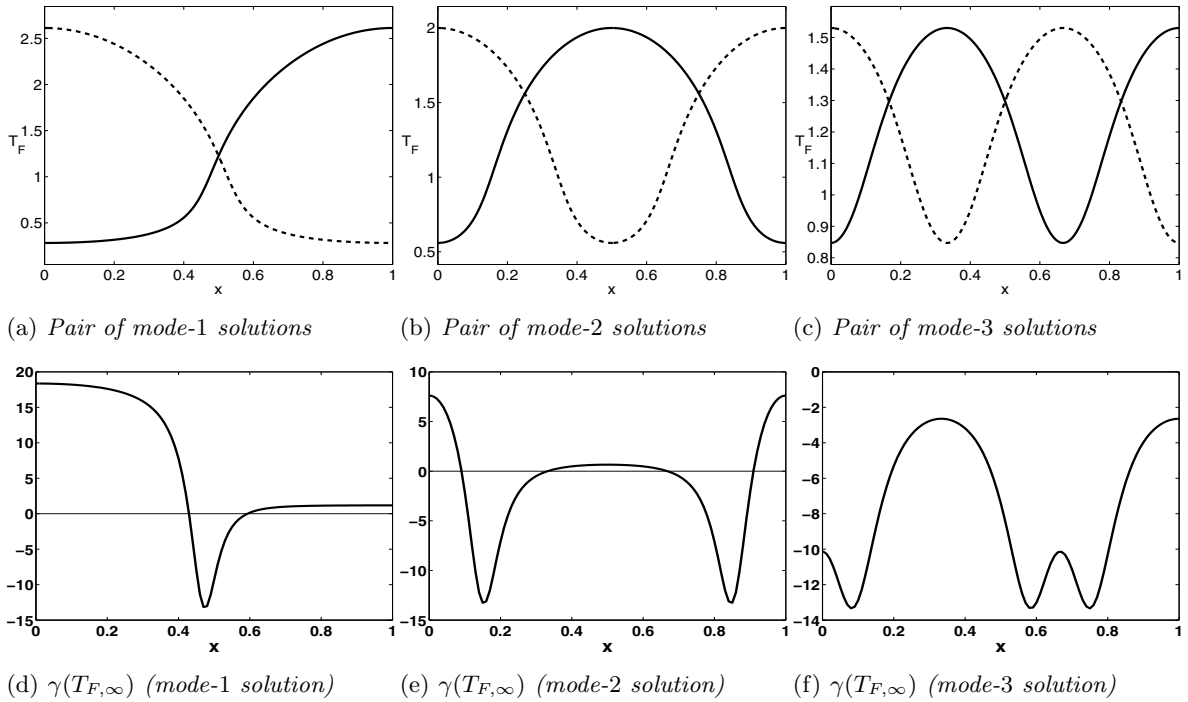


Figure 3. *Bifurcation diagram for the nonlinearity parameter P. Heavy curves correspond to homogeneous solutions; normal curves correspond to heterogeneous solutions. Filled circles represent bifurcations. The leftmost bifurcation is the tangent bifurcation that leads to multiple homogeneous solutions; the bifurcations from which the heterogeneous branches emerge are pitchfork bifurcations. Included also is the corresponding index Σ_J (section 4.1.3).*

eigenvector $\boldsymbol{v}_n$ determines the form of the bifurcating heterogeneous solution and equals the nth Fourier mode. These heterogeneous solutions that originate from the bifurcation point corresponding to λ_n are called "mode-n" solutions. The lower and upper legs in a pitchfork bifurcation correspond to $T_{F,\infty}$ and its conjugate $T_{F,\infty}^*$, respectively; cf. (3.40).

The final states ($P = 1$) in the bifurcation diagram (Figure 3) correspond to the physically meaningful steady-state solutions to the boiling problem (2.1). Figure 4 (top row) shows the boundary profiles $T_{F,\infty}$ associated with the pairs of steady-state mode-n solutions, where solid and dashed lines indicate $T_{F,\infty}$ and its conjugate $T_{F,\infty}^*$, respectively. Figure 4 (bottom row) gives the profiles of the function $\gamma(x)$ from (3.10) corresponding to $T_{F,\infty}$.

4.1.3. Qualitative linear stability properties. Theorems 3.3 and 3.5 state that the qualitative stability (i.e., stable or unstable) of the heterogeneous mode-n solutions depends on γ_{min} and the index Σ_J. The boundary profiles (Figure 4) show that for all cases we have $\gamma < 0$—and hence transition boiling—on one or more sections of the boundary Γ_F (Figures 4(d)–(f)). This implies that $\gamma_{min} < 0$ and thus rules out unconditional stability (Theorem 3.3); stability properties depend on the quantity Σ_J as defined in Theorem 3.5. We generalize the definition of Σ_J by defining Σ_J as the P-dependent number of negative eigenvalues of the



(a) *Pair of mode-1 solutions* (b) *Pair of mode-2 solutions* (c) *Pair of mode-3 solutions*

(d) $\gamma(T_{F,\infty})$ *(mode-1 solution)* (e) $\gamma(T_{F,\infty})$ *(mode-2 solution)* (f) $\gamma(T_{F,\infty})$ *(mode-3 solution)*

Figure 4. *Heterogeneous steady-state (mode-n) solutions of the boiling problem. The top row gives the boundary profiles $T_{F,\infty}$ (solid) and $T^*_{F,\infty}$ (dashed) of the pairs of mode-n solutions. The bottom row gives coefficient γ corresponding with $T_{F,\infty}$; $\gamma < 0$ indicates regions of transition boiling.*

Jacobian $\hat{\boldsymbol{J}}_P$ from (4.1). The identity $\hat{\boldsymbol{J}}_1 = \boldsymbol{J}_0$ means that for $P = 1$ this generalized Σ_J coincides with the Σ_J as defined in Theorem 3.5 (provided that the resolution is high enough to determine the correct number of negative eigenvalues; cf. Remark 3). Thus the linear stability properties of the steady-state solutions follow directly from the bifurcation analysis visualized in Figure 3. On the homogeneous branches Σ_J can be easily computed using (4.2). Computational analysis reveals that on the heterogeneous solution branches no further bifurcations occur, apart from the pitchfork bifurcation points at the intersection with the homogeneous branch. Thus Σ_J remains constant along a heterogeneous branch, meaning that Σ_J for $P = 1$ (the relevant quantity for Theorem 3.5) equals Σ_J at the underlying bifurcation. The value for Σ_J at $P = 1$ is computed numerically. The values for Σ_J corresponding with each solution branch are indicated in Figure 3. For all heterogeneous solutions we have $\Sigma_J > 0$ at $P = 1$, and, thus, due to Theorem 3.5, *all heterogeneous solutions are unstable.* Moreover, Σ_J—and thereby the number of unstable modes (Theorem 3.5)—increases with each bifurcation. In this sense mode-n solutions become more unstable for higher n. The change of the value of Σ_J at the bifurcation points and its effect on the linear stability properties are closely related to the "exchange of stability principle"; cf. [23].

The linear stability analysis shows that the pool-boiling problem is linearly bistable. The homogeneous nucleate-boiling ($T_F^{(1)}$) and film-boiling ($T_F^{(3)}$) states are the only stable states; other steady states are always unstable. The bistability implies that the system has two basins of attraction, associated with the two stable states $T_F^{(1)}$ and $T_F^{(3)}$, divided by a separatrix

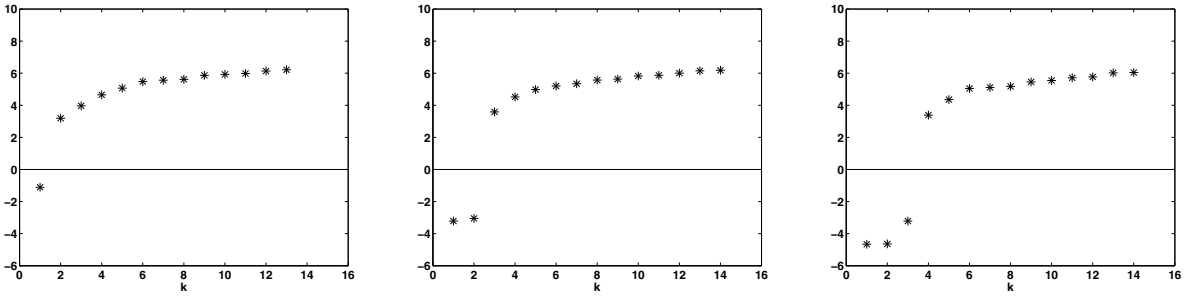
formed by the *stable* manifolds of the *unstable* solutions [19].

The bistability established above is entirely consistent with laboratory experiments on physical boiling systems [6, 25, 26]. The homogeneous nucleate-boiling and film-boiling states are the only hitherto known stable states of such systems and thus are the only steady-state solutions that may occur naturally. The bistability admits the boiling system (for specific perturbations) to transit from nucleate to film boiling and vice versa via intermediate—and highly unstable—heterogeneous states. The triggering mechanisms behind such transitions and the ensuing evolution of the boiling states remain ill understood to date, however. Consequently, prevention of, in particular, the transition from nucleate to film boiling, or "burnout" [16], remains *the* key challenge in industrial boiling processes (section 1). The bistability furthermore implies the absence of stable homogeneous boiling states outside the nucleate-boiling and film-boiling regimes. This poses formidable challenges for reproducible measurements of the boiling curve in the transition-boiling regime; namely, boiling conditions are assumed homogeneous along the entire boiling curve (section 2). Active stabilization of transition boiling during boiling-curve measurements offers a way for attainment of such conditions [27]. However, it appears that accomplishment of homogeneous transition boiling is not always possible, despite successful stabilization. This significantly complicates boiling-curve measurements, which implicitly assume a homogeneous state.

4.1.4. Quantitative linear stability properties. The *unstable* heterogeneous solutions develop into essentially unsteady states in the case of some nonzero initial perturbation $v_0(\boldsymbol{x})$. In practice, such perturbations are always present. Hence, the heterogeneous steady-state solutions cannot be sustained by the system and must undergo some evolution in time. This evolution depends largely on the unstable eigenmode(s) $\psi_k(\boldsymbol{x})$, i.e., for which the corresponding eigenvalue μ_k is negative, according to (3.17).

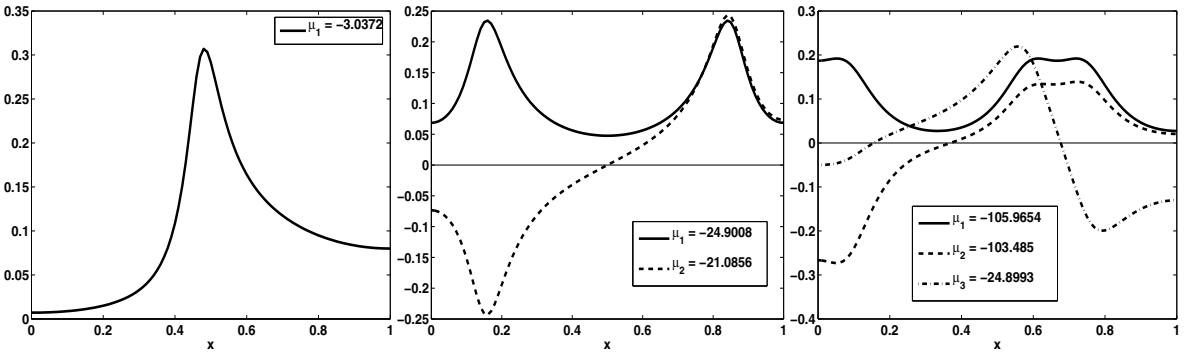
Figure 5 gives the sequence of eigenvalues $\mu_1 < \mu_2 < \cdots$ corresponding with the mode-n solutions ($n = 1, 2, 3$). The number of negative eigenvalues equals $n = \Sigma_J$, consistent with the values of the index Σ_J in Figure 3. Furthermore, the magnitude $|\mu_k|$ of the negative eigenvalues grows—implying higher growth rates of perturbations—with increasing n. Figure 6 shows the boundary profiles of the unstable eigenmodes $\psi_k(\boldsymbol{x})$, with corresponding μ_k-values as indicated, associated with the mode-n solutions. Figure 7 gives the three unstable modes of the mode-3 solution in the whole domain $\mathcal{D}$. Note that due to the maximum principle the extrema of ψ_k occur at the boundary of $\mathcal{D}$. The mode-1 solution has only one unstable eigenmode. This mode dominates the evolution of the instability, largely independent of the initial perturbation. The mode-2 and mode-3 solutions have multiple unstable eigenmodes with associated eigenvalues of comparable magnitude, and, consequently, the space spanned by these modes has dimension larger than one. Thus the evolution of the instability becomes essentially dependent upon the initial perturbation and allows a much richer spatial structure compared to the mode-1 case.

4.1.5. The role of heater properties in the stability behavior. The inherent instability of transition boiling greatly hampers reliable and reproducible laboratory experiments on boiling heat transfer under transition conditions. The heater properties are important design parameters for such transition experiments [14]. Our pool-boiling model enables examination



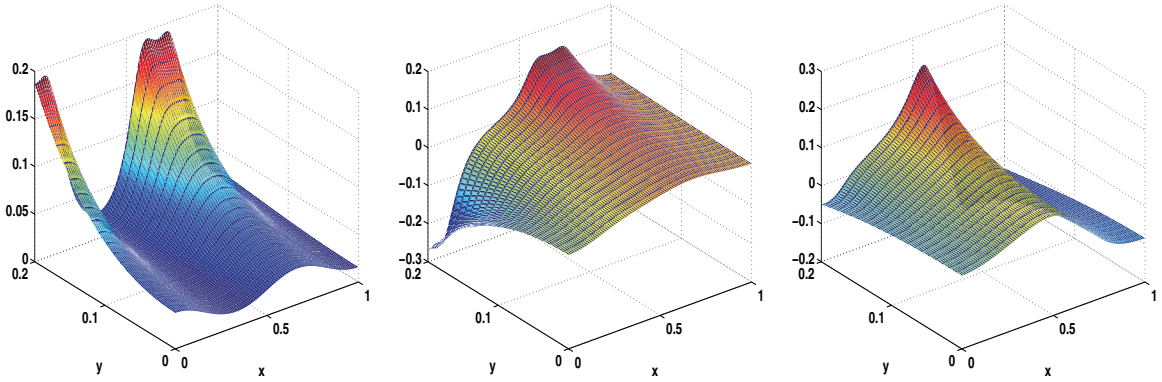
(a) *Mode-1 solutions* (b) *Mode-2 solutions* (c) *Mode-3 solutions*

Figure 5. *Sequence of eigenvalues $\mu_1 < \mu_2 < \cdots$ corresponding to each of the mode-n solutions. The eigenvalues are represented as $\beta_k = sign(\mu_k)\ln|\mu_k|$ so as to enhance legibility.*



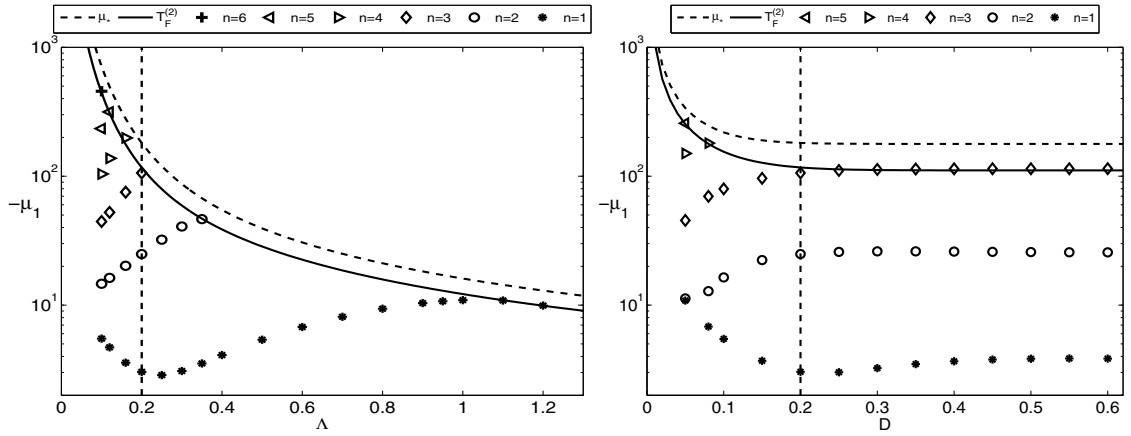
(a) *Mode-1 solutions* (b) *Mode-2 solutions* (c) *Mode-3 solutions*

Figure 6. *Unstable modes (eigenfunctions $\psi_k(\boldsymbol{x})$ with $\mu_k < 0$) restricted to the boundary Γ_F, corresponding to each of the mode-n solutions.*



(a) $\psi_1(\boldsymbol{x})$, $\mu_1 = -106.0$ (b) $\psi_2(\boldsymbol{x})$, $\mu_2 = -103.5$ (c) $\psi_3(\boldsymbol{x})$, $\mu_3 = -24.9$

Figure 7. *The three unstable modes of the mode-3 solution in the domain $\mathcal{D}$.*



(a) *Thermal conductivity* Λ (b) *Thickness* D

Figure 8. *Effect of varying heater properties* (Λ *and* D) *upon the stability properties. Shown is the eigenvalue* $\mu_1 < 0$ *of the most unstable eigenmode for the mode-*n *solutions (symbols), the homogeneous transition solution* $T_F^{(2)}$*, and the lower bound* μ_* *according to* (3.26)*. The dashed vertical line indicates the parameter value used in the case study.*

of the role of the heater properties in the stability behavior of pool-boiling systems under representative conditions (i.e., heterogeneous boiling states on realistic heaters) and may thus be beneficial for the design of laboratory experiments. To this end a short exploratory study, as a prelude to future work, is given below.

In the present model the heater properties are described by the parameters Λ (thermal conductivity) and D (heater thickness). In the case study we investigate the changes in stability properties of the heterogeneous solutions induced by variation of Λ or D. The instability is quantified in terms of the eigenvalue $\mu_1 < 0$ of the most unstable eigenmode. Figure 8 gives μ_1 as a function of Λ (panel (a)) and D (panel (b)) for the mode-n heterogeneous steady-state solutions (symbols), the homogeneous transition solution $T_F^{(2)}$ (with $\mu_1 = z_1^-$ as in Theorem 3.2), and the lower bound μ_* according to (3.26). The dashed vertical line indicates the parameter value used in the case study. (The plots actually show $-\mu$ so as to facilitate the logarithmic scale.) Both graphs reveal that the number of mode-n solutions decreases with both increasing Λ and increasing D. Beyond $\Lambda \approx 1.2$ mode-n solutions have vanished altogether; beyond $D \approx 0.1$ the system settles for three mode-n solutions and effectively becomes independent of the heater thickness for $D \gtrapprox 0.5$. Thus increasing thermal conductivity and/or the heater thickness promotes homogeneity and uniqueness of boiling states. This is consistent with results in [1].

In Figure 8(a) one can observe the bifurcation of mode-n profiles from the $T_F^{(2)}$-profile with decreasing thermal conductivity Λ. (For D essentially the same happens. This is less apparent in Figure 8(b), though.) The instability of the mode-n solutions is stronger (in the sense that $|\mu_1|$ increases) for larger n, and $T_F^{(2)}$ appears to be the most unstable solution. Physically, this may be explained by the fact that the portion of the interface on which the temperature is in the transition regime (i.e., where $\gamma < 0$) grows with n and is maximal for the homogeneous state $T_F^{(2)}$.

The mode-n solutions for $n > 1$ become more unstable (larger $|\mu_1|$) for increasing Λ and/or D. For the homogeneous solution $T_F^{(2)}$, on the other hand, this lessens the instability. For the mode-1 solution $|\mu_1|$ exhibits a nonmonotonic dependence on Λ and on D. These observations reveal that the dependence of the stability behavior on changes in the heater properties is related to the kind of steady-state solution. For the mode-n ($n > 1$) solutions, increasing heater thickness D and/or thermal conductivity Λ amplifies instability, whereas for the homogeneous solution $T_F^{(2)}$ this dampens instability.

The above results strongly suggest that, despite significant quantitative variations in μ_1, the instability itself remains under all conditions and that, in consequence, the existence of stable heterogeneous solutions for specific heater properties is therefore highly unlikely. Laboratory experiments support this assertion (consult [14] for a survey). This consolidates the widely accepted observation that active stabilization, via, e.g., the methodology by [6], is essential for detailed experimental studies on homogeneous boiling states in the transition region and on any heterogeneous boiling state. Moreover, this confirms that the stability analysis of [28] is erroneous, as already pointed out by [15]. Haramura [28] studied the stability of homogeneous boiling states on the fluid-heater interface of a 3D heater with constant heating in the transition boiling regime (i.e., $\gamma < 0$) and derived critical conditions for which the homogeneous boiling state supposedly becomes unstable. These critical conditions are in contradiction to findings in previous studies (e.g., [14, 8]) as well as to those in the present study.

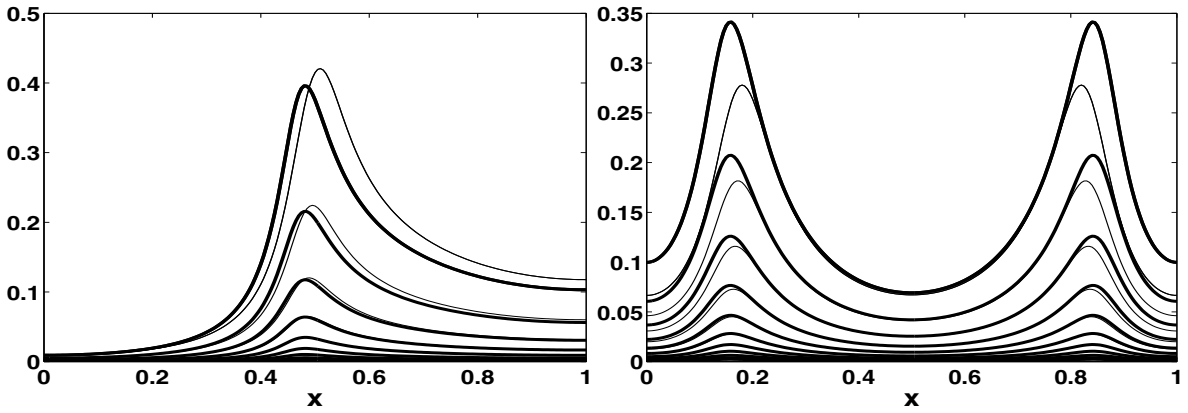
4.2. Evolution of perturbed unstable steady-state solutions. The *nonlinear* long-term evolution of perturbed unstable steady-state solutions, i.e., the *nonlinear* (in)stability behavior, can be determined via numerical simulation of (2.1). We used a spectral tau method based on Fourier (x) and Chebyshev (y) expansion of $T(\boldsymbol{x}, t)$ for spatial discretization in combination with a second-order Crank–Nicolson time-marching scheme [24]. The nonlinearity on the interface Γ_F has been dealt with by Picard iteration [29].

Numerical studies of (2.1) serve two purposes, namely, validation of the linear stability analysis (section 3.2) and gaining first insight into the nonlinear stability behavior of the pool-boiling system. These two topics are addressed in section 4.2.1 and section 4.2.2, respectively.

4.2.1. Validation of the linear stability analysis. The spectral scheme proposed above is used for the numerical simulation of the linearized heat-transfer model (3.8). Tests with various identical initial conditions $v_0(\boldsymbol{x})$ for each of the steady-state solutions T_∞ reveal that solutions $v(\boldsymbol{x}, t)$ obtained through the linear model (3.8) and the expansion (3.17) coincide within machine accuracy. This validates the eigenmode decomposition (3.17).

A second issue is a comparison of the evolution of the perturbation $v(\boldsymbol{x}, t)$ in the *non*linear model (2.1) and in the eigenmode decomposition (3.17). Tests reveal that in both cases the stability properties are qualitatively the same: heterogeneous steady-state solutions are unstable, and the homogeneous nucleate-boiling and film-boiling states are stable. This provides strong evidence that the pool-boiling system is *non*linearly bistable as well.

Quantitative validation of the linear stability analysis and establishment of a range of validity of the linear approximation follow from investigation of two representative heterogeneous cases, namely, the steady-state mode-1 and mode-2 solutions. As initial perturbation we take $v_0(\boldsymbol{x}) = \epsilon\psi_1(\boldsymbol{x})$, with $\epsilon = 0.01$ and ψ_1 the first eigenfunction (normalized) of (3.12)–(3.14).



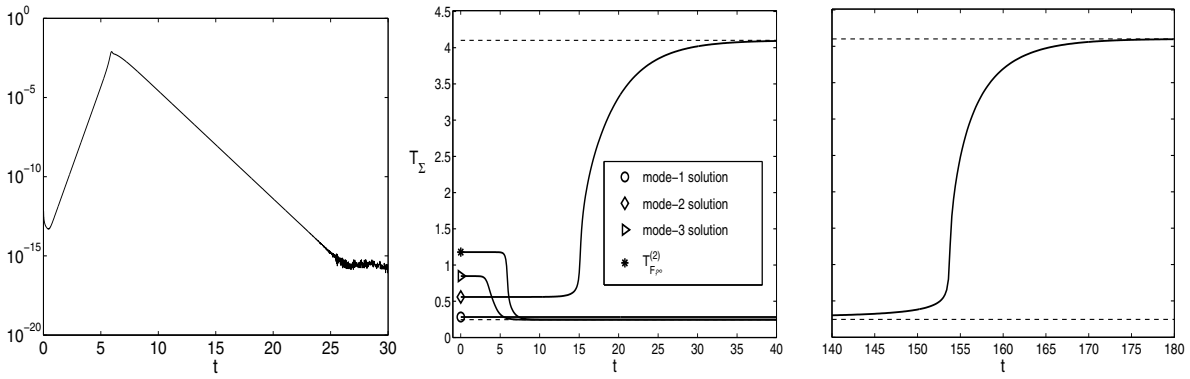
(a) *Mode-1 solution* (b) *Mode-2 solution*

Figure 9. *Solutions $v(\boldsymbol{x},t)$ of the linear approximation (3.11) (heavy lines) and of the nonlinear model (thin lines) for the mode-1 (panel (a)) and mode-2 (panel (b)) steady-state solutions. Initial perturbation is $v_0(\boldsymbol{x}) = 0.01\psi_1(\boldsymbol{x})$. The curves correspond to time steps $\Delta t = 5$ (mode-1) and $\Delta t = 0.5$ (mode-2).*

Figure 9 shows the profiles at the interface Γ_F at several time instances t corresponding with the solutions to the nonlinear model and the linearized problem. (For the nonlinear model we give the departure $v(\boldsymbol{x},t) = T(\boldsymbol{x},t) - T_\infty(\boldsymbol{x})$ from the initial state.) The more pronounced peaks correspond with more advanced time levels. The results show a good agreement between linear approximation and nonlinear evolutions for a significant time interval.

4.2.2. Nonlinear stability analysis. The present case study involves the following unstable steady-state solutions: the homogeneous solution $T^{(2)}_{F,\infty}$ in the transition-boiling regime and the three pairs of mode-n solutions. Perturbations are not imposed explicitly but are due to rounding errors in the machine representation of the initial condition. This implies a machine-dependent yet reproducible perturbation. These small perturbations are sufficient to trigger evolution of the unstable states. We use the functionals $T_\Sigma = \sum_n \widetilde{T}_n$ and $\mathcal{K}(k\Delta t) = \|\mathbb{T}^k - \mathbb{T}^{k-1}\|/\|\mathbb{T}^k\|$, where $\mathbb{T}^k$ is the matrix consisting of all coefficients in the discrete Fourier–Chebyshev series of the solution at $t = k\Delta$, as measures for quantifying the evolution. The mode-n solutions always occur as pairs (3.40). We consider only one solution of this pair (lower legs of the heterogeneous branches in Figure 3).

Figure 10(a) demonstrates the dynamical behavior of the system during transition from unstable to stable steady states via the measure of unsteadiness $\mathcal{K}(t)$ for the unstable homogeneous solution $T^{(2)}_{F,\infty}$. The progression clearly reveals that the evolution of the temperature field accelerates (i.e., $\mathcal{K}(t)$ grows continuously) up to a turning point at $t \approx 6$, where the situation reverses and a deceleration sets in that continues until a stable steady state is reached at $t \approx 25$. (The erratic evolution beyond $t \approx 25$ is due to fluctuations around the steady state at machine-accuracy level.) The mode-n solutions exhibit essentially similar behavior as that shown in Figure 10(a); differences are entirely quantitative in that turning points and attainment of stable steady states occur at different instances in time. Figure 10(b) gives the evolution from unstable to stable steady states in terms of the functional $T_\Sigma(t)$ (split into two frames.) The unstable homogeneous solution $T^{(2)}_{F,\infty}$ and the mode-3 solutions progress towards



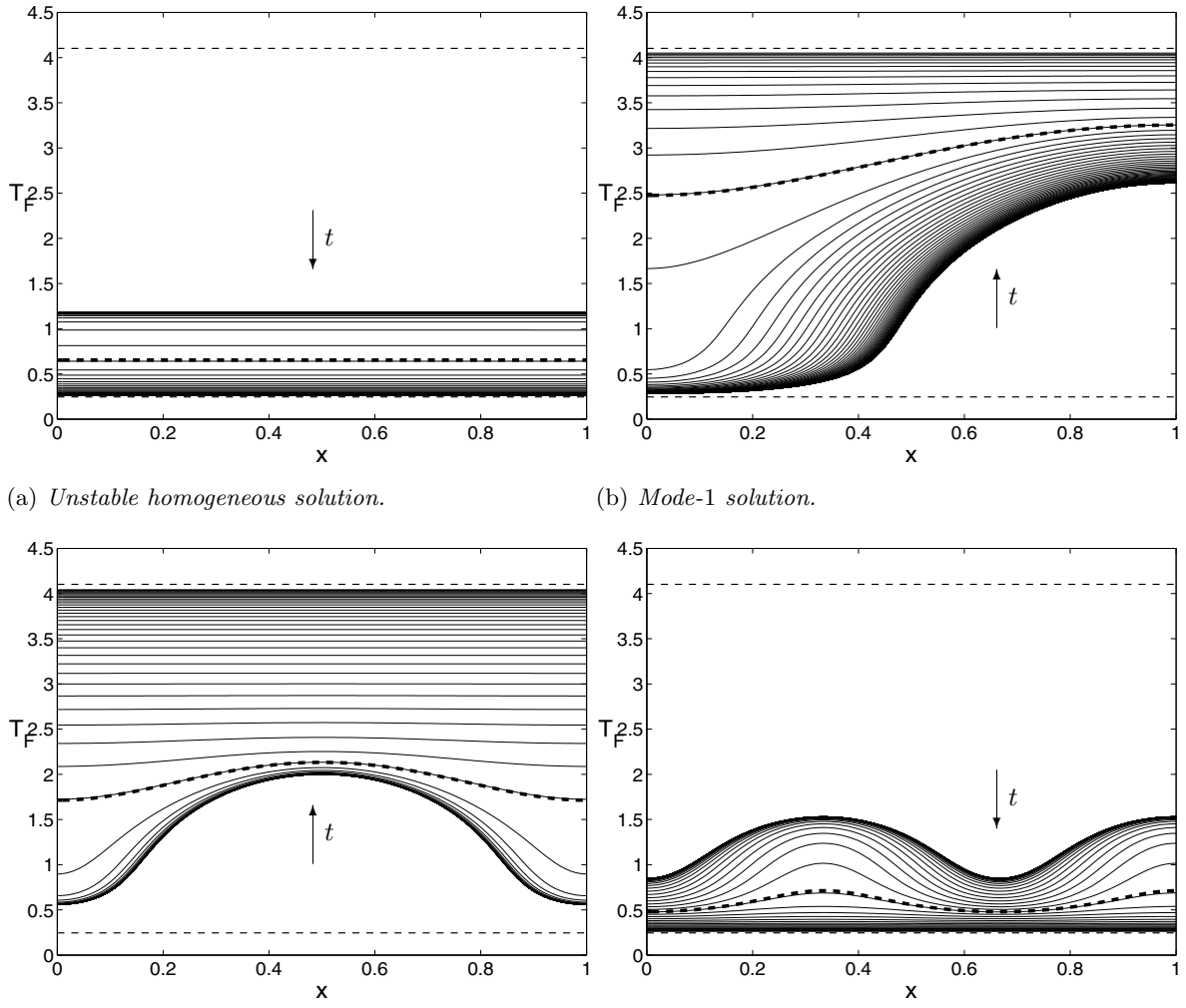
(a) $\mathcal{K}(t)$ *for* $T^{(2)}_{F,\infty}$. (b) *Progression towards one of the two stable steady states.*

Figure 10. *Evolution of perturbed unstable steady-state solutions towards a stable state. Panel* (a) *shows the initial acceleration and subsequent deceleration of the evolution with the measure of unsteadiness* $\mathcal{K}(t)$ *for the unstable homogeneous solution* $T^{(2)}_{F,\infty}$. *Panel* (b) *gives the progression of the unstable states (as indicated) towards one of the two stable steady states (dashed lines) in terms of the functional* $T_\Sigma(t)$. *(Split into two frames, the right frame concerns only the mode-*1 *solution.) Only the parent solution* T_F *of each conjugate pair* (3.40) *is included.*

the stable steady-state solution $T^{(1)}_{F,\infty}$ in the nucleate-boiling regime (lower dashed line); the mode-1 and mode-2 solutions progress towards the stable steady-state solution $T^{(3)}_{F,\infty}$ in the film-boiling regime (upper dashed line). The sharp transitions of the evolutions occur around the aforementioned turning points and reflect the fact that the changeover from unstable to stable states happens rather abruptly within a relatively narrow time window. Moreover, this changeover takes place earlier, suggesting stronger instability, with decreasing length scales of the heterogeneous features (higher n) of the mode-n solutions. This is in qualitative agreement with experimental observations [9].

Figure 11 gives the evolutions of the profiles of the interface temperature T_F for each of the unstable steady-state solutions in Figure 10(b) at equidistant time intervals, where the arrows indicate progression in time. (Time intervals are different for each case and are set proportional to the duration of the entire evolution.) The heavy dashed profiles correspond with the intermediate state at the respective turning points; the lower and upper dashed lines indicate the stable nucleate-boiling and film-boiling states, respectively. The evolution of the profiles nicely illustrates the progression towards either the nucleate-boiling state (panels (a) and (d)) or the film-boiling state (panels (b) and (c)). Note that for the unstable homogeneous case $T^{(2)}_{F,\infty}$ (panel (a)) the profile remains homogeneous throughout the evolution. Moreover, the expansion and subsequent condensation of the profiles before and after the turning point (heavy dashed lines) demonstrate the initial acceleration and the subsequent deceleration of the evolution. The acceleration phase of the mode-n solutions is characterized by rapid smoothing of the heterogeneous features (panels (b)–(d)); during the subsequent deceleration phase the (approximately) homogeneous intermediate state gradually tends to the ultimate stable state.

5. Conclusions. In this paper we consider a 2D nonlinear heat-transfer problem as a model for pool-boiling systems. The model problem involves only the temperature distribu-



(a) *Unstable homogeneous solution.*

(b) *Mode-*1 *solution.*

(c) *Mode-*2 *solution.*

(d) *Mode-*3 *solution.*

Figure 11. *Evolution of the perturbed unstable steady-state solutions (heavy lines) on the interface. Shown are the progressions of the boundary profiles (equidistant time intervals) towards one of the two stable steady states. The arrow indicates progression in time. The heavy dashed profiles indicate the intermediate state at the turning points; the lower and upper dashed lines indicate the stable nucleate-boiling and film-boiling states, respectively. Only one solution T_F of each conjugate pair* (3.40) *is included.*

tion within the heater and models the heat exchange with the boiling medium via a nonlinear boundary condition imposed at the fluid-heater interface. This results in a linear parabolic partial differential equation (heat equation) with a nonlinear Neumann boundary condition at the fluid-heater interface. Important information about the (qualitative) behavior of this dynamical system can be obtained from its steady-state solutions and the corresponding stability properties. The steady-state behavior has been studied in [1]. The main topic of the present study is the corresponding stability behavior. To this end a linear (short-term) and a nonlinear (long-term) stability analysis are performed.

In the linear stability analysis the linearized heat-transfer model is, by separation of space

and time, reduced to a nonlinear eigenvalue problem that depends only on the two spatial variables. Separation of the two spatial variables subsequently leads to a nonlinear spatially *one*-dimensional problem for the eigenvalues. Analysis of the latter problem yields generic stability properties for steady-state solutions. These are demonstrated and validated by numerical simulations in a representative case study. One of the main conclusions is that the (linearized as well as nonlinear) system is bistable: all steady-state solutions, except the homogeneous nucleate-boiling and film-boiling states, are inherently unstable. Perturbed unstable states always progress towards one of these two stable states. Our study furthermore strongly suggests that these stability properties are qualitatively independent of heater properties (thermal conductivity Λ and thickness D). Changes in heater properties affect the stability properties of the system only in a quantitative manner. Thus the present study rigorously demonstrates the bistability of pool-boiling systems, which is consistent with laboratory experiments [6] and other theoretical studies [10, 12, 14, 13].

The numerical simulations for the case study provide evidence that there is a strong analogy between the nonlinear heat-transfer problem and generic nonlinear parabolic evolution equations, which typically have a nonlinearity in the partial differential equation. This analogy suggests the fundamental property that the dynamical behavior of the system is dominated by a global attractor made up of steady-state solutions and corresponding heteroclinic connections (section 3.1). Both the linear and nonlinear stability behavior that we observe in the pool-boiling system indicate that this property holds. Although many stability results, based on both theoretical analysis and numerical experiments, are derived in this paper, a complete rigorous mathematical analysis of the dynamics of the pool-boiling model is not yet available.

Recent studies revealed that the steady-state behavior and the mathematical structure of 3D pool-boiling problems is essentially similar to that of the simplified 2D case considered here [2]. This means that the stability behavior found in the present work in principle extends to the 3D case. Moreover, the present analysis may form the basis for future research on active stabilization of unstable heterogeneous boiling states by extending the model with a temperature-control loop similar to that proposed in [14].

Appendix. Heat-flux function. The heat-flux function $q_F(Z;\Pi_2,\Pi_3)$ is given by

$$q_F(Z) = h(Z)Z, \tag{A.1}$$

with

$$h(Z) = C_D\left\{F_1 - F_2 H(C_D Z - 1)\right\}, \qquad H(\zeta) = \frac{1}{2}\left[\tanh\left(\frac{2\zeta}{W}\right) + 1\right]. \tag{A.2}$$

The function $H(\zeta)$ is a smoothed Heaviside function. The parameter $W > 0$ controls the width of the transient (from $H = 0$ to $H = 1$) around $\zeta = 0$ and is specified a priori. The value of W indirectly sets the physical parameter Π_3. The coefficient C_D rescales the argument Z such that the single deflection point of q_F coincides with $Z = 1$; i.e., $q_F''(1) = 0$. Its value is defined implicitly through

$$2\frac{dH}{d\zeta}(C_D - 1) + C_D\frac{d^2H}{d\zeta^2}(C_D - 1) = 0 \tag{A.3}$$

and thus depends only on W. It can be shown that q_F as in (A.1) possesses a local maximum and minimum at $Z_{max} < 1$ and $Z_{min} > 1$, respectively. Introduction of the scaling factors F_1 and F_2, which scale q_F such that the conditions

$$q_F'(Z_{max}) = 0, \quad q_F'(Z_{min}) = 0, \quad q_F(Z_{max}) = 1, \quad q_F(Z_{min}) = \Pi_2^{-1} \tag{A.4}$$

are fulfilled, then results in a heat-flux function that is consistent with the physical boiling curve. For given Π_2 and W the conditions (A.4) result in four nonlinear equations that can be solved for the four unknowns $(F_1, F_2, Z_{min}, Z_{max})$.

REFERENCES

[1] M. Speetjens, A. Reusken, and W. Marquardt, *Steady-state solutions in a nonlinear pool-boiling model*, Commun. Nonlinear Sci. Numer. Simul., 13 (2008), pp. 1475–1494.

[2] M. Speetjens, A. Reusken, and W. Marquardt, *Steady-state solutions in a three-dimensional nonlinear pool-boiling heat-transfer model*, Commun. Nonlinear Sci. Numer. Simul., 13 (2008), pp. 1518–1537.

[3] J. R. Thome, *Boiling*, in Handbook of Heat Transfer, A. Bejan and A. D. Krause, eds., Wiley & Sons, New York, 2003, pp. 635–717.

[4] I. Mudawar, *Assessment of high-heat-flux thermal management schemes*, IEEE Trans. Compon. Packag. Technol., 24 (2001), pp. 122–141.

[5] E. F. Adiutori, *New theory of thermal stability in boiling systems*, Nucleonics, 22 (1964), pp. 92–101.

[6] H. Auracher and W. Marquardt, *Heat transfer characteristics and mechanisms along entire boiling curves under steady-state and transient conditions*, J. Heat Fluid Flow, 25 (2004), pp. 223–242.

[7] S. A. Kovalev, *On methods of studying heat transfer in transition boiling*, Int. J. Heat Mass Transfer, 11 (1968), pp. 279–283.

[8] K. Stephan, *Stabilität beim Sieden*, Brennst.-Wäme-Kraft, 17 (1965), pp. 571–578.

[9] V. K. Dhir, *Boiling heat transfer*, Annu. Rev. Fluid Mech., 30 (1998), pp. 365–401.

[10] A. V. Gurevich and R. G. Mints, *Self-heating in normal metals and superconductors*, Rev. Modern Phys., 59 (1987), pp. 941–999.

[11] S. A. Kovalev and G. B. Rybchinskaya, *Prediction of the stability of pool boiling heat transfer to finite disturbances*, Int. J. Heat Mass Transfer, 21 (1978), pp. 691–700.

[12] S. A. Kovalev and S. V. Usitakov, *Analysis of the stability of boiling modes involving the use of stability diagrams*, High Temp., 41 (2003), pp. 68–78.

[13] J. Blum, T. Lüttich, and W. Marquardt, *Temperature wave propagation as a route from nucleate to film boiling?*, in Proceedings of the Second International Symposium on Two-Phase Flow Modelling and Experimentation, Rome, Vol. 1, G. P. Celata, P. DiMarco, and R. K. Shah, eds., Edizioni ETS, Pisa, 1999.

[14] J. Blum, W. Marquardt, and H. Auracher, *Stability of boiling systems*, Int. J. Heat Mass Transfer, 39 (1996), pp. 3021–3033.

[15] J. Blum and W. Marquardt, *Objection to Haramura's Criterions for Temperature Uniformity across the Surface in Transition Boiling*, Technical report LPT-1998-14, RWTH Aachen, Aachen, Germany, 1998.

[16] H. van Ouwekerk, *Burnout in pool boiling. The stability of boiling mechanisms*, Int. J. Heat Mass Transfer, 15 (1972), pp. 25–33.

[17] P. C. Fife, *Mathematical Aspects of Reacting and Diffusing Systems*, Springer, Berlin, 1979.

[18] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Springer, New York, 1997.

[19] K. Mischaikow, *Global asymptotic dynamics of gradient-like bistable equations*, SIAM J. Math. Anal., 26 (1995), pp. 1199–1224.

[20] S. Maier-Paape and U. Miller, *Path-following the equilibria of the Cahn-Hilliard equation on the square*, Comput. Vis. Sci., 5 (2002), pp. 115–138.

[21] S. Maier-Paape, K. Mischaikow, and T. Wanner, *Structure of the attractor of the Cahn-Hilliard equation on a square*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 1221–1263.
[22] G. Iooss and D. D. Joseph, *Elementary Stability and Bifurcation Theory*, Springer, Berlin, 1990.
[23] H. Kielhöfer, *Bifurcation Theory. An Introduction with Applications to PDEs*, Springer, New York, 2004.
[24] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer, New York, 1988.
[25] M. Buchholz, H. Auracher, T. Lüttich, and W. Marquardt, *A study of local heat transfer mechanisms along the entire boiling curve by means of microsensors*, Int. J. Heat Fluid Flow, 25 (2006), pp. 243–261.
[26] M. Buchholz, T. Lüttich, H. Auracher, and W. Marquardt, *Experimental investigation of local processes in pool boiling along the entire boiling curve*, Int. J. Therm. Sci., 45 (2004), pp. 269–293.
[27] H. Auracher and W. Marquardt, *Experimental studies of boiling mechanisms in all boiling regimes under steady-state and transient conditions*, Int. J. Therm. Sci., 41 (2002), pp. 586–598.
[28] Y. Haramura, *Temperature uniformity across the surface in transition boiling*, ASME J. Heat Transfer, 113 (1991), pp. 980–984.
[29] E. Kreyszig, *Advanced Engineering Mathematics*, Wiley, Chichester, UK, 1999.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 962–987
© 2008 Society for Industrial and Applied Mathematics

Regularity Properties of Critical Invariant Circles of Twist Maps, and Their Universality*

Arturo Olvera† and Nikola P. Petrov‡

Abstract. We accurately compute the golden and silver critical invariant circles of several area-preserving twist maps of the cylinder. We define some functions related to the invariant circle and to the dynamics of the map restricted to the circle (for example, the conjugacy between the circle map giving the dynamics on the invariant circle and a rigid rotation on the circle). The global Hölder regularities of these functions are low (some of them are not even once differentiable). We present several conjectures about the universality of the regularity properties of the critical circles and the related functions. Using a Fourier analysis method developed by de la Llave and one of the authors, we compute numerically the Hölder regularities of these functions. Our computations show that—within their numerical accuracy—these regularities are the same for the different maps studied. We discuss how our findings are related to some previous results: (a) to the constants giving the scaling behavior of the iterates on the critical invariant circle (discovered by Kadanoff and Shenker) and (b) to some characteristics of the singular invariant measures connected with the distribution of iterates. Some of the functions studied have pointwise Hölder regularity that has different values at different points. Our results give convincing numerical support to the fact that the points with different Hölder exponents of these functions are interspersed in the same way for different maps, which is a strong indication that the underlying twist maps belong to the same universality class. In particular, the numerical results on the regularity of the so-called big conjugacies imply that the Hölder spectra of the functions conjugating the dynamics on the critical invariant circle to a rigid rotation are the same. This, in turn, shows that the invariant measures on the critical circles have the same singularity spectra.

Key words. twist maps, critical invariant circles, universality, regularity of functions, Fourier analysis

AMS subject classifications. 37-04, 37C15, 37E20, 37E40, 42-04

DOI. 10.1137/070687967

1. Introduction.

It has been known since the late 1970s and early 1980s that many objects at the boundary of chaotic behavior exhibit remarkable scaling properties and that, furthermore, these properties are "universal." Such properties are exhibited by unimodal maps of the interval [1, 2] and [3] (reprinted in [59]), critical maps of the circle [4] (reprinted in [59]), critical KAM tori [5, 6], and other systems. These observations have been explained in terms of a renormalization group analysis, following a methodology that had been developed earlier in the study of critical phenomena in statistical mechanics and field theory [7, 8, 9], [10]

*Received by the editors April 11, 2007; accepted for publication (in revised form) by J. Meiss April 11, 2008; published electronically August 6, 2008.
http://www.siam.org/journals/siads/7-3/68796.html

†IIMAS-UNAM, FENOMEC, Apdo. Postal 20–726, México D. F. 01000, Mexico (aoc@uxmym1.iimas.unam.mx). This author's visits to the University of Oklahoma have been supported by NSF and UNAM.

‡Department of Mathematics, University of Oklahoma, Norman, OK 73019 (npetrov@ou.edu). The research of this author was partially supported by National Science Foundation grant DMS-0405903 and by the Michigan Center for Theoretical Physics (where part of this research was conducted). His visits to Mexico City have been supported by UNAM.

(reprinted as a book [11]), and [12] (reprinted in [61]).

The scale invariance of the critical objects affects many of their properties. Notably, the Hölder regularity of the critical objects (or some functions related to them) tends to have a low and fractional value. Presumably the values of the regularities are related to exponents and geometric properties of the renormalization group fixed points which describe the critical objects.

Furthermore, the observation that critical objects can be divided into "universality classes" such that all objects in a given class "look the same" can be tested numerically. One way to do this is to define certain functions related to the critical objects—typically these functions are not very regular (in some cases not even once differentiable)—and to test numerically whether the regularities of these functions are the same for different objects. Another—even more sensitive—test for universality is to take two functions, say h_1 and h_2, from the same class, and to study the regularity of functions such as $H_{1,2} := h_1 \circ h_2^{-1}$—for h_1 and h_2 belonging to the same universality class, one can expect $h_1 \circ h_2^{-1}$ to be more regular than h_1 and h_2^{-1}. From the fact that $H_{1,2}$ is more regular than h_1 and h_2^{-1} one can also draw important conclusions about the pointwise Hölder regularity of the functions h_i. If the Hölder regularity of the functions h_i has different values at different points (as in the case we consider), then one can conclude that the points at which the function h_1 has certain values of the pointwise Hölder exponent are interspersed in the same way as the corresponding points for the function h_2 (for more precise statements, see sections 2.6 and 5.3).

The idea of using the regularity of a function as an indicator for the universality class was tested in [13] in the case of noncritical and critical (with different degree of criticality) circle maps, in which the empirical results are accompanied by an extensive mathematical theory. A substantial part of the effort in [13] was to develop implementations of methods known in harmonic analysis (finite differences, Littlewood–Paley theory, wavelet analysis) to assess the regularity of the objects numerically.

In the present paper, we extend the methodology of [13] to the study of critical invariant circles of area-preserving twist maps. Invariant circles in dynamical systems are among the most important objects that organize the long-term behavior of the system, and the critical ones are especially important because of their role as "last barriers to chaos" (for readable reviews see, e.g., [14] or, with more emphasis on the mathematical aspects, the recent book [15]). Critical invariant circles have been studied extensively since the early 1980s [5, 6, 10, 12].

We accurately compute the golden critical invariant circles of several standard-like area-preserving twist maps and some functions related to the dynamics of the iterates of the maps on these circles. Then we apply methods developed in [13] to study the Hölder regularity of these functions and some universality aspects. We also perform some computations in the case of critical invariant circles with rotation number equal to the silver mean.

In section 2 we give some background on twist maps and their critical invariant circles, define the functions that are the objects of our numerical study, and present several precise conjectures concerning the properties of the critical invariant circles and the functions introduced. Section 3 is devoted to a discussion of the numerical methods used to compute critical invariant circles and to assess Hölder regularity of functions. We collect our results in section 4, and in section 5 we discuss their significance and relationship with previous studies.

2. Critical invariant circles of twist maps.

2.1. Twist maps. Let $\mathbb{T} := \mathbb{R}/\mathbb{Z}$ stand for the circle. We will be concerned with maps F of the (infinite) cylinder $\mathbb{T} \times \mathbb{R}$,

$$F : \mathbb{T} \times \mathbb{R} \to \mathbb{T} \times \mathbb{R} : (\theta, r) \mapsto F(\theta, r) =: (\theta', r') \ ,$$

which satisfy the following properties:

- *Area preservation.* The map F preserves the oriented area: $\det DF = 1$.
- *Zero-flux.* The oriented area between a homotopically nontrivial circle and its image under F is 0. (In our situation, this is equivalent to saying that every nontrivial circle intersects its image.)
- *Twist condition.* For any fixed value of θ, $\frac{\partial \theta'}{\partial r} > 0$.

A map of the cylinder can be identified with a map $\tilde{F} : \mathbb{R}^2 \to \mathbb{R}^2$ (called a *lift* of F) which satisfies

$$\tilde{F}(\theta + 1, r) = \tilde{F}(\theta, r) + (1, 0) \ .$$

Often one does not need to keep the distinction.

The maps which we will use in our numerical studies are of the form $(\theta', r') := F(\theta, r)$ with

$$\begin{aligned} \theta' &= (\theta + r') \bmod 1 \ , \\ r' &= r + \lambda V(\theta) \ , \end{aligned} \tag{2.1}$$

where λ is a parameter and $V : \mathbb{T} \to \mathbb{R}$ is a function satisfying $\int_0^1 V(\theta)\, d\theta = 0$. In particular, many numerical studies have been devoted to studying (2.1) with

$$V(\theta) = -\frac{1}{2\pi} \sin 2\pi\theta \ , \tag{2.2}$$

in which case we will call the map F the *Taylor–Chirikov* map. In our studies we used six functions V (given in section 4.1), all of them odd. The fact that the functions V are odd allows us to use the technique of the symmetry lines described in section 3.1 in order to compute periodic orbits of F. Of course, one would expect that the same results hold for any V, not only for odd ones.

Given an orbit $\mathcal{X} = \{(\theta_n, r_n) = F^n(\theta_0, r_0) \mid n = 0, 1, 2, \ldots\}$, we define its *rotation number*, $\rho(\mathcal{X})$, as the limit

$$\rho(\mathcal{X}) := \lim_{n \to \pm\infty} \frac{\tilde{\theta}_n - \theta_0}{n}$$

whenever this limit exists; here $\tilde{\theta}_n$ is the projection to the first argument of the nth iterate of the point (θ_0, r_0) under a lift $\tilde{F}$ of F. In contrast to the situation for circle maps, the rotation number depends on the orbit (and it may happen that some orbits do not have a rotation number).

We say that an orbit is *well-ordered* when, for every k and l, the function of n defined as $e(n) = \theta_{n+k} - l - \theta_n$ has the same sign. Every well-ordered orbit has a rotation number (the converse, however, is not true).

It is also easy to see that if a bounded orbit is well-ordered and $\rho(\theta_0, r_0)$ is irrational, the closure of the orbit, $\overline{\{(\theta_n, r_n)\}_{n=0}^{\infty}}$, is a perfect set (i.e., every point is an accumulation point of points in the set); in other words, in this case the orbit is either a homotopically nontrivial circle or a Cantor set.

A set $U \subseteq \mathbb{T} \times \mathbb{R}$ is *invariant* if $U = F(U)$.

The following result plays an important role [15, Chap. 2].

Theorem 2.1. *If F is as above, for every $\rho \in \mathbb{R}$ there exists a well-ordered orbit with rotation number ρ.*

2.2. Invariant circles of twist maps—rigorous results. The proof of the following theorem can be found in [16] (reprinted in [60]) and [17] (reprinted in [61]). We refer to [15] and [18] for a detailed exposition.

Theorem 2.2. *Let U be an open simply connected invariant set containing one of the ends of the cylinder. Then the boundary, ∂U, of the set U is an invariant circle which is the graph of a Lipschitz function. In other words, ∂U can be written as $r = R(\theta)$, where $R : \mathbb{T} \to \mathbb{R}$ is a Lipschitz function.*

For the map (2.1), the Lipschitz constant of the function R can be bounded by an expression which involves only the Lipschitz constant of the function F in a neighborhood of the circle ∂U.

In particular, we have the following corollary.

Corollary 2.1. *Any homotopically nontrivial invariant circle is the graph of a Lipschitz function R.*

A number ρ is said to be *Diophantine* if, for each $m, n \in \mathbb{N} \setminus \{0\}$, for some $C > 0$, and for some $d > 2$, it satisfies

$$\left|\rho - \frac{m}{n}\right| > \frac{C}{n^d} \ .$$

In the case when the map F is close to integrable and its rotation number ρ is Diophantine, one can apply the Kolmogorov–Arnold–Moser theory to obtain that there exists an analytic invariant circle such that the orbits on it have rotation number ρ.

Golden, respectively, *silver*, invariant circles are those with rotation number equal to the *golden mean*,

$$\sigma_{\mathrm{G}} := [1, 1, 1, \ldots] = \frac{\sqrt{5} - 1}{2} \ , \tag{2.3}$$

respectively, to the *silver mean*, $\sigma_{\mathrm{S}} := [2, 2, 2, \ldots.] = \sqrt{2} - 1$. Here we have used the notation $\rho = [a_1, a_2, a_3, \ldots] = 1/(a_1 + 1/(a_2 + 1/(a_3 + \cdots)))$ for the continued fraction expansion of $\rho \in (0, 1)$ [19].

There are rigorous results that guarantee the nonexistence of invariant circles of F of the form (2.1).

Theorem 2.3.

(i) *If $\sup_\theta |\lambda V(\theta)| > 1$, then (2.1) has no invariant circles.*

(ii) *If $\sup_\theta |V'(\theta)| = 1$ (which holds for the function (2.2)), then for $|\lambda| > \frac{4}{3}$ the map (2.1) has no invariant circles.*

(iii) *For V given by (2.2), the map (2.1) has no golden invariant circles for $|\lambda| > \frac{63}{64} = 0.984375\ldots.$*

(iv) *For V given by (2.2), the map (2.1) has no golden invariant circles for $|\lambda| > 0.9718$.*

Part (i) of Theorem 2.3 is elementary: if $\lambda \sup_\theta |V(\theta)| > 1$, then there will exist points $(\theta^*, r^*) \in \mathbb{T}\times\mathbb{R}$ such that $F(\theta^*, r^*) = (\theta^*, r^*+1)$, which, when iterated, gives that $F^n(\theta^*, r^*) = (\theta^*, r^*+n)$—the unbounded growth of the second coordinate of $F^n(\theta^*, r^*)$ with n implies that a topologically nontrivial invariant circle cannot exist. (These orbits are called "Chirikov accelerator modes" [20].)

Part (ii) can be found in [17], and parts (iii) and (iv) are proved by computer-assisted methods in [21] and [22], respectively.

It is widely believed that the following conjecture holds.

Conjecture 2.1. *For a Diophantine number ρ and for a map F of the form (2.1), there is a number $\Lambda(\rho)$ such that when $|\lambda| > \Lambda(\rho)$, there is no invariant circle with rotation number ρ, and when $|\lambda| < \Lambda(\rho)$, there exists an analytic invariant circle with rotation number ρ. The invariant circle becomes critical when $|\lambda| = \Lambda(\rho)$.*

Since our paper is devoted to homotopically nontrivial invariant circles, we will usually omit the words "homotopically nontrivial."

2.3. Functions related to the critical invariant circles. We are interested in describing the critical invariant circles with rotation number ρ which are in the boundary of existence. Postponing for the moment issues about how these objects can actually be computed, we point out that, to a given critical invariant circle γ of rotation number ρ, we can associate

- the function $R : \mathbb{T} \to \mathbb{R}$ such that the critical invariant circle γ is the graph of R:

$$\gamma = \{(\theta, r) \in \mathbb{T} \times \mathbb{R} \, : \, r = R(\theta)\} \; ; \tag{2.4}$$

- the *advance map* $g : \mathbb{T} \to \mathbb{T}$ defined by

$$F(\theta, R(\theta)) = (g(\theta), R \circ g(\theta)) \; ; \tag{2.5}$$

- the *hull map* $\Psi : \mathbb{T} \to \mathbb{T} \times \mathbb{R}$, which gives a representation of the invariant circle γ in such a way that the dynamics on γ becomes a rotation by ρ, i.e.,

$$F \circ \Psi(\theta) = \Psi(\theta + \rho) \; ; \tag{2.6}$$

- the map $h = \pi_1 \circ \Psi : \mathbb{T} \to \mathbb{T}$ (where $\pi_1 : \mathbb{T} \times \mathbb{R} \to \mathbb{T}$ is the projection onto $\mathbb{T}$), which conjugates the advance map to a rotation by ρ:

$$g \circ h(\theta) = h(\theta + \rho) \; ; \tag{2.7}$$

- the map $h^{-1} : \mathbb{T} \to \mathbb{T}$, which is the inverse of the map h defined in (2.7).

We note the following rigorous results.

Theorem 2.2 guarantees that the function R is Lipschitz. It is an easy consequence of the implicit function theorem that g should be as regular as R. Nevertheless, it is useful to compute the regularities of both g and R independently to assess the reliability of the numerical methods used.

Because of (2.7), it is clear that the regularity of g is not smaller than the minimum of the regularities of h and h^{-1}.

2.4. The "big" conjugacies. Let ρ be a Diophantine number, F_i $(i = 1, 2)$ be area-preserving twist maps, and γ_i be the critical invariant circle of F_i with rotation number ρ. Let g_{γ_i} and h_{γ_i} be the associated advance map (2.5) and conjugacy (2.7), respectively. We introduce the conjugating functions

$$G_{\gamma_1,\gamma_2} := g_{\gamma_1} \circ g_{\gamma_2}^{-1} : \mathbb{T} \to \mathbb{T} ,$$
$$H_{\gamma_1,\gamma_2} := h_{\gamma_1} \circ h_{\gamma_2}^{-1} : \mathbb{T} \to \mathbb{T} .$$

We will call these functions "big" conjugacies to distinguish them from the "small" conjugacies h that conjugate the projected dynamics on the critical circles to a rigid rotation (2.7). Note that the "big" conjugacies satisfy

$$G_{\gamma_1,\gamma_2} \circ G_{\gamma_2,\gamma_3} = G_{\gamma_1,\gamma_3} , \quad H_{\gamma_1,\gamma_2} \circ H_{\gamma_2,\gamma_3} = H_{\gamma_1,\gamma_3} .$$

Below we discuss one aspect of the definition of the big conjugacies that will be important in our computations.

Since there is no "origin" on the circle $\mathbb{T}$, one has a certain amount of freedom in the definition of some maps. For example, if the function Ψ is a hull map (i.e., satisfies (2.6)), then the function $\tilde{\Psi}$ defined as $\tilde{\Psi}(\theta) = \Psi(\theta + \zeta)$ will also satisfy (2.6) for any choice of the constant ζ. Similarly, the map h (2.7) that conjugates the advance map g to a rigid rotation can be redefined by composing it on the right with a rotation, and the resulting map, $\tilde{h}(\theta) = h(\theta + \zeta)$, will also conjugate g to a rigid rotation. Naturally, all important properties of the maps h and $\tilde{h}$—in particular, their Hölder regularity—will be the same. However, one cannot use this freedom liberally when studying the big conjugacies. To understand the reason for this, consider the map h defined by (2.7) for some twist map F. Naturally, the map $h \circ h^{-1}$ is the identity map, so it is C^∞. However, for any nonzero ζ in the definition of $\tilde{h}$, there is no guarantee that the map $h \circ \tilde{h}^{-1}$ will be C^∞. This is due to the fact that the regularity of h may be different at different points, and while in $h \circ h^{-1}$ these "irregularities" cancel out, in $h \circ \tilde{h}^{-1}$ the action of h does not necessarily "undo" the irregularities caused by $\tilde{h}^{-1}$. In section 2.5 we explain in detail how we choose ζ in order to avoid the "spurious" irregularities of the big conjugacy.

In section 2.6 we formulate some conjectures about the regularity of the big conjugacies, and in the conclusion (section 5.3) we discuss the implications for the pointwise Hölder spectra of the functions h_i.

2.5. Big conjugacies and symmetries. Consider two functions h_{γ_1} and h_{γ_2} corresponding to the critical circles γ_1 and γ_2 of the twist maps F_1 and F_2. If F_1 and F_2 happen to belong to the same "universality class" (see section 2.6), then one would expect that the big conjugacy H_{γ_1,γ_2} will be more regular than the functions h_{γ_1} and $h_{\gamma_2}^{-1}$. To avoid introducing spurious irregularities in H_{γ_1,γ_2}, we use the symmetries of the map h that come from the symmetries of the F [23], [24] (reprinted in [61]), and [25].

It is well known that if the function V is odd, then the map F given by (2.1) can be written as a composition of two involutions:

$$F = I_1 \circ I_0 , \qquad I_0^2 = I_1^2 = \mathrm{Id} , \tag{2.8}$$

where

$$I_0(\theta, r) = (-\theta, r + \lambda V(\theta)) \ , \quad I_1(\theta, r) = (-\theta + r, r) \ . \tag{2.9}$$

From (2.8) we have $I_0 \circ F = F^{-1} \circ I_0$ and $I_1 \circ F = F^{-1} \circ I_1$. Acting on (2.6) with I_0 from the left, we obtain

$$F^{-1} \circ (I_0 \circ \Psi)(\theta) = (I_0 \circ \Psi)(\theta + \rho) \ . \tag{2.10}$$

On the other hand, if we define the function $L : \mathbb{T} \to \mathbb{T} \times \mathbb{R}$ by $L(\theta) := \Psi(-\theta)$, then (2.6) can be written as

$$F^{-1} \circ L(\theta) = L(\theta + \rho) \ . \tag{2.11}$$

Comparing (2.10) and (2.11), we see that L and $I_0 \circ \Psi$ can differ only by a shift in the argument; i.e., there has to exist a constant ζ such that $I_0 \circ \Psi(\theta) = L(\theta + \zeta) = \Psi(-\theta - \zeta)$. This, together with (2.9) and $h = \pi_1 \circ \Psi$, implies

$$h(\theta) = -h(-\theta - \zeta) \ .$$

This implies that $h(-\frac{\zeta}{2}) = 0$, and the numerical value of ζ can be found from the computed values of h. Setting $\tilde{h}(\theta) := h(\theta - \frac{\zeta}{2})$, we obtain that $\tilde{h}$ is an odd function. In what follows, we will assume that the appropriate value of ζ has been subtracted, and we will omit the tilde over h.

2.6. Universality. In this section, we formulate precisely some conjectures on the behavior of critical invariant circles described by a nontrivial fixed point of the renormalization group. It seems quite possible that these conjectures can be proved as conditional theorems assuming existence and certain properties of this fixed point.

One of the most striking predictions of the renormalization group theory is that many characteristics of the critical invariant circles are largely independent of the details of the map. This is captured by the notion of universality.

Definition 2.1. *We say that a numerical characteristic is* universal *when it takes the same value in an open set of maps. We say that a property is* universal *when it holds for an open set of maps.*

The open sets alluded to in Definition 2.1 are called *domains of universality.*

For the case that we will be concerned with, the description of the domains of universality in terms of properties of the nontrivial fixed points of the renormalization operator is still debated, but there are indications that the domain of universality is not the whole space [26, 27, 28].

Conjecture 2.2. *The existence of one and only one nontrivial fixed point of the renormalization operator is a universal property.*

This conjecture has been known for a long time [12]. Recently in [29] it has been shown how this conjecture can be formulated in terms of some other conjectured properties related to the transversal intersection of some manifolds with the unstable invariant manifold of the nontrivial fixed point. Even the formulation of the subsequent conjectures depends on Conjecture 2.2.

An important new development is the computer-assisted proof [30] of the existence of a nontrivial fixed point of the renormalization group associated with the breakup of the golden invariant torus (the proof in [30] is based on renormalization of time-dependent Hamiltonians developed in [31] and earlier papers of the same author).

The concept of universality is rather natural when one wants to study properties that depend on the speed at which the set of maps converges to the fixed point under the renormalization operator. In particular, regularity of conjugacies depends on this speed of convergence and, hence, should be a universal quantity (more precise formulations are given in [32]). To formulate the conjectures below, we will need the following definition.

Definition 2.2. *For $\kappa = n + \chi$ with $n \in \mathbb{Z}$, $\chi \in (0,1)$, we say that the function $K : \mathbb{T} \to \mathbb{R}$ has* (global) Hölder exponent *κ and write $K \in \Lambda_\kappa(\mathbb{T})$ when K is n times differentiable and, for some constant $C > 0$,*

$$|D^n K(\theta) - D^n K(\tilde{\theta})| \le C|\theta - \tilde{\theta}|^\chi$$

for all $\theta, \tilde{\theta} \in \mathbb{T}$.

For the case of an integer value of κ, this definition is more complicated, but we will omit it since in the applications considered in this paper κ is not an integer.

Now we formulate several conjectures concerning the regularities of the functions introduced in sections 2.3 and 2.4.

Conjecture 2.3. *The regularity, $\kappa(R)$, of the critical invariant circle is a universal number.*

Conjecture 2.4. *The regularities $\kappa(g)$, $\kappa(h)$, and $\kappa(h^{-1})$ are universal numbers.*

Conjecture 2.5. *For pairs of critical circles γ_1 and γ_2, the regularities $\kappa(G_{\gamma_1,\gamma_2})$ and $\kappa(H_{\gamma_1,\gamma_2})$ are universal numbers.*

Directly from the definition of Hölder regularity, one can see that if $\kappa(\phi)$ and $\kappa(\psi)$ are between 0 and 1, then $\kappa(\phi \circ \psi) \ge \kappa(\phi)\,\kappa(\psi)$. This implies that

$$\kappa(H_{\gamma_1,\gamma_2}) = \kappa(h_{\gamma_1} \circ h_{\gamma_2}^{-1}) \ge \kappa(h_{\gamma_1})\,\kappa(h_{\gamma_2}^{-1})\ . \tag{2.12}$$

For all critical invariant circles γ_i that we studied, we obtained numerically that $\kappa(h_{\gamma_i}) < 1$ and $\kappa(h_{\gamma_i}^{-1}) < 1$, so (2.12) yields that H_{γ_1,γ_2} is not less regular than $\kappa(h_{\gamma_1})\,\kappa(h_{\gamma_2}^{-1})$. For γ_1 and γ_2 in the same universality class, however, we expect more—because of "cancellation" of the "singularities" of h_{γ_1} and $h_{\gamma_2}^{-1}$, we state our final conjecture.

Conjecture 2.6. *The following inequalities hold for $i = 1, 2$:*

$$\kappa(h_{\gamma_i}) < \kappa(H_{\gamma_1,\gamma_2})\ , \qquad \kappa(h_{\gamma_i}^{-1}) < \kappa(H_{\gamma_1,\gamma_2})\ .$$

3. Description of the numerical methods.

In this section we first describe the methods used for numerical computation of invariant circles and the related functions described in sections 2.3 and 2.4. Then we briefly discuss the method we use to compute the Hölder regularity of the functions.

3.1. Computing critical invariant circles.

We need to compute (homotopically nontrivial) critical invariant circles of twist maps of the form (2.1) with a Diophantine rotation number. We approximate such invariant circles by well-ordered periodic orbits (whose existence is guaranteed by Theorem 2.1). Consider a sequence $\{\mathcal{X}^{(j)}\}_{j\in\mathbb{N}}$ of well-ordered periodic orbits whose rotation numbers, $\{\rho_j\}_{j\in\mathbb{N}}$, constitute a sequence of rational numbers which converge

to a Diophantine number ρ. Then the limit of these periodic orbits will be a well-ordered invariant set $\mathcal{X}_\rho$ of rotation number ρ; the existence of this set is guaranteed by the Aubry–Mather theory [33, Chap. 13], [15, Chap. 2]. The set $\mathcal{X}_\rho$ can be a continuous curve which is a graph of a Lipschitz function under appropriate conditions (Theorem 2.2) or an orbit homeomorphic to a Cantor set (Cantorus). In the former case, the rotation number uniquely determines the invariant circle (see Remark 4.2), while in the latter case there may exist different Cantor sets with the same rotation number.

We approximate a Diophantine number ρ by the rational numbers given by finite truncations of the continued fraction expansion of ρ. In the case of the golden mean σ_{G} (2.3), these rational approximants are ratios $\rho_m = Q_{m-1}/Q_m$ of consecutive Fibonacci numbers Q_m. The limit of the periodic orbits with rotation numbers ρ_m is the invariant set $\mathcal{X}_\rho$ we are looking for [24].

The problem of computing well-ordered orbits with a prescribed rational rotation number ρ_m is greatly simplified if the function $V(\theta)$ in (2.1) is odd. In this case the task of finding a periodic orbit is reduced to a one-dimensional problem because the map F can be written as the composition of two involutions as in (2.8); if such a decomposition is possible, the map F is said to be *reversible*. If F is reversible, there exists a set of lines in the (θ, r) space—called *symmetry lines*—that are invariant with respect to the involutions I_0 and I_1 (2.9). It can be shown that any periodic orbit has two points that belong to one of the symmetry lines; hence we can find these points (and, therefore, the periodic orbits that contain them) by using a one-dimensional root finder [24]. Using the fact that the periodic orbits computed in this way are well ordered, we can implement a numerical procedure to compute periodic orbits of several million points that approximate the invariant set $\mathcal{X}_\rho$.

We are interested in studying the properties of area-preserving twist maps of the form (2.1). When the parameter λ in (2.1) is equal to 0, the corresponding twist map acts on each point (θ, r) as a rigid rotation in the θ-direction, $F(\theta, r) = (\theta + r, r)$, and hence the phase space is foliated by invariant circles of the form $\{r = \text{const}\}$. For small values of $|\lambda|$, KAM theory guarantees the existence of invariant circles with Diophantine rotation numbers. According to Conjecture 2.1, there is an upper bound $\Lambda(\rho)$ on the values of $|\lambda|$ such that for $|\lambda| < \Lambda(\rho)$ there exists an invariant circle with rotation number ρ (some rigorous upper bounds on $\Lambda(\rho)$ are given in Theorem 2.3). To find an accurate numerical approximation of the critical value, $\Lambda(\rho)$, of λ for which the invariant circle of rotation number ρ disintegrates, we applied an empirical method known as the "residue criterion" proposed in [24], developed in [34], and partially justified rigorously in [35, 36]. The main idea of this method is to determine the value of λ such that the residue of all the approximating periodic orbits reaches the same value. Let R_m be the residue of a periodic orbit which is the mth approximant to an invariant circle with rotation number ρ. If $\lim_{m\to\infty} R_m = 0$, then there exists an invariant circle with rotation number ρ; if $\lim_{m\to\infty} R_m = \infty$, then the invariant set $\mathcal{X}_\rho$ is a Cantor set. A critical invariant circle is obtained at the value of λ for which R_m tends to a finite value as $m \to \infty$ (see Remark 4.2).

3.2. Studying Hölder regularity numerically. In this section we describe briefly the method we employed to study Hölder regularity, referring the reader to [13] for details, additional references, and an assessment of the numerical accuracy of various numerical methods for computing regularity.

In this paper, we will use only the method developed in [13] that was found to be the most accurate for studying Hölder regularity—the so-called continuous Littlewood–Paley (CLP) method. Here we do not use the wavelet-based methods implemented in [13]. The CLP method has been used in [37, 25, 38].

3.2.1. Theoretical basis of the CLP method. The CLP method is based on the following result (which can be found in [39, Chap. 5, Lemma 5]).

Theorem 3.1 (CLP). *The function $K \in L^\infty(\mathbb{T})$ is in $\Lambda_\kappa(\mathbb{T})$ if and only if for some integer $\eta > \kappa$ there exists a constant $C > 0$ such that for any $t > 0$*

$$\left\| \left(\frac{\partial}{\partial t} \right)^\eta \mathrm{e}^{-t\sqrt{-\Delta}} K \right\|_{L^\infty(\mathbb{T})} \leq C t^{\kappa - \eta} , \tag{3.1}$$

where Δ is the one-dimensional Laplace operator: $\Delta K(\theta) = K''(\theta)$.

Remark 3.1. If the above result holds for some integer $\eta > \alpha$, then it holds for all integers $\tilde{\eta} > \alpha$.

Remark 3.2. The operator $\mathrm{e}^{-t\sqrt{-\Delta}}$ is a convolution with the Poisson kernel: $\mathrm{e}^{-t\sqrt{-\Delta}} K = P_{\exp(-2\pi t)} * K$. The function $u(\theta, t) := \mathrm{e}^{-t\sqrt{-\Delta}} K(\theta)$ is a solution of Laplace's equation, $u_{\theta\theta} + u_{tt} = 0$, on the half-cylinder $(\theta, t) \in \mathbb{T} \times (0, \infty)$, with Dirichlet boundary condition $u(\theta, 0) = K(\theta)$.

Remark 3.3. The mathematical theory requires only that (3.1) be an upper bound. In our numerical experiments, however, this bound is saturated for a significant range of values of t. This fact is very possibly a consequence of the self-similarity at small scales of the functions we consider (which is at the basis of the renormalization group description). This saturation was also observed for the functions considered in [13, 37, 25, 38].

3.2.2. Remarks on the numerical implementation. To use the CLP method, we need to repeatedly apply fast Fourier transform (FFT), which is easiest to do if the values of the function K in (3.1) are known at 2^N equally spaced points in the interval $[0, 1)$ for some positive integer N. However, as we describe in section 4, we do not have control over the set of points at which the values of K can be computed (where K stands for any of the functions R, g, h, h^{-1}, H, G). Hence, the first step in applying the CLP method would be the computation of the values of K on an evenly spaced grid. If we accurately know the values of K at M points in $[0, 1)$, we can expect that, by using some interpolation method, we will be able to obtain the approximate values of K on $2^N \approx M$ equidistant points, $\{2^{-N} j\}_{j=0}^{2^N - 1}$. To compute the approximate values of K on the equidistant grid, we used cubic spline interpolation. Using interpolation poses the question of whether the interpolated values faithfully represent the true values of K. Naturally, the answer to this question is no, but, practically, if M is large enough, the interpolated values will be very close to the true values, which will allow us to accurately compute many Fourier coefficients of K. The degree of "contamination" of the Fourier spectra due to the interpolation depends on the uniformity of the distribution of the M points at which the value of K is accurately known (see Remark 4.3).

To apply the CLP method numerically, we observe that the operator $(\frac{\partial}{\partial t})^\eta \mathrm{e}^{-t\sqrt{-\Delta}}$ used in Theorem 3.1 is diagonal in a Fourier series representation: if $K(\theta) = \sum_{k \in \mathbb{Z}} \hat{K}_k \mathrm{e}^{-2\pi i k \theta}$, then

$$\left(\frac{\partial}{\partial t}\right)^{\eta} \mathrm{e}^{-t\sqrt{-\Delta}} K(\theta) = \sum_{k\in\mathbb{Z}} (-2\pi|k|)^{\eta}\, \mathrm{e}^{-2\pi t|k|}\, \hat{K}_k\, \mathrm{e}^{-2\pi i k\theta} \ . \tag{3.2}$$

Having computed the values of the spline interpolant to the function K on an equally spaced grid, applying (3.1) is easy. Namely, we fix some values of the parameters η and t, perform FFT to find $\hat{K}_k$, and compute the Fourier coefficients of $(\frac{\partial}{\partial t})^{\eta}\, \mathrm{e}^{-t\sqrt{-\Delta}} K$. Then we apply inverse FFT to find the values of $(\frac{\partial}{\partial t})^{\eta}\, \mathrm{e}^{-t\sqrt{-\Delta}} K$ at the equally spaced set of points $\{2^{-N} j\}_{j=0}^{2^N-1}$; among these values we find the one with maximum absolute value—this value we take for the numerical value of the left-hand side of (3.1). For a fixed value of η, we repeat this procedure for many values of t (we used several hundred values of t in our computations). According to (3.1), if we plot

$$\log \left\| \left(\frac{\partial}{\partial t}\right)^{\eta} \mathrm{e}^{-t\sqrt{-\Delta}} K \right\|_{L^{\infty}(\mathbb{T})} \quad \text{versus} \quad \log t \ , \tag{3.3}$$

the points should lie below a straight line of slope $\kappa - \eta$. As pointed out in Remark 3.3 (see also Remark 4.4) the points on the log-log plot should not only be below this straight line, but should also be close to it. We perform linear regression to find the slope of this line, from which we find κ.

4. Numerical results.

4.1. Twist maps studied. We study numerically a set of one-parameter families of area-preserving twist maps of the form (2.1), each family having a different function V. Within each family we find numerically the value $\Lambda(\sigma_{\mathrm{G}})$ of the parameter λ for which the golden (resp., silver) invariant circle is critical. The set of functions V that we selected—all of them odd (so that we can use the symmetry lines technique as explained in section 3.1)—consists of the following:

1. The standard (Taylor–Chirikov) map:

$$V_1(\theta) = -\frac{1}{2\pi} \sin 2\pi\theta \ . \tag{4.1}$$

2. The "standard map with two harmonics":

$$V_2(\theta) = -\frac{1}{2\pi} \left[\sin(2\pi\theta) - 0.03 \sin(6\pi\theta)\right] . \tag{4.2}$$

3. The "critical standard map with two harmonics":

$$V_3(\theta) = -\frac{1}{2\pi} \left[\sin(2\pi\theta) - \frac{1}{2} \sin(6\pi\theta)\right] . \tag{4.3}$$

   For this choice of coefficients, the first three derivatives of $V(\theta)$ at $\theta = 0$ are zero.

4. The "0.2-analytic map":

$$V_4(\theta) = -\frac{1}{2\pi} \frac{\sin(2\pi\theta)}{1 - 0.2\cos(2\pi\theta)} \ . \tag{4.4}$$

   This map has infinitely many nonzero Fourier coefficients. It would be very interesting to study this map when the coefficient of the cosine function in the denominator is close to 1, but then it would be extremely difficult to compute periodic orbits.

5. The "0.4-analytic map":

$$V_5(\theta) = -\frac{1}{2\pi}\frac{\sin(2\pi\theta)}{1-0.4\cos(2\pi\theta)} \ . \tag{4.5}$$

6. The "tent map":

$$V_6(\theta) = \sum_{j=1}^{17} c_j \sin(2\pi j\theta) \ , \tag{4.6}$$

where $c_j = (-1)^{\frac{j+1}{2}}\frac{4}{\pi^2 j^2}$ for j odd and $c_j = 0$ for j even are the Fourier coefficients of the function

$$\mathcal{V}(\theta) = \begin{cases} -4\theta & \text{for } 0 \le \theta < \frac{1}{4} \ , \\ 4\theta - 2 & \text{for } \frac{1}{4} \le \theta < \frac{3}{4} \ , \\ 4 - 4\theta & \text{for } \frac{3}{4} \le \theta < 1 \ . \end{cases}$$

The function V_6 is close to the piecewise linear continuous function $\mathcal{V}$.

Our numerical experiments were performed with the twist maps coming from the above six functions $V(\theta)$ and the corresponding values $\Lambda(\sigma_{\mathrm{G}})$, following the six steps below. All numerical values here are given for rotation number golden mean; for silver mean the algorithm is analogous.

1. As discussed in section 3.1, the invariant circle of rotation number σ_{G} can be obtained as a limit of periodic orbits of rotation numbers equal to ratios of consecutive Fibonacci numbers, $\rho_m = Q_{m-1}/Q_m$. We chose to compute hyperbolic periodic orbits and found the values of $\Lambda(\sigma_{\mathrm{G}})$ by applying Greene's residue criterion [24].
2. The highest approximant to the critical invariant circle that we computed was a periodic orbit with rotation number $Q_{29}/Q_{30} = 832040/1346269$. The value of $\Lambda(\sigma_{\mathrm{G}})$ was determined by using the condition that the difference $|R_{30} - R_{29}|$ of the residues of the periodic orbits with periods Q_{29} and Q_{30} be zero (in practice, we wanted this difference to be smaller than 10^{-10}); this procedure is related to Greene's residue criterion (see Remark 4.2 below). The periodic orbits were computed with an error not exceeding 10^{-23}.
3. We computed the hyperbolic periodic orbit $\{(\theta_m, r_m)\}_{m=0}^{M-1}$ of period $M = Q_{30}$. The values of the advance map g (2.5) at the points θ_m ($m = 0, 1, \ldots, M-1$) were then computed by $g(\theta_m) = \theta_{m+1}$ (here and below, we take mod 1 wherever needed). The values of the conjugacy h at the points $m\sigma_{\mathrm{G}}$ (these points correspond to m applications of the rigid rotation by σ_{G} to 0) are given by $h(m\sigma_{\mathrm{G}}) = \theta_m$ and, similarly, $h^{-1}(\theta_m) = m\sigma_{\mathrm{G}}$.
4. In our Fourier analysis–based CLP method we need to deal with periodic functions, so we compute the "periodized" versions, $g - \mathrm{Id}$, $h - \mathrm{Id}$, and $h^{-1} - \mathrm{Id}$, of the functions g, h, and h^{-1}. Then we sort the periodized functions with respect to their argument; the function R is already periodic, so we just sort its values.
5. The periodic functions are passed to the cubic spline interpolation routine to find approximations to the values of the corresponding functions on a uniformly spaced grid of 2^N points; we used $N = 20$ (so that $2^{20} = 1048576$ is roughly equal to the length, Q_{30}, of the periodic orbit).

6. The interpolated values of the functions are given to the CLP algorithm to compute their Hölder regularity. We used integer values of η in (3.1) from 1 to 5, and for each analyzed function chose the value of η that gave the best straight line on the log-log plot (3.3). The log-log plots for the other values of η were used as a consistency check.

In sections 4.4.1 and 4.4.2 we give the results on the regularity of the functions related to the critical invariant circles with rotation numbers σ_{G} and σ_{S}, respectively (for σ_{S}, we used only the functions V_1 and V_2).

Remark 4.1. In computing the big conjugacies H_{γ_1,γ_2}, we had to take special care to preserve the symmetries of the maps h. For each critical circle we studied, we needed to find the appropriate value of the constant ζ and shift the argument of the corresponding function h as explained in section 2.5.

Remark 4.2. Greene's residue criterion [24] was defined using elliptic periodic orbits, but one can obtain similar results using hyperbolic orbits (in this case the residue's value is negative; for rotation number golden mean it is $\lim_{m\to\infty} R_m \approx -0.255426$); see [34]. According to the Aubry–Mather theory, the hyperbolic periodic orbits are minimizing, while the elliptic ones are minimax orbits [15]. Each invariant circle is a minimizing orbit, and it is the limit set of a sequence of minimizing periodic orbits whose rotation numbers converge to the rotation number of the invariant circle. For any sequence of periodic orbits with rotation number $\frac{p_n}{q_n}$ that converges to ω (the rotation number of the invariant circle), the hyperbolic and the elliptic orbits that belong to this sequence are interleaved: any elliptic orbit with rotation number $\frac{p_{n+1}}{q_{n+1}}$ is bounded by two hyperbolic orbits with rotation number $\frac{p_n}{q_n}$ and $\frac{p_{n-1}}{q_{n-1}}$, respectively. Hence, the sequence of elliptic periodic orbits and the sequence of hyperbolic periodic orbits converge to the same invariant circle. We conjectured that the regularity of invariant circles computed with elliptic and hyperbolic periodic orbits is the same, and performed some numerical experiments that supported this conjecture. We use hyperbolic periodic orbits to compute the regularity of critical invariant curves because our numerical methods for computing the orbits are more robust for hyperbolic orbits than for the elliptic ones.

4.2. Critical invariant circles—visual explorations. In Figure 1 we show the critical invariant circles which, by the definition (2.4), are graphs of the functions R corresponding to the six twist maps studied. The graphs of the "periodized versions" of the advanced maps, $g - \mathrm{Id}$; the conjugacies, $h - \mathrm{Id}$; and their inverses, $h^{-1} - \mathrm{Id}$, are plotted in Figures 2, 3, and 4, respectively.

Figure 5 illustrates the self-similar nature of the functions h; needless to say, the insets are true zooms of parts of the graph of the function.

Figure 6 shows the graphs of several periodized big conjugacies $H - \mathrm{Id}$; it is obvious that these functions are smoother than the "small" conjugacies h.

4.3. Fourier spectra, CLP method. Figure 7 depicts $\log_{10}$ of the modulus of the kth Fourier coefficient of a periodized conjugacy $(h - \mathrm{Id})$ versus $\log_{10} k$; here h is the conjugacy corresponding to the twist map F with V_3 (4.3). The horizontal distance between two adjacent high peaks is approximately equal to $|\log_{10} \sigma_{\mathrm{G}}| \approx 0.209$, which is a manifestation of the self-similarity at small scales. The $\log_{10}$-$\log_{10}$ plots of the Fourier spectra of the functions $(g - \mathrm{Id})$ and $(h^{-1} - \mathrm{Id})$ for the same map F are given in Figure 8.

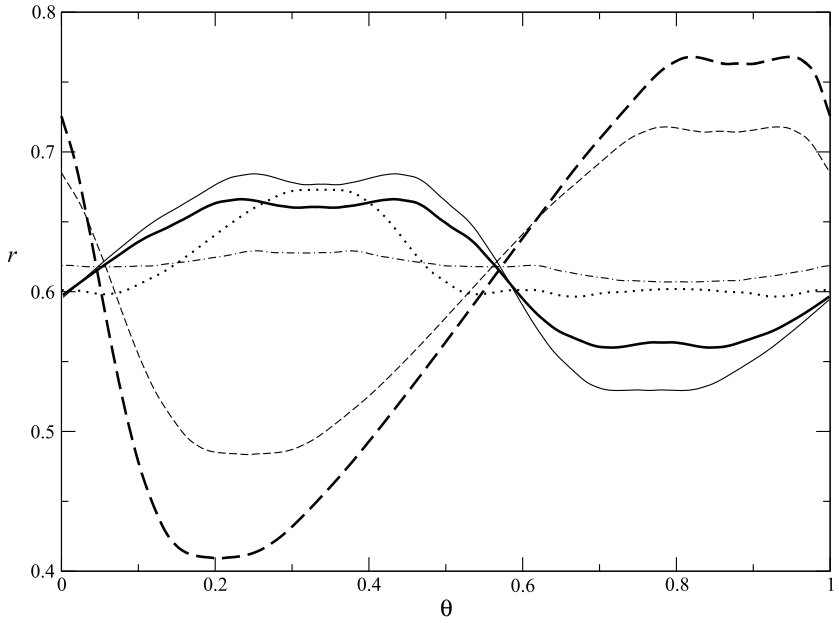


Figure 1. *Critical invariant circles, $r = R(\theta)$, of the maps corresponding to the maps $V_1, V_2, \ldots, V_6$ given by (4.1)–(4.6) (V_1 = thin solid line, V_2 = thick solid line, V_3 = dotted line, V_4 = thin dashed line, V_5 = thick dashed line, V_6 = dotted-dashed line).*

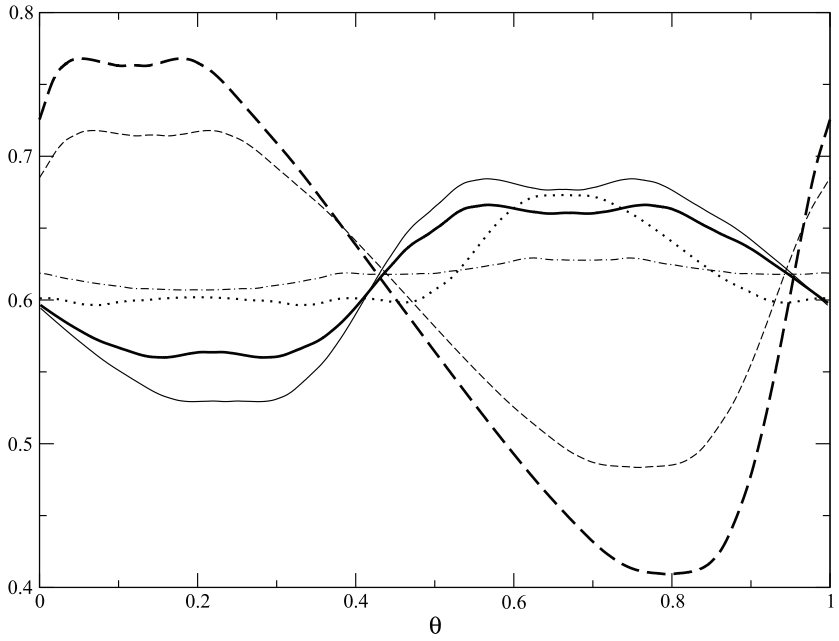


Figure 2. *"Periodized" advance maps $g - \mathrm{Id}$ (notation same as in Figure 1).*

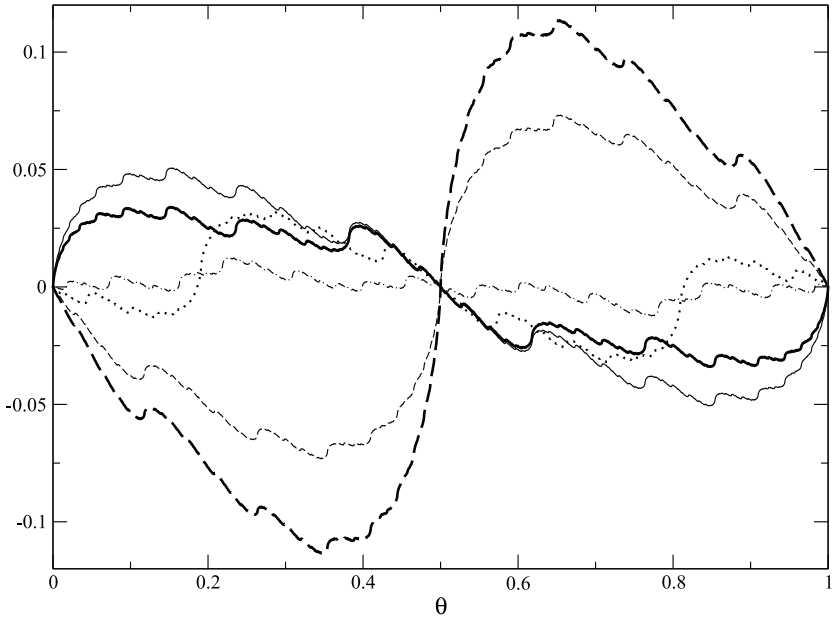


Figure 3. *"Periodized" conjugacies* $h - \mathrm{Id}$ *(notation same as in Figure 1).*

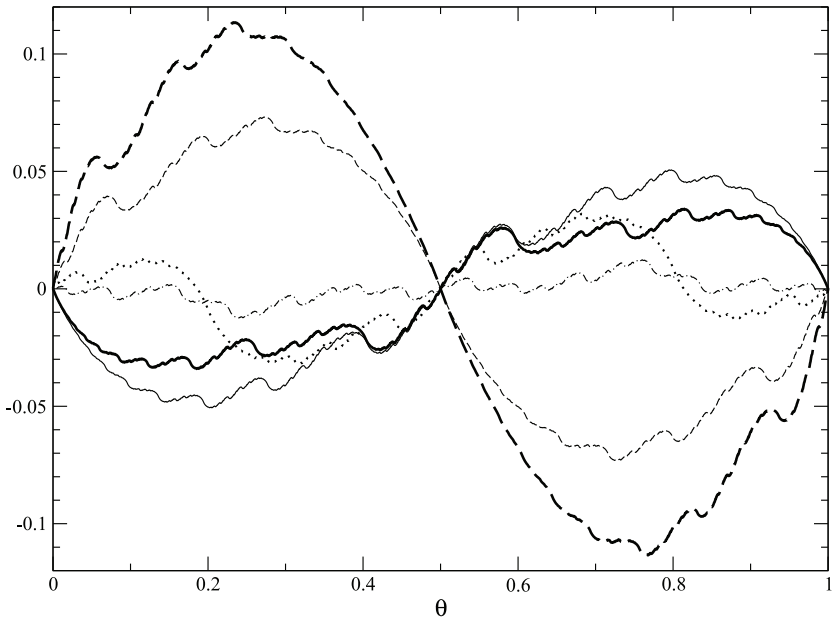


Figure 4. *"Periodized" inverse conjugacies* $h^{-1} - \mathrm{Id}$ *(notation same as in Figure 1).*

Remark 4.3. Note that the spectrum of h is very accurate even at length-scales $\sim 10^{-6}$, while the spectrum of h^{-1} is quite noisy. As explained in sections 3.2.2 and 4.1, the main reason for this is that the exact values of h are known at the points $(m\sigma_{\mathrm{G}})$ mod 1, which are almost uniformly distributed on $\mathbb{T}$. On the other hand, we know the exact values of g and h^{-1} at the

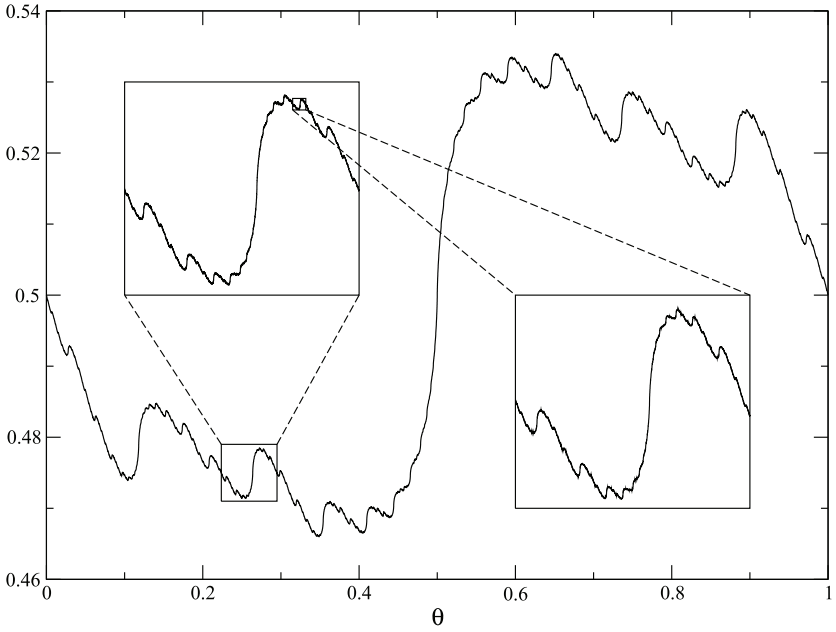


Figure 5. *Zooming in the graph of the function* $h - \mathrm{Id}$ *corresponding to the map* V_2 (4.2).

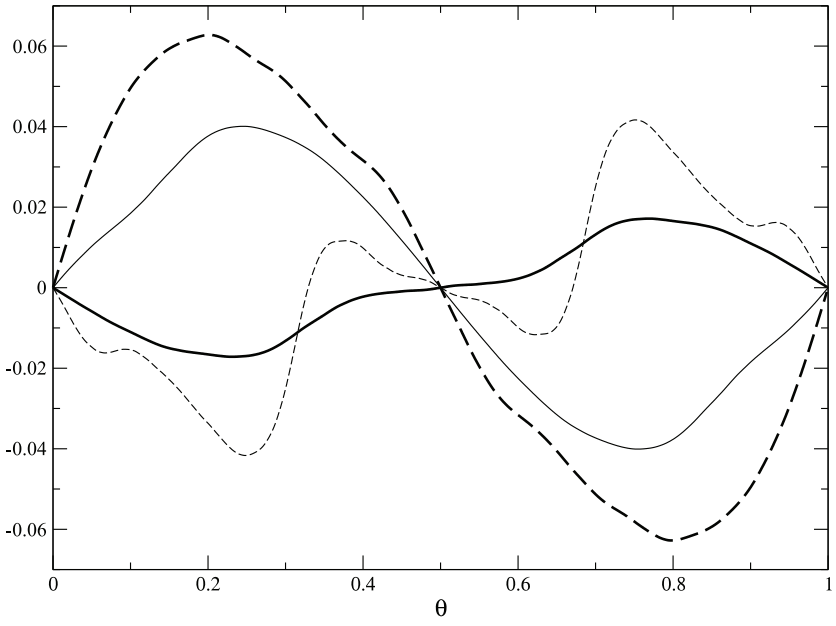


Figure 6. *"Periodized" big conjugacies* $H - \mathrm{Id}$.

very nonuniformly distributed points of the form $g^m(\theta_0)$ (because the underlying invariant measure is singular; see section 5.2), which results in the presence of big gaps between these points and, hence, distorted values of the spline interpolant.

In Figure 9 we show several plots of $\log_{10}$ of the left-hand side of (3.1) versus $\log_{10} t$.

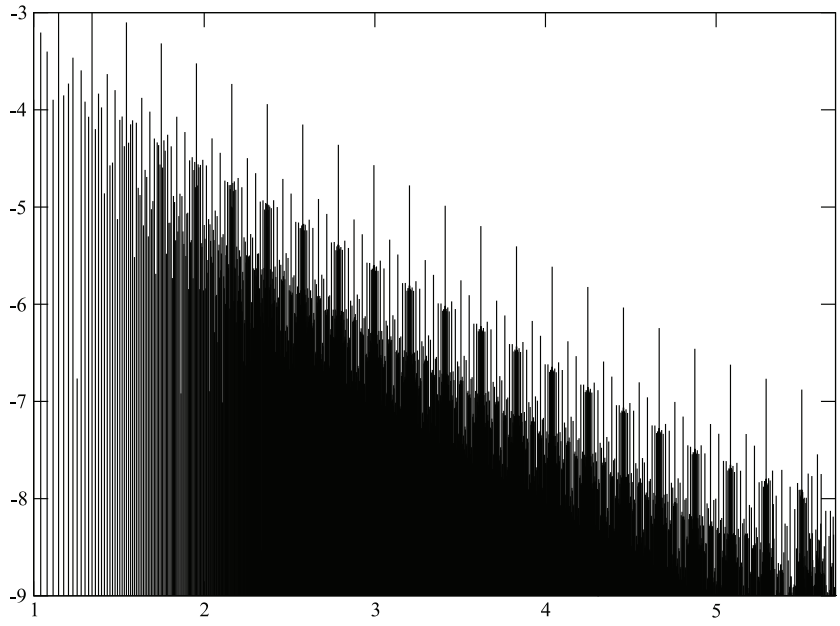


Figure 7. *Plot of* $\log_{10}|(\widehat{h-\mathrm{Id}})_k|$ *versus* $\log_{10} k$*, where* h *corresponds to the map* F *coming from the function* V_3 (4.3).

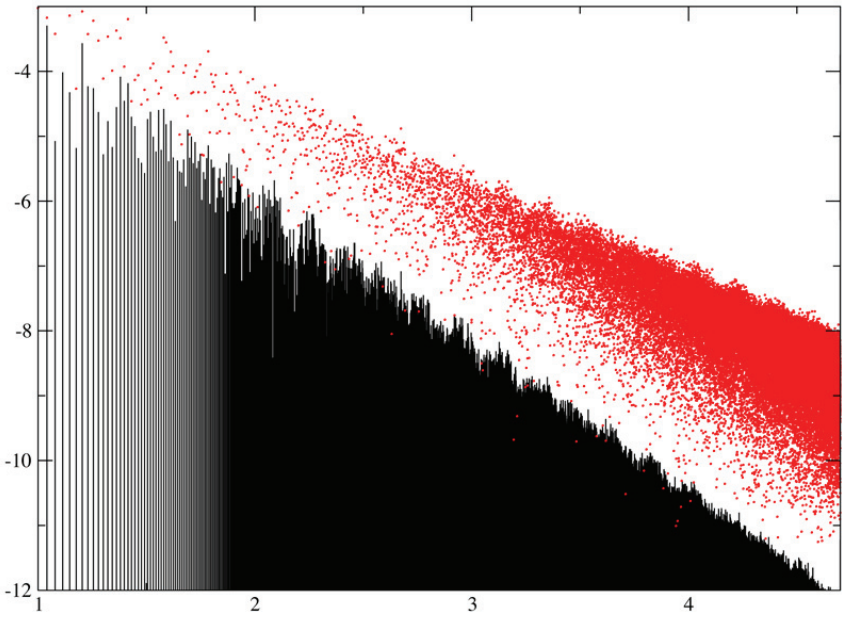


Figure 8. *Plot of* $\log_{10}|(\widehat{g-\mathrm{Id}})_k|$ *and* $\log_{10}|(\widehat{h^{-1}-\mathrm{Id}})_k|$ *versus* $\log_{10} k$*, for the same map* F *as in Figure 7. The impulses correspond to* $(g-\mathrm{Id})$*, and the dots above them to* $(h^{-1}-\mathrm{Id})$.

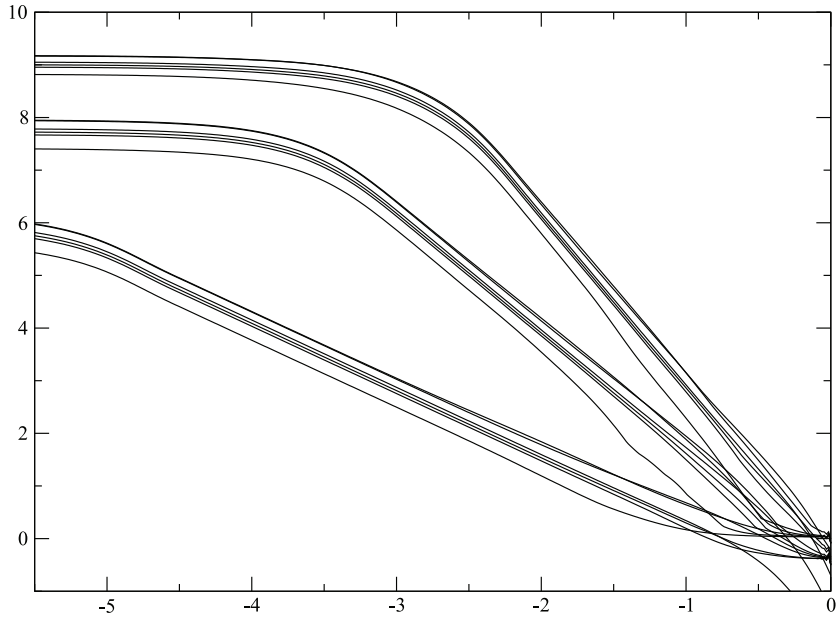


Figure 9. *Plots of* $\log_{10} \|(\frac{\partial}{\partial t})^\eta \, \mathrm{e}^{-t\sqrt{-\Delta}} K\|_{L^\infty(\mathbb{T})}$ *versus* $\log_{10} t$ *for the functions* $K = (h - \mathrm{Id})$ *for the twist maps coming from* $V_1, \ldots, V_6$ *for* $\eta = 2$ *(shallowest lines),* $\eta = 3$*, and* $\eta = 4$ *(steepest lines).*

The six lines in each group of lines of similar slope correspond to the six different choices (4.1)–(4.6) of functions V, and the lines in each group come from the same value of η in (3.1). Each of the "lines" in the figure in fact consists of 400 points (visually indistinguishable). The computer time spent on the CLP analysis is of the order of one minute per point (we used 2^{20} Fourier coefficients to compute each of these points).

We computed the regularity by performing linear regression on the points on graphs like the one in Figure 9, in the regions where the points follow more or less a straight line. As one can see from this figure, for t close to 1 (i.e., $\log_{10} t \approx 0$), the graphs for different functions are not straight lines, then as t decreases they form more or less straight lines, and as t decreases further, these lines level out. This behavior can be understood intuitively from (3.2)—for $t \approx 1$ the high-k Fourier coefficients are strongly suppressed by the factor $(-2\pi|k|)^\eta \, \mathrm{e}^{-2\pi t|k|}$, so the CLP method still does not "feel" the asymptotic self-similarity of the functions at small length-scales; at the other extreme, the leveling out of the lines for very small t comes from the fact that in our computations we use a finite—albeit very large—number of Fourier coefficients.

Remark 4.4. The "straight lines" in Figure 9 are not really straight (which has been noticed in different contexts in [13, 25]). We show this effect in Figure 10, which was created as follows. We took the six lines for $\eta = 2$ from Figure 9, and for each of them we computed the slope of the line as a function of the horizontal coordinate in the figure, $\log_{10} t$. To compute this slope, we took each pair of adjacent points on the line and found the slope of the straight line connecting these points. The distance between two consecutive peaks in Figure 9 is $|\log_{10} \sigma_{\mathrm{G}}|$; more interestingly, as $\log_{10} t$ becomes more negative, the lines tend to the same wavy line, until all lines reach saturation around $\log_{10} t \approx -4.5$.

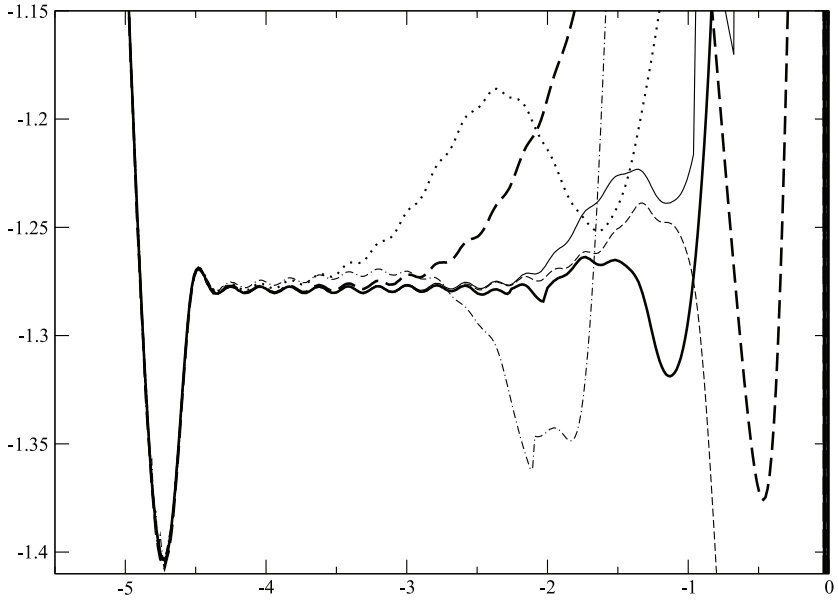


Figure 10. *Slope of the lines on Figure 9 as a function of* $\log_{10} t$ *(see text). The notation is the same as in Figure 1.*

Table 1

Regularities of the functions R, g, h, and h^{-1} *for the golden critical invariant circles of maps F corresponding to different functions V.*

F with	$\kappa(R)$	$\kappa(g)$	$\kappa(h)$	$\kappa(h^{-1})$
V_1	1.83(9)	1.83(9)	0.722(1)	0.92(1)
V_2	1.79(6)	1.75(9)	0.721(1)	0.92(1)
V_3	1.83(4)	1.84(3)	0.724(2)	0.93(2)
V_4	1.86(8)	1.86(8)	0.722(1)	0.92(1)
V_5	1.85(5)	1.85(5)	0.724(2)	0.93(1)
V_6	1.85(15)	1.88(12)	0.726(3)	0.93(2)

4.4. Hölder regularities—numerical results.

4.4.1. Hölder regularities for rotation number golden mean. Table 1 summarizes our numerical results. The first column gives the map V used in the numerical computations (for the six functions V given by (4.1)–(4.6)). In the other columns we give the values of the (global) Hölder exponent κ of the function R (representing the invariant circle as a graph in the (θ, r)-plane), the advance map g, the conjugacy h, and its inverse h^{-1}, coming from the (dynamics on) the golden critical invariant circle of the corresponding area-preserving twist map F. The notation used is the following: 1.85(15) stands for 1.85 ± 0.15, and 0.726(3) for 0.726 ± 0.003. Note that, within the numerical error, $\kappa(R) = \kappa(g)$, as expected.

We also computed the Hölder regularities of all big conjugacies H between each of the six functions $h_1, \ldots, h_6$ (coming from $V_1, \ldots, V_6$) with all other h_j's. We applied the CLP method to find that the regularity of all thirty functions H studied is

$$\kappa(H) = 1.80 \pm 0.15 \ . \tag{4.7}$$

4.4.2. Hölder regularities for rotation number silver mean. We computed the regularity of the maps R_j, g_j, h_j, and h_j^{-1} $(j = 1, 2)$ for the twist maps with the functions V_1 and V_2, as well as the regularity of the "big conjugacies" $H_{1,2}$ and $H_{2,1}$, for the critical invariant circle with rotation number the *silver mean*, $\sigma_S = [2, 2, 2, \ldots]$. The regularities for V_1 and V_2 were numerically the same, so we give only one value; similarly for $H_{1,2}$ and $H_{2,1}$. Here are the results for the regularities (the subscript S stands for "silver"):

$$\begin{aligned}
\kappa(R_S) &= 1.70 \pm 0.15 \ , \\
\kappa(g_S) &= 1.75 \pm 0.15 \ , \\
\kappa(h_S) &= 0.715 \pm 0.015 \ , \\
\kappa(h_S^{-1}) &= 0.87 \pm 0.02 \ , \\
\kappa(H_S) &= 1.80 \pm 0.15 \ .
\end{aligned}$$

5. Discussion and conclusion. In sections 5.1 and 5.2 we point out some relationships between our results and previous studies related to universal scaling factors and singular measures. In the final section 5.3, we recapitulate our findings.

5.1. Hölder regularity and scaling factors. Here we will explain how the scaling of the distances of closest returns of the iterates of a point gives bounds on the Hölder regularity of some of the functions we study. Our analysis here is reminiscent of the analysis in [13, sect. 8.2]. The numerical values below are for rotation number the golden mean.

We start by recalling the crucial observation of Kadanoff and Shenker [5, 6] (see also [10, sect. 4.4]) of the existence of universal scalings in the distribution of the iterates of the Taylor–Chirikov map on the critical invariant circle γ in neighborhoods of certain points of γ. Let $\theta_{\rm rar} \in \mathbb{T}$ stand for the value around which the iterates of the function g are most rarefied (in our notation $\theta_{\rm rar} = \frac{1}{2}$, while in [6] $\theta_{\rm rar} = 0$). Let $\theta_{\rm den} \in \mathbb{T}$ stand for the value around which the iterates of the function g are most dense (in our notation $\theta_{\rm den} = 0$, while in [6] it is $\theta_{\rm den} = \frac{1}{2}$). Since by Theorem 2.2 the function R is Lipschitz, around the points $(\theta_{\rm rar}, R(\theta_{\rm rar}))$ and $(\theta_{\rm den}, R(\theta_{\rm den}))$, the iterates of any point on γ under F are most rarefied, respectively, dense. Shenker and Kadanoff found that the critical invariant circle in a neighborhood of $\theta_{\rm rar}$ is asymptotically invariant under simultaneous scalings in both the θ- and r-directions, with scaling factors

$$\alpha_0 \approx -1.414836 \quad (\text{in } \theta) \ , \qquad \beta_0 \approx -3.0668882 \quad (\text{in } r)$$

(see also the bounds on these values in Stirnemann [40]). This implies that, for large n,

$$\frac{g^{Q_{n+1}}(\theta_{\rm rar}) - \theta_{\rm rar}}{g^{Q_n}(\theta_{\rm rar}) - \theta_{\rm rar}} \approx \alpha_0^{-1} \ , \qquad \frac{R(g^{Q_{n+1}}(\theta_{\rm rar})) - R(\theta_{\rm rar})}{R(g^{Q_n}(\theta_{\rm rar})) - R(\theta_{\rm rar})} \approx \beta_0^{-1} \ . \tag{5.1}$$

The scaling around $\theta_{\rm den}$ is a bit more complicated—it is called "step-3" scaling for obvious reasons:

$$\frac{g^{Q_{n+3}}(\theta_{\rm den}) - \theta_{\rm den}}{g^{Q_n}(\theta_{\rm den}) - \theta_{\rm den}} \approx \alpha_3^{-1} \ , \qquad \frac{R(g^{Q_{n+3}}(\theta_{\rm den})) - R(\theta_{\rm den})}{R(g^{Q_n}(\theta_{\rm den})) - R(\theta_{\rm den})} \approx \beta_3^{-1} \ , \tag{5.2}$$

where the "step-3" scaling factors are

$$\alpha_3 \approx -4.84581 \quad (\text{in } \theta) \ , \qquad \beta_3 \approx -16.8597 \quad (\text{in } r) \ .$$

To understand heuristically why these scalings give restrictions on the Hölder regularity of R, set $\Delta\theta := g^{Q_{n+1}}(\theta_{\rm rar}) - \theta_{\rm rar}$, $\Delta r := R(g^{Q_{n+1}}(\theta_{\rm rar})) - R(\theta_{\rm rar})$ for some large value of n. Then if the local Hölder exponent of R at $\theta = \theta_{\rm rar}$ is κ, we will have $|\Delta r| \sim |\Delta\theta|^\kappa$. If the graph of R is asymptotically invariant around $(\theta_{\rm rar}, R(\theta_{\rm rar}))$ with respect to the scalings (5.1), we will have $|\beta_0\,\Delta r| \sim |\alpha_0\,\Delta\theta|^\kappa$. "Dividing out" the last two relationships, we obtain $|\beta_0| \sim |\alpha_0|^\kappa$, i.e., $\kappa \sim \frac{\log|\beta_0|}{\log|\alpha_0|}$. This argument (which can easily be made rigorous) implies that the (global) Hölder exponent of R does not exceed $\frac{\log|\beta_0|}{\log|\alpha_0|} \approx 3.22945$. The scaling (5.2) yields a tighter bound on the Hölder regularity of R:

$$\kappa(R) \le \frac{\log|\beta_3|}{\log|\alpha_3|} \approx 1.7901\ . \tag{5.3}$$

Note that the fact that the scaling (5.2) is "step-3" (as opposed to "step-1") is irrelevant for the bounds on the Hölder regularity.

To obtain bounds on $\kappa(h)$ and $\kappa(h^{-1})$, we use Lemma 8.1 from [13], which says that if the function h conjugates f_1 and f_2, $h\circ f_1 = f_2 \circ h$, and if for some sequence of positive integers Q_n the functions f_j $(j = 1, 2)$ behave in a neighborhood of the fixed point $\theta_{\rm fix} = h(\theta_{\rm fix})$ of h as

$$f_j^{Q_n}(\theta_{\rm fix}) = \theta_{\rm fix} + C_j\eta_j^{-n} + o(\eta_j^{-n})$$

for some constants η_j and C_j, then $\kappa(h) \le \frac{\log|\eta_2|}{\log|\eta_1|}$. Applying this to the definition of h and using the well-known fact that $(Q_n\,\sigma_{\rm G}) \bmod 1 \le C\sigma_{\rm G}^n$, we obtain the bounds

$$\kappa(h) \le \frac{\log|\alpha_0^{-1}|}{\log|\sigma_{\rm G}|} \approx 0.721125\ , \qquad \kappa(h^{-1}) \le \frac{\log|\sigma_{\rm G}^3|}{\log|\alpha_3^{-1}|} \approx 0.91478\ . \tag{5.4}$$

A comparison with Table 1 suggests that these bounds are saturated.

5.2. Conjugacies and singular measures. The functions whose Hölder regularity we study are defined through high iterates of maps. For example, the graph of the function R defined by (2.4) is nothing but the critical invariant circle γ of F which is densely filled by the iterates $F^n(\theta_0, r_0)$ of some point $(\theta_0, r_0) \in \gamma$. Here we discuss how some characterizations of the singularities in the distribution of the iterates of F on γ are related to the Hölder regularity of some of the functions considered.

Hentschel and Procaccia [41] pointed out the importance of the *generalized (Rényi) dimensions* $D(q)$ of a singular measure for dynamical systems; these quantities have been defined previously in the context of probability theory by Rényi [42]. Halsey et al. in their seminal paper [50] related heuristically the Rényi dimension of a singular measure to the *spectrum of singularities* $f(\alpha)$. We recall that $f(\alpha)$ is the Hausdorff dimension of the set E_α of points where the measure has singularity of strength α. The spectrum $f(\alpha)$ is a function supported on the interval $[\alpha_{\min}, \alpha_{\max}]$, where $\alpha_{\min} = D(\infty)$ (resp., $\alpha_{\max} = D(-\infty)$) describes the scaling behavior of the measure in the region where the measure is most dense, respectively, most rarefied.

Let (θ_0, r_0) be an arbitrary point on the critical invariant circle γ of the area-preserving twist map F. Then the distribution of the iterates in a very long orbit, $\{F^n(\theta_0, r_0)\}_{n=0}^K$,

approaches as $K \to \infty$ the "density" of the measure on γ that is invariant with respect to the restriction of the map F onto γ. (We put "density" in quotation marks because for singular measures this is not a function but a set of Dirac δ-distributions.) This invariant measure on γ induces an invariant measure μ_g of the map g on $\mathbb{T}$. It is easy to see that (2.7) implies that

$$h^{-1}(\theta) = \int_0^\theta d\mu_g$$

(for an appropriately chosen ζ in the redefinition of h as in section 2.5). This relationship implies that the spectrum of singularities $f(\alpha)$ of the measure μ_g is the same as the *Hölder spectrum* $f_{\mathrm{H}}(\alpha)$ of the function h^{-1}. By definition, $f_{\mathrm{H}}(\alpha)$ is the Hausdorff dimension of the set where the local Hölder exponent of the function is equal to α; for a readable account we refer the reader to Jaffard [43, 44]. The (global) Hölder regularity $\kappa(\phi)$ of a function ϕ is equal to the lowest end, $\alpha_{\min}$, of the support of the Hölder spectrum, $f_{\mathrm{H}}(\alpha)$, of ϕ.

Osbaldestin and Sarkis [45] applied the method of [50] to determine numerically the functions $f(\alpha)$ and $D(q)$ of the invariant measure μ_g coming from the distribution of iterates of the Taylor–Chirikov map F on the golden invariant circle. They found that

$$\alpha_{\min} = D(\infty) \approx 0.915\ , \qquad \alpha_{\max} = D(-\infty) \approx 1.387 \approx \frac{1}{0.720}\ .$$

Comparing these with the values in Table 1, the reader should recognize that their $\alpha_{\min}$ is nothing but our $\kappa(h^{-1})$, while $\alpha_{\max}$ is equal to the inverse of the regularity of the conjugacy h.

Burić, Mudrinić, and Todorović [46, 47] studied numerically the Taylor–Chirikov map and the map (2.1) with $V(\theta) = \frac{1}{2}\sin 2\pi\theta + \frac{1}{4}\sin 4\pi\theta$ for rotation numbers with continued fraction expansions of the form $[S, 1^\infty] := [S, 1, 1, 1, \ldots]$, $[S, 2^\infty]$, $[S, 3^\infty]$, $[S, 4^\infty]$, where S stands for some short string of positive integers. They found that $f(\alpha)$ and $D(q)$ depend only on the tail but do not depend on the initial part S or on whether the Taylor–Chirikov map or the other map was used in their numerics.

Other papers related to numerical computations of singular measures on critical invariant circles of area-preserving twist maps are Shi and Hu [48, 49], where the methods of [50] (reprinted in [59]) were used, and Hunt et al. [51], where the authors used the thermodynamic formalism developed in [52] (reprinted in [59]) to compute the information dimension $D(1)$ of the standard map for different rotation numbers.

5.3. Conclusion. We accurately computed the golden critical invariant circles for six twist maps of the form (2.1) and the global Hölder regularity κ of some functions related to the dynamics on these circles, as well as regularities of the functions associated with the silver critical invariant circles of two twist maps. Our numerical experiments lend credibility to Conjectures 2.3, 2.4, and 2.5 concerning the universality of the regularities of the functions R, g, h, h^{-1}, and H (see Table 1 and (4.7)). Yamaguchi and Tanikawa [53] found numerically that the golden invariant circle (given by the function R) of the Taylor–Chirikov map is differentiable, but R' is not of bounded variation; our studies significantly narrow the numerical bounds on $\kappa(R)$ for this and other maps.

Our results seem to indicate that the regularities of R, h, and h^{-1} saturate the upper bounds (5.3) and (5.4) coming from previous studies of scaling exponents.

Our finding that $\kappa(H)$ is greater than $\kappa(h)$ and $\kappa(h^{-1})$ by a comfortable margin (cf. Conjecture 2.6) has an interesting consequence. As discussed in section 5.2, the Hölder regularity of h and h^{-1} is different at different points, and for each $\alpha \in (\alpha_{\min}, \alpha_{\max})$, the set E_α (where the pointwise Hölder exponent of h^{-1} is α) has Hausdorff dimension $f_H(\alpha)$ strictly between 0 and 1. Previous numerical studies indicated that $f_H(\alpha)$ are the same for different maps F. Our finding shows that the "irregularities" of functions h coming from different maps F are interspersed in the same way in $[0, 1]$ for all twist maps studied. Note that this does not mean that for a certain value of α the sets E_α are *the same* for different F in the same universality class—only the way all sets E_α for different α are interwoven is universal.

Perhaps it would be interesting for the reader to compare the results of the present paper with those on regularity of nontwist maps [25]. One has to keep in mind, however, that, while—according to Theorem 2.2—the critical invariant circles for twist maps are graphs (in polar coordinates), in the case of nontwist maps this is not so. Therefore, some of the objects we studied here (i.e., the functions R and g) do not have analogues for nontwist maps.

It would be interesting to apply wavelet-maxima methods for pointwise regularity [54, 55] (see also the rigorous analysis in [43, 44]) to the problem studied in this paper and to compare the results of the wavelet analysis with the results about the singular invariant measures.

The case of more general Bryuno numbers also deserves attention (see, e.g., [56, 57]). However, we do not think that numerical studies of the regularities of functions related to twist maps with such rotation numbers are feasible at present for several reasons. Most importantly, the accuracy of the results on regularity behaves like the logarithm of the computational effort and of the size of the data arrays needed in the computations; note that Figure 9 is in log-log scale. Also, the inherent "oscillations" around the straight line in that figure (shown in Figure 10) contribute to the numerical error in the determination of the averaged slopes. Achieving higher precision in computing the regularity will require computing the parameters of the twist map with a very high accuracy, which in turn will necessitate very long runs of the programs.

As a by-product of our studies, we have computed millions of Fourier coefficients of the functions h and noticed some self-similarity properties that to the best of our knowledge have not been observed before. Currently we are working on understanding these properties.

Acknowledgments. We would like to express our gratitude to Rafael de la Llave, who introduced the authors of the present paper to each other, suggested the problem, and took an active part in the early stages of this research. We have profited immensely from his expert advice and friendly prodding throughout our work on the paper.

We also thank the referees whose constructive suggestions helped us clarify some important points.

Our computations were carried out on the computers of IIMAS-UNAM and the Department of Mathematics of the University of Texas. A.O. would like to thank Ana Pérez for computational support. We used the doubledouble software developed by Keith Briggs, and the convenient plotting tool Grace [58] (a descendant of ACE/gr developed by Paul J. Turner). We express our thanks to all these people and organizations.

REFERENCES

[1] C. Tresser and P. Coullet, *Itérations d'endomorphismes et groupe de renormalisation*, C. R. Acad. Sci. Paris Sér. A-B, 287 (1978), pp. A577–A580.

[2] M. J. Feigenbaum, *Quantitative universality for a class of nonlinear transformations*, J. Statist. Phys., 19 (1978), pp. 25–52.

[3] M. J. Feigenbaum, *The universal metric properties of nonlinear transformations*, J. Statist. Phys., 21 (1979), pp. 669–706.

[4] S. Shenker, *Scaling behavior in a map of a circle onto itself: Empirical results*, Phys. D, 5 (1982), pp. 405–411.

[5] L. Kadanoff, *Scaling for a critical Kolmogorov-Arnold-Moser trajectory*, Phys. Rev. Lett., 47 (1981), pp. 1641–1643.

[6] S. Shenker and L. P. Kadanoff, *Critical behavior of a KAM surface.* I. *Empirical results*, J. Statist. Phys., 27 (1982), pp. 631–656.

[7] P. Collet, J. P. Eckmann, and O. E. Lanford, III, *Universal properties of maps on an interval*, Comm. Math. Phys., 76 (1980), pp. 211–254.

[8] M. J. Feigenbaum, L. P. Kadanoff, and S. J. Shenker, *Quasiperiodicity in dissipative systems: A renormalization group analysis*, Phys. D, 5 (1982), pp. 370–386.

[9] S. Ostlund, D. Rand, J. Sethna, and E. Siggia, *Universal properties of the transition from quasi-periodicity to chaos in dissipative systems*, Phys. D, 8 (1983), pp. 303–342.

[10] R. S. MacKay, *Renormalization in Area Preserving Maps*, Ph.D. thesis, Princeton University, Princeton, NJ, 1982.

[11] R. S. MacKay, *Renormalization in Area-Preserving Maps*, World Scientific, Singapore, 1993.

[12] R. S. MacKay, *A renormalisation approach to invariant circles in area-preserving maps*, Phys. D, 7 (1983), pp. 283–300.

[13] R. de la Llave and N. Petrov, *Regularity of conjugacies between critical circles maps: An experimental study*, Experiment. Math., 11 (2002), pp. 219–241.

[14] J. D. Meiss, *Symplectic maps, variational principles, and transport*, Rev. Modern Phys., 64 (1992), pp. 795–848.

[15] C. Golé, *Symplectic Twist Maps*, World Scientific, River Edge, NJ, 2001.

[16] G. D. Birkhoff, *Surface transformations and their dynamical applications*, Acta Math., 43 (1922), pp. 1–119.

[17] J. Mather, *Nonexistence of invariant circles*, Ergodic Theory Dynam. Systems, 4 (1984), pp. 301–309.

[18] J. J. Mather and G. Forni, *Action minimizing orbits in Hamiltonian systems*, in Transition to Chaos in Classical and Quantum Mechanics (Montecatini Terme, 1991), J. Bellissard et al., eds., Springer, Berlin, 1994, pp. 92–186.

[19] A. Ya. Khinchin, *Continued Fractions*, Dover, Mineola, NY, 1997.

[20] B. V. Chirikov, *A universal instability of many-dimensional oscillator systems*, Phys. Rep., 52 (1979), pp. 264–379.

[21] R. S. MacKay and I. C. Percival, *Converse KAM: Theory and practice*, Comm. Math. Phys., 98 (1985), pp. 469–512.

[22] I. Jungreis, *A method for proving that monotone twist maps have no invariant circles*, Ergodic Theory Dynam. Systems, 11 (1991), pp. 79–84.

[23] R. De Vogelaere, *On the structure of symmetric periodic solutions of conservative systems, with applications*, in Contributions to the Theory of Nonlinear Oscillations, S. Lefschetz, ed., Princeton University Press, Princeton, NJ, 1958, pp. 53–84.

[24] J. M. Greene, *A method for determining a stochastic transition*, J. Math. Phys., 20 (1979), pp. 1183–1201.

[25] A. Apte, R. de la Llave, and N. P. Petrov, *Regularity of critical invariant circles of the standard nontwist map*, Nonlinearity, 18 (2005), pp. 1173–1187.

[26] J. A. Ketoja, *Renormalisation in a circle map with two inflection points*, Phys. D, 55 (1992), pp. 45–68.

[27] J. A. Ketoja and R. S. MacKay, *Rotationally-ordered periodic orbits for multiharmonic area-preserving twist maps*, Phys. D, 73 (1994), pp. 388–398.

[28] C. Falcolini and R. de la Llave, *Numerical calculation of domains of analyticity perturbation theories in the presence of small divisors*, J. Statist. Phys., 67 (1992), pp. 645–666.

[29] R. de la Llave and A. Olvera, *The obstruction criterion for non-existence of invariant circles and renormalization*, Nonlinearity, 19 (2006), pp. 1907–1937.
[30] H. Koch, *A renormalization group fixed point associated with the breakup of golden invariant tori*, Discrete Contin. Dyn. Syst., 11 (2004), pp. 881–909.
[31] H. Koch, *On the renormalization of Hamiltonian flows, and critical invariant tori*, Discrete Contin. Dyn. Syst., 8 (2002), pp. 633–646.
[32] R. de la Llave and R. P. Schafer, *Rigidity Properties of One Dimensional Expanding Maps and Applications to Renormalization*, manuscript, 1996.
[33] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.
[34] A. Olvera and C. Simó, *An obstruction method for the destruction of invariant curves*, Phys. D, 26 (1987), pp. 181–192.
[35] C. Falcolini and R. de la Llave, *A rigorous partial justification of Greene's criterion*, J. Statist. Phys., 67 (1992), pp. 609–643.
[36] R. S. MacKay, *Greene's residue criterion*, Nonlinearity, 5 (1992), pp. 161–187.
[37] T. Carletti, *The 1/2-complex Bruno function and the Yoccoz function: A numerical study of the Marmi-Moussa-Yoccoz conjecture*, Experiment. Math., 12 (2003), pp. 491–506.
[38] K. Fuchss, A. Wurm, A. Apte, and P. J. Morrison, *Breakup of shearless meanders and "outer" tori in the standard nontwist map*, Chaos, 16 (2006), 033120.
[39] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[40] A. Stirnemann, *Towards an existence proof of MacKay's fixed point*, Comm. Math. Phys., 188 (1997), pp. 723–735.
[41] H. G. E. Hentschel and I. Procaccia, *The infinite number of generalized dimensions of fractals and strange attractors*, Phys. D, 8 (1983), pp. 435–444.
[42] A. Rényi, *On measures of entropy and information*, in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, 1960, University of California Press, Berkeley, CA, 1961, pp. 547–561.
[43] S. Jaffard, *Multifractal formalism for functions Part* I: *Results valid for all functions*, SIAM J. Math. Anal., 28 (1997), pp. 944–970.
[44] S. Jaffard, *Multifractal formalism for functions Part* II: *Self-similar functions*, SIAM J. Math. Anal., 28 (1997), pp. 971–998.
[45] A. H. Osbaldestin and M. Y. Sarkis, *Singularity spectrum of a critical KAM torus*, J. Phys. A, 20 (1987), pp. L953–L958.
[46] N. Burić, M. Mudrinić, and K. Todorović, *Equivalent classes of critical circles*, J. Phys. A, 30 (1997), pp. L161–L165.
[47] N. Burić, M. Mudrinić, and K. Todorović, *Universal scaling of critical quasiperiodic orbits in a class of twist maps*, J. Phys. A, 31 (1998), pp. 7848–7854.
[48] J. Shi and B. Hu, *Crossover phenomena in the multifractal behavior of invariant circles*, Phys. Lett. A, 156 (1991), pp. 267–271.
[49] B. Hu and J. Shi, *Nonanalytic twist maps and Frenkel-Kontorova model*, Phys. D, 71 (1994), pp. 23–38.
[50] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, *Fractal measures and their singularities: The characterization of strange sets*, Phys. Rev. A (3), 33 (1986), pp. 1141–1151.
[51] B. R. Hunt, K. M. Khanin, Ya. G. Sinai, and J. A. Yorke, *Fractal properties of critical invariant curves*, J. Statist. Phys., 85 (1996), pp. 261–276.
[52] E. B. Vul, Ya. G. Sinai, and K. M. Khanin, *Feigenbaum universality and thermodynamic formalism*, Russian Math. Surveys, 39 (1984), pp. 1–40.
[53] Y. Yamaguchi and K. Tanikawa, *A remark on the smoothness of critical KAM curves in the standard mapping*, Progr. Theoret. Phys., 101 (1999), pp. 1–24.
[54] A. Arneodo, E. Bacry, and J. F. Muzy, *The thermodynamics of fractals revisited with wavelets*, Phys. A, 213 (1995), pp. 232–275.
[55] J. F. Muzy, E. Bacry, and A. Arneodo, *The multifractal formalism revisited with wavelets*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 245–302.

[56] A. Berretti and G. Gentile, *Bryuno function and the standard map*, Comm. Math. Phys., 220 (2001), pp. 623–656.
[57] A. Berretti and G. Gentile, *Scaling of the critical function for the standard map: Some numerical results*, Nonlinearity, 17 (2004), pp. 649–670.
[58] The Grace team, Grace homepage, http://plasma-gate.weizmann.ac.il/Grace/.
[59] P. Cvitanović, *Universality in Chaos*, Adam Hilger, Bristol, UK, 1989.
[60] G. D. Birkhoff, *Collected Works*, Vol. II, Dover, New York, 1968.
[61] R. S. MacKay and J. D. Meiss, *Hamiltonian Dynamical Systems*, Adam Hilger, Bristol, UK, 1993.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 988–1031
© 2008 Society for Industrial and Applied Mathematics

Chaotic Braided Solutions via Rigorous Numerics: Chaos in the Swift–Hohenberg Equation*

Jan Bouwe van den Berg† and Jean-Philippe Lessard‡

Abstract. We prove that the stationary Swift–Hohenberg equation has chaotic dynamics on a critical energy level for a large (continuous) range of parameter values. The first step of the method relies on a computer assisted, rigorous, continuation method to prove the existence of a periodic orbit with certain geometric properties. The second step is topological: we use this periodic solution as a skeleton, through which we braid other solutions, thus forcing the existence of infinitely many braided periodic orbits. A semiconjugacy to a subshift of finite type shows that the dynamics is chaotic.

Key words. forcing, chaos, pattern formation, computer-assisted proof, Swift–Hohenberg

AMS subject classifications. 92C15, 54H20, 37Mxx, 65P20

DOI. 10.1137/070709128

1. Introduction.

Finding analytic solutions of nonlinear, parameter dependent, ordinary differential equations (ODEs) is in general an extremely difficult task—most of the time impossible. The use of numerical techniques then becomes a useful path to adopt in order to understand the dynamics of a given nonlinear ODE. One may obtain insight not just through simulations; today the numerical output can also be used to rigorously extract coarse topological information from the systems, often revealing complicated dynamics. In particular, proving the existence of chaos in nonlinear dynamical systems in such a way has become quite popular (see [1, 14, 17, 26, 33, 35, 36]). One may interpret these results as forcing-type theorems, since a finite number of computable objects can be used to draw conclusions about the existence of infinitely many other objects. In this paper we propose a novel approach along those lines to prove existence of chaos for a class of problems with a special structure, namely, so-called second-order Lagrangian dynamical systems with the Twist property. This is a class of variational problems that lead to fourth-order ODEs. In particular, the well-known Swift–Hohenberg equation, one of standard models for pattern formation, falls into this class of problems.

A common feature of the proofs in [1, 17, 26, 35, 36] is the use of interval arithmetic to integrate the flow over sets and look for images of these rigorously integrated sets on some

*Received by the editors November 23, 2007; accepted for publication (in revised form) by H. Kokubu April 18, 2008; published electronically September 11, 2008. This research was supported in part by NWO grants 639.031.204 and 639.032.202, by NSF grant DMS 0511115, and by grants from D.O.E. and DARPA.
http://www.siam.org/journals/siads/7-3/70912.html

†Department of Mathematics, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands (janbouwe@few.vu.nl).

‡Department of Mathematics, Rutgers University, Hill Center-Busch Campus, 110 Frelinghuysen Rd, Piscataway, NJ 08854-8019, and Department of Mathematics, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands (lessard@math.rutgers.edu).

prescribed Poincaré sections. In contrast, our proof requires only proving the existence of *a single periodic solution* of a certain type. This will be done via validated continuation (cf. [15, 18]). One important advantage of this validated continuation method is that it becomes natural to prove the existence of chaos for a continuous range of parameter values.

We focus our attention on the Swift–Hohenberg equation, a fourth-order parabolic partial differential equation (PDE), traditionally written as

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

which is widely used as a model for pattern formation due to a finite wavelength instability, such as in Rayleigh–Bénard convection (see, e.g., [11, 32]). The onset of instability is at $\alpha = 0$. The Swift–Hohenberg equation was originally devised to describe the behavior of systems at, and just beyond, the onset of a supercritical finite wavelength instability, and it has since served as a universal model equation in the study of pattern formation, where the parameter α is not necessarily taken to be small (see, e.g., [9, 21, 28, 31]). Stationary profiles satisfy the ODE

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

which has a constant of integration, called the *energy*

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

which has been normalized so that, for $\alpha > 1$, the nontrivial homogeneous states $U = \pm\sqrt{\alpha - 1}$ have energy $E = 0$. The dynamics of (2) has been studied extensively, especially for small $\alpha > 0$, but many questions remain open for larger values of the parameter. Numerical simulations suggest chaotic behavior for most $\alpha > 0$, but this has so far not been verified rigorously. In particular, although both shooting methods (e.g., [2, 7, 8, 27]) and variational methods (e.g., [6, 22, 24, 29]) have been used extensively to study (2) and related fourth-order equations, they have not succeeded in revealing chaos for the Swift–Hohenberg ODE.

The energy level $E = 0$ is special in the sense that it is a singular energy level, and it contains the nontrivial homogeneous states $U = \pm\sqrt{\alpha - 1}$. Those equilibria are stable solutions of the PDE (1) for $\alpha > \frac{3}{2}$, and saddle-foci for the ODE (2) in the same parameter range. It is well known that saddle-foci may act as organizing centers for complicated dynamics [16, 23], and this inspires us to focus our attention on the dynamics in the energy level $E = 0$. Our main result is to establish rigorously that the Swift–Hohenberg ODE has chaotic dynamics in the energy level $E = 0$ for a large continuous range of parameter values.

Proposition 1. *The dynamics of the Swift–Hohenberg ODE* (2) *on the energy level* $E = 0$ *is chaotic for all* $\alpha \geq 2$.

Before we discuss the method of proof, let us comment on some generalizations of this result. First, the method is amenable to a larger class of equations, namely, second-order Lagrangians with the Twist property; see [3]. Second, the result is stable in the sense that for energy levels arbitrarily close to 0, chaos can be proved via a few adjustments of the arguments. We will comment on both generalizations when appropriate but will keep the

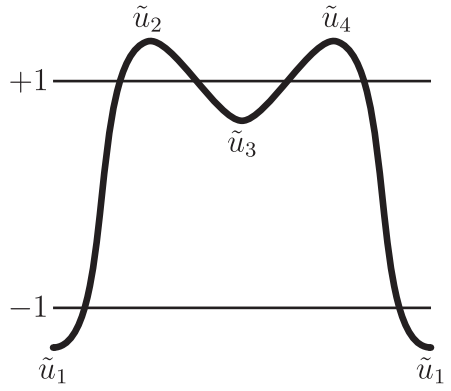


Figure 1. *Sketch of a periodic solution $\widetilde{u}$ satisfying the geometric properties $\mathcal{H}$.*

focus firmly on Proposition 1 to reduce technical details. Finally, the parameter range $\alpha \geq 2$ can be extended somewhat using our method but certainly not to cover the entire range $\alpha > 0$, as will be explained below.

We now turn to the method behind Proposition 1. Rather than working directly with (2), we first perform a change of coordinates that compactifies the parameter range as well as makes the notation more convenient. The new variables are

$$y = \frac{X}{\sqrt[4]{\alpha - 1}}, \qquad u(y) = \frac{U(X)}{\sqrt{\alpha - 1}}, \qquad \nu = \frac{2}{\sqrt{\alpha - 1}}. \tag{3}$$

The parameter range $\alpha \geq 2$ corresponds to $\nu \in (0, 2]$, and the differential equation becomes

$$-u'''' - \nu u'' + u - u^3 = 0, \tag{4}$$

while the expression for the energy is now

$$E = u'''u' - \frac{1}{2}u''^2 + \frac{\nu}{2}u'^2 + \frac{1}{4}(u^2 - 1)^2. \tag{5}$$

Equation (4), and variants with different nonlinearities, have been thoroughly investigated (see [10] and [28] and the references therein), but the parameter range under scrutiny here, namely, $\nu \geq 0$, remains much less explored than the range $\nu < 0$, mainly because most methods are somewhat less powerful for positive ν. An exception are the braid invariants introduced in [19], which are especially suited to deal with positive ν, and which we will indeed exploit in section 2.

In the method presented in this paper, chaos is forced by the existence of a single periodic solution $\widetilde{u}$ with specific geometric properties, much like a period-3 solution of an interval map implies chaos [25] (or a pseudo-Anosov braid in the context of surface homeomorphisms [34]). The periodic solution we are looking for needs to satisfy the following geometric properties (see also Figure 1):

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

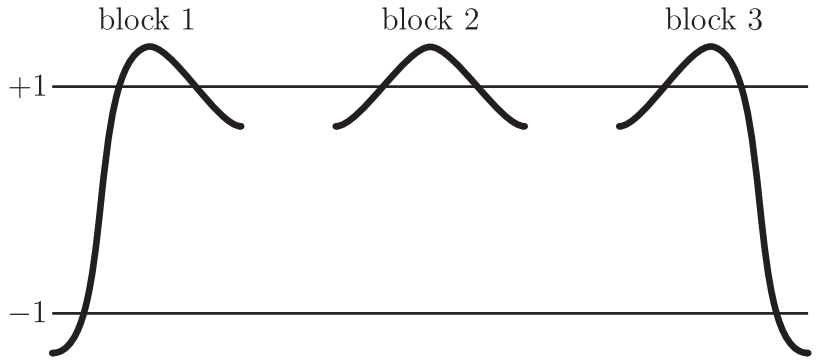


Figure 2. *Building blocks for the solutions that lead to the chaos of Theorem 2.*

Let the extrema $\widetilde{u}_i$ be attained in $\widetilde{y}_i$; then the last condition can be reformulated as

$$\widetilde{u}(\widetilde{y}_1 + y) = \widetilde{u}(\widetilde{y}_1 - y) \qquad \text{and} \qquad \widetilde{u}(\widetilde{y}_3 + y) = \widetilde{u}(\widetilde{y}_3 - y).$$

In particular, this implies that $\widetilde{u}_2 = \widetilde{u}_4$. We should note that condition H_4 is in fact not necessary for the results below to hold, but it simplifies the exposition.

As said before, such periodic solutions can be used to prove chaos when the equilibria $u = \pm 1$ are saddle-foci, i.e., when $\nu < \sqrt{8}$.

Theorem 2 (forcing). *Let $\nu \in [0, \sqrt{8})$, and suppose there exists a periodic solution $\widetilde{u}$ of (4) at the energy level $E = 0$ satisfying the geometric conditions $\mathcal{H}$. Then (4) 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 construction of the chaotic invariant set hinges on an application of Conley index theory for discretized braids [19] which will be adapted to our specific situation. The formulation in terms of discretized braids and the computation of the Conley index for well-chosen neighborhoods, whose construction involves the special periodic solution $\widetilde{u}$, is presented in section 2, together with all details of the proof of Theorem 2. Let us briefly discuss some intuition behind the result. The set of solutions of (4) that leads us to chaotic dynamics is obtained by putting the three building blocks in Figure 2 together. The order of the blocks should follow the intuition coming from Figure 2; i.e., blocks 1 and 2 may be followed by block 2 or 3, while block 3 can only be followed by block 1. The sequence of building blocks may be chosen arbitrarily as long as these rules are obeyed, and the different possibilities are sufficiently complicated to lead to chaos. The final technical step in proving chaos is then to find a semiconjugacy to a subshift of finite type; see section 2.1.

It is important to note that the only hypothesis that needs to be verified in order to prove the existence of chaos in (4) at $E = 0$ is the existence of the periodic solution $\widetilde{u}$ satisfying $\mathcal{H}$. This will be done via rigorous numerics, or computer assisted (interval arithmetic) calculations, together with a set of analytic estimates of the "tail" terms, i.e., the remainder terms not covered by the finite dimensional reduction. The construction leads to the existence of the periodic solution with the required geometric properties for a large range of parameter values.

Theorem 3 (rigorous computation). *For every $\nu \in [0,2]$ (4) has a periodic solution at energy level $E = 0$ satisfying the geometric properties $\mathcal{H}$.*

The change of variables (3) directly converts Theorems 2 and 3 into Proposition 1.

Numerical simulations suggest that although the parameter range in Theorem 3 (and hence Proposition 1) can be increased somewhat, the solution $\widetilde{u}$ with the described geometric behavior in fact disappears in a saddle-node bifurcation at some critical value $\nu_* > 2$ ($\nu_* \approx 2.03$). Hence, one has to find a different mechanism to force chaos if one wants to prove a similar result for the parameter range $\nu > 2$ (or, e.g., $\alpha \in (0,2]$).

We are going to employ Fourier transformation, a finite dimensional reduction, and a Newton-like operator, which we will prove is a contraction map via rigorous estimation of the tail term. This method has been successfully used in [15] and [18], but here we need to extend it considerably in three crucial aspects. First, the requirement $E = 0$ means that, besides satisfying the differential equation, the solution must obey an additional requirement. This means that the period of the periodic solution cannot be fixed a priori, and instead is another unknown. The extra equation leads, at a more technical level, to the need for better convolution estimates (see Appendix A), as will be explained later. Second, rigorous continuation is required in order to obtain results not for isolated values of ν (cf. [15, 18, 38]) but for the entire parameter interval $\nu \in [0,2]$. Note that in [13], a result about an entire parameter interval was also obtained, but at a much more computationally expensive price. Third, the geometric condition $\mathcal{H}$ needs to be verified rigorously to be able to combine the computational effort with the topological argument from Theorem 2, so that $\widetilde{u}$ forces chaotic dynamics.

We give a brief outline of the arguments here; full details can be found in section 3. Let $\frac{2\pi}{L}$ be the a priori unknown period of the solution $\widetilde{u}$, and let the local minima be attained at $y = 0$ and $y = \frac{\pi}{L}$. The symmetry condition H_4 implies that $u'(0) = 0$; hence evaluating the energy constraint (5) at $y = 0$ reduces (5) to

$$u''(0) = \frac{1}{\sqrt{2}}\big(u(0)^2 - 1\big), \tag{6}$$

where we have used that, since we look for solutions satisfying $\mathcal{H}$, we may assume that $u(0) < 1$ is a nondegenerate minimum; hence $u''(0) > 0$. In view of the symmetries, the Ansatz

$$u(y) = a_0 + 2\sum_{l=1}^{\infty} a_l \cos(lLy)$$

is natural, and it reduces (4) to (with $a_{-k} \equiv a_k$)

$$g_k \overset{\text{def}}{=} \left[1 + \nu L^2 k^2 - L^4 k^4\right] a_k - \sum_{\substack{k_1+k_2+k_3=k \\ k_i \in \mathbb{Z}}} a_{k_1} a_{k_2} a_{k_3} = 0 \qquad \text{for all } k \geq 0,$$

while (6) becomes

$$e \overset{\text{def}}{=} -2L^2 \sum_{l=1}^{\infty} l^2 a_l - \frac{1}{\sqrt{2}} \left[a_0 + 2\sum_{l=1}^{\infty} a_l\right]^2 + \frac{1}{\sqrt{2}} = 0.$$

The first sum in the above expression requires a faster than cubic decay of the sequence a_l, which leads to the need for improved convolution estimates (proved in Appendix A), since the bounds previously used in [12, 13, 14, 15, 18] turn out to be impractical in the current context. With the notation $x = (L, a_0, a_1, a_2, \ldots)$ and $f(x, \nu) = (e, g_0, g_1, g_2, \ldots)$, we are thus looking for a solution of $f(x, \nu) = 0$. In this formulation, and using a finite dimensional reduction, we may now use the classical predictor-corrector algorithm for following a continuous branch. However, we need to add a validation step (see [15]) and interval arithmetic to make this into a mathematically rigorous proof. We stress that the interval arithmetic, although necessary and time consuming, is of much less practical importance than the analytic error estimates due to the finite dimensional reduction. Using this finite dimensional reduction, we can, with the help of a computer, for fixed $\bar{\nu}$ find an approximate solution $\bar{x}$ of $f(x, \bar{\nu}) = 0$, as well as an approximate solution $\dot{x}$ of $\partial_x f(\bar{x}, \bar{\nu})\dot{x} + \partial_\nu f(\bar{x}, \bar{\nu}) = 0$. We also compute an approximate inverse J of $\partial_x f(\bar{x}, \bar{\nu})$. Via rigorous estimates on the remainder terms we show that

$$T(x, \Delta_\nu) = x - Jf(x, \bar{\nu} + \Delta_\nu)$$

is a contraction map on a small ball around $\bar{x} + \Delta_\nu \dot{x}$ in an appropriate Banach space for all sufficiently small Δ_ν. Repeating this for many small parameter intervals leads us to the existence of a symmetric periodic solution for all $\nu \in [0, 2]$, and, once we have verified the geometric conditions $\mathcal{H}$, to a proof of Theorem 3. It should be clear from the reformulation above that it is quite natural to do parameter continuation. In fact, we expect to find a *continuous* branch of solutions parametrized by ν. Although continuity is easy to verify for each continuation step separately, and indeed this property is used in section 4 to reduce the number of computations required, we do not need a globally (i.e., for all $\nu \in [0, 2]$) continuous branch for our proof. We refer the reader to [5] for a general overview of obtaining globally continuous branches of solutions using these techniques in the more general context of pseudo–arc-length continuation.

Let us comment on further developments. We prove here that the Poincaré return map from Theorem 2, which is in fact the map that follows solutions from one local minimum to the next (see section 2.1), has topological entropy of at least 0.48. It is possible to obtain better bounds on the entropy, still based on the existence of the periodic solution $\widetilde{u}$, using the $u \to -u$ mirror symmetry, but we will not go into the details here. Furthermore, an analysis along the lines of [4] may lead to statements about the size of the attractor for boundary value problems associated to the PDE (1), all enabled by the rigorous establishment of a single periodic solution with the geometric properties $\mathcal{H}$. This is currently under investigation.

The outline of this paper is as follows. As already noted, it suffices to prove Theorems 2 and 3, which together imply Proposition 1. The forcing theorem, Theorem 2, is proved in section 2. The method that leads to the existence of the special solution described in Theorem 3 is explained in detail in section 3. Furthermore, section 4 deals with the verification of the geometric properties $\mathcal{H}$. The analytic estimates in sections 3 and 4 lead to an algorithm, in which a finite set of inequalities needs to be checked, which is left to a computer program (with interval arithmetic). The estimates together with the output from the computer program prove Theorem 3. The appendix contains some general and rather sharp convolution estimates needed for the analytic bounds on the remainder terms.

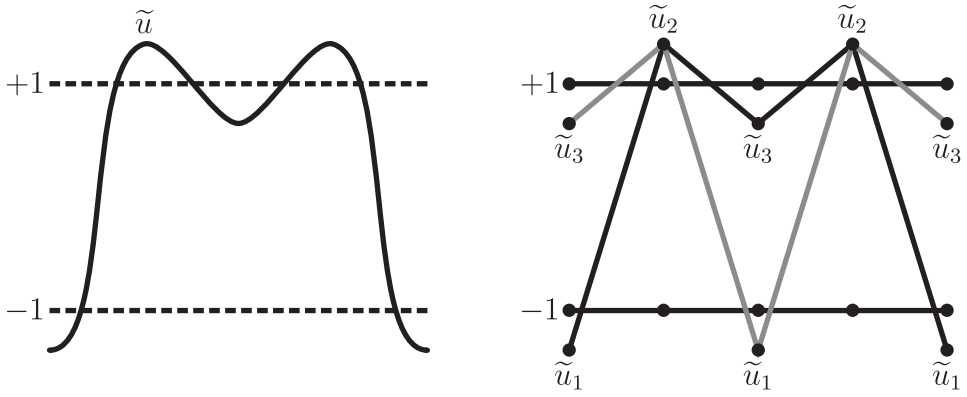


Figure 3. *Left: Sketch of the solution $\widetilde{u}$. Right: Discretized version $\{\widetilde{u}_i\}_{i=1}^4$ and a shift $\{\widetilde{u}_{i+2}\}_{i=1}^4$.*

Additional files come with the paper. The MATLAB functions *SH_continuation.m*, *SH_rigorous_continuation.m*, *SH_mesh_generator.m*, *SH_geometric_properties.m*, and *SH_run_proof.m* (70912_01.zip [14KB]) rigorously verify Procedures 16 and 21. Furthermore, the accompanying animation (70912_02.gif [1.32MB]) shows the evolution of the rigorously computed periodic solution, as we move the parameter from $\nu = 0$ to $\nu = 2$. In section 3.1, details about the computer implementation of the rigorous verification of Procedures 16 and 21 are given, together with a brief description of the MATLAB functions.

2. Forcing theorem. In this section we assume, as in Theorem 2, that $\nu \in [0, \sqrt{8})$ and that there exists a periodic solution $\widetilde{u}$ of (4) at $E = 0$ with the geometric properties $\mathcal{H}$. The idea behind the proof is that we code periodic solutions u of (4) at $E = 0$ by their extrema (see Figure 3). This leads to a discretization of the problem. If $u' = 0$, then, by (5), $u'' = \pm\frac{1}{\sqrt{2}}(u^2 - 1)$. Hence, extrema are nondegenerate except at $u = \pm 1$, and we are going to avoid those values, so we may for the moment assume all extrema to be nondegenerate. We denote the sequence of extrema of u by $\{u_i\}_{i\in\mathbb{Z}}$, where u_i represents a local minimum for odd i and a local maximum for even i (see also Figure 3).

For $\nu \geq 0$ our system is a so-called Twist system on $E = 0$, as defined and proved in [3]. The fact that the energy level is singular (contains equilibria) leads to some technical complications, but we shall overcome them relatively easily in our present context. We can therefore use the braid theory for discretized parabolic equations from [19]. The results on Twist systems that are needed in this paper are summarized in the next lemma; its proof can be found in [3] and [19, Thm. 37].

Lemma 4. *Let $\nu \geq 0$. There exist functions $\mathcal{R}_i \in C^1(\Omega_i; \mathbb{R})$ with domains*

$$\Omega_i = \left\{(u, v, w) \in \mathbb{R}^3 \;\middle|\; (-1)^i u < (-1)^i v,\ (-1)^i w < (-1)^i v,\ \text{and } u, v, w \neq \pm 1\right\},$$

with the following properties:

(a) *$\mathcal{R}_{i+2} = \mathcal{R}_i$, so there are really only two different functions in play.*

(b) *$(\mathcal{R}_i)_{i\in\mathbb{Z}}$ is a* parabolic recurrence relation*; i.e., it has the monotonicity property*

$$\partial_{u_{i-1}}\mathcal{R}_i > 0 \qquad \text{and} \qquad \partial_{u_{i+1}}\mathcal{R}_i > 0. \tag{7}$$

(c) *Define*

$$\Omega = \big\{(u_i)_{i\in\mathbb{Z}} \;\big|\; (u_{i-1}, u_i, u_{i+1}) \in \Omega_i \text{ for all } i\big\}.$$

A sequence $(u_i)_{i\in\mathbb{Z}} \in \Omega$ *solves*

$$\mathcal{R}_i(u_{i-1}, u_i, u_{i+1}) = 0 \qquad \text{for all } i$$

if and only if *it corresponds to solution of* (4) *at* $E = 0$ *with nondegenerate extrema* u_i. *An analogous statement holds for semi-infinite sequences* $(u_i)_{i\geq i_0}$ *which solve* $\mathcal{R}_i(u_{i-1}, u_i, u_{i+1}) = 0$ *for all* $i \geq i_0 + 1$.

The shapes of the domains Ω_i reflect the fact that minima are preceded and followed by maxima (and vice versa). The lemma thus implies that instead of looking for (periodic) solutions of (4) at $E = 0$ with nondegenerate extrema, we may try to find (periodic) sequences u_i that solve $\mathcal{R}_i = 0$. We remark that Lemma 4 (and the method in this paper) extends to a more general class of fourth-order ODEs, namely, those derived from a second-order Lagrangian satisfying the Twist property; see [3]. The Twist property, in essence, means that there are unique monotone solutions of the ODE between extremal values u_i and u_{i+1}.

We want to exploit the fact that the energy level $E = 0$ contains the equilibria $u = \pm 1$. However, these solutions do not correspond to a proper sequence of extrema. The linearization around the equilibria is going to help us resolve this issue. Namely, for $-\sqrt{8} < \nu < \sqrt{8}$ the equilibria ± 1 are saddle-foci, and this leads to the following fact (formulated here for the equilibrium $+1$).

Lemma 5. *Let* $-\sqrt{8} < \nu < \sqrt{8}$. *For any* $\varepsilon > 0$ *there exists a sequence* $\{u_i^\varepsilon\}_{i=1}^\infty$,

$$0 < (-1)^i(u_i^\varepsilon - 1) < \varepsilon,$$

which satisfies

$$\mathcal{R}_i(u_{i-1}^\varepsilon, u_i^\varepsilon, u_{i+1}^\varepsilon) = 0 \qquad \text{for } i \geq 2.$$

Notice that we do not claim that $\mathcal{R}_1(u_0^\varepsilon, u_1^\varepsilon, u_2^\varepsilon) = 0$; we did not even define u_0^ε.

Proof. The idea is that the u_i^ε are the extrema of an orbit in the stable manifold of $+1$, which is contained in the energy level $E = 0$. That $u_i^\varepsilon - 1$ alternates sign follows from the fact that the equilibrium $+1$ is a saddle-focus: it is easy to check that for $-\sqrt{8} < \nu < \sqrt{8}$ the *linearized* equation (i.e., $u = 1 + v$ with $v'''' + \nu v'' + 2v + O(v^2) = 0$) has solutions of the form

$$1 + Ce^{-\lambda_r x}\cos(\lambda_i x + \phi),$$

with C and ϕ arbitrary (with $\lambda_r, \lambda_i > 0$ depending on ν). In particular, the stable manifold of the linearized problem intersects the hyperplane $\{u' = 0\}$ in the line

$$\ell = \big\{(1 + v, 0, -\sqrt{2}\,v, 2\sqrt{2}\lambda_r v) \;\big|\; v \in \mathbb{R}\big\}.$$

For the nonlinear equation we need to invoke the stable manifold theorem. Let us denote the stable manifold by $W^s(+1)$ and the local stable manifold by $W^s_{\text{loc}} = W^s(+1) \cap B_{\varepsilon_0}(+1)$ for $\varepsilon_0 > 0$ chosen sufficiently small (so that the following arguments hold). By the stable manifold theorem, the local stable manifold intersects the hyperplane $\{u' = 0\}$ in a curve tangent to ℓ, and thus

$$W^s_{\text{loc}} \cap \{u' = 0\} \subset \big\{(1 + v, 0, -\sqrt{2}\,v + O(v^2), 2\sqrt{2}\lambda_r v + O(v^2)) \;\big|\; v \in \mathbb{R}\big\} \cap B_{\varepsilon_0}(+1).$$

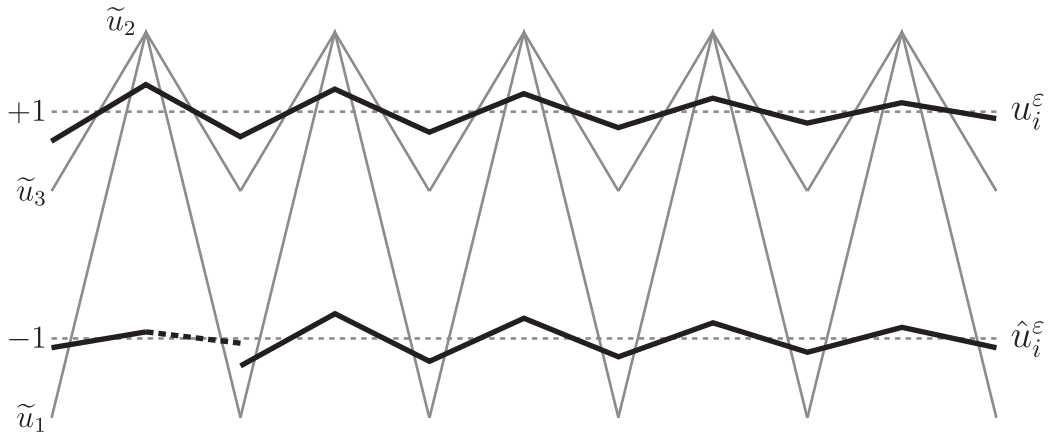


Figure 4. *The "up-down" setting including the oscillating tails in the local stable manifolds of* ± 1.

In particular, for ε_0 sufficiently small, for solutions u in the local stable manifold it holds that if $u' = 0$ and $u > 1$, then $u'' < 0$, whereas if $u' = 0$ and $u < 1$, then $u'' > 0$. This shows that all solutions in the local stable manifold have successive extrema on alternating sides of $u = 1$. Now pick one orbit in the local stable manifold and denote its extrema by $\{u_i^{\varepsilon_0}\}_{i=1}^{\infty}$, with $u_1^{\varepsilon_0}$ a local minimum. Then $0 < (-1)^i(u_i^{\varepsilon_0} - 1) < \varepsilon_0$, and $u_i^{\varepsilon_0} \to 1$ as $i \to \infty$ (exponentially fast, in fact). For $\varepsilon < \varepsilon_0$ we may choose $u_i^\varepsilon = u_{i+2n(\varepsilon)}^{\varepsilon_0}$ for some $n(\varepsilon) \in \mathbb{N}$ sufficiently large. ■

We can use the symmetry to obtain an analogous result near -1. To be explicit, $\bar{u}_i^\varepsilon = -u_{i+1}^\varepsilon$ satisfies $0 < (-1)^i(\bar{u}_i^\varepsilon + 1) < \varepsilon$. For "technical" reasons to become clear later, we will need to shift this solution, modulo the $2p$-periodicity:

$$\hat{u}_i^\varepsilon = \bar{u}_{i-2 \bmod 2p}^\varepsilon. \tag{8}$$

In fact, we have not yet chosen the period of the sequences/solutions under scrutiny, but we will do so shortly. See Figure 4 for an illustration of u_i^ε and $\hat{u}_i^\varepsilon$. Notice that $\hat{u}_i^\varepsilon$ does not "close" at $i = 3$. Nevertheless, this will not stop us from putting it to use below.

To study solutions of $\mathcal{R}_i = 0$ we introduce an artificial new time variable s and consider $u_i(s)$ evolving according to the flow $u_i' = \mathcal{R}_i$. Clearly, we want to find stationary points, and we are going to construct isolating neighborhoods for the flow (any $p \in \mathbb{N}$)

$$\frac{du_i}{ds} = \mathcal{R}_i(u_{i-1}, u_i, u_{i+1}), \qquad i = 1, \dots, 2p, \tag{9}$$

where we identify $u_0 = u_{2p}$ and $u_1 = u_{2p+1}$. The monotonicity property (7) implies that this flow has the decreasing intersection-number property: if two solutions are represented as piecewise linear functions (as in most of the figures), then the number of intersections can only decrease as time s increases.

To build the isolating neighborhoods for (9), consider first the solution $\widetilde{u}$ with geometric properties $\mathcal{H}$. Since it is a periodic solution of (4) at $E = 0$, it follows from Lemma 4(c) that

(recall that $\widetilde{u}_2 = \widetilde{u}_4$)

$$\begin{aligned}\mathcal{R}_1(\widetilde{u}_2, \widetilde{u}_1, \widetilde{u}_2) &= 0,\\ \mathcal{R}_2(\widetilde{u}_1, \widetilde{u}_2, \widetilde{u}_3) &= 0,\\ \mathcal{R}_1(\widetilde{u}_2, \widetilde{u}_3, \widetilde{u}_2) &= 0,\\ \mathcal{R}_2(\widetilde{u}_3, \widetilde{u}_2, \widetilde{u}_1) &= 0.\end{aligned}$$

Next, we choose

$$\varepsilon = \frac{1}{2}\max\{-1 - \widetilde{u}_1,\ \widetilde{u}_2 - 1,\ 1 - \widetilde{u}_3\}.$$

Although not strictly necessary for understanding the arguments that follow, it is worth mentioning that in the setting of discretized braids described in [19], we are going to use a skeleton consisting of four strands (see Figure 3, right, and Figure 4): $v_i^1 = \widetilde{u}_i$ and $v_i^2 = \widetilde{u}_{i+2}$, and $v_i^3 = u_i^\varepsilon$ and $v_i^4 = \hat{u}_i^\varepsilon$. To be precise, both v^1 and v^2 are defined for all $i \in \mathbb{Z}$ and are 4-periodic. Furthermore, v^3 is defined for all $i \geq 1$ (though not periodic), while v^4 is defined for $i = 0, \ldots, 2p+1$, with $v_0^4 = v_{2p}^4$ and $v_{2p+1}^4 = v_1^4$. All four strands satisfy

$$\mathcal{R}_i(v_{i-1}, v_i, v_{i+1}) = 0 \qquad \text{for } i = 1, \ldots, 2p,$$

with the *exception* of v^3 at $i = 1$ and v^4 at $i = 2, 3$. Below we will make sure that these points do not come into play in the construction of isolating neighborhoods.

Consider a finite, but arbitrarily long, sequence

$$\mathbf{a} = \{\mathbf{a}_j\}_{j=1}^N, \qquad \mathbf{a}_j \geq 2. \tag{10}$$

Let the period of the sequences (u_i) be $p = \sum_{j=1}^N \mathbf{a}_j$. Now that p is fixed, the meaning of $\hat{u}_i^\varepsilon$ in (8) is settled. We define the set of partial sums

$$\mathcal{A} = \left\{ \sum_{j=1}^{n-1} \mathbf{a}_j \;\middle|\; n = 1, \ldots, N \right\}.$$

Note that $0 \in \mathcal{A}$. Now define the set (neighborhood) $U_\mathbf{a} \subset \mathbb{R}^{2p}$ as a product of intervals

$$U_\mathbf{a} = \{u_i \in I_i,\ i = 1, \ldots, 2p\},$$

where the intervals are given by

$$\begin{aligned} I_i &= [u_i^\varepsilon, \widetilde{u}_2] && \text{if } i \text{ is even;}\\ I_i &= [\widetilde{u}_3, u_i^\varepsilon] && \text{if } i \text{ is odd and } \frac{i-1}{2} \notin \mathcal{A};\\ I_i &= [\widetilde{u}_1, \hat{u}_i^\varepsilon] && \text{if } i \text{ is odd and } \frac{i-1}{2} \in \mathcal{A}. \end{aligned}$$

Notice that $U_\mathbf{a}$ is contained in the domain of definition Ω of $\mathcal{R}$, since ± 1 are not in any of the intervals I_i, and the "up-down" criterion is also satisfied (the intervals I_i for odd i lie strictly

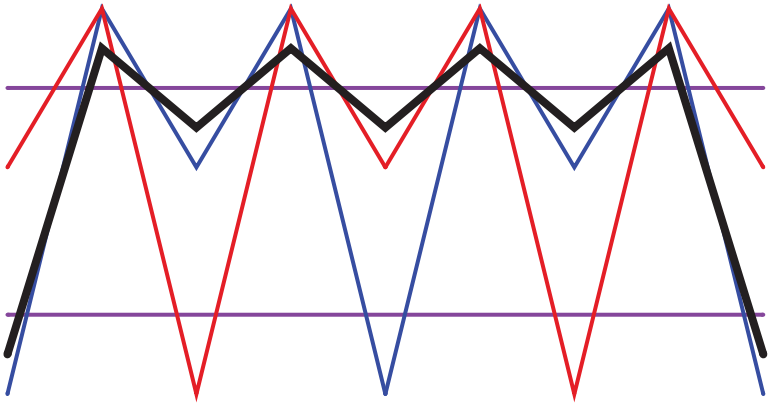


Figure 5. *The thin colored lines denote the skeleton, where we represent u^ε and $\hat{u}^\varepsilon$ by constants for convenience. The thick black lines represent the free strand, which is in $U_{\mathbf{a}}$ for $\mathbf{a} = (4)$, $p = 4$. One can check that on the boundary of $U_{\mathbf{a}}$ the number of crossings with at least one of the skeletal strands decreases; hence the flow points outward on the boundary $\partial U_{\mathbf{a}}$.*

below the ones for even i). It is useful to review the intervals in the context of Figure 4 and to look at Figure 6 for an example with $\mathbf{a} = 243$.

We now prove that every $U_{\mathbf{a}}$ contains an equilibrium of (9), still under the assumption that $\widetilde{u}$ is a periodic solution of (4) at $E = 0$ with geometric properties $\mathcal{H}$.

Lemma 6. *For any* $\mathbf{a}$ *defined in* (10) *the set* $U_{\mathbf{a}}$ *contains an equilibrium, corresponding to a periodic solution of* (4) *on* $E = 0$.

Proof. The case $\mathbf{a} = 222\ldots2 = 2^q$ is exceptional, since the point $(\widetilde{u}_1, \widetilde{u}_2, \widetilde{u}_3, \widetilde{u}_2)^q$, corresponding to the periodic solution $\widetilde{u}$, lies on the boundary of (the closed set) $U_{222\ldots2}$.

For all other $\mathbf{a}$, the corresponding solution lies in the interior of $U_{\mathbf{a}}$. The proof follows from the general theory in [19]. Namely, suppose from now on that $\mathbf{a} \neq 222\ldots2$. Then $U_{\mathbf{a}}$ is an isolating block for the flow in the sense of the Conley index, and the flow points outward everywhere on the boundary. This is relatively easy to check on the codimension 1 boundaries of $U_{\mathbf{a}}$, i.e., exactly one of the u_i lies on the boundary of I_i, while all the others are in the interior; for the higher codimension boundaries, see [19]. For the following arguments it may be helpful for the reader to consult Figure 5.

Let us consider one of the sides of the $2p$-cube $U_{\mathbf{a}}$. For example, $u_i = u_i^\varepsilon$ for some even i; i.e., u_i is on the lower boundary of I_i. Since $u_{i-1} < u_{i-1}^\varepsilon$, and $u_{i+1} < u_{i+1}^\varepsilon$ on the codimension 1 piece of this side, we infer from the monotonicity (7) that

$$\frac{du_i}{ds} = \mathcal{R}_i(u_{i-1}, u_i, u_{i+1}) < \mathcal{R}_i(u_{i-1}^\varepsilon, u_i^\varepsilon, u_{i+1}^\varepsilon) = 0.$$

Hence the flow points outward. And when $u_i = \widetilde{u}_2$ for some even i (the upper boundary point of I_i), then, since $\mathbf{a}_j \geq 2$, either $\frac{i-1}{2} \notin \mathcal{A}$ or $\frac{i+1}{2} \notin \mathcal{A}$, or both. Let us consider the case $\frac{i-1}{2} \notin \mathcal{A}$ (the other case is analogous); then $u_{i-1} > \widetilde{u}_3$ and $u_{i+1} > \widetilde{u}_1$ (assuming again that $(u_i)_{i=1}^{2p}$ is in a codimension 1 boundary). Hence

$$\frac{du_i}{ds} = \mathcal{R}_i(u_{i-1}, u_i, u_{i+1}) > \mathcal{R}_2(\widetilde{u}_3, \widetilde{u}_2, \widetilde{u}_1) = 0,$$

and thus the flow points outward again. All other (codimension 1) boundaries can be dealt with analogously.

We should note that, by construction of the neighborhoods in combination with the definition of u^ε and $\hat{u}^\varepsilon$, we avoid the three points where the skeleton does not satisfy the recurrence relation. In particular, no part of the boundary $\partial U_{\mathbf{a}}$ lies in the hyperplanes $u_1 = u_1^\varepsilon$ (since $u_1 < -1$) or $u_2 = \hat{u}_2^\varepsilon$ or $u_3 = \hat{u}_3^\varepsilon$ (since $\mathbf{a}_1 \geq 2$; hence $u_2, u_3 > \widetilde{u}_3$). We leave the remaining details to the reader.

As said before, for the higher codimension boundaries we refer the reader to [19, Prop. 11, Thm. 15]. We can now conclude that since $U_{\mathbf{a}}$ is a $2p$-cube and the flow points outward on $\partial U_{\mathbf{a}}$, its Conley index is homotopic to a $2p$-sphere, and the nonvanishing of its Euler characteristic implies that there has to be a stationary point in the interior of $U_{\mathbf{a}}$ [19, Lem. 36] corresponding to a solution of (4) by Lemma 4(c). ■

Remark 7. A similar result holds for energy levels close to $E = 0$. The main difference is that the infinite sequences u_i^ε, consisting of the extrema of a solution in the stable manifold of 1, is not available in energy levels $E \neq 0$. Nevertheless, an analogous construction can be set up for E sufficiently close to 0, *provided* the sequences $\mathbf{a} = \{\mathbf{a}_j\}_{j=1}^N$ are now chosen with $2 \leq \mathbf{a}_j \leq N_E < \infty$, where N_E tends to infinity as E approaches 0. For $E > 0$ (and small) an additional difficulty arises, because the Twist property (and hence Lemma 4) does not immediately follow from the results in [3]. However, perturbation methods can be used to show that the Twist property persists for small $E > 0$, at least away from the equilibria $u = \pm 1$. The details are beyond the scope of the current paper.

Remark 8. When the symmetry condition H_4 is dropped, then the definition of $U_{\mathbf{a}}$ needs to be modified to accommodate for the fact that $\widetilde{u}_2 \neq \widetilde{u}_4$. The shape of the set $U_{\mathbf{a}}$ will be more complicated than just a single $2p$-cube. Namely, one needs to consider the appropriate discretized braid class; see [19]. Nevertheless, the results in [19] show that the Conley index of this braid class is again homotopic to a sphere, and Lemma 6 and the results in the next section remain valid in the nonsymmetric setting.

2.1. Topological entropy. In this section we construct a semiconjugacy from a Poincaré section of the flow to a subshift of finite type with positive entropy, and thereby we finish the proof of Theorem 2. This process involves a couple of somewhat technical steps.

First, we look at an alternative coding, which is more convenient when examining the entropy. We extend any sequence $\mathbf{a} = \{\mathbf{a}_j\}_{j=1}^N$, $\mathbf{a}_j \geq 2$, periodically to a bi-infinite sequence: $\mathbf{a}_{j+N} = \mathbf{a}_j$. To such a periodic sequence we associate a bi-infinite sequence of 0's and 1's:

$$\mathbf{b} = \cdots 01^{\mathbf{a}_{-2}-1}01^{\mathbf{a}_{-1}-1}01^{\mathbf{a}_0-1}.01^{\mathbf{a}_1-1}01^{\mathbf{a}_2-1}0\cdots .$$

Notice, in particular, that $\mathbf{b}_0 = 0$. This coding is also indicated in Figure 6. It is not hard to see that the sequences $\mathbf{b}$ are in the symbol space Σ_B generated by the adjacency matrix

$$B = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

We now interpret $U_{\mathbf{a}}$ as an *infinite* product of intervals, and in terms of $\mathbf{b}$ the intervals making up the neighborhood $U_{\mathbf{a}} = U_{\mathbf{b}}$ are given by ($i \in \mathbb{Z}$)

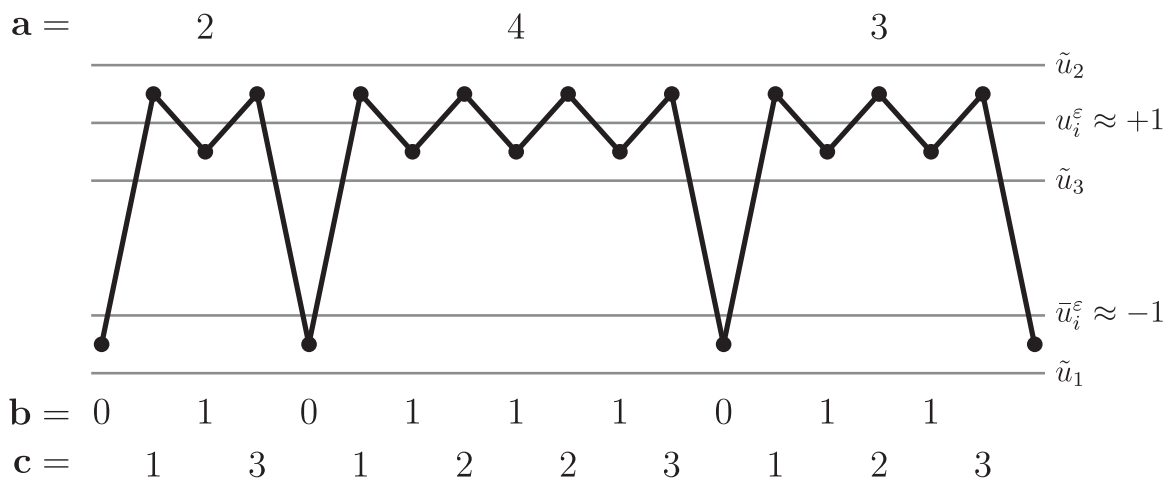


Figure 6. *A schematic example of a pattern in $U_{\mathbf{a}}$. At the top is the coding* $\mathbf{a}$*, and below are the corresponding codings* $\mathbf{b}$ *and* $\mathbf{c}$*, which are used in the discussion of the entropy.*

$$
\begin{aligned}
I_i &= [u_i^\varepsilon, \widetilde{u}_2] && \text{if } i \text{ is even,}\\
I_i &= [\widetilde{u}_3, u_i^\varepsilon] && \text{if } i \text{ is odd and } b_{\frac{i+1}{2}} = 0,\\
I_i &= [\widetilde{u}_1, \hat{u}_i^\varepsilon] && \text{if } i \text{ is odd and } b_{\frac{i+1}{2}} = 1.
\end{aligned}
$$

Let $u_{\mathbf{a}} = u_{\mathbf{b}}$ be the periodic solutions of (4) at $E = 0$ corresponding to the stationary points in $U_{\mathbf{a}} = U_{\mathbf{b}}$, which was found in Lemma 6.

An arbitrary periodic sequence in Σ_B might not have a $b_0 = 0$, but for any periodic sequence $\mathbf{b} \neq 1^\infty$ in Σ_B we can find a periodic solution of (4) at $E = 0$ in $U_{\mathbf{b}}$ by using an appropriate shift.

It is now time to set up the semiconjugacy from a Poincaré map of the flow to the shift-map σ on Σ_B. It is not difficult to check that the sets $\mathcal{C} \subset \{E = 0\} \subset \mathbb{R}^4$ of all orbits, varying over all possible periodic $\mathbf{a}$'s or $\mathbf{b}$'s, is uniformly bounded. Taking the closure of this set, we obtain a compact invariant set $\overline{\mathcal{C}}$ for the ODE. Notice that it may include the equilibrium solution $u \equiv 1$.

Next we choose a Poincaré section. The energy level $E = 0$ is a three-dimensional subset of the phase space $\mathbb{R}^4$. A local minimum in $E = 0$ is defined by the values of u and u''', since $u'' = \frac{1}{\sqrt{2}}|u^2 - 1|$. The Poincaré section is therefore defined as the two-dimensional subset

$$\mathcal{P} = \left\{ (u, 0, \tfrac{1}{\sqrt{2}}|u^2 - 1|, u''') \;\middle|\; u, u''' \in \mathbb{R} \right\},$$

and the return map $\mathcal{T} : \mathcal{P} \to \mathcal{P}$ follows solutions from one minimum to the next. For the special point $\pm\mathbf{1} = (\pm 1, 0, 0, 0) \in \mathcal{P}$ we define $\mathcal{T}(\pm\mathbf{1}) = \pm\mathbf{1}$.

Lemma 9. *For any $\nu > -\sqrt{8}$ the Poincaré return map $\mathcal{T}$ is well defined on $\mathcal{P}$, and $\mathcal{T}$ is continuous.*

Proof. That $\mathcal{T}$ is well defined follows from the fact that any solution $u \not\equiv 1$ in $E = 0$ has infinitely many extrema, which was proved in, e.g., [28, Lem. 3.1.2]. That $\mathcal{T}$ is continuous

away from $\pm\mathbf{1}$ follows from the methods in the proof of Lemmas 3.1.3 and 3.1.4 in [28]. There, only symmetric settings were considered, but the method prevails in the nonsymmetric setting. Continuity of $\mathcal{T}$ at $\pm\mathbf{1}$ follows from the fact that the equilibria are either saddle-foci ($|\nu| < \sqrt{8}$) or centers ($\nu \geq \sqrt{8}$). Roughly speaking, the closer an orbit starts to ± 1, the more extrema near ± 1 it has. It is not hard to deduce continuity at $\pm\mathbf{1}$ from this. ■

By construction and Lemma 9, the return map $\mathcal{T}$, defined on $\mathcal{P}$, has a compact invariant set $\Lambda = \mathcal{P} \cap \overline{\mathcal{C}}$. For $\lambda \in \Lambda$, we denote by $u^\lambda(y)$ the solution with *initial data* λ.

Lemma 10. *Let $\lambda \in \Lambda$, and let u^λ be the associated solution of (4). Suppose $u^\lambda \not\equiv 1$. Then u^λ has only nondegenerate extrema, with the maxima in $(1, \widetilde{u}_2]$ and the minima either in $[\widetilde{u}_3, 1)$ or in $[\widetilde{u}_1, -1)$.*

Proof. We write $u = u^\lambda$. All solutions in $\overline{\mathcal{C}}$ can be approximated arbitrarily closely (on compact intervals) by the periodic solutions found in Lemma 6. Nondegenerate extrema persist, whereas degenerate extrema ($u' = u'' = 0$) in $E = 0$ must lie on the lines $u = \pm 1$. Hence, the bounds on the extrema follow immediately from the definitions of $U_{\mathbf{a}}$ and I_i, once we have excluded degenerate extrema, i.e., inflection points on $u = \pm 1$.

To rule out inflection points we argue by contradiction (see also [28, Lem. 3.1.5]). Suppose $u \in \overline{\mathcal{C}}$ has an inflection point, say, $u'(0) = u''(0) = 0$. Hence $u(0) = \pm 1$ and $u'''(0) \neq 0$ (if $u'''(0) = 0$, then $u \equiv 1$ by uniqueness of the initial value problem). We consider the case $u(0) = 1$ and $u'''(0) > 0$; the other three cases are ruled out in an analogous manner. Let $\{u_n\}_{n=1}^\infty$ be a sequence of periodic solutions found in Lemma 6, such that $u_n \to u$ in C^3 on any bounded interval. Then by the implicit function theorem, for large enough n, there exist points y_n such that $\lim_{n\to\infty} y_n = 0$ and $u_n''(y_n) = 0$. We know that $u_n'(y_n) \neq 0$, since the u_n have only nondegenerate extrema.

We now consider two cases: either $u_n'(y_n) > 0$ or $u_n'(y_n) < 0$ for infinitely many $n \in \mathbb{N}$. In the former case we argue as follows. Taking a subsequence we may assume that $u_n'(y_n) > 0$ for all n. We conclude from $E = 0$ and $u_n''(y_n) = 0$ that

$$u_n'''(y_n) + \frac{\nu}{2} u_n'(y_n) = -\frac{(u_n(y_n)^2 - 1)^2}{4u_n'(y_n)} \leq 0.$$

Taking the limit $n \to \infty$ in the above inequality leads to $u'''(0) + \frac{\nu}{2} u'(0) \leq 0$, which contradicts the assumption that u has an inflection point at $y = 0$ with $u'''(0) > 0$.

In the latter case, we may assume that $u_n'(y_n) < 0$ for all n. Since in the inflection point $u'''(0) > 0$, it follows that $u'(y) > 0$ for $y \neq 0$ sufficiently small. This means that for n large enough there are sequences $y_n^1 < y_n < y_n^2$ of local maxima and minima of u_n, respectively, such that $y_n^{1,2} \to 0$ as $n \to \infty$. Since the periodic solutions u_n have their extrema on alternating sides of $+1$ by construction, there is a sequence $y_n^3 \in (y_n^1, y_n^2)$ such that $u_n(y_n^3) = 1$ and $u_n'(y_n^3) < 0$. We conclude from $E = 0$ and $u_n(y_n^3) = 1$ that

$$u_n'''(y_n^3) + \frac{\nu}{2} u_n'(y_n^3) = \frac{u_n''(y_n^3)^2}{u_n'(y_n^3)} \leq 0,$$

and we reach a contradiction as before by taking the limit $n \to \infty$ in this inequality, thereby concluding the proof. ■

To define the semiconjugacy $\rho : \Lambda \to \Sigma_B$ we consider the solution $u^\lambda(y)$ of (4) with initial data $\lambda \in \Lambda$. If $u^\lambda \equiv 1$, we define $\rho(\lambda) = 1^\infty$. Otherwise, let u_i^λ be the sequence of extrema

of u^λ, indexed such that u_1^λ is the minimum corresponding to $\lambda \in \mathcal{P}$. By Lemma 10, this sequence lies in $U_{\mathbf{b}_\lambda}$ for some $\mathbf{b}_\lambda \in \Sigma_B$, and we define $\rho(\lambda) = \mathbf{b}_\lambda$.

It follows from Lemmas 9 and 10 that the map ρ is continuous (also in the point $+\mathbf{1}$, if $+\mathbf{1}$ happens to lie in Λ). Moreover, $\rho \circ \mathcal{T} = \sigma \circ \rho$ by construction and the properties of the return map (where σ is the shift map). Finally, ρ is surjective. Namely, all *periodic* sequences in Σ_B have corresponding solutions in $\overline{\mathcal{C}}$ (and thus in Λ), and since the set of periodic sequences is dense in Σ_B and Λ is compact, surjectivity follows. Hence, ρ defines a semiconjugacy, and it follows (e.g., [30]) that the topological entropy of the map $\mathcal{T}$ on Λ is positive:

$$h_{\text{top}}(\mathcal{T}|_\Lambda) \geq h_{\text{top}}(\sigma|_{\Sigma_B}) = \log\left(\frac{1+\sqrt{5}}{2}\right).$$

This finishes the proof of Theorem 2.

Remark 11. There is another way to make a coding that was already alluded to in the introduction. The profile between two successive minima can, qualitatively, have three shapes (see Figure 2):

$$\begin{aligned} u_i < -1 \quad &\text{and} \quad u_{i+2} > -1, \qquad \text{coded by } \mathbf{c}_i = 1; \\ u_i > -1 \quad &\text{and} \quad u_{i+2} > -1, \qquad \text{coded by } \mathbf{c}_i = 2; \\ u_i > -1 \quad &\text{and} \quad u_{i+2} < -1, \qquad \text{coded by } \mathbf{c}_i = 3. \end{aligned}$$

This new coding $\mathbf{c}$ is also indicated in Figure 6. The corresponding adjacency matrix is

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

This of course does not improve the bound on the entropy of $\mathcal{T}$, but the slightly more complicated coding $\mathbf{c}$, using "up-down" building blocks, is more intuitively related to the shape of solutions, as expressed in Figure 2.

3. Rigorous continuation. We are going to restrict our attention to symmetric periodic solutions u satisfying $\mathcal{H}$. Hence, let

$$u(y) = a_0 + 2\sum_{l=1}^{\infty} a_l \cos(lLy), \tag{11}$$

with L an a priori unknown variable. Since $u'(0) = 0$, and the energy (5) is a conserved quantity along the orbits of (4), we get that

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

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

$$u''(0) - \frac{1}{\sqrt{2}}[u(0)^2 - 1] = 0. \tag{12}$$

Substituting the expansion (11) of $u(y)$ into (12), we obtain

$$e(L,a) \stackrel{\text{def}}{=} -2L^2 \sum_{l=1}^{\infty} l^2 a_l - \frac{1}{\sqrt{2}} \left[a_0 + 2 \sum_{l=1}^{\infty} a_l \right]^2 + \frac{1}{\sqrt{2}} = 0. \tag{13}$$

Plugging the expansion (11) into (4) and computing the inner product of the resulting equations with each $\cos(kLx)$, $k \geq 0$, we get

$$g_k(L,a,\nu) \stackrel{\text{def}}{=} [1 + \nu L^2 k^2 - L^4 k^4] a_k - \sum_{\substack{k_1+k_2+k_3=k \\ k_i \in \mathbb{Z}}} a_{k_1} a_{k_2} a_{k_3} = 0, \tag{14}$$

where $a = (a_0, a_1, \ldots)$. Define $x = (x_{-1}, x_0, x_1, \ldots) = (L, a_0, a_1, a_2, \ldots)$, and $g = (g_0, g_1, g_2, \ldots)^T$, as well as

$$f(x,\nu) = \begin{bmatrix} e(x) \\ g(x,\nu) \end{bmatrix}. \tag{15}$$

To simplify the presentation, we use the notation $f_{-1} = e$ and $f_k = g_k$ for $k \geq 0$. Now, since we want to use rigorous numerical methods to find pairs (x,ν) such that $f(x,\nu) = 0$, we need to consider a finite dimensional projection of (15). Define $x_F = (x_{-1}, x_0, \ldots, x_{m-1}) = (L, a_0, \ldots, a_{m-1}) \in \mathbb{R}^{m+1}$,

$$e^{(m)}(x_F) \stackrel{\text{def}}{=} -2L^2 \sum_{l=1}^{m-1} l^2 a_l - \frac{1}{\sqrt{2}} \left[a_0 + 2 \sum_{l=1}^{m-1} a_l \right]^2 + \frac{1}{\sqrt{2}},$$

and

$$g^{(m)}(x_F,\nu) \stackrel{\text{def}}{=} [g_0(x_F,\nu), \ldots, g_{m-1}(x_F,\nu)]^T.$$

The *Galerkin projection* of (15) is defined by

$$f^{(m)}(x_F,\nu) \stackrel{\text{def}}{=} \begin{bmatrix} e^{(m)}(x_F) \\ g^{(m)}(x_F,\nu) \end{bmatrix}.$$

It is important to note that $f^{(m)}$ has both a finitely truncated domain and a finitely truncated codomain.

We now describe how we can modify the classical, numerical, predictor-corrector algorithm for following a continuous branch of solutions to fit our setting. Suppose that, at parameter value $\nu = \nu_0$, we have used a Newton-like iteration method to numerically find $\bar{x}_F$ such that

$$f^{(m)}(\bar{x}_F, \nu_0) \approx 0. \tag{16}$$

Throughout this paper, Df represents the derivative of f with respect to the x_F- or x-variable. If $(\bar{x}_F, \nu_0)$ is a solution of $f^{(m)}(x_F,\nu) = 0$ such that the Jacobian matrix $Df^{(m)}(\bar{x}_F, \nu_0)$ is invertible, then, by the implicit function theorem, there exists a unique one-dimensional local continuum of solutions (x_F,ν) such that the solution x_F is locally a function of the parameter

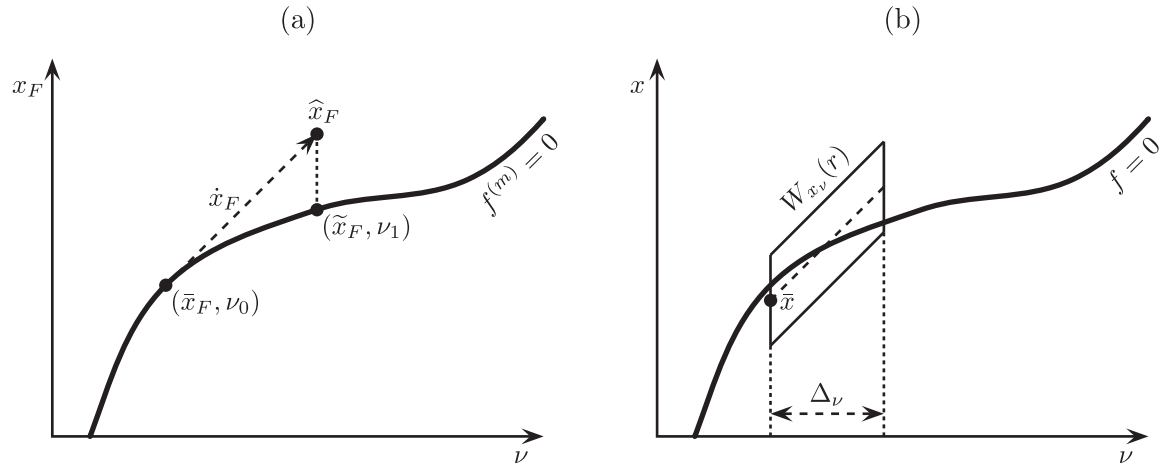


Figure 7. (a) *Sketch of the predictor-corrector algorithm for the truncated problem* $f^{(m)}(x_F, \nu) = 0$. (b) *The neighborhood* $W_{x_\nu}(r)$ *in which we find solutions of the full problem* $f(x, \nu) = 0$.

ν (near ν_0). Moreover, as a computational counterpart, we may numerically find a *tangent* vector $\dot{x}_F$ to the solution curve at $(\bar{x}_F, \nu_0)$, i.e.,

$$Df^{(m)}(\bar{x}_F, \nu_0)\dot{x}_F + \frac{\partial f^{(m)}}{\partial \nu}(\bar{x}_F, \nu_0) \approx 0. \tag{17}$$

We can then use this tangent vector to obtain a *predictor* $\widehat{x}_F = \bar{x}_F + (\nu_1 - \nu_0)\dot{x}_F$ for the solution at a parameter value ν_1 close to ν_0. The *corrector* then consists of iterating a Newton-like map, with initial point $\widehat{x}_F$, to converge to the zero $\widetilde{x}_F$ of $f^{(m)}(x_F, \nu_1)$; see also Figure 7(a). There are two essential problems to overcome in this scheme. First, the result for the finite dimensional truncation needs to be "lifted" to the infinite dimensional setting. This lifting is carried out via "validated continuation" [15], with a final interval arithmetic step. Second, the described method leads to a discrete set of solutions (x, ν), whereas we aim for solutions for a *continuous* range of parameter values, and we describe our approach below.

Denote by $\bar{x}_F = (\bar{L}, \bar{a}_0, \bar{a}_1, \ldots, \bar{a}_{m-1})$ and $\dot{x}_F = (\dot{L}, \dot{a}_0, \dot{a}_1, \ldots, \dot{a}_{m-1})$ the approximate solutions of (16) and (17), respectively, and define their infinite extensions $\bar{x} = (\bar{x}_F, 0, 0, 0, \ldots)$ and $\dot{x} = (\dot{x}_F, 0, 0, 0, \ldots)$. We define the "linear part" of g_k as

$$\mu_k(L, \nu) \stackrel{\text{def}}{=} 1 + \nu L^2 k^2 - L^4 k^4.$$

Furthermore, let the $(m+1) \times (m+1)$ matrix J_F be the numerically computed inverse of $Df^{(m)}(\bar{x}_F, \nu_0)$, and let 0_F be the $1 \times (m+1)$ row vector $(0, 0, \ldots, 0)$. We define the linear operator on sequence spaces

$$A \stackrel{\text{def}}{=} \begin{bmatrix} J_F & 0_F^T & 0_F^T & 0_F^T & \cdots \\ 0_F & \mu_m(\bar{L}, \nu_0)^{-1} & 0 & 0 & \cdots \\ 0_F & 0 & \mu_{m+1}(\bar{L}, \nu_0)^{-1} & 0 & \cdots \\ 0_F & 0 & 0 & \mu_{m+2}(\bar{L}, \nu_0)^{-1} & \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix}, \tag{18}$$

which acts as an *approximate inverse* of the linear operator $Df(\bar{x}, \nu_0)$. We shall always make sure that m is sufficiently large, so that $\mu_m(\bar{L}, \nu_0) < 0$. For ν close to ν_0, we consider the Newton-like operator

$$T_\nu(x) \stackrel{\text{def}}{=} x - A \cdot f(x, \nu). \tag{19}$$

To formalize this approach, it is convenient to use a functional analytic setting. As weight functions we define, for $s > 0$,

$$\omega_k^s = \begin{cases} 1, & k = -1, 0, \\ k^s, & k \geq 1. \end{cases} \tag{20}$$

In general, one can play around with these weight functions (and the norm below); for the problem in this paper, we have found the choice (20) to be appropriate, because it leads to a proof. These weight functions are used in the norms

$$\|x\|_s = \sup_{k=-1,0,1,\ldots} |x_k \omega_k^s| \tag{21}$$

and the sequence spaces

$$\Omega^s = \{x, \|x\|_s < \infty\},$$

consisting of sequences with algebraically decaying tails. For the first sum in (13) to make sense, we shall require that $s > 3$.

Lemma 12. *We have the following:*

(a) *The sequence space Ω^s with norm $\|\cdot\|_s$ is a Banach space for all s. The injections $\Omega^{s_1} \hookrightarrow \Omega^{s_2}$ are compact for all $s_1 > s_2$.*

(b) *Let $s > 3$. The map $T(x, \nu)$ is continuous from $\Omega^s \times \mathbb{R}$ to Ω^{s+2}, and $T(x, \nu)$ is compact from $\Omega^s \times \mathbb{R}$ to Ω^s.*

(c) *Let $s_0 > 3$ and fix ν. Zeros of $f(x, \nu)$, or, equivalently, fixed points of $T(x, \nu)$, that are in Ω^{s_0}, are in Ω^s for all $s \geq s_0$.*

(d) *Let $s > 3$. A sequence $x = (L, a_0, a_1, \ldots) \in \Omega^s$ is a zero of f, or a fixed point of T,* if and only if *u given by (11) is a periodic solution of (4) at energy level $E = 0$, with period $\frac{2\pi}{L}$, and symmetric in $y = 0$ and $y = \frac{\pi}{L}$.*

Proof. We only outline the proofs; the reader may quite easily fill in the details.

Part (a) follows from standard functional analytic arguments. For part (b), observe that, for any $s > 1$, the weighted discrete convolution $k^s |\sum_{k_1+k_2+k_3=k} a_{k_1} a_{k_2} a_{k_3}|$ is uniformly bounded if $k^s |a_k|$ is uniformly bounded; see, for example, Appendix A. It then follows that $g_k = -L^4 k^4 a_k + O(k^2)$. Since $\mu_k = -L^4 k^4 + O(k^2)$, the composition $[A \cdot f(x, \nu)]_k = x_k + O(k^{-2})$ as $k \to \infty$, and we conclude that T maps $\Omega^s \times \mathbb{R}$ to Ω^{s+2} for $s > 3$ (one needs $s > 3$ for the first sum in (13) to be well defined). Continuity of T is straightforward to verify, and the compact injections from part (a) then imply the compactness statement in (b).

Concerning part (c), the equivalence of zeros of f and fixed points of T is obvious, and the remainder of the statement follows immediately from parts (a) and (b) by using that $x = T(x)$. Finally, since the tail of a fixed point of T decays faster than any algebraic rate, all sums may be differentiated term by term; hence u defined by (11) is a solution of

the differential equation (periodic, with energy $E = 0$). On the other hand, any (periodic) solution of the differential equation is C^∞; hence the tail of its Fourier transform decays faster than any algebraic rate, and thus, by standard arguments, the Fourier transform solves $g = 0$, and part (d) follows. ∎

Lemma 12(d) shows that the problem of finding (symmetric) periodic solutions of (4) at $E = 0$ is equivalent to studying fixed points of T. We will find balls in Ω^s on which T, for fixed ν, is a contraction mapping, thus leading to solutions of (4). Let us define the ball of radius r, centered at the origin,

$$W(r) \stackrel{\text{def}}{=} [-r, r]^2 \times \prod_{k=1}^{\infty} \left[-\frac{r}{k^s}, \frac{r}{k^s}\right]. \tag{22}$$

For $\Delta_\nu = \nu - \nu_0$ small, we define the *predictors based at* ν_0 by

$$x_\nu = \bar{x} + \Delta_\nu \dot{x}.$$

For ν close to ν_0 we define the ball centered at x_ν by

$$W_{x_\nu}(r) = x_\nu + W(r).$$

We look for fixed points of T inside these balls/neighborhoods; see also Figure 7(b). To show that T is a contraction mapping, we need bounds Y_k and Z_k for all $k = -1, 0, 1, 2, \ldots$, such that, with $\Delta_\nu = \nu - \nu_0$,

$$\left|[T_\nu(x_\nu) - x_\nu]_k\right| \le Y_k(\Delta_\nu), \tag{23}$$

and

$$\sup_{w, w' \in W(r)} \left|[DT_\nu(x_\nu + w')w]_k\right| \le Z_k(r, \Delta_\nu). \tag{24}$$

We will find such bounds in sections 3.2 and 3.3, respectively. Notice that $Y_k \ge 0$ and $Z_k \ge 0$. Although this is not a restriction, we will, in this paper, consider only $\Delta_\nu \ge 0$, since we will initiate the continuation at the parameter value $\nu = 0$ and finish at $\nu = 2$; hence we do continuation in one direction only.

Variants of the following lemma were also used in [13, 15, 18, 37].

Lemma 13. *Fix $s > 3$ and $\nu = \nu_0 + \Delta_\nu$. If there exists an $r > 0$ such that $\|Y + Z\|_s < r$, with $Y = (Y_{-1}, Y_0, Y_1, \ldots)$ and $Z = (Z_{-1}, Z_0, Z_1, \ldots)$ the bounds as defined in (23) and (24), then there is a unique $\widetilde{x}_\nu \in W_{x_\nu}(r)$ such that $f(\widetilde{x}_\nu, \nu) = 0$.*

Proof. We outline the proof, which can be found in more detail in [15] and [37]. The mean value theorem (applied componentwise), combined with the assumption $\|Y + Z\|_s < r$, implies that $T(\cdot, \nu)$ maps $W_{x_\nu}(r)$ into itself. Since $Y_k \ge 0$ and $Z_k \ge 0$, it follows that $\|Z\|_s \le \|Y + Z\|_s < r$. We infer from the mean value theorem that the Lipschitz constant of $T(\cdot, \nu)$ on $W_{x_\nu}(r)$ can be estimated above by $\|Z\|_s / r < 1$, so that T is a contraction mapping. Finally, zeros of f correspond to fixed point of T; hence an application of the Banach fixed point theorem concludes the proof. ∎

In order to verify the hypotheses of Lemma 13 in a computationally efficient way, we introduce the notion of *radii polynomials*. Namely, as will become clear in sections 3.2 and 3.3, the functions $Y_k(\Delta_\nu)$ and $Z_k(r,\Delta_\nu)$ are polynomials in their independent variables. Also, for sufficiently large k, say, $k \geq M$, one may choose

$$Y_k = 0 \qquad \text{and} \qquad Z_k = \widehat{Z}_M \left(\frac{M}{k}\right)^s$$

for some $\widehat{Z}_M(r,\Delta_\nu) > 0$. This leads us to the following definition.

Definition 14. *Let $Y_k(\Delta_\nu) = 0$ and $Z_k(r,\Delta_\nu) = \widehat{Z}_M(r,\Delta_\nu)\left(\frac{M}{k}\right)^s$ for all $k \geq M$. We define the $M+2$* radii polynomials *$\{p_{-1}, p_0, \ldots, p_{M-1}, p_M\}$ by*

$$p_k(r,\Delta_\nu) \stackrel{\text{def}}{=} \begin{cases} Y_k(\Delta_\nu) + Z_k(r,\Delta_\nu) - \frac{r}{\omega_k^s}, & k = -1, 0, \ldots, M-1, \\ \widehat{Z}_M(r,\Delta_\nu) - \frac{r}{\omega_M^s}, & k = M. \end{cases}$$

The usefulness of the radii polynomials p_k follows from the observation that the polynomials Y_k and Z_k have a few exceptionally small terms. Namely, it turns out that they are *roughly* of the form (to be made precise in sections 3.2 and 3.3)

$$\begin{aligned} Y_k &\sim \delta_1 + \delta_2 \Delta_\nu + O(\Delta_\nu^2), \\ Z_k &\sim \delta_3 r + O(\Delta_\nu r, r^2), \end{aligned}$$

where the δ_1, δ_2, and δ_3 are small, because of the choice of $\bar{x}$, the choice of $\dot{x}$, and the choice of the linear operator A in the Newton-like map T, respectively. It is easy to see that the zeroth-order term of Z_k vanishes. Hence, the radii polynomials are roughly of the form

$$p_k(r,\Delta_\nu) \sim (\delta_1 + \Delta_\nu \delta_2) - \left(\frac{1}{\omega_k^s} - \delta_3\right) r + O(r^2, \Delta_\nu r, \Delta_\nu^2),$$

so that one may anticipate them to be negative for small r (but not too small) for a reasonably large range of Δ_ν.

Lemma 15. *Let $s > 3$ and let $Y_k(\Delta_\nu) = 0$ and $Z_k(r,\Delta_\nu) = \widehat{Z}_M(r,\Delta_\nu)\left(\frac{M}{k}\right)^s$ for all $k \geq M$. Suppose that there exists an $r > 0$ such that $p_k(r,\Delta_\nu) < 0$ for all $k = -1, \ldots, M$; then the hypotheses of Lemma 13 are satisfied for $\nu = \nu_0 + \Delta_\nu$.*

Proof. Let $s > 3$, $r > 0$, and $\nu = \nu_0 + \Delta_\nu$ such that $p_k(r,\Delta_\nu) < 0$ for all $k = -1, 0, \ldots, M$. Since $Y_k + Z_k = \widehat{Z}_M\left(\frac{M}{k}\right)^s$ for $k \geq M$, by definition of the radii polynomials, we get that

$$\begin{aligned} \|Y + Z\|_s &= \sup_{k=-1,0,1,\ldots} |[Y_k(\Delta_\nu) + Z_k(r,\Delta_\nu)]\omega_k^s| \\ &= \max_{k=-1,0,\ldots,M} \{p_k(r,\Delta_\nu)\omega_k^s + r\} < r. \end{aligned}$$

∎

Combining Lemmas 13 and 15, it should now become clear that proving the existence of zeros of f, and hence periodic solutions of (4) at $E = 0$, is *computable*, since only a finite number of polynomial inequalities need to be verified. There is one final observation to be made. The Y_k and the Z_k are monotonically increasing in the variable $\Delta_\nu \geq 0$, that is,

$Y_k(\Delta_\nu^0) \le Y_k(\Delta_\nu^1)$ and $Z_k(r,\Delta_\nu^0) \le Z_k(r,\Delta_\nu^1)$ for $0 \le \Delta_\nu^0 \le \Delta_\nu^1$. As a consequence, the same property holds for the radii polynomials: if $0 \le \Delta_\nu^0 \le \Delta_\nu^1$, then $p_k(r,\Delta_\nu^0) \le p_k(r,\Delta_\nu^1)$ for all $k = -1,\dots,M$. Hence, if the hypotheses in Lemma 13 are satisfied for some $\Delta_\nu^0 > 0$, then they are satisfied for all $\Delta_\nu \in [0,\Delta_\nu^0]$; hence there are corresponding periodic solutions of (4) for all $\nu \in [\nu_0, \nu_0 + \Delta_\nu^0]$.

For what follows we recall that the construction of the radii polynomials p_k involves Y_k and Z_k, defined in (26) and (31), respectively, as well as M_0 and $\widehat{Z}_M$, defined in (30) and (32), respectively. Furthermore, because the coefficients of the polynomials Y_k and Z_k are all positive, there is at most one interval in the positive half line on which p_k is negative. With these considerations in mind, the following procedure leads to a proof of the existence part of Theorem 3. The geometric properties $\mathcal{H}$ will be checked in Procedure 21 in section 4.

Procedure 16. *To check the hypotheses in Lemma 15 on the interval $\nu \in [0,2]$ we proceed as follows.*

1. *Choose minimum and maximum step-sizes $0 < \Delta_{\min} < \Delta_{\max}$. Initiate $s > 3$, $m \in \mathbb{N}$, $M \ge \max\{3m-2, 6\}$, $\nu_0 = 0$, $\Delta_\nu \in [\Delta_{\min}, \Delta_{\max}]$, $\Delta_\nu^0 = 0$, and an approximate zero $\widehat{x}_F$ of $f^{(m)}(x_F, 0)$. Calculate the analytic estimates $(\alpha_k,\ k = 0,\dots,M)$ that are independent of everything.*
2. *With a classical Newton iteration, find near $\widehat{x}_F$ an approximate solution $\bar{x}_F$ of $f^{(m)}(x_F, \nu_0) = 0$. Calculate an approximate solution $\dot{x}_F$ of (17). Use the first component of $\bar{x}_F$ to calculate with interval arithmetic $M_0(\bar{L}, \nu_0)$ and check that $M_0 \le M$ (this is never a problem in practice).*
3. *Compute, using interval arithmetic, the coefficients of the radii polynomials p_k, $k = -1,\dots,M$. This is the computationally most expensive step, since it involves the coefficients in Tables 1, 3, and 4 and in particular requires the calculation of convolution terms.*
4. *Calculate numerically $I = [I_-, I_+] \overset{\text{def}}{=} \bigcap_{k=-1}^{M} \{r \ge 0 \mid p_k(r,\Delta_\nu) \le 0\}$.*
   - *If $I = \varnothing$, then go to Step 6.*
   - *If $I \ne \varnothing$, then let $r = \min\{\frac{11}{10} I_-, \frac{I_- + I_+}{2}\}$. Compute with interval arithmetic $p_k(r,\Delta_\nu)$. If $p_k(r,\Delta_\nu) < 0$ for all $k = -1,\dots,M$, then go to Step 5; else go to Step 6.*
5. *Update $\Delta_\nu^0 \leftarrow \Delta_\nu$ and $r_0 \leftarrow r$. If $\frac{10}{9}\Delta_\nu \le \Delta_{\max}$, then update $\Delta_\nu \leftarrow \frac{10}{9}\Delta_\nu$ and go to Step 4; else go to Step 7.*
6. *If $\Delta_\nu^0 > 0$, then go to Step 7; else if $\frac{9}{10}\Delta_\nu \ge \Delta_{\min}$, then update $\Delta_\nu \leftarrow \frac{9}{10}\Delta_\nu$ and go to Step 4; else go to Step 8.*
7. *The continuation step has succeeded. Store, for future reference, $\bar{x}_F$, $\dot{x}_F$, r_0, ν_0, and Δ_ν^0. Determine ν_1 approximately equal to, but interval arithmetically less than, $\nu_0 + \Delta_\nu^0$. If $\nu_1 \ge 2$, then terminate the procedure* successfully*; else make the updates $\nu_0 \leftarrow \nu_1$, $\Delta_\nu \leftarrow \Delta_\nu^0$, $\widehat{x}_F \leftarrow \bar{x}_F + \Delta_\nu^0 \dot{x}_F$, and $\Delta_\nu^0 \leftarrow 0$, and go to Step 2 for the next continuation step.*
8. *The continuation step has failed. Either decrease $\Delta_{\min}$ and return to Step 6 or increase s or M and return to Step 3; or increase m and return to Step 2. Alternatively, terminate the procedure* unsuccessfully *at $\nu = \nu_0$ (although with success on $[0,\nu_0]$).*

Note that computing I using interval arithmetic would be expensive, since we would need to compute each set $\{r \ge 0 \mid p_k(r,\Delta_\nu) \le 0\}$ $(k = -1,\dots,M-1)$ using interval arithmetic.

Since we need only to get one radius $r \in I$, we chose to approximate I with floating point arithmetic, pick a point r in the approximation of I, and finally prove that $r \in I$ using interval arithmetic. During the procedure, a sequence of intervals $[\nu_0, \nu_0 + \Delta_\nu^0]$ covering $[0, 2]$ is stored, together with the variables defining the neighborhoods $W_{x_\nu}(r_0)$. In section 4, the balls $W_{x_{\nu_0}}(r_0)$ will be used in Procedure 21 (and Lemma 22) to check the geometric conditions $\mathcal{H}$. Concerning the usefulness of the procedure in practice, the proof of the pudding is in the eating.

Lemma 17. *Let $s = 4$, $m = 43$, $M = 127$, $\Delta_{\min} = 10^{-10}$, and $\Delta_{\max} = 2$. Then we can choose an approximate zero $\widehat{x}_F = \widehat{x}_F^*$ of $f^{(m)}(x_F, 0)$ such that Procedure 16 terminates* successfully. *Hence there are periodic solutions of (4) at $E = 0$ for all $\nu \in [0, 2]$.*

The choice of $\widehat{x}_F^*$ is made in such a way that the solutions found satisfy the geometric conditions $\mathcal{H}$. This is checked using Procedure 21.

Proof. A MATLAB computer program successfully performing Procedure 16 accompanies the paper. In particular, we never end up in Step 8 of the procedure. Concerning the implementation, the only difficult evaluations are the convolution terms, which can be computed in a very efficient way using the fast Fourier transform combined with interval arithmetic; see [18]. More details about the implementation are given in section 3.1. ∎

3.1. Implementation. In this section, we discuss in detail the implementation of the rigorous verification of Procedure 16 and Procedure 21 (which checks the geometric properties $\mathcal{H}$; see section 4). First, as seen in [15], the errors induced by the floating point computations of the coefficients of the radii polynomials are small. Hence, finding a positive r at which all radii polynomials are negative without interval arithmetic gives significant confidence about the success of Procedure 16. On top of that, the computational efficiency of floating point arithmetic in MATLAB allows for fast computations. With this in mind, we wrote a preliminary function called *SH_continuation.m*, which verifies, without interval arithmetic, that Procedure 16 performs successfully. Using the values given in Lemma 17, we obtained 13068 successful nonrigorous steps in a bit more than 7 minutes. In Figure 8, we plot the evolution of Δ_ν as the parameter ν runs from 0 to 2.

This being done, and armed with confidence, we then aimed for the proof. First we wrote *SH_rigorous_continuation.m*, the equivalent of *SH_continuation.m*, in the MATLAB interval arithmetic package *Intlab* (see [20]). We did not try to optimize for speed in the interval arithmetic setting, since we preferred to keep the changes with respect to the floating point version *SH_continuation.m* limited. The speed of the interval arithmetic is thus rather slow, and we decided to distribute the computations over 20 different computers, each running 3 simultaneous calculations. Hence, we used the 13068 output points from *SH_continuation.m* to generate a nonuniform mesh $\{(\nu_j, x_j) \mid j = 1, \dots, 61\}$ of the branch under study, where $\nu_1 = 0$ and $\nu_{61} = 2$. The function *SH_mesh_generator.m* picks 61 points out of the 13068 defining the discrete branch. Note that we do not need to define x_{61}, since at this point we have already reached $\nu = 2$. This mesh is stored in the file *SH_mesh_points.mat*. Then, for each $j = 1, \dots, 60$, we called the function *SH_run_proof*(j). This function first starts *Intlab*, loads *SH_mesh_points.dat*, and then rigorously verifies Procedure 16 between the parameter values ν_j and ν_{j+1}, using *SH_rigorous_continuation.m* with the initial point x_j as input. Finally, it verifies, by running the interval arithmetic function *SH_geometric_properties.m*, that the periodic orbits rigorously generated by *SH_rigorous_continuation*(x_j, ν_j, ν_{j+1}) satisfy the geo-

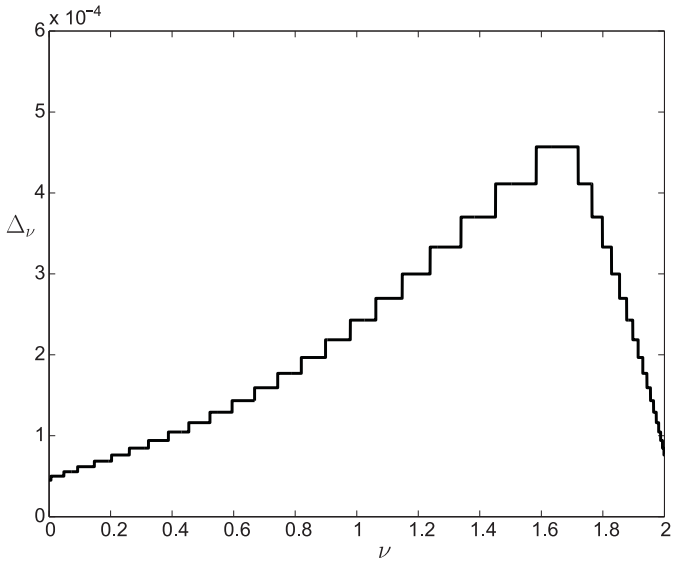


Figure 8. *The step size Δ_ν as a function of the parameter $\nu \in [0,2]$. The step size increases at first, but it then decreases as we approach a saddle-node bifurcation.*

metric properties $\mathcal{H}$, as described in section 4. Thus, the proof of Lemma 17 (and Lemma 22) was finished when all 60 runs ended successfully. The total running time was around 12 hours.

3.2. The bounds $Y_k(\Delta_\nu)$.

Recalling (19) and (23), in this section we want to find bounds

$$\left|[-A \cdot f(x_\nu, \nu)]_k\right| \leq Y_k(\Delta_\nu).$$

We use the following notation. For an arbitrary vector $y_F = (y_{-1}, y_0, y_1, \ldots, y_{m-1})$, the infinite extension is $y = (y_{-1}, y_0, y_1, \ldots, y_{m-1}, 0, 0, 0, \ldots)$. For an infinite sequence $z = (z_{-1}, z_0, z_1, \ldots)$, the finite restriction is $z_F = (z_{-1}, z_0, z_1, \ldots, z_{m-1})$, whereas $z_I = (0, 0, \ldots, 0, z_m, z_{m+1}, z_{m+2}, \ldots)$ denotes the infinite tail. Similar notation is used for vectors/sequences of which the index starts at 0 rather than -1—in particular, the vectors $\bar{a}_F = (\bar{a}_0, \bar{a}_1, \ldots, \bar{a}_{m-1})$ and $\dot{a}_F = (\dot{a}_0, \dot{a}_1, \ldots, \dot{a}_{m-1})$. Also, absolute values of vectors, infinite sequences, and matrices are taken componentwise, e.g., $|x| = (|x_{-1}|, |x_0|, |x_1|, |x_2|, \ldots)$. Furthermore, we use the convolutions

$$(a * b * c)_k = \sum_{\substack{k_1+k_2+k_3=k \\ k_1,k_2,k_3 \in \mathbb{Z}}} a_{|k_1|} b_{|k_2|} c_{|k_3|},$$

which is of course the standard convolution when taking $a_{-k} \equiv a_k$, and (with the extension convention)

$$(a_F * b_F * c_F)_k = (a * b * c)_k = \sum_{\substack{k_1+k_2+k_3=k \\ |k_i|<m}} a_{|k_1|} b_{|k_2|} c_{|k_3|},$$

which vanishes for $k \geq 3m - 2$.

Table 1

The nonzero coefficients in the expansion (25) of f_k. In particular, $d^4_{-1} = d^5_{-1} = 0$. Notice that for $k \geq m$ the only nonvanishing terms are the convolution terms. Furthermore, for $k \geq 3m-2$ all coefficients are 0.

$k = -1$	
d^1_{-1}	$-2\bar{L}^2 \sum_{l=1}^{m-1} l^2 \dot{a}_l - 4\bar{L}\dot{L} \sum_{l=1}^{m-1} l^2 \bar{a}_l - \sqrt{2}\left(\bar{a}_0 + 2\sum_{l=1}^{m-1} \bar{a}_l\right)\left(\dot{a}_0 + 2\sum_{l=1}^{m-1} \dot{a}_l\right)$
d^2_{-1}	$-4\bar{L}\dot{L} \sum_{l=1}^{m-1} l^2 \dot{a}_l - 2\dot{L}^2 \sum_{l=1}^{m-1} l^2 \bar{a}_l - \frac{1}{2}\sqrt{2}\left(\dot{a}_0 + 2\sum_{l=1}^{m-1} \dot{a}_l\right)^2$
d^3_{-1}	$-2\dot{L}^2 \sum_{l=1}^{m-1} l^2 \dot{a}_l$
$k = 0, 1, 2, \ldots$	
d^1_k	$\left(1 + \nu_0 \bar{L}^2 k^2 - \bar{L}^4 k^4\right) \dot{a}_k + \left((2\nu_0 \bar{L}\dot{L} + \bar{L}^2)k^2 - 4\bar{L}^3 \dot{L} k^4\right) \bar{a}_k - 3(\bar{a} * \bar{a} * \dot{a})_k$
d^2_k	$\left((2\nu_0 \bar{L}\dot{L} + \bar{L}^2)k^2 - 4\bar{L}^3 \dot{L} k^4\right) \dot{a}_k + \left((\nu_0 \dot{L}^2 + 2\bar{L}\dot{L})k^2 - 6\bar{L}^2 \dot{L}^2 k^4\right) \bar{a}_k - 3(\bar{a} * \dot{a} * \dot{a})_k$
d^3_k	$\left((\nu_0 \dot{L}^2 + 2\bar{L}\dot{L})k^2 - 6\bar{L}^2 \dot{L}^2 k^4\right) \dot{a}_k + \left(\dot{L}^2 k^2 - 4\bar{L}\dot{L}^3 k^4\right) \bar{a}_k - (\dot{a} * \dot{a} * \dot{a})_k$
d^4_k	$\left(\dot{L}^2 k^2 - 4\bar{L}\dot{L}^3 k^4\right) \dot{a}_k - \dot{L}^4 k^4 \bar{a}_k$
d^5_k	$-\dot{L}^4 k^4 \dot{a}_k$

Exploiting that f is a vector of polynomials in the components of x and ν, we write

$$f_k(x_\nu, \nu) = f_k(\bar{x} + \Delta_\nu \dot{x},\ \nu_0 + \Delta_\nu) = f_k(\bar{x}, \nu_0) + \sum_{i=1}^{5} d^i_k(\bar{x}, \dot{x}, \nu_0)\Delta^i_\nu. \tag{25}$$

Here the constants d^i_k are listed in Table 1.

For the zeroth-order term we have for the finite part $f_F(\bar{x}, \nu_0) = f^{(m)}(\bar{x}_F, \nu_0)$, which is very small, since $\bar{x}_F$ is a numerical zero of $f^{(m)}(x_F, \nu_0)$. The choice of $\dot{x}_F$ given by (17) implies that the first-order term d^1_F is also small, since

$$f_F(x_\nu, \nu) = f^{(m)}(\bar{x}_F, \nu_0) + \left[Df^{(m)}(\bar{x}_F, \nu_0)\dot{x}_F + \frac{\partial f}{\partial \nu}^{(m)}(\bar{x}_F, \nu_0)\right]\Delta_\nu + O(\Delta^2_\nu).$$

For the tail ($k \geq m$) we have $f_k(\bar{x}, \nu_0) = -(\bar{a} * \bar{a} * \bar{a})_k = -(\bar{a}_F * \bar{a}_F * \bar{a}_F)_k$, which vanishes for $k \geq 3m - 2$.

Using the vectors d^i from Table 1, this leads to bounds $Y_k(\Delta_\nu)$ as listed below, with $\Delta_\nu \geq 0$:

$$Y_F = |J_F \cdot f^{(m)}(\bar{x}_F, \nu_0)| + \sum_{i=1}^{5} \left|J_F \cdot d^i_F\right| \Delta^i_\nu \tag{26a}$$

for $k = -1, 0, 1, \ldots, m-1$;

$$Y_k = \frac{|(\bar{a} * \bar{a} * \bar{a})_k| + 3|(\bar{a} * \bar{a} * \dot{a})_k|\Delta_\nu + 3|(\bar{a} * \dot{a} * \dot{a})_k|\Delta^2_\nu + |(\dot{a} * \dot{a} * \dot{a})_k|\Delta^3_\nu}{|\mu_k(\bar{L}, \nu_0)|} \tag{26b}$$

for $m \le k \le 3m-3$; and

$$Y_k = 0 \tag{26c}$$

for $k \ge 3m-2$.

3.3. The bounds $Z_k(r, \Delta_\nu)$. In this section we construct bounds

$$\sup_{w,w' \in W(r)} \left| [DT_\nu(x_\nu + w')w]_k \right| \le Z_k(r, \Delta_\nu).$$

We will use the notation introduced at the start of section 3.2. Furthermore, at several instances, we employ a computational parameter M, and although not necessary, we choose the same value of M every time for simplicity.

Recall that J_F is a numerical inverse of $Df^{(m)}(\bar{x}_F, \nu_0)$. To simplify the exposition, we introduce an almost inverse of the operator A defined in (18):

$$A^\dagger \stackrel{\text{def}}{=} \begin{bmatrix} Df^{(m)}(\bar{x}_F, \nu_0) & 0_F^T & 0_F^T & 0_F^T & \cdots \\ 0_F & \mu_m(\bar{L}, \nu_0) & 0 & 0 & \cdots \\ 0_F & 0 & \mu_{m+1}(\bar{L}, \nu_0) & 0 & \cdots \\ 0_F & 0 & 0 & \mu_{m+2}(\bar{L}, \nu_0) & \\ \vdots & \vdots & \vdots & & \ddots \end{bmatrix}.$$

We split $Df_\nu(x_\nu + w')w$ into two pieces:

$$Df_\nu(x_\nu + w')w = A^\dagger w + \left[Df_\nu(x_\nu + w') - A^\dagger\right]w$$

and hence

$$DT_\nu(w' + x_\nu)w = [I - AA^\dagger]w - A\left[Df_\nu(x_\nu + w') - A^\dagger\right]w, \tag{27}$$

where the first term will be very small. For $w, w' \in W(r)$ we consider $v, v' \in W(1)$ defined by $w = rv$ and $w' = rv'$. Similar to section 3.2, we expand the expression $[Df_\nu(x_\nu + w') - A^\dagger]w$ in terms of r and Δ_ν:

$$\left([Df_\nu(x_\nu + w') - A^\dagger]w\right)_k = \sum_{i=1}^{5} \sum_{j=0}^{5-i} c_k^{i,j}(\bar{x}, \dot{x}, v, v', \nu_0) r^i \Delta_\nu^j. \tag{28}$$

Here the constants $c_k^{i,j}$ are listed in Table 2. Since $A^\dagger$ does not depend on r or Δ_ν, it is involved only in the calculation of the coefficient $c_k^{1,0}$. In particular, for the finite part,

$$c_F^{1,0} = Df_F(\bar{x}, \nu_0)v - Df^{(m)}(\bar{x}_F, \nu_0)v_F,$$

and for the tail ($k \ge m$),

$$c_k^{1,0} = Df_k(\bar{x}, \nu_0)v - \mu_k(\bar{L}, \nu_0)v_k.$$

The other coefficients $c_k^{i,j}$ can be easily generated with the help of a computer (e.g., with Maple). Here we have rearranged the terms in the output somewhat to make the formulas in

Table 2

The nonzero coefficients in the expansion (28). In the expression for $c_k^{1,0}$ we have used the notation $v_I = (0, 0, \dots, 0, v_m, v_{m+1}, v_{m+2}, \dots)$.

$k = -1$	
$c_{-1}^{1,0}$	$-2\bar{L}^2 \sum_{l=m}^{\infty} l^2 v_l - \sqrt{2}\left(\bar{a}_0 + 2\sum_{l=1}^{m-1} \bar{a}_l\right)\left(2\sum_{l=m}^{\infty} v_l\right)$
$c_{-1}^{1,1}$	$-4\left(\dot{L}\sum_{l=1}^{m-1} l^2 \bar{a}_l + \bar{L}\sum_{l=1}^{m-1} l^2 \dot{a}_l\right) v_{-1} - 4\bar{L}\dot{L}\sum_{l=1}^{\infty} l^2 v_l - \sqrt{2}\left(\dot{a}_0 + 2\sum_{l=1}^{m-1} \dot{a}_l\right)\left(v_0 + 2\sum_{l=1}^{\infty} v_l\right)$
$c_{-1}^{1,2}$	$-4\dot{L} v_{-1} \sum_{l=1}^{m-1} l^2 \dot{a}_l - 2\dot{L}^2 \sum_{l=1}^{\infty} l^2 v_l$
$c_{-1}^{2,0}$	$-4\left(v'_{-1}\sum_{l=1}^{m-1} l^2 \bar{a}_l + \bar{L}\sum_{l=1}^{\infty} l^2 v'_l\right) v_{-1} - 4\bar{L} v'_{-1} \sum_{l=1}^{\infty} l^2 v_l - \sqrt{2}\left(v'_0 + 2\sum_{l=1}^{\infty} v'_l\right)\left(v_0 + 2\sum_{l=1}^{\infty} v_l\right)$
$c_{-1}^{2,1}$	$-4\left(v'_{-1}\sum_{l=1}^{m-1} l^2 \dot{a}_l + \dot{L}\sum_{l=1}^{\infty} l^2 v'_l\right) v_{-1} - 4\dot{L} v'_{-1} \sum_{l=1}^{\infty} l^2 v_l$
$c_{-1}^{3,0}$	$-4 v'_{-1} v_{-1} \sum_{l=1}^{\infty} l^2 v'_l - 2(v'_{-1})^2 \sum_{l=1}^{\infty} l^2 v_l$
$k = 0, 1, 2, \dots$	
$c_k^{1,0}$	$\begin{cases} -3\,(\bar{a} * \bar{a} * v_I)_k & \text{for } 0 \le k \le m-1, \\ -3\,(\bar{a} * \bar{a} * v)_k & \text{for } k \ge m \end{cases}$
$c_k^{1,1}$	$-k^2\left[(4k^2\bar{L}^3\dot{L} - \bar{L}^2 - 2\nu_0\bar{L}\dot{L})v_k + 2\big((6k^2\bar{L}^2\dot{L} - \bar{L} - \nu_0\dot{L})\bar{a}_k + (2k^2\bar{L}^3 - \nu_0\bar{L})\dot{a}_k\big)v_{-1}\right] - 6\,(\bar{a} * \dot{a} * v)_k$
$c_k^{1,2}$	$-k^2\left[(6k^2\bar{L}^2\dot{L}^2 - 2\bar{L}\dot{L} - \nu_0\dot{L}^2)v_k + 2\big((6k^2\bar{L}\dot{L}^2 - \dot{L})\bar{a}_k + (6k^2\bar{L}^2\dot{L} - \bar{L} - \nu_0\dot{L})\dot{a}_k\big)v_{-1}\right] - 3\,(\dot{a} * \dot{a} * v)_k$
$c_k^{1,3}$	$-k^2\dot{L}\left[(4k^2\bar{L}\dot{L}^2 - \dot{L})v_k + 2\big(2k^2\dot{L}^2\bar{a}_k + (6k^2\bar{L}\dot{L} - 1)\dot{a}_k\big)v_{-1}\right]$
$c_k^{1,4}$	$-k^4\dot{L}^3\left[\dot{L}v_k + 4\dot{a}_k v_{-1}\right]$
$c_k^{2,0}$	$-2k^2\left[(2k^2\bar{L}^3 - \nu_0\bar{L})(v'_{-1}v_k + v'_k v_{-1}) + (6k^2\bar{L}^2 - \nu_0)\bar{a}_k v'_{-1} v_{-1}\right] - 6\,(\bar{a} * v' * v)_k$
$c_k^{2,1}$	$-2k^2\left[(6k^2\bar{L}^2\dot{L} - \bar{L} - \nu_0\dot{L})(v'_{-1}v_k + v'_k v_{-1}) + ((12k^2\bar{L}\dot{L} - 1)\bar{a}_k + (6k^2\bar{L}^2 - \nu_0)\dot{a}_k)v'_{-1}v_{-1}\right] - 6\,(\dot{a} * v' * v)_k$
$c_k^{2,2}$	$-2k^2\left[(6k^2\bar{L}\dot{L}^2 - \dot{L})(v'_{-1}v_k + v'_k v_{-1}) + (6k^2\dot{L}^2\bar{a}_k + (12k^2\bar{L}\dot{L} - 1)\dot{a}_k)v'_{-1}v_{-1}\right]$
$c_k^{2,3}$	$-4k^4\dot{L}^2\left[\dot{L}(v'_{-1}v_k + v'_k v_{-1}) + 3\dot{a}_k v'_{-1} v_{-1}\right]$
$c_k^{3,0}$	$-k^2 v'_{-1}\left[(6k^2\bar{L}^2 - \nu_0)(v'_{-1}v_k + 2v'_k v_{-1}) + 12k^2\bar{L}\bar{a}_k v'_{-1} v_{-1}\right] - 3\,(v' * v' * v)_k$
$c_k^{3,1}$	$-k^2 v'_{-1}\left[(12k^2\bar{L}\dot{L} - 1)(v'_{-1}v_k + 2v'_k v_{-1}) + 12k^2(\dot{L}\bar{a}_k + \bar{L}\dot{a}_k)v'_{-1}v_{-1}\right]$
$c_k^{3,2}$	$-6k^4\dot{L} v'_{-1}\left[\dot{L}(v'_{-1}v_k + 2v'_k v_{-1}) + 2\dot{a}_k v'_{-1} v_{-1}\right]$
$c_k^{4,0}$	$-4k^4(v'_{-1})^2\left[\bar{L}(v'_{-1}v_k + 3v'_k v_{-1}) + \bar{a}_k v'_{-1} v_{-1}\right]$
$c_k^{4,1}$	$-4k^4(v'_{-1})^2\left[\dot{L}(v'_{-1}v_k + 3v'_k v_{-1}) + \dot{a}_k v'_{-1} v_{-1}\right]$
$c_k^{5,0}$	$-k^4(v'_{-1})^3\left[v'_{-1}v_k + 4v'_k v_{-1}\right]$

Table 2 more aesthetically pleasing. However, such cosmetic changes are of course not needed for any *practical* purposes.

We now compute uniform upper bounds for the $c_k^{i,j}$, i.e., $C_k^{i,j} \ge 0$, such that

$$\left|c_k^{i,j}(\bar{x}, \dot{x}, v, v', \nu_0)\right| \le C_k^{i,j}(\bar{x}, \dot{x}, \nu_0) \qquad \text{for all } v, v' \in W(1). \tag{29}$$

The most involved are the convolution terms, and we have a dedicated lemma to estimate

those. Although the formulas are quite cumbersome, the numbers defined below are easily calculated with a computer. We introduce the computational parameter $M \in \mathbb{N}$, arbitrary for now, and we define

$$\gamma_M \stackrel{\text{def}}{=} 2\left[\frac{M}{M-1}\right]^s + \left[\frac{4\ln(M-2)}{M} + \frac{\pi^2-6}{3}\right]\left[\frac{2}{M}+\frac{1}{2}\right]^{s_*-2},$$

where s_* is the largest integer such that $s_* \le s$, and

$$\beta_k \stackrel{\text{def}}{=} \begin{cases} 4 + \frac{1}{2^{2s-1}(2s-1)}, & k=0, \\ 2\left[2+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{3^{s-1}(s-1)}\right] + \sum_{k_1=1}^{k-1} \frac{k^s}{k_1^s(k-k_1)^s}, & k=1,2,\ldots,M-1, \\ 2\left[2+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{3^{s-1}(s-1)}\right] + \gamma_M, & k=M. \end{cases}$$

Using this, we set

$$\alpha_k = \frac{6\beta_M}{(M+k)^s(M-1)^{s-1}(s-1)} + 3\sum_{j=M}^{M+k-1} \frac{\beta_{j-k}}{j^s(j-k)^s} \qquad \text{for } k=0,1,\ldots,M-1,$$

while for $k=M$ we define

$$\alpha_M \stackrel{\text{def}}{=} \beta_0 + \sum_{j=1}^{M-1} \frac{\beta_j}{j^s}\left[1+\frac{1}{\left[1-\frac{j}{M}\right]^s}\right] + \beta_M\left[2+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{3^{s-1}(s-1)}+\frac{1}{(M-1)^{s-1}(s-1)}+\gamma_M\right].$$

We introduce, for infinite sequence $a=(a_0,a_1,a_2,\ldots)$, the notation

$$|a|_M = (|a|_0, |a|_1, \ldots, |a|_{M-1}).$$

Analogous to (21) and (22), but now for sequences with index starting at 0 rather than -1, we define

$$\|a\|_s^0 \stackrel{\text{def}}{=} \sup_{k=0,1,\ldots} |a_k\omega_k^s| = \sup\{|a_0|, |a_1|, 2^s|a_2|, 3^s|a_3|, 4^s|a_4|, \ldots\}$$

and

$$W^0(r) \stackrel{\text{def}}{=} \{a,\ \|a\|_s^0 \le r\} = [-r,r] \times \prod_{k=1}^{\infty}\left[-\frac{r}{k^s}, \frac{r}{k^s}\right].$$

Lemma 18. *Let $M \ge 6$, and let a, b, and c lie in the balls $W^0(A_a)$, $W^0(A_b)$, and $W^0(A_c)$. Then for $k=0,1,\ldots,M-1$ we have*

$$(a*b*c)_k \in \left\{\left(|a|_M * |b|_M * |c|_M\right)_k + A_aA_bA_c\alpha_k\right\}[-1,1],$$

while for $k \ge M$ we have

$$(a*b*c)_k \in A_aA_bA_c\frac{\alpha_M}{k^s}[-1,1].$$

Proof. The proof is a special case of the general convolution estimates in Appendix A, with $p = 3$, $M_1 = M$, and the notation $\beta_k = \alpha_k^{(2)}$, $\alpha_k = 3\varepsilon_k^{(3)}$, and $\alpha_M = \alpha_M^{(3)}$. ∎

We are now ready to estimate the coefficients $c_k^{i,j}(\bar{x}, \dot{x}, v, v', \nu_0)$, but we first introduce a bit more notation, namely, in view of Lemma 18,

$$Q_k(a,b,c) \stackrel{\text{def}}{=} \begin{cases} \left(|a|_M * |b|_M * |c|_M\right)_k + \|a\|_s^0 \|b\|_s^0 \|c\|_s^0 \alpha_k, & k = 0, 1, \dots M-1, \\ \|a\|_s^0 \|b\|_s^0 \|c\|_s^0 \frac{\alpha_M}{k^s}, & k \geq M. \end{cases}$$

Furthermore, for $s > 1$, we use the notation

$$\zeta(s, l_0) \stackrel{\text{def}}{=} \sum_{l=l_0}^{\infty} \frac{1}{l^s} \qquad \text{and} \qquad \zeta(s) \stackrel{\text{def}}{=} \zeta(s,1) = \sum_{l=1}^{\infty} \frac{1}{l^s}$$

and their estimates (which require only a finite computation) for $l_0 \leq M$,

$$\zeta_M(s, l_0) \stackrel{\text{def}}{=} \sum_{l=l_0}^{M} \frac{1}{l^s} + \frac{1}{(M-1)^{s-1}(s-1)} \qquad \text{and} \qquad \zeta_M(s) \stackrel{\text{def}}{=} \zeta_M(s,1)$$

so that $\zeta(s, l_0) \leq \zeta_M(s, l_0)$ and $\zeta(s) \leq \zeta_M(s)$. Finally, let

$$\mathbb{I} \stackrel{\text{def}}{=} \left(1, 1, \frac{1}{2^s}, \frac{1}{3^s}, \frac{1}{4^s}, \dots\right) \qquad \text{and} \qquad \mathbb{I}_I \stackrel{\text{def}}{=} \left(0, 0, \dots, 0, \frac{1}{m^s}, \frac{1}{(m+1)^s}, \dots\right).$$

With this notation in place, and using Lemma 18, the bounds $C_k^{i,j}(\bar{x}, \dot{x}, \nu_0)$ satisfying (29), listed in Table 3, are now straightforward to derive. For fixed k, these constants $C_k^{i,j}$ each involve only a finite computation, but there are of course still infinitely many values of k to consider. Notice first that for $k \geq m$ many terms in Table 3 vanish, since only the first m elements of $\bar{a}$ and $\dot{a}$ are nonzero. For the same reason, calculating $\|\bar{a}\|_s^0$ and $\|\dot{a}\|_s^0$ is a finite computation. Moreover, many terms can be estimated using the fact that, for any $A_1, A_2 \in \mathbb{R}$,

$$\left|A_1 + \frac{A_2}{k^2}\right| \leq \max\left\{\left|A_1 + \frac{A_2}{M^2}\right|, |A_1|\right\} \qquad \text{for all } k \geq M.$$

It follows from these considerations and Lemma 18 that

$$C_k^{i,j} \leq \widehat{C}_M^{i,j}\, k^{4-s} \qquad \text{for } k \geq M \geq \min\{m, 6\},$$

where the $\widehat{C}_M^{i,j}$ are listed in Table 4.

To conclude the calculation of Z_k we need an estimate on

$$|\mu_k(\bar{L}, \nu_0)| = \bar{L}^4 k^4 \left|1 - \frac{\nu_0}{\bar{L}^2 k^2} + \frac{1}{\bar{L}^4 k^4}\right|$$

for large k. Let

$$M_0(\bar{L}, \nu_0) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{for } \nu_0 \leq 0, \\ \sqrt{2\nu_0}/\bar{L} & \text{for } \nu_0 > 0. \end{cases} \tag{30}$$

Table 3

The uniform bounds $C_k^{i,j}(\bar{x},\dot{x},\nu_0)$ on the coefficients $c_k^{i,j}(\bar{x},\dot{x},v,v',\nu_0)$. For $k=0$ one should read $k^{2-s}=0$ and $k^{4-s}=0$, irrespective of s.

$k=-1$	
$C_{-1}^{1,0}$	$2\,\bar{L}^2\,\zeta_M(s-2,m)+2\sqrt{2}\left\lvert\bar{a}_0+2\sum_{l=1}^{m-1}\bar{a}_l\right\rvert\zeta_M(s,m)$
$C_{-1}^{1,1}$	$4\left\lvert\dot{L}\sum_{l=1}^{m-1}l^2\bar{a}_l+\bar{L}\sum_{l=1}^{m-1}l^2\dot{a}_l\right\rvert+4\lvert\bar{L}\dot{L}\rvert\,\zeta_M(s-2)+\sqrt{2}\left\lvert\dot{a}_0+2\sum_{l=1}^{m-1}\dot{a}_l\right\rvert\left[1+2\,\zeta_M(s)\right]$
$C_{-1}^{1,2}$	$4\left\lvert\dot{L}\sum_{l=1}^{m-1}l^2\dot{a}_l\right\rvert+2\dot{L}^2\,\zeta_M(s-2)$
$C_{-1}^{2,0}$	$8\lvert\bar{L}\rvert\,\zeta_M(s-2)+4\left\lvert\sum_{l=1}^{m-1}l^2\bar{a}_l\right\rvert+\sqrt{2}\left[1+2\,\zeta_M(s)\right]^2$
$C_{-1}^{2,1}$	$4\left\lvert\sum_{l=1}^{m-1}l^2\dot{a}_l\right\rvert+8\lvert\dot{L}\rvert\,\zeta_M(s-2)$
$C_{-1}^{3,0}$	$6\,\zeta_M(s-2)$
$k=0,1,2,\ldots$	
$C_k^{1,0}$	$\begin{cases}3\,Q_k(\bar{a},\bar{a},\mathbb{I}_I) & \text{for } 0\le k\le m-1,\\ 3\,Q_k(\bar{a},\bar{a},\mathbb{I}) & \text{for } k\ge m\end{cases}$
$C_k^{1,1}$	$\left\lvert4k^2\bar{L}^3\dot{L}-\bar{L}^2-2\nu_0\bar{L}\dot{L}\right\rvert k^{2-s}+2\left\lvert2k^2\left(3\bar{L}^2\dot{L}\bar{a}_k+\bar{L}^3\dot{a}_k\right)-\left(\bar{L}\bar{a}_k+\nu_0\dot{L}\bar{a}_k+\nu_0\bar{L}\dot{a}_k\right)\right\rvert k^2+6\,Q_k(\bar{a},\dot{a},\mathbb{I})$
$C_k^{1,2}$	$\left\lvert6k^2\bar{L}^2\dot{L}^2-2\bar{L}\dot{L}-\nu_0\dot{L}^2\right\rvert k^{2-s}+2\left\lvert6k^2\left(\bar{L}\dot{L}^2\bar{a}_k+\bar{L}^2\dot{L}\dot{a}_k\right)-\left(\dot{L}\bar{a}_k+\bar{L}\dot{a}_k+\nu_0\dot{L}\dot{a}_k\right)\right\rvert k^2+3\,Q_k(\dot{a},\dot{a},\mathbb{I})$
$C_k^{1,3}$	$\left\lvert4k^2\bar{L}\dot{L}^3-\dot{L}^2\right\rvert k^{2-s}+2\left\lvert2k^2\left(\dot{L}^3\bar{a}_k+3\bar{L}\dot{L}^2\dot{a}_k\right)-\dot{L}\dot{a}_k\right\rvert k^2$
$C_k^{1,4}$	$\dot{L}^4k^{4-s}+4\lvert\dot{L}^3\dot{a}_k\rvert k^4$
$C_k^{2,0}$	$4\left\lvert2k^2\bar{L}^3-\nu_0\bar{L}\right\rvert k^{2-s}+2\left\lvert6k^2\bar{L}^2\bar{a}_k-\nu_0\bar{a}_k\right\rvert k^2+6\,Q_k(\bar{a},\mathbb{I},\mathbb{I})$
$C_k^{2,1}$	$4\left\lvert6k^2\bar{L}^2\dot{L}-\bar{L}-\nu_0\dot{L}\right\rvert k^{2-s}+2\left\lvert6k^2\left(2\bar{L}\dot{L}\bar{a}_k+\bar{L}^2\dot{a}_k\right)-\bar{a}_k-\nu_0\dot{a}_k\right\rvert k^2+6\,Q_k(\dot{a},\mathbb{I},\mathbb{I})$
$C_k^{2,2}$	$4\left\lvert6k^2\bar{L}\dot{L}^2-\dot{L}\right\rvert k^{2-s}+2\left\lvert6k^2\left(\dot{L}^2\bar{a}_k+2\bar{L}\dot{L}\dot{a}_k\right)-\dot{a}_k\right\rvert k^2$
$C_k^{2,3}$	$8\lvert\dot{L}^3\rvert k^{4-s}+12\dot{L}^2\lvert\dot{a}_k\rvert k^4$
$C_k^{3,0}$	$3\left\lvert6k^2\bar{L}^2-\nu_0\right\rvert k^{2-s}+12\lvert\bar{L}\bar{a}_k\rvert k^4+3\,Q_k(\mathbb{I},\mathbb{I},\mathbb{I})$
$C_k^{3,1}$	$3\left\lvert12k^2\bar{L}\dot{L}-1\right\rvert k^{2-s}+12\left\lvert\dot{L}\bar{a}_k+\bar{L}\dot{a}_k\right\rvert k^4$
$C_k^{3,2}$	$18\dot{L}^2k^{4-s}+12\lvert\dot{L}\dot{a}_k\rvert k^4$
$C_k^{4,0}$	$16\lvert\bar{L}\rvert k^{4-s}+4\lvert\bar{a}_k\rvert k^4$
$C_k^{4,1}$	$16\lvert\dot{L}\rvert k^{4-s}+4\lvert\dot{a}_k\rvert k^4$
$C_k^{5,0}$	$5\,k^{4-s}$

Table 4
The uniform bounds $\widehat{C}_M^{i,j}$ on $k^{s-4}C_k^{i,j}(\bar{x},\dot{x},\nu_0)$ for $k \geq M$.

$\widehat{C}_M^{1,0}$	$\frac{3(\|\bar{a}\|_s^0)^2\alpha_M}{M^4}$
$\widehat{C}_M^{1,1}$	$\max\left\{\left\|4\bar{L}^3\dot{L} - \frac{\bar{L}^2+2\nu_0\bar{L}\dot{L}}{M^2}\right\|, 4\left\|\bar{L}^3\dot{L}\right\|\right\} + \frac{6\|\bar{a}\|_s^0\|\dot{a}\|_s^0\alpha_M}{M^4}$
$\widehat{C}_M^{1,2}$	$\max\left\{\left\|6\bar{L}^2\dot{L}^2 - \frac{2\bar{L}\dot{L}+\nu_0\dot{L}^2}{M^2}\right\|, 6\bar{L}^2\dot{L}^2\right\} + \frac{3(\|\dot{a}\|_s^0)^2\alpha_M}{M^4}$
$\widehat{C}_M^{1,3}$	$\max\left\{\left\|4\bar{L}\dot{L}^3 - \frac{\dot{L}^2}{M^2}\right\|, 4\left\|\bar{L}\dot{L}^3\right\|\right\}$
$\widehat{C}_M^{1,4}$	$\dot{L}^4$
$\widehat{C}_M^{2,0}$	$4\max\left\{\left\|2\bar{L}^3 - \frac{\nu_0\bar{L}}{M^2}\right\|, 2\left\|\bar{L}^3\right\|\right\} + \frac{6\|\bar{a}\|_s^0\alpha_M}{M^4}$
$\widehat{C}_M^{2,1}$	$4\max\left\{\left\|6\bar{L}^2\dot{L} - \frac{\bar{L}+\nu_0\dot{L}}{M^2}\right\|, 6\bar{L}^2\left\|\dot{L}\right\|\right\} + \frac{6\|\dot{a}\|_s^0\alpha_M}{M^4}$
$\widehat{C}_M^{2,2}$	$4\max\left\{\left\|6\bar{L}\dot{L}^2 - \frac{\dot{L}}{M^2}\right\|, 6\left\|\bar{L}\right\|\dot{L}^2\right\}$
$\widehat{C}_M^{2,3}$	$8\left\|\dot{L}^3\right\|$
$\widehat{C}_M^{3,0}$	$3\max\left\{\left\|6\bar{L}^2 - \frac{\nu_0}{M^2}\right\|, 6\bar{L}^2\right\} + \frac{3\alpha_M}{M^4}$
$\widehat{C}_M^{3,1}$	$3\max\left\{\left\|12\bar{L}\dot{L} - \frac{1}{M^2}\right\|, 12\left\|\bar{L}\dot{L}\right\|\right\}$
$\widehat{C}_M^{3,2}$	$18\dot{L}^2$
$\widehat{C}_M^{4,0}$	$16\left\|\bar{L}\right\|$
$\widehat{C}_M^{4,1}$	$16\left\|\dot{L}\right\|$
$\widehat{C}_M^{5,0}$	5

Then it is not hard to check that

$$|\mu_k(\bar{L},\nu_0)| \geq \frac{\bar{L}^4k^4}{2} \qquad \text{for } k \geq M_0.$$

Using the vectors $C^{i,j}$ and the numbers $\widehat{C}_M^{i,j}$ from Tables 3 and 4, and in view of (27), this leads to bounds $Z_k(r,\Delta_\nu)$ as listed below, with $M \geq \min\{M_0, m, 6\}$ and $\Delta_\nu \geq 0$:

$$Z_F = |I - J_F \cdot Df^{(m)}(\bar{x}_F,\nu_0)| \cdot \mathbb{I}_F\, r + \sum_{i=1}^{5}\sum_{j=0}^{5-i}\left|J_F\right| \cdot C_F^{i,j}\, r^i\, \Delta_\nu^j \tag{31a}$$

for $k = -1, 0, 1, \ldots, m-1$;

$$Z_k = \frac{1}{|\mu_k(\bar{L},\nu_0)|}\sum_{i=1}^{5}\sum_{j=0}^{5-i} C_k^{i,j}\, r^i\, \Delta_\nu^j \tag{31b}$$

for $m \le k \le M-1$; and

$$Z_k = \frac{2}{\bar{L}^4}\frac{1}{k^s}\sum_{i=1}^{5}\sum_{j=0}^{5-i} \widehat{C}_M^{i,j}\, r^i\, \Delta_\nu^j \tag{31c}$$

for $k \ge M$. Finally, for the purpose of Definition 14 and Lemma 15, we set

$$\widehat{Z}_M \overset{\text{def}}{=} \frac{2}{\bar{L}^4 M^s}\sum_{i=1}^{5}\sum_{j=0}^{5-i} \widehat{C}_M^{i,j}\, r^i\, \Delta_\nu^j, \tag{32}$$

so that $Z_k = \widehat{Z}_M\big(\frac{M}{k}\big)^s$.

4. Verification of the geometric properties $\mathcal{H}$. We now put ourselves in the situation of a single, successful, rigorous continuation step, where we have found an $r > 0$ such that the set ($s > 3$)

$$W_{x_\nu}(r) = x_\nu + W(r) \qquad \text{with } x_\nu = \bar{x} + (\nu-\nu_0)\dot{x}, \tag{33}$$

centered at the predictor based at ν_0, contains a unique fixed point $\widetilde{x}^\nu$ of $T(x,\nu)$ for each parameter value $\nu \in [\nu_0,\nu_1]$. We write $\widetilde{x}^\nu = (\widetilde{L}^\nu, \widetilde{a}_0^\nu, \widetilde{a}_1^\nu, \widetilde{a}_2^\nu, \ldots)$. The functions $\widetilde{u}^\nu$ defined via (11) are periodic solutions of (4) with period $2\pi/\widetilde{L}^\nu$, which are symmetric in $y=0$ and $y=\pi/\widetilde{L}^\nu$. For convenience, we incorporate the period of the periodic solution in the definition of the geometric condition as follows:

$$\mathcal{H}_{\widetilde{L}}\ \begin{cases} (H_1) & \widetilde{u} \text{ has exactly four monotone laps and extrema } \{\widetilde{u}_i\}_{i=1}^4 \text{ on } [0, 2\pi/\widetilde{L}], \\ (H_2) & \widetilde{u}_1 \text{ and } \widetilde{u}_3 \text{ are minima, and } \widetilde{u}_2 \text{ and } \widetilde{u}_4 \text{ are maxima,} \\ (H_3) & \widetilde{u}_1 < -1 < \widetilde{u}_3 < 1 < \widetilde{u}_2, \widetilde{u}_4, \\ (H_4) & \widetilde{u}(x) \text{ is symmetric in its minima } \widetilde{u}_1 \text{ and } \widetilde{u}_3. \end{cases}$$

We need to make sure that the unique zero of f in W_{x_ν} satisfies these properties. The following lemma will help us in the verification process, since it shows that we need only to check the conditions for *one* parameter value along any continuous branch of solutions.

Lemma 19. *Let $\widetilde{u}^\nu$, $\nu_0 \le \nu \le \nu_1$, be periodic solutions of (4) at the energy level $E=0$ with period $2\pi/\widetilde{L}^\nu$, which are symmetric in $y=0$ and $y=\pi/\widetilde{L}^\nu$. Suppose that $\widetilde{u}^\nu$ and $\widetilde{L}^\nu$ depend continuously on ν; i.e., $\widetilde{u}^\nu$ depends continuously on ν as a C^3-function on compact intervals. If $\widetilde{u}^{\nu_0}$ satisfies $\mathcal{H}_{\widetilde{L}^{\nu_0}}$, then $\widetilde{u}^\nu$ satisfies $\mathcal{H}_{\widetilde{L}^\nu}$ for all $\nu \in [\nu_0,\nu_1]$.*

Proof. To reduce clutter, we remove all tildes from the notation. By symmetry, we need only to consider the interval $[0,\pi/L^\nu]$. Let

$$N = \big\{\nu \in [\nu_0,\nu_1] \;\big|\; u^\nu \text{ satisfies } \mathcal{H}_{L^\nu}\big\}.$$

By assumption, $\nu_0 \in N$. We will show that N is both open and closed (in the relative topology), i.e., connected; hence $N = [\nu_0,\nu_1]$ as asserted.

It is relatively easy to see that N is open. Namely, for $\nu \in N$ the extrema of u^ν do not lie on the lines $u = \pm 1$. It then follows from the energy identity $E = 0$ that the extrema of u^ν are all nondegenerate. Hence, the conditions $H_{1,2,3,4}$ are open conditions (under the symmetry assumption in the lemma).

To prove that N is closed is a bit more involved. Let $\{\nu_n\}_{n=1}^\infty \subset N$ be a sequence converging to $\nu_* \in [\nu_0, \nu_1]$. Our goal is to show that $\nu_* \in N$. We denote $u^n = u^{\nu_n}$ and $u_* = u^{\nu_*}$. Let the extrema $u^n_{1,2,3}$ be attained in $y^n_{1,2,3}$. Clearly, we have that $y^n_1 = 0$ and $y^n_3 = \pi/L^\nu_n$, while, taking a subsequence, we may additionally assume that y^n_2 converges to some y^*_2 as $n \to \infty$. Denote also $y^*_1 = 0$ and $y^*_3 = \pi/L^{\nu_*}$. By C^3-continuity, we have $u'_*(y^*_{1,2,3}) = 0$, and

$$u_*(y^*_1) \le -1 \le u_*(y^*_3) \le 1 \le u_*(y^*_2). \tag{34}$$

In fact, the inequalities are strict. We prove this for the last inequality $u_*(y^*_2) > 1$; the other cases are analogous. Suppose, by contradiction, that $u_*(y^*_2) = 1$. Since $u'_*(y^*_2) = 0$, it follows from $E = 0$ that $u''_*(y^*_2) = 0$. By continuity,

$$\max_y u_*(y) = \lim_{n\to\infty} \max_y u^n(y) = \lim_{n\to\infty} u^n(y^n_2) = u_*(y^*_2) = 1.$$

This implies that $u'''_*(y^*_2) = 0$. Uniqueness of the initial value problem for the ODE then says that $u_*(y) = 1$ for all y, which contradicts $u_*(y^*_1) \le -1$. Similarly one can show that all the other inequalities in (34) are strict; hence u_* has at least four extrema on $[0, 2\pi/L^{\nu_*}]$, and those satisfy H_3.

The final step is to prove that u_* does not have more than four monotone laps. We argue once more by contradiction. Recall that $y^*_2 = \lim_{n\to\infty} y^n_2$. Suppose there is a point $z \in (0, \pi/L^{\nu_*})$ with $u'_*(z) = 0$, and $z \ne y^*_2$. If $u''_*(z) \ne 0$, then, by the implicit function theorem, this extremum persists for u^n for n sufficiently large, leading to more than four monotone laps of u^n, contradicting the fact that u^n satisfies the geometric conditions. Hence, it must be that $u''_*(z) = 0$, and thus $u_*(z) = \pm 1$, since $E = 0$. Moreover, since $u_* \not\equiv \pm 1$, we must have $u'''_*(z) \ne 0$. Let us consider the case $u_*(z) = 1$ and $u'''_*(z) > 0$; all other (three) cases are analogous.

We thus have

$$u_*(z) = 1, \quad u'_*(z) = 0, \quad u''_*(z) = 0, \quad u'''_*(z) > 0. \tag{35}$$

Clearly $u'_*(y) > 0$ for y sufficiently close but not equal to z. By continuity, we have $(u^n)'(z \pm \varepsilon) > 0$ for ε sufficiently small and n large enough. By the implicit function theorem, for large enough n, there exist points $z_n \in [z-\varepsilon, z+\varepsilon]$ such that $\lim_{n\to\infty} z_n = z$ and $(u^n)''(z_n) = 0$, and $(u^n)'(z_n) \ne 0$, since u^n has no additional extrema in $(0, L^{\nu_n})$ besides y^n_2. In fact, $(u^n)'(z_n) > 0$, since, if $(u^n)'(z_n) < 0$, then u^n would have two extrema in $[z-\varepsilon, z+\varepsilon]$, leading to more than four monotone laps of u^n, which is a contradiction. Hence $(u^n)'(z_n) > 0$.

We conclude from $E = 0$ and $(u^n)''(z_n) = 0$ that

$$\Big[(u^n)'''(z_n) + \frac{\nu_n}{2}(u^n)'(z_n)\Big](u^n)'(z_n) = -\frac{1}{4}(u^n(z_n)^2 - 1)^2.$$

Since $(u^n)'(z_n) > 0$, this means that

$$(u^n)'''(z_n) + \frac{\nu_n}{2}(u^n)'(z_n) \le 0.$$

Finally, we take the limit $n \to \infty$ in the above inequality to obtain

$$u_*'''(z) + \frac{\nu_*}{2} u_*'(z) \leq 0,$$

which contradicts (35). Hence u_* indeed has exactly four monotone laps on $[0, 2\pi/L^{\nu_*}]$, implying that $\nu_* \in N$ and that N is closed. ∎

We now need only to show that the geometric properties are satisfied at $\nu = \nu_0$, since the solutions depend continuously on ν.

Lemma 20. *The fixed point $\widetilde{x}_\nu \in \Omega^s$ depends continuously on ν for $\nu \in [\nu_0, \nu_1]$. Similarly, the corresponding periodic solutions $\widetilde{u}_\nu$ depend continuously on ν as C^3-functions on compact intervals.*

Proof. Recall that we are dealing with a single continuation step $\nu \in [\nu_0, \nu_1]$, so that the neighborhoods on which T is a contraction mapping are given by (33). The assertion now follows from the continuity and compactness properties of the map T, described in Lemma 12, using standard functional analytic arguments. ∎

To check that $\widetilde{u}^{\nu_0}$ has the properties $\mathcal{H}_{\widetilde{L}^{\nu_0}}$, we follow the procedure outlined below. To reduce clutter, we often drop ν_0 from the notation. We introduce the variables $z = \widetilde{L}^{\nu_0} y$ and $v(z) = \widetilde{u}^{\nu_0}(y)$, so that

$$v(z) = \widetilde{a}_0 + 2 \sum_{k=1}^{\infty} \widetilde{a}_k \cos(kz).$$

This way, we separate the shape of the solution from the period; only the shape is important for the geometric conditions. Clearly, $v'(0) = v'(\pi) = 0$, and v is symmetric in those extrema.

We recall that $\bar{x} = (\bar{L}, \bar{a}_0, \bar{a}_1, \ldots, \bar{a}_{m-1}, 0, 0, \ldots)$, while the fixed point is given by $\widetilde{x} = (\widetilde{L}, \widetilde{a}_0, \widetilde{a}_1, \widetilde{a}_2, \ldots)$. We have $\widetilde{a}_k \in \mathbf{a}_k$, where the intervals are given by

$$\mathbf{a}_k \stackrel{\text{def}}{=} \begin{cases} [\bar{a}_0 - r,\, \bar{a}_0 + r], & k = 0, \\ [\bar{a}_k - \frac{r}{k^s},\, \bar{a}_k + \frac{r}{k^s}], & k = 1, \ldots, m-1, \\ [-\frac{r}{k^s}, \frac{r}{k^s}], & k \geq m. \end{cases}$$

Consider $z \in \mathbf{z} \stackrel{\text{def}}{=} [z^-, z^+] \subset \mathbb{R}$. Then, using interval arithmetic, we can compute rigorous interval enclosures of $v(z)$, $v'(z)$, and $v''(z)$:

$$\begin{aligned} v(z) \in\ & \mathbf{v}[\mathbf{z}] \stackrel{\text{def}}{=} \mathbf{a}_0 + 2 \sum_{k=1}^{m-1} \mathbf{a}_k \cos(k\mathbf{z}) + \frac{2r}{(m-1)^{s-1}(s-1)}[-1,1], \\ v'(z) \in\ & \mathbf{v}'[\mathbf{z}] \stackrel{\text{def}}{=} -2 \sum_{k=1}^{m-1} \mathbf{a}_k\, k \sin(k\mathbf{z}) + \frac{2r}{(m-1)^{s-2}(s-2)}[-1,1], \\ v''(z) \in\ & \mathbf{v}''[\mathbf{z}] \stackrel{\text{def}}{=} -2 \sum_{k=1}^{m-1} \mathbf{a}_k\, k^2 \cos(k\mathbf{z}) + \frac{2r}{(m-1)^{s-3}(s-3)}[-1,1]. \end{aligned}$$

We now use the following procedure (see also Figure 9). Note that we know a priori that $v'(0) = v'(\pi) = 0$.

Procedure 21. *Checking that $\widetilde{u}^{\nu_0}$ satisfies $\mathcal{H}_{\widetilde{L}^{\nu_0}}$ is equivalent to verifying that v satisfies $\mathcal{H}_\pi$. We proceed as follows.*

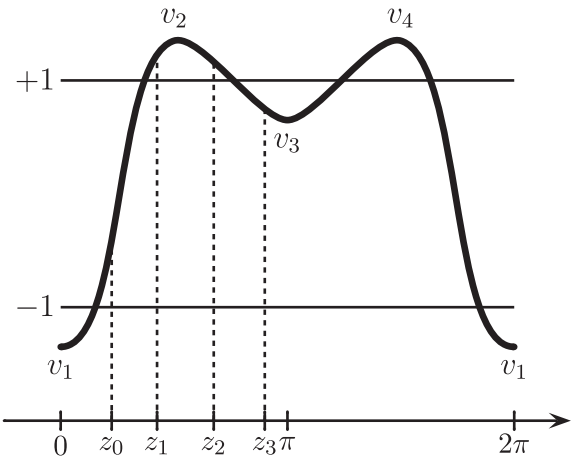


Figure 9. *Illustration of the procedure to make sure that v satisfies $\mathcal{H}_\pi$, i.e., the periodic solution $\widetilde{u}^{\nu_0}$ satisfies the geometric conditions $\mathcal{H}_{\tilde{L}_0^\nu}$. The extrema are denoted by $v_i = \widetilde{u}_i^{\nu_0}$.*

1. *Verify that* $\mathbf{v}[0] \subset (-\infty, -1)$. *That implies that* $\widetilde{u}_1^{\nu_0} = v(0) < -1$.
2. *Find an (approximately largest)* $z_0 > 0$ *such that* $\mathbf{v}''[0, z_0] \subset (0, \infty)$. *Hence, there is a unique extremum in* $[0, z_0]$, *namely, a minimum, at* $z = 0$.
3. *Find an (approximately largest)* $z_1 > z_0$ *such that* $\mathbf{v}'[z_0, z_1] \subset (0, \infty)$. *Hence, the interval* $[z_0, z_1]$ *does not contain any extremum.*
4. *Verify that* $\mathbf{v}[z_1] \subset (1, \infty)$.
5. *Find an (approximately largest)* $z_2 > z_1$ *such that both* $\mathbf{v}[z_1, z_2] \subset (1, \infty)$ *and* $\mathbf{v}''[z_1, z_2] \subset (-\infty, 0)$.
6. *Verify that* $\mathbf{v}'[z_2] \subset (-\infty, 0)$. *That implies that there is a unique extremum* z_* *in* $[z_1, z_2]$, *namely, a maximum* $\widetilde{u}_2^{\nu_0} = v(z_*) > 1$.
7. *Find an (approximately largest)* $z_3 > z_2$ *such that* $\mathbf{v}'[z_2, z_3] \subset (-\infty, 0)$. *Hence, the interval* $[z_2, z_3]$ *does not contain any extremum.*
8. *Verify that* $\mathbf{v}''[z_3, \pi] \subset (0, \infty)$ *and* $\mathbf{v}[\pi] \subset (-1, 1)$. *That implies that there is a unique extremum in* $[z_3, \pi]$, *namely, a minimum* $\widetilde{u}_3^{\nu_0} = v(\pi) \in (-1, 1)$ *at* $z = \pi$.

Combining Lemma 19 with Procedure 21 leads to the required result.

Lemma 22. *The choice of the approximate zero* $\widehat{x}_F^*$ *of* $f^{(m)}(x_F, 0)$ *in Lemma 17 can be made such that for each of the resulting intervals* $[\nu_0, \nu_0 + \Delta_\nu^0]$ *covering* $[0, 2]$, *which were extracted in Procedure 16, Procedure 21 is successful at* ν_0. *Hence the solutions found in Lemma 17 via Procedure 16 satisfy the geometric conditions* $\mathcal{H}$.

The animation (70912_02.gif [1.32MB]) accompanying this paper shows the changing shape of the periodic solution $\widetilde{u}^\nu$ as the parameter ν increases from 0 to 2.

Proof. A MATLAB computer program (*SH_geometric_properties.m*; see also section 3.1) successfully performing Procedure 21 accompanies the paper. The numerical implementation of Procedure 21 is rather straightforward. Steps 1, 4, and 6 and the second part of Step 8 are mere evaluations of a function using interval arithmetic. Steps 2, 3, 5, and 7 and the first part of Step 8 are all implemented in the same fashion. For instance, we describe here what is done in the implementation of Step 3. First, we consider a mesh $\{x_0, \dots, x_n\}$ of the interval

$[z_0, \pi]$, and we find the largest $k \in \{1, \dots, n\}$ for which we have that $\mathbf{v}'[x_{i-1}, x_i] \subset (1, \infty)$ for all $i \in \{1, \dots, k\}$. We then let $z_1 = x_i$. Note that the smaller the mesh size, the nearer z_1 will be to a zero of v'. Every verification thus requires a series of evaluations of $\mathbf{v}$, $\mathbf{v}'$, and $\mathbf{v}''$ using interval arithmetic. In the implementation, we chose the mesh size to be 0.01. ■

In conclusion, Theorem 3 is a consequence of Lemmas 17 and 22, which are based on Procedures 16 and 21, respectively.

Appendix A. Estimates for infinite convolution sums with power decay. In this section, we present two lemmas that are fundamental in the construction of the radii polynomials. Let $s \geq 2$ be a real number and $M \geq 6$ a natural number.

We introduce an improvement of general estimates for infinite convolution sums with power decay of the form

$$\sum_{k_1+\cdots+k_p=k} a^{(1)}_{k_1} \cdots a^{(p)}_{k_p}, \tag{36}$$

introduced in [12, 14] and used in [13, 15, 18] (the special case $p = 2$ was considered earlier in [38]). Most of the estimates used in the above papers are corollaries of Lemma 5.8 in [14].

Lemma 23 (from [14]). *Let $A > 0$ and $s \geq 2$. Let $\{a_k\}_{k\in\mathbb{Z}}$ be such that $a_{-k} = a_k$, $a_0 \in A[-1, 1]$, and $a_k \in \frac{A}{|k|^s}$ for all $k \in \mathbb{Z} \setminus \{0\}$. Let $\alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$. Then*

$$\sum_{\sum n_i=k} a_{n_1} \cdots a_{n_p} \subseteq \begin{cases} \alpha^{p-1} A^p[-1, 1], & k = 0, \\ \frac{\alpha^{p-1}A^p}{|k|^s}[-1, 1], & k \neq 0. \end{cases}$$

Observe that the coefficient α provided by Lemma 23 grows exponentially in s. One reason for being interested in getting tighter analytic estimates for sums of the form (36) comes from the fact that, in solving (13) and (14), we need $p = 3$ and $s \geq 4$. If we use the bounds given by Lemma 23, the computational cost of the rigorous continuation will dramatically increase, since we will need to use a very large computational parameter M. A lower bound on M (depending on the α of Lemma 23) can actually be found in [18, section 2.2].

In this appendix we consider general values of the degree p of the convolution and the decay power s, since the specific case is hardly any simpler than the general one. Moreover, the general convolution estimates may be of use for future applications of the method laid out in this paper.

Throughout this appendix we assume that $a_{-k} = a_k$ for all $k \in \mathbb{Z}$. Since

$$\sum_{\substack{k_1+\cdots+k_p=-k \\ k_i\in\mathbb{Z}}} a^{(1)}_{k_1} \cdots a^{(p)}_{k_p} = \sum_{\substack{k_1+\cdots+k_p=k \\ k_i\in\mathbb{Z}}} a^{(1)}_{k_1} \cdots a^{(p)}_{k_p},$$

we consider only the cases $k \in \mathbb{N}$. Note that the estimates are also applicable to the situation where $a_{-k} = -a_k$ for all k.

Before introducing the new general estimates, we need the following result.

Lemma 24. *Let $s \geq 2$, and let s_* be the largest integer such that $s_* \leq s$. Let, for $k \geq 4$,*

$$\gamma_k \overset{\text{def}}{=} 2\left[\frac{k}{k-1}\right]^s + \left[\frac{4\ln(k-2)}{k} + \frac{\pi^2 - 6}{3}\right]\left[\frac{2}{k} + \frac{1}{2}\right]^{s_*-2}. \tag{37}$$

Then, for $k \geq 4$,

$$\sum_{k_1=1}^{k-1} \frac{k^s}{k_1^s (k-k_1)^s} \leq \gamma_k.$$

Proof. First observe that

$$\begin{aligned}
\sum_{k_1=1}^{k-1} \frac{k^s}{k_1^s (k-k_1)^s} &= 2\left[\frac{k}{k-1}\right]^s + \sum_{k_1=2}^{k-2} \frac{k^s}{k_1^s (k-k_1)^s} \\
&= 2\left[\frac{k}{k-1}\right]^s + k^{s-1} \sum_{k_1=2}^{k-2} \frac{(k-k_1)+k_1}{k_1^s (k-k_1)^s} \\
&= 2\left[\frac{k}{k-1}\right]^s + k^{s-1} \left[\sum_{k_1=2}^{k-2} \frac{1}{k_1^s (k-k_1)^{s-1}} + \sum_{k_1=2}^{k-2} \frac{1}{k_1^{s-1} (k-k_1)^s}\right] \\
&= 2\left[\frac{k}{k-1}\right]^s + 2\sum_{k_1=2}^{k-2} \frac{k^{s-1}}{k_1^{s-1} (k-k_1)^s}.
\end{aligned}$$

We now set, using the above,

$$\begin{aligned}
\phi_k^{(s)} &\stackrel{\text{def}}{=} \sum_{k_1=2}^{k-2} \frac{k^{s-1}}{k_1^{s-1} (k-k_1)^s} \\
&= \frac{1}{2} \sum_{k_1=2}^{k-2} \frac{k^s}{k_1^s (k-k_1)^s}.
\end{aligned}$$

We obtain the recurrence inequality

$$\begin{aligned}
\phi_k^{(s)} &= \sum_{k_1=2}^{k-2} \frac{k^{s-1}}{k_1^{s-1} (k-k_1)^s} = k^{s-2} \sum_{k_1=2}^{k-2} \frac{(k-k_1)+k_1}{k_1^{s-1} (k-k_1)^s} \\
&= k^{s-2} \left[\sum_{k_1=2}^{k-2} \frac{1}{k_1^{s-1} (k-k_1)^{s-1}} + \sum_{k_1=2}^{k-2} \frac{1}{k_1^{s-2} (k-k_1)^s}\right] \\
&= \frac{1}{k} \sum_{k_1=2}^{k-2} \frac{k^{s-1}}{k_1^{s-1} (k-k_1)^{s-1}} + \sum_{k_1=2}^{k-2} \frac{k^{s-2}}{k_1^{s-2} (k-k_1)^s} \\
&\leq \frac{1}{k} \sum_{k_1=2}^{k-2} \frac{k^{s-1}}{k_1^{s-1} (k-k_1)^{s-1}} + \frac{1}{2} \sum_{k_1=2}^{k-2} \frac{k^{s-2}}{k_1^{s-2} (k-k_1)^{s-1}} \\
&= \left[\frac{2}{k} + \frac{1}{2}\right] \phi_k^{(s-1)}.
\end{aligned}$$

Hence, since $\frac{k}{k_1(k-k_1)} \leq 1$ for $2 \leq k_1 \leq k-2$ and $k \geq 4$,

$$\phi_k^{(s)} \leq \phi_k^{(s_*)} \leq \phi_k^{(2)} \left[\frac{2}{k} + \frac{1}{2}\right]^{s_*-2},$$

where s_* is the largest integer such that $s_* \leq s$, and

$$\begin{aligned}\phi_k^{(2)} &= \sum_{k_1=2}^{k-2} \frac{k}{k_1(k-k_1)^2} = \sum_{k_1=2}^{k-2} \frac{1}{k_1(k-k_1)} + \sum_{k_1=2}^{k-2} \frac{1}{(k-k_1)^2} \\ &= \frac{2}{k}\sum_{k_1=2}^{k-2} \frac{1}{k_1} + \sum_{k_1=2}^{k-2} \frac{1}{k_1^2} \leq \frac{2}{k}\ln(k-2) + \frac{\pi^2}{6} - 1.\end{aligned}$$

By combining the above inequalities, we conclude that

$$\sum_{k_1=1}^{k-1} \frac{k^s}{k_1^s(k-k_1)^s} \leq 2\left[\frac{k}{k-1}\right]^s + \left[\frac{4\ln(k-2)}{k} + \frac{\pi^2-6}{3}\right]\left[\frac{2}{k}+\frac{1}{2}\right]^{s_*-2} = \gamma_k. \qquad \blacksquare$$

Note that the estimates will be given via a recurrent definition in p, i.e., the power of the nonlinearity. Hence, we begin by getting explicitly the estimates for the case $p = 2$. Throughout this note, we use $M \geq 6$ as a computational parameter; its use is primarily to make all the estimates computable in practice.

A.1. Estimates for the quadratic nonlinearity.

Lemma 25 (quadratic estimates). *Let $s \geq 2$ and $M \geq 6$. Define*

$$\alpha_k^{(2)} \overset{\text{def}}{=} \begin{cases} 4 + \frac{1}{2^{2s-1}(2s-1)} & \text{for } k = 0, \\ 2\left[2 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{3^{s-1}(s-1)}\right] + \sum_{k_1=1}^{k-1} \frac{k^s}{k_1^s(k-k_1)^s} & \text{for } 1 \leq k \leq M-1, \\ 2\left[2 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{3^{s-1}(s-1)}\right] + \gamma_k & \text{for } k \geq M. \end{cases}$$

Let $A_1, A_2 > 0$ such that $a_0^{(i)} \in A_i[-1,1]$ and $a_k^{(i)} \in \frac{A_i}{|k|^s}[-1,1]$ for all $k \neq 0$ and for $i = 1, 2$. Suppose that $a_{-k}^{(i)} = a_k^{(i)}$. Then

$$\sum_{\substack{k_1+k_2=k \\ k_i \in \mathbb{Z}}} a_{k_1}^{(1)} a_{k_2}^{(2)} \in \begin{cases} \alpha_0^{(2)} A_1 A_2\,[-1,1], & k = 0, \\ \frac{\alpha_k^{(2)} A_1 A_2}{|k|^s}\,[-1,1], & k \neq 0. \end{cases}$$

Proof. Let $k = 0$. Then

$$\begin{aligned}\sum_{\substack{k_1+k_2=0 \\ k_i \in \mathbb{Z}}} a_{k_1}^{(1)} a_{k_2}^{(2)} &= \sum_{k_1<0} a_{k_1}^{(1)} a_{-k_1}^{(2)} + a_0^{(1)} a_0^{(2)} + \sum_{k_1>0} a_{k_1}^{(1)} a_{-k_1}^{(2)} \\ &= a_0^{(1)} a_0^{(2)} + 2\sum_{k_1=1}^{\infty} a_{k_1}^{(1)} a_{k_1}^{(2)} \\ &\in A_1 A_2 \left[1 + 2\sum_{k_1=1}^{\infty} \frac{1}{k_1^{2s}}\right][-1,1] \\ &\subseteq A_1 A_2 \left[4 + \frac{1}{2^{2s-1}(2s-1)}\right][-1,1] \\ &= \alpha_0^{(2)} A_1 A_2\,[-1,1].\end{aligned}$$

Now consider $k \in \{1, \dots, M-1\}$. Then

$$\begin{aligned}\sum_{\substack{k_1+k_2=k\\ k_i\in\mathbb{Z}}} a^{(1)}_{k_1}a^{(2)}_{k_2} &= \sum_{k_1=-\infty}^{-1} a^{(1)}_{k_1}a^{(2)}_{k-k_1} + a^{(1)}_0 a^{(2)}_k + \sum_{k_1=1}^{k-1} a^{(1)}_{k_1}a^{(2)}_{k-k_1} + a^{(1)}_k a^{(2)}_0 + \sum_{k_1=k+1}^{\infty} a^{(1)}_{k_1}a^{(2)}_{k-k_1}\\ &\in A_1A_2\left[\frac{2}{k^s} + 2\sum_{k_1=1}^{\infty}\frac{1}{k_1^s(k+k_1)^s} + \frac{1}{k^s}\sum_{k_1=1}^{k-1}\frac{k^s}{k_1^s(k-k_1)^s}\right][-1,1]\\ &\subset A_1A_2\left[\frac{2}{k^s} + \frac{2}{k^s}\sum_{k_1=1}^{\infty}\frac{1}{k_1^s} + \frac{1}{k^s}\sum_{k_1=1}^{k-1}\frac{k^s}{k_1^s(k-k_1)^s}\right][-1,1]\\ &\subseteq \frac{\alpha^{(2)}_k A_1A_2}{k^s}[-1,1],\end{aligned}$$

where we, quite arbitrarily, have bound the infinite sum $\sum_{k_1=1}^{\infty}\frac{1}{k_1^s}$ using an integral estimate after the third term. For the case $k \geq M$, we do the same analysis as in the case $k \in \{1, \dots, M-1\}$, and we use the upper bound γ_k from Lemma 24. ■

Remark 26. For any $k \geq M \geq 6$, we have that $\alpha^{(2)}_k \leq \alpha^{(2)}_M$.

Proof. For $k \geq 6$, the fact that $\frac{\ln(k-1)}{(k+1)} \leq \frac{\ln(k-2)}{k}$ implies that $\gamma^{(s)}_{k+1} \leq \gamma^{(s)}_k$. The conclusion then follows from the definition $\alpha^{(2)}_k$ for $k \geq M \geq 6$. ■

A.2. Estimates for a general nonlinearity. Let $p \geq 3$ be the degree of the nonlinearity, $s \geq 2$ the decay of the coefficients, and $M \geq 6$ a natural number. We compute the general estimates recursively. Hence, we first suppose that for every $k \geq 0$, we know explicitly $\alpha^{(p-1)}_k > 0$ such that

$$\sum_{\substack{k_1+\cdots+k_{p-1}=k\\ k_i\in\mathbb{Z}}} a^{(1)}_{k_1}\cdots a^{(p-1)}_{k_{p-1}} \in \begin{cases}\alpha^{(p-1)}_0\left(\prod_{i=1}^{p-1}A_i\right)[-1,1], & k=0,\\ \frac{\alpha^{(p-1)}_k}{|k|^s}\left(\prod_{i=1}^{p-1}A_i\right)[-1,1], & k\neq 0,\end{cases}$$

and such that $\alpha^{(p-1)}_k \leq \alpha^{(p-1)}_M$ for all $k \geq M$. We define

$$\alpha^{(p)}_k \overset{\text{def}}{=} \begin{cases}\alpha^{(p-1)}_0 + 2\sum_{k_p=1}^{M-1}\frac{\alpha^{(p-1)}_{k_p}}{k_p^{2s}} + \frac{2\alpha^{(p-1)}_M}{(M-1)^{2s-1}(2s-1)} & \text{for } k=0,\\ \sum_{k_p=1}^{M-k-1}\frac{\alpha^{(p-1)}_{k+k_p}k^s}{k_p^s(k+k_p)^s} + \alpha^{(p-1)}_M\left(1+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{3^{s-1}(s-1)}\right)\\ \quad + \alpha^{(p-1)}_k + \sum_{k_p=1}^{k-1}\frac{\alpha^{(p-1)}_{k_p}k^s}{k_p^s(k-k_p)^s} + \alpha^{(p-1)}_0 + \sum_{k_p=1}^{M-1}\frac{\alpha^{(p-1)}_{k_p}k^s}{(k+k_p)^s k_p^s}\\ \quad + \frac{\alpha^{(p-1)}_M}{(M-1)^{s-1}(s-1)} & \text{for } 1\leq k\leq M-1,\\ \alpha^{(p-1)}_M\left[2+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{3^{s-1}(s-1)}+\frac{1}{(M-1)^{s-1}(s-1)}+\gamma_k\right]\\ \quad + \alpha^{(p-1)}_0 + \sum_{k_p=1}^{M-1}\frac{\alpha^{(p-1)}_{k_p}}{k_p^s}\left[1+\frac{1}{\left[1-\frac{k_p}{M}\right]^s}\right] & \text{for } k\geq M.\end{cases}$$

Lemma 27. *For $i = 1, \dots, p$, let $A_i > 0$ such that $a_0^{(i)} \in A_i[-1,1]$ and $a_k^{(i)} \in \frac{A_i}{|k|^s}[-1,1]$ for all $k \neq 0$. Suppose that $a_{-k}^{(i)} = a_k^{(i)}$. Then*

$$\sum_{\substack{k_1+\cdots+k_p=k \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_p}^{(p)} \in \begin{cases} \alpha_0^{(p)} \left(\prod_{i=1}^p A_i\right)[-1,1], & k = 0, \\ \frac{\alpha_k^{(p)}}{|k|^s} \left(\prod_{i=1}^p A_i\right)[-1,1], & k \neq 0. \end{cases}$$

Proof. Several times throughout the proof, we use that $\alpha_k^{(p-1)} \leq \alpha_M^{(p-1)}$ for all $k \geq M$. For $k = 0$,

$$\begin{aligned} \sum_{\substack{k_1+\cdots+k_p=0 \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_p}^{(p)} &= \sum_{k_p=-\infty}^{-1} a_{k_p}^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=-k_p \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} \\ &\quad + a_0^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=0 \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} \\ &\quad + \sum_{k_p=1}^{\infty} a_{k_p}^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=-k_p \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} \\ &\in \left(\prod_{i=1}^p A_i\right) \left[\sum_{k_p=1}^{\infty} \frac{\alpha_{k_p}^{(p-1)}}{k_p^{2s}} + \alpha_0^{(p-1)} + \sum_{k_p=1}^{\infty} \frac{\alpha_{k_p}^{(p-1)}}{k_p^{2s}}\right][-1,1] \\ &\subseteq \left(\prod_{i=1}^p A_i\right) \left[\alpha_0^{(p-1)} + 2\sum_{k_p=1}^{M-1} \frac{\alpha_{k_p}^{(p-1)}}{k_p^{2s}} + \frac{2\alpha_M^{(p-1)}}{(M-1)^{2s-1}(2s-1)}\right][-1,1] \\ &= \alpha_0^{(p)} \left(\prod_{i=1}^p A_i\right)[-1,1]. \end{aligned}$$

For any $k \geq 1$,

$$\begin{aligned} \sum_{\substack{k_1+\cdots+k_p=k \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_p}^{(p)} &= \sum_{k_p=-\infty}^{-1} a_{k_p}^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=k-k_p \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} \\ &\quad + a_0^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=k \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} + \sum_{k_p=1}^{k-1} a_{k_p}^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=k-k_p \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} \\ &\quad + a_k^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=0 \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} + \sum_{k_p=k+1}^{\infty} a_{k_p}^{(p)} \sum_{\substack{k_1+\cdots+k_{p-1}=k-k_p \\ k_i\in\mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_{p-1}}^{(p-1)} \end{aligned}$$

$$\in \left(\prod_{i=1}^{p} A_i\right)\left[\sum_{k_p=1}^{\infty}\frac{\alpha_{k+k_p}^{(p-1)}}{k_p^s(k+k_p)^s}+\frac{\alpha_k^{(p-1)}}{k^s}+\sum_{k_p=1}^{k-1}\frac{\alpha_{k_p}^{(p-1)}}{k_p^s(k-k_p)^s}+\frac{\alpha_0^{(p-1)}}{k^s}+\sum_{k_p=1}^{\infty}\frac{\alpha_{k_p}^{(p-1)}}{(k+k_p)^s k_p^s}\right][-1,1].$$

Consider $k\in\{1,\dots,M-1\}$. Since $\alpha_{k_p}^{(p-1)}\le\alpha_M^{(p-1)}$ for all $k_p\ge M$, we have

$$\begin{aligned}\sum_{k_p=1}^{\infty}\frac{\alpha_{k+k_p}^{(p-1)}}{k_p^s(k+k_p)^s}&=\sum_{k_p=1}^{M-k-1}\frac{\alpha_{k+k_p}^{(p-1)}}{k_p^s(k+k_p)^s}+\sum_{k_p=M-k}^{\infty}\frac{\alpha_{k+k_p}^{(p-1)}}{k_p^s(k+k_p)^s}\\&\le\sum_{k_p=1}^{M-k-1}\frac{\alpha_{k+k_p}^{(p-1)}}{k_p^s(k+k_p)^s}+\alpha_M^{(p-1)}\sum_{k_p=M-k}^{\infty}\frac{1}{k_p^s(k+k_p)^s}\\&\le\sum_{k_p=1}^{M-k-1}\frac{\alpha_{k+k_p}^{(p-1)}}{k_p^s(k+k_p)^s}+\alpha_M^{(p-1)}\sum_{k_p=1}^{\infty}\frac{1}{k_p^s(k+k_p)^s}\\&\le\frac{1}{k^s}\left[\sum_{k_p=1}^{M-k-1}\frac{\alpha_{k+k_p}^{(p-1)}k^s}{k_p^s(k+k_p)^s}+\alpha_M^{(p-1)}\left(1+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{3^{s-1}(s-1)}\right)\right].\end{aligned}$$

Similarly,

$$\sum_{k_p=1}^{\infty}\frac{\alpha_{k_p}^{(p-1)}}{(k+k_p)^s k_p^s}\le\frac{1}{k^s}\left[\sum_{k_p=1}^{M-1}\frac{\alpha_{k_p}^{(p-1)}k^s}{(k+k_p)^s k_p^s}+\frac{\alpha_M^{(p-1)}}{(M-1)^{s-1}(s-1)}\right].$$

Recalling the definition of $\alpha_k^{(p)}$ for the cases $k\in\{1,\dots,M-1\}$, we get that

$$\sum_{\substack{k_1+\cdots+k_p=k\\k_i\in\mathbb{Z}}}a_{k_1}^{(1)}\cdots a_{k_p}^{(p)}\in\frac{\alpha_k^{(p)}}{k^s}\left(\prod_{i=1}^{p}A_i\right)[-1,1].$$

Consider now $k\ge M$; then

$$\sum_{k_p=1}^{\infty}\frac{\alpha_{k+k_p}^{(p-1)}}{k_p^s(k+k_p)^s}+\frac{\alpha_k^{(p-1)}}{k^s}\le\frac{\alpha_M^{(p-1)}}{k^s}\left[2+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{3^{s-1}(s-1)}\right].$$

Using Lemma 24, we get that

$$\sum_{k_p=1}^{k-1} \frac{\alpha_{k_p}^{(p-1)}}{k_p^s(k-k_p)^s} = \sum_{k_p=1}^{M-1} \frac{\alpha_{k_p}^{(p-1)}}{k_p^s(k-k_p)^s} + \frac{1}{k^s} \sum_{k_p=M}^{k-1} \frac{k^s \alpha_{k_p}^{(p-1)}}{k_p^s(k-k_p)^s}$$
$$\leq \frac{1}{k^s} \sum_{k_p=1}^{M-1} \frac{\alpha_{k_p}^{(p-1)}}{k_p^s\left(1-\frac{k_p}{k}\right)^s} + \frac{\alpha_M^{(p-1)}}{k^s} \sum_{k_p=M}^{k-1} \frac{k^s}{k_p^s(k-k_p)^s}$$
$$\leq \frac{1}{k^s} \left[\sum_{k_p=1}^{M-1} \frac{\alpha_{k_p}^{(p-1)}}{k_p^s\left(1-\frac{k_p}{M}\right)^s} + \alpha_M^{(p-1)} \gamma_k \right].$$

Also,

$$\sum_{k_p=1}^{\infty} \frac{\alpha_{k_p}^{(p-1)}}{(k+k_p)^s k_p^s} \leq \frac{1}{k^s} \left[\sum_{k_p=1}^{M-1} \frac{\alpha_{k_p}^{(p-1)}}{k_p^s} + \frac{\alpha_M^{(p-1)}}{(M-1)^{s-1}(s-1)} \right].$$

Combining the three above inequalities, we finally have that

$$\sum_{\substack{k_1+\cdots+k_p=k \\ k_i \in \mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_p}^{(p)}$$
$$\in \frac{1}{k^s} \left(\prod_{i=1}^{p} A_i \right) \left[\alpha_0^{(p-1)} + \sum_{k_p=1}^{M-1} \frac{\alpha_{k_p}^{(p-1)}}{k_p^s} \left(1 + \frac{1}{\left(1-\frac{k_p}{M}\right)^s} \right) \right.$$
$$\left. + \alpha_M^{(p-1)} \left(2 + \tfrac{1}{2^s} + \tfrac{1}{3^s} + \tfrac{1}{3^{s-1}(s-1)} + \tfrac{1}{(M-1)^{s-1}(s-1)} + \gamma_k \right) \right] [-1,1]$$
$$= \frac{\alpha_k^{(p)}}{k^s} \left(\prod_{i=1}^{p} A_i \right) [-1,1]. \quad \blacksquare$$

Remark 28. For any $k \geq M \geq 6$, we have that $\alpha_k^{(p)} \leq \alpha_M^{(p)}$.

Proof. The proof is identical to that of Remark 26. $\blacksquare$

A.3. Comparison of the general estimates. We now compare the new estimates with the ones given by Lemma 23 for different values of p and s. Since the only difference in the estimates is α^{p-1} versus $\alpha_k^{(p)}$, these are the quantities we compare in Table 5. In particular, the new estimates lead to an improvement of a factor 10^2 for the values $p = 3$ and $s = 4$ used in this paper, while for higher values of p and s they become even more beneficial. For the computation, we fixed $M = 100$; in the case $p \geq 3$, increasing M would make the $\alpha_k^{(p)}$ smaller still.

A.4. Refinement for $k \in \{0, \ldots, M-1\}$. We present a corollary of Lemma 27 which gives better bounds for $0 \leq k \leq M-1$.

Corollary 29. *Let $p \geq 3$ be the degree of the nonlinearity, $s \geq 2$ the decay of the coefficients, and $M \geq 6$ a natural number. Consider another computational number $M_1 \geq M$. Let the $\{\alpha_k^{(p-1)}\}_{k \in \{0,\ldots,M_1\}}$ be defined in Lemma 27. For $i = 1, \ldots, p$, let $A_i > 0$ be such that*

Table 5
Comparison of the estimates α^{p-1} versus $\alpha_k^{(p)}$ used in Lemmas 23 and 27.

p	s	k	α^{p-1}	$\alpha_k^{(p)}$
2	4	10	$5.87 \cdot 10^{1}$	$6.65 \cdot 10^{0}$
3	4	30	$3.44 \cdot 10^{3}$	$4.44 \cdot 10^{1}$
3	4	90	$3.44 \cdot 10^{3}$	$4.31 \cdot 10^{1}$
3	5	30	$1.31 \cdot 10^{4}$	$4.20 \cdot 10^{1}$
3	7	30	$2.03 \cdot 10^{5}$	$4.12 \cdot 10^{1}$
3	10	30	$1.29 \cdot 10^{7}$	$4.26 \cdot 10^{1}$
3	50	30	$1.55 \cdot 10^{31}$	$1.68 \cdot 10^{2}$
4	4	10	$2.02 \cdot 10^{5}$	$3.28 \cdot 10^{2}$
4	5	10	$1.50 \cdot 10^{6}$	$3.17 \cdot 10^{2}$
4	7	10	$9.13 \cdot 10^{7}$	$3.59 \cdot 10^{2}$
5	10	10	$1.65 \cdot 10^{14}$	$3.98 \cdot 10^{3}$
5	20	10	$1.81 \cdot 10^{26}$	$1.82 \cdot 10^{5}$
10	25	20	$4.25 \cdot 10^{72}$	$1.75 \cdot 10^{8}$
20	50	90	$2.07 \cdot 10^{296}$	$5.01 \cdot 10^{15}$

$a_0^{(i)} \in A_i[-1,1]$ *and* $a_k^{(i)} \in \frac{A_i}{|k|^s}[-1,1]$ *for all* $k \neq 0$, *and let* $|a|_{M_1}^{(i)} = \left(|a_0^{(i)}|, \dots, |a_{M_1-1}^{(i)}|\right)$. *Suppose that* $a_{-k}^{(i)} = a_k^{(i)}$. *For* $k \in \{0, \dots, M-1\}$, *define*

$$\varepsilon_k^{(p)} = \frac{2\alpha_{M_1}^{(p-1)}}{(M_1+k)^s(M_1-1)^{s-1}(s-1)} + \sum_{k_p=M_1}^{M_1+k-1} \frac{\alpha_{k_p-k}^{(p-1)}}{k_p^s(k_p-k)^s}.$$

Then we have that, for $k \in \{0, \dots, M-1\}$,

$$\left(a^{(1)} * \cdots * a^{(p)}\right)_k \in \left[\left(|a|_{M_1}^{(1)} * \cdots * |a|_{M_1}^{(p)}\right)_k + \left(\prod_{i=1}^{p} A_i\right) p\varepsilon_k^{(p)}\right][-1,1].$$

Proof. First notice that

$$\begin{aligned}\left(a^{(1)} * \cdots * a^{(p)}\right)_k &= \sum_{\substack{k_1+\cdots+k_p=k \\ k_i \in \mathbb{Z}}} a_{k_1}^{(1)} \cdots a_{k_p}^{(p)} \\ &= \sum_{\substack{k_1+\cdots+k_p=k \\ |k_i|<M_1}} a_{k_1}^{(1)} \cdots a_{k_p}^{(p)} + \sum_{\substack{k_1+\cdots+k_p=k \\ \max\{|k_1|,\dots,|k_p|\} \geq M_1}} a_{k_1}^{(1)} \cdots a_{k_p}^{(p)}.\end{aligned}$$

We have that

$$\sum_{\substack{k_1+\cdots+k_p=k\\ |k_p|\geq M_1}} a^{(1)}_{k_1}\cdots a^{(p)}_{k_p} = \sum_{k_p=-\infty}^{-M_1} a^{(p)}_{k_p} \sum_{k_1+\cdots+k_{p-1}=k-k_p} a^{(1)}_{k_1}\cdots a^{(p-1)}_{k_{p-1}} + \sum_{k_p=M_1}^{\infty} a^{(p)}_{k_p} \sum_{k_1+\cdots+k_{p-1}=k-k_p} a^{(1)}_{k_1}\cdots a^{(p-1)}_{k_{p-1}}$$

$$\in \left(\prod_{i=1}^{p} A_i\right) \sum_{k_p=M_1}^{\infty} \left[\frac{\alpha^{(p-1)}_{k+k_p}}{k_p^s(k+k_p)^s} + \frac{\alpha^{(p-1)}_{k_p-k}}{k_p^s(k_p-k)^s}\right][-1,1]$$

$$\subseteq \left(\prod_{i=1}^{p} A_i\right) \left[2\alpha^{(p-1)}_{M_1} \sum_{k_p=M_1}^{\infty} \frac{1}{k_p^s(k+k_p)^s} + \sum_{k_p=M_1}^{M_1+k-1} \frac{\alpha^{(p-1)}_{k_p-k}}{k_p^s(k_p-k)^s}\right][-1,1]$$

$$\subseteq \left(\prod_{i=1}^{p} A_i\right) \left[\frac{2\alpha^{(p-1)}_{M_1}}{(M_1+k)^s(M_1-1)^{s-1}(s-1)} + \sum_{k_p=M_1}^{M_1+k-1} \frac{\alpha^{(p-1)}_{k_p-k}}{k_p^s(k_p-k)^s}\right][-1,1].$$

Recalling the definition of $\varepsilon^{(p)}_k$, we can conclude that

$$\left(a^{(1)} * \cdots * a^{(p)}\right)_k \in \left[\left(|a|^{(1)}_{M_1} * \cdots * |a|^{(p)}_{M_1}\right)_k + \left(\prod_{i=1}^{p} A_i\right) p\varepsilon^{(p)}_k\right][-1,1]. \qquad \blacksquare$$

Acknowledgments. The authors would like to thank Marcio Gameiro and Rob van der Vorst for helpful discussions.

REFERENCES

[1] G. Arioli and P. Zgliczyński, *Symbolic dynamics for the Hénon-Heiles Hamiltonian on the critical level*, J. Differential Equations, 171 (2001), pp. 173–202.
[2] J. B. van den Berg, L. A. Peletier, and W. C. Troy, *Global branches of multi-bump periodic solutions of the Swift-Hohenberg equation*, Arch. Ration. Mech. Anal., 158 (2001), pp. 91–153.
[3] J. B. van den Berg and R. C. van der Vorst, *Second order Lagrangian twist systems: Simple closed characteristics*, Trans. Amer. Math. Soc., 354 (2002), pp. 1393–1420.
[4] J. B. van den Berg and R. C. van der Vorst, *Stable patterns for fourth-order parabolic equations*, Duke Math. J., 115 (2002), pp. 513–558.
[5] J. B. van den Berg, J.-P. Lessard, and K. Mischaikow, *Rigorous Branch Following*, in preparation.
[6] B. Buffoni, *Periodic and homoclinic orbits for Lorentz-Lagrangian systems via variational methods*, Nonlinear Anal., 26 (1996), pp. 443–462.
[7] B. Buffoni, A. R. Champneys, and J. F. Toland, *Bifurcation and coalescence of a plethora of homoclinic orbits for a Hamiltonian system*, J. Dynam. Differential Equations, 8 (1996), pp. 221–279.
[8] B. Buffoni and J. F. Toland, *Global existence of homoclinic and periodic orbits for a class of autonomous Hamiltonian systems*, J. Differential Equations, 118 (1995), pp. 104–120.
[9] J. Burke and E. Knobloch, *Localized states in the generalized Swift-Hohenberg equation*, Phys. Rev. E (3), 73 (2006), 056211.
[10] A. R. Champneys, *Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics*, Phys. D, 112 (1998), pp. 158–186.
[11] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.

[12] S. Day, *A Rigorous Numerical Method in Infinite Dimensions*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 2003.
[13] S. Day, Y. Hiraoka, K. Mischaikow, and T. Ogawa, *Rigorous numerics for global dynamics: A study of the Swift–Hohenberg equation*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 1–31.
[14] S. Day, O. Junge, and K. Mischaikow, *A rigorous numerical method for the global analysis of infinite-dimensional discrete dynamical systems*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 117–160.
[15] S. Day, J.-P. Lessard, and K. Mischaikow, *Validated continuation for equilibria of PDEs*, SIAM J. Numer. Anal., 45 (2007), pp. 1398–1424.
[16] R. L. Devaney, *Homoclinic orbits in Hamiltonian systems*, J. Differential Equations, 21 (1976), pp. 431–438.
[17] M. Gameiro, T. Gedeon, W. Kalies, H. Kokubu, K. Mischaikow, and H. Oka, *Topological horseshoes of travelling waves for a fast-slow predator-prey system*, J. Dynam. Differential Equations, 19 (2007), pp. 623–654.
[18] M. Gameiro, J.-P. Lessard, and K. Mischaikow, *Validated continuation over large parameter ranges for equilibria of PDEs*, Math. Comput. Simulation, to appear.
[19] R. W. Ghrist, J. B. van den Berg, and R. C. van der Vorst, *Morse theory on spaces of braids and Lagrangian dynamics*, Invent. Math., 152 (2003), pp. 369–432.
[20] G. I. Hargreaves, *Interval Analysis in MATLAB*, Numerical Analysis Report, University of Manchester, Manchester, UK, 2002.
[21] M. F. Hilali, S. Métens, P. Borckmans, and G. Dewel, *Pattern selection in the generalized Swift-Hohenberg model*, Phys. Rev. E (3), 51 (1995), pp. 2046–2052.
[22] W. D. Kalies, J. Kwapisz, J. B. van den Berg, and R. C. A. M. van der Vorst, *Homotopy classes for stable periodic and chaotic patterns in fourth-order Hamiltonian systems*, Comm. Math. Phys., 214 (2000), pp. 573–592.
[23] W. D. Kalies, J. Kwapisz, and R. C. A. M. van der Vorst, *Homotopy classes for stable connections between Hamiltonian saddle-focus equilibria*, Comm. Math. Phys., 193 (1998), pp. 337–371.
[24] W. D. Kalies and R. C. A. M. van der Vorst, *Multitransition homoclinic and heteroclinic solutions of the extended Fisher-Kolmogorov equation*, J. Differential Equations, 131 (1996), pp. 209–228.
[25] T. Y. Li and J. A. Yorke, *Period three implies chaos*, Amer. Math. Monthly, 82 (1975), pp. 985–992.
[26] K. Mischaikow and M. Mrozek, *Chaos in the Lorenz equations: A computer-assisted proof*, Bull. Amer. Math. Soc. (N.S.), 32 (1995), pp. 66–72.
[27] L. A. Peletier and W. C. Troy, *Chaotic spatial patterns described by the extended Fisher-Kolmogorov equation*, J. Differential Equations, 129 (1996), pp. 458–508.
[28] L. A. Peletier and W. C. Troy, *Spatial Patterns: Higher Order Models in Physics and Mechanics*, Progr. Nonlinear Differential Equations Appl. 45, Birkhäuser Boston, Boston, 2001.
[29] M. A. Peletier, *Sequential buckling: A variational analysis*, SIAM J. Math. Anal., 32 (2001), pp. 1142–1168.
[30] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, Stud. Adv. Math., 2nd ed., CRC Press, Boca Raton, FL, 1999.
[31] W. van Saarloos, *Front propagation into unstable states*, Phys. Rep., 386 (2003), pp. 29–222.
[32] J. Swift and P. C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. A, 15 (1977), pp. 319–328.
[33] A. Szymczak, *The Conley index and symbolic dynamics*, Topology, 35 (1996), pp. 287–299.
[34] W. P. Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. (N.S.), 19 (1988), pp. 417–431.
[35] W. Tucker, *The Lorenz attractor exists*, C. R. Acad. Sci. Paris Sér. I Math., 328 (1999), pp. 1197–1202.
[36] D. Wilczak, *Chaos in the Kuramoto-Sivashinsky equations—a computer-assisted proof*, J. Differential Equations, 194 (2003), pp. 433–459.
[37] N. Yamamoto, *A numerical verification method for solutions of boundary value problems with local uniqueness by Banach's fixed-point theorem*, SIAM J. Numer. Anal., 35 (1998), pp. 2004–2013.
[38] P. Zgliczyński and K. Mischaikow, *Rigorous numerics for partial differential equations: The Kuramoto-Sivashinsky equation*, Found. Comput. Math., 1 (2001), pp. 255–288.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 1032–1048
© 2008 Society for Industrial and Applied Mathematics

Existence of a Reversible T-Point Heteroclinic Cycle in a Piecewise Linear Version of the Michelson System*

Victoriano Carmona†, Fernando Fernández-Sánchez†, and Antonio E. Teruel‡

Abstract. The proof of the existence of a global connection in differential systems is generally a difficult task. Some authors use numerical techniques to show this existence, even in the case of continuous piecewise linear systems. In this paper we give an analytical proof of the existence of a reversible T-point heteroclinic cycle in a continuous piecewise linear version of the widely studied Michelson system. The principal ideas of this proof can be extended to other piecewise linear systems.

Key words. piecewise linear systems, heteroclinic orbits, invariant manifolds

AMS subject classifications. 34C23, 34C37, 37G99

DOI. 10.1137/070709542

1. Introduction.

The existence of global connections in a differential system usually forces a complex dynamical behavior in a neighborhood of such connections. For instance, under the presence of a homoclinic cycle to a saddle-focus equilibrium point satisfying an eigenvalue ratio condition, the celebrated works of Shil'nikov [26, 27] ensure the existence of infinitely many periodic orbits of saddle type accumulating to the homoclinic cycle.

Moreover, the existence of a global connection in a differential system implies the appearance of subsidiary connections for certain perturbations of the system. For example, an analysis of the bifurcation structure of homoclinic cycles and subsidiary connections can be found in [15].

Heteroclinic cycles are also organizing centers of a very complex dynamic [10, 11, 12, 14]. In particular, Dumortier, Ibañez, and Kokubu [11] conjecture the existence of an infinite set of bifurcation phenomena, called a *cocoon* bifurcation [20], accumulating at a reversible T-point, that is, a point of the parameter space where a special kind of heteroclinic cycle satisfying some nondegeneracy condition appears. Furthermore, they explain the occurrence of such bifurcation phenomena as a consequence of the presence of this global connection.

Unfortunately, for nonlinear differential systems it is not easy to guarantee the existence of a global connection. Even though this is possible, some other extra conditions, for instance,

*Received by the editors November 29, 2007; accepted for publication (in revised form) by J. Meiss May 1, 2008; published electronically September 25, 2008.
http://www.siam.org/journals/siads/7-3/70954.html

†Departamento de Matemática Aplicada II, Universidad de Sevilla, Escuela Superior de Ingenieros, Camino de los Descubrimientos s/n, 41092 Seville, Spain (vcarmona@us.es, fefesan@us.es). These authors were partially supported by the Ministerio de Ciencia y Tecnología, Plan Nacional I+D+I, under the projects MTM2004-04066, MTM2006-00847, and MTM2007-64193 and by the Conserjería de Educación y Ciencia de la Junta de Andalucía (TIC-0130, EXC/2005/FQM-872).

‡Departament de Matemàtiques i Informàtica, Universitat de les Illes Balears, Carretera de Valldemossa km. 7.5, 07122 Palma de Mallorca, Spain (antonioe.teruel@uib.es). This author was partially supported by the MCYT grant MTM2005-06098-C02-1 and by UIB grant UIB2005/6.

the nondegeneracy conditions in the case of a reversible T-point heteroclinic cycle, cannot always be verified in a rigorous way. For example, the existence of a heteroclinic cycle in the Michelson system [25, 28], having an explicit expression for one of its heteroclinic orbits, is known [16, 18]. Nevertheless, the genericity conditions that determine if such a heteroclinic cycle is a reversible T-point heteroclinic cycle remain to be verified [11].

On the other hand, there are a lot of papers devoted to the existence of global connections in piecewise linear differential systems [2, 8, 9, 22, 23, 24]. Many of these works require numerical arguments to show the existence of the global connections. However, in [22], the authors provide an analytical proof of the existence of a homoclinic cycle in a three-dimensional piecewise linear system. In a similar way, in the present work we pay attention to the existence of a reversible T-point heteroclinic cycle in piecewise linear systems.

Now, some ideas for establishing the main result of the paper are introduced. The Michelson system is the one-parameter family of autonomous three-dimensional differential systems

$$\begin{cases} \dot{x} = y, \\ \dot{y} = z, \\ \dot{z} = d^2 - y - \dfrac{1}{2}x^2, \end{cases} \tag{1.1}$$

where the dot stands for the derivative with respect to t and, without loss of generality, we can assume that $d \geq 0$. This family appears in the study of traveling wave solutions of the one-dimensional Kuramoto–Sivashinsky equation [25]. It also arises in the analysis of the unfolding of the nilpotent singularity of codimension three [10, 13].

For $d \neq 0$ the Michelson system has two equilibrium points, $\mathbf{p}_\pm = (\pm d\sqrt{2}, 0, 0)$, which are of saddle-focus type. The stable manifold $W^s(\mathbf{p}_+)$ of $\mathbf{p}_+$ (respectively, the unstable manifold $W^u(\mathbf{p}_-)$ of $\mathbf{p}_-$) is one-dimensional and the unstable manifold $W^u(\mathbf{p}_+)$ of $\mathbf{p}_+$ (respectively, the stable manifold $W^s(\mathbf{p}_-)$ of $\mathbf{p}_-$) is two-dimensional.

The vector field $\mathbf{f}$ associated to the Michelson system (1.1) satisfies the following important properties that affect the solution set:

- The divergence of $\mathbf{f}$ is identically zero. Therefore, the family is volume-preserving.
- The vector field $\mathbf{f}$ is invariant under the linear involution $\mathbf{R}(x, y, z) = (-x, y, -z)$ and sign reverse; that is,
$$\mathbf{R}(\mathbf{f}(\mathbf{x})) = -\mathbf{f}(\mathbf{R}(\mathbf{x})),$$
where $\mathbf{x} = (x, y, z)^T$.

From the second property, the following dynamical consequences are obtained:

- Let $\phi(t; \mathbf{p})$ denote the flow of system (1.1). The equality $\phi(t; \mathbf{R}(\mathbf{p})) = \mathbf{R}(\phi(-t; \mathbf{p}))$ holds, and we say that the family is time-reversible with respect to the linear involution $\mathbf{R}$.
- If $\mathbf{p}$ is a point on the y-axis, then $\mathbf{R}(\mathbf{p}) = \mathbf{p}$. Therefore, the orbit through $\mathbf{p}$ is reversible with respect to $\mathbf{R}$, that is, $\phi(t; \mathbf{p}) = \mathbf{R}(\phi(-t; \mathbf{p}))$, and the y-axis is called the reversibility axis.
- By the reversibility, the stable and unstable manifolds of $\mathbf{p}_-$ satisfy $W^s(\mathbf{p}_-) = \mathbf{R}\left(W^u(\mathbf{p}_+)\right)$ and $W^u(\mathbf{p}_-) = \mathbf{R}\left(W^s(\mathbf{p}_+)\right)$, respectively.

An interesting object of the dynamics of the Michelson system which has been widely studied (see [11, 16, 19, 20] and the references therein) is the so-called reversible T-point

heteroclinic cycle. Following [11], we say that the Michelson system (1.1) has a reversible T-point heteroclinic cycle $\Gamma = \{\mathbf{p}_+\} \cup \rho_\pm \cup \{\mathbf{p}_-\} \cup \rho_\mp$ for the parameter value d_0 if the following hold:

(a) The two-dimensional manifolds $W^u(\mathbf{p}_+)$ and $W^s(\mathbf{p}_-)$ have a transversal intersection along the heteroclinic orbit $\rho_\pm$.

(b) As the parameter d is varied around d_0 the heteroclinic orbit $\rho_\mp$ unfolds generically. That is, the one-dimensional manifolds $W^u(\mathbf{p}_-)$ and $W^s(\mathbf{p}_+)$ intersect at $\rho_\mp$ for $d = d_0$ and the "distance" between $W^u(\mathbf{p}_-)$ and $W^s(\mathbf{p}_+)$ measured in a transverse plane is diffeomorphic to $\mu = d - d_0$.

The value d_0 is usually called a reversible T-point. Note that some authors [11] use the name Bykov cycle to refer to a T-point heteroclinic cycle [3, 4, 5].

By abuse of notation, we can call the orbit $\rho_\mp$ a one-dimensional heteroclinic orbit, because this orbit corresponds to the one-dimensional invariant manifolds of the singular points, and the orbit $\rho_\pm$ a two-dimensional heteroclinic orbit because it is contained in the intersection set of the two-dimensional invariant manifolds of the singular points.

From several recent works it is possible to discern that piecewise linear systems are able to reproduce the dynamics of differentiable systems. Thus, it is natural to wonder if a suitable continuous piecewise linear version of the Michelson system with a reversible T-point exists.

An easy way to obtain a continuous piecewise linear system from the Michelson system is to perform, for $d \neq 0$, the change of variables $x \to x/d^2$, $y \to y/d^2$, $z \to z/d^2$ followed by the change of function $x^2 \to |x|$. This procedure transforms system (1.1) into

$$\begin{cases} \dot{x} = y, \\ \dot{y} = z, \\ \dot{z} = 1 - y - c|x|, \end{cases} \tag{1.2}$$

where $c = \frac{d^2}{2}$. Note that this system is also volume-preserving and time-reversible with respect to the involution $\mathbf{R}$.

Due to the lack of differentiability of the piecewise linear vector fields, the generic tools of the analysis of differentiable systems cannot be applied. Therefore, the techniques used in [11] for the Michelson system are useless for our piecewise linear continuous version. Nevertheless, we also show that some dynamical aspects of the Michelson system remain in our piecewise linear version.

System (1.2) is formed by two linear systems separated by the plane $\{x = 0\}$, called the separation plane, and it can be written in a matricial form as

$$\dot{\mathbf{x}} = \begin{cases} A^+\mathbf{x} + \mathbf{e}_3 & \text{if } x \geq 0, \\ A^-\mathbf{x} + \mathbf{e}_3 & \text{if } x \leq 0, \end{cases} \tag{1.3}$$

with

$$A^+ = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -c & -1 & 0 \end{pmatrix}, \quad A^- = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ c & -1 & 0 \end{pmatrix}, \quad \text{and} \quad \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

In the half-space $\{x < 0\}$, the system has exactly one equilibrium point $\mathbf{p}_- = (-1/c, 0, 0)^T$ which is a saddle-focus point. Let $\lambda > 0$ and $\alpha \pm \mathrm{i}\,\beta$ be the eigenvalues of the Jacobian matrix

at $\mathbf{p}_-$. This clearly implies that

$$c = \lambda(1+\lambda^2), \quad \alpha = -\frac{\lambda}{2}, \quad \beta = \frac{\sqrt{4+3\lambda^2}}{2}. \tag{1.4}$$

By the reversibility with respect to $\mathbf{R}$, there exists exactly one saddle-focus equilibrium $\mathbf{p}_+ = (1/c, 0, 0)^T$ in the half-space $\{x > 0\}$ whose eigenvalues are given by $-\lambda$ and $-\alpha \pm \mathrm{i}\,\beta$.

Using the expression of the parameter c given in (1.4), system (1.2) can be written as

$$\begin{cases} \dot{x} = y, \\ \dot{y} = z, \\ \dot{z} = 1 - y - \lambda(1+\lambda^2)|x|, \end{cases} \tag{1.5}$$

and the parameter $\lambda > 0$ can be chosen as the fundamental parameter of the family.

In the particular case of piecewise linear systems, global connections can be classified attending to the number of intersections with the separation plane. Using the notation introduced above, we say that a reversible T-point heteroclinic cycle Γ of system (1.5) is (n, m) if the one-dimensional heteroclinic orbit $\rho_\mp$ intersects the separation plane $\{x = 0\}$ at exactly n points and the two-dimensional heteroclinic orbit $\rho_\pm$ intersects $\{x = 0\}$ at exactly m points.

Obviously, the reversibility of system (1.5) forces n to be odd. As we will see later, due to the local linear shape of the one-dimensional invariant manifolds, n has to be different than one. So, the $(3, 1)$ reversible T-point heteroclinic cycle can be considered to be the simplest one and its existence will be the main goal of this work, as it is summarized in the following theorem.

Theorem 1.1. *There exists a value $\lambda_1 \in (1/2, 1)$ such that the piecewise linear version (1.5) of the Michelson system has a $(3, 1)$ reversible T-point heteroclinic cycle Γ for $\lambda = \lambda_1$.*

Some numerical computations allow us to obtain $\lambda_1 \approx 0.65153556$. In fact, its boundary values $1/2$ and 1 do not have any dynamical meaning, and they have been chosen for the sake of simplicity of the handmade calculations involved in the proof.

Piecewise linear system (1.5) has been obtained from the Michelson system by using a natural transformation. This transformation preserves not only the look of the equations but also the properties of reversibility and volume-preservation. We must emphasize that this is not casual. In fact, considering a piecewise linear continuous system with separation plane $\{x = 0\}$ and two saddle-focus equilibria, and assuming the reversibility, volume-preservation, and nondegeneracy conditions, the unique system that can be obtained is (1.5) except for linear changes of variables and time. More precisely, a piecewise linear continuous system with separation plane $\{x = 0\}$ under the hypothesis of observability can be written as

$$\begin{cases} \dot{x} = t^{\pm} x - y, \\ \dot{y} = m^{\pm} x - z, \\ \dot{z} = d^{\pm} x - 1, \end{cases} \tag{1.6}$$

where parameters t^+, m^+, and d^+ correspond to the linear system in the half-space $\{x > 0\}$ while t^-, m^-, and d^- correspond to the linear system in the half-space $\{x < 0\}$. Observability is a nondegeneracy condition, which means that the dynamics of the system cannot be uncoupled [6, 7]. The reversibility condition with respect to $\mathbf{R}$ implies that $t^- = -t^+$,

$m^- = m^+$, and $d^- = -d^+$. On the other hand, system (1.6) is volume-preserving if and only if $t^+ = t^- = 0$. Now, the existence of two saddle-focus equilibria forces the parameters d^+ and m^+ to be positive. Hence, system (1.6) is really

$$\begin{cases} \dot{x} = -y, \\ \dot{y} = mx - z, \\ \dot{z} = d|x| - 1, \end{cases}$$

where $d = d^+ > 0$ and $m = m^+ > 0$. The trivial linear change of variables and time,

$$X = -m^{3/2}x, \quad Y = my, \quad Z = m^{3/2}x - m^{1/2}z, \quad \tau = m^{1/2}t,$$

transforms the system into

$$\begin{cases} X' = Y, \\ Y' = Z, \\ Z' = 1 - Y - \frac{d}{m^{3/2}}|X|, \end{cases}$$

where the prime stands for the derivative with respect to τ. This is the piecewise linear version of the Michelson system previously obtained.

The rest of the paper is devoted to the proof of Theorem 1.1, and it is organized as follows. In section 2 we describe the basic geometric elements of the problem. In section 3 we prove the existence, for every λ in a semi-infinite interval, of a two-dimensional heteroclinic orbit with exactly one intersection point with the plane $\{x = 0\}$. For one of these values of λ, there exists a one-dimensional heteroclinic orbit with exactly three intersection points with $\{x = 0\}$, as is proved in section 4. From this follows the existence of a simple heteroclinic cycle Γ. In section 5 we prove that Γ satisfies the nondegeneracy conditions of a reversible T-point heteroclinic cycle.

2. Some geometric elements of the flow. In this section we describe the behavior of the flow crossing the plane $\{x = 0\}$ and the basic elements of the linear dynamics locally contained in the half-spaces $\{x < 0\}$ and $\{x > 0\}$.

For every point $\mathbf{p} = (x_\mathbf{p}, y_\mathbf{p}, z_\mathbf{p})^T \in \mathbb{R}^3$ we denote by $\mathbf{x}_\mathbf{p}(t;\lambda) = (x_\mathbf{p}(t;\lambda), y_\mathbf{p}(t;\lambda), z_\mathbf{p}(t;\lambda))^T$ the solution of the system (1.5) with parameter λ and initial condition $\mathbf{x}_\mathbf{p}(0;\lambda) = \mathbf{p}$. The corresponding orbit is denoted by $\gamma_\mathbf{p}$.

If $x_\mathbf{p} = 0$ and $y_\mathbf{p} > 0$, then the orbit $\gamma_\mathbf{p}$ crosses transversally the plane $\{x = 0\}$ with $x_\mathbf{p}(-t;\lambda) < 0$ and $x_\mathbf{p}(t;\lambda) > 0$ for $t > 0$ small enough. If $x_\mathbf{p}(t;\lambda)$ vanishes in $(0, +\infty)$, then we define the flying time $t^+_\mathbf{p}$ as the positive value such that $x_\mathbf{p}(t^+_\mathbf{p};\lambda) = 0$ and $x_\mathbf{p}(t;\lambda) > 0$ in $(0, t^+_\mathbf{p})$. In such a case, we define the Poincaré map Π_+ at the point $\mathbf{p}$ as $\Pi_+(\mathbf{p}) = (0, y_\mathbf{p}(t^+_\mathbf{p};\lambda), z_\mathbf{p}(t^+_\mathbf{p};\lambda))^T$. Note that the Poincaré map Π_+ depends only on the linear system $\dot{\mathbf{x}} = A^+\mathbf{x} + \mathbf{e}_3$ given in (1.3).

If $x_\mathbf{p} = 0$ and $y_\mathbf{p} < 0$, then the orbit $\gamma_\mathbf{p}$ crosses transversally the plane $\{x = 0\}$ with $x_\mathbf{p}(-t;\lambda) > 0$ and $x_\mathbf{p}(t;\lambda) < 0$ for $t > 0$ small enough. If $x_\mathbf{p}(t;\lambda)$ vanish in $(0, +\infty)$, then we define the flying time $t^-_\mathbf{p}$ as the positive value such that $x_\mathbf{p}(t^-_\mathbf{p};\lambda) = 0$ and $x_\mathbf{p}(t;\lambda) < 0$ in $(0, t^-_\mathbf{p})$. In such a case, we define the Poincaré map Π_- at the point $\mathbf{p}$ as $\Pi_-(\mathbf{p}) = (0, y_\mathbf{p}(t^-_\mathbf{p};\lambda), z_\mathbf{p}(t^-_\mathbf{p};\lambda))^T$. This map depends only on the linear system $\dot{\mathbf{x}} = A^-\mathbf{x} + \mathbf{e}_3$.

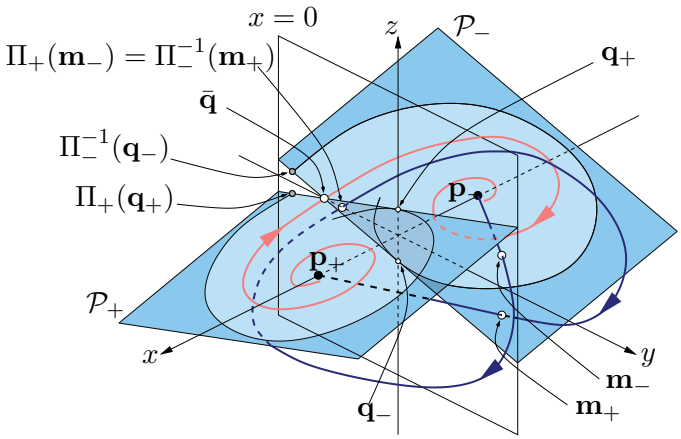


Figure 1. *Some geometric elements of the flow.*

If $\mathbf{p}$ belongs to the z-axis, i.e., $x_\mathbf{p} = 0$ and $y_\mathbf{p} = 0$, then $\mathbf{p}$ is called a contact point of the flow of system (1.5) with the plane $\{x = 0\}$ because the vector field at this point is tangent to the plane. Following [21], the first coordinate of the Taylor expansion of $\mathbf{x}_\mathbf{p}(t;\lambda) - \mathbf{p}$ at $t = 0$ is

$$\mathbf{e}_1^T\left(\mathbf{x}_\mathbf{p}(t;\lambda) - \mathbf{p}\right) = z_\mathbf{p}\frac{t^2}{2} + \frac{t^3}{3!} + \mathbf{e}_1^T\mathbf{x}_\mathbf{p}^{(4)}(\xi;\lambda)\frac{t^4}{4!}.$$

Hence, if $z_\mathbf{p} < 0$, then orbit $\gamma_\mathbf{p}$ is locally contained in the half-space $\{x \le 0\}$; if $z_\mathbf{p} > 0$, then $\gamma_\mathbf{p}$ is locally contained in the half-space $\{x \ge 0\}$; and if $z_\mathbf{p} = 0$, then $\gamma_\mathbf{p}$ crosses the plane $\{x = 0\}$ from the half-space $\{x < 0\}$ to the half-space $\{x > 0\}$.

Now we describe the basic elements of the linear dynamics in every half-space; all this information is summarized in Figure 1.

The stable manifold $W^s(\mathbf{p}_+)$ of $\mathbf{p}_+$ contains the half-line $\mathcal{L}_+ = \{\mathbf{p}_+ + \mu(1, -\lambda, \lambda^2) : -\frac{1}{\lambda(1+\lambda^2)} \le \mu < \infty\}$ generated by the eigenvector associated to the eigenvalue $-\lambda$ of the matrix A^+. The half-line $\mathcal{L}_+$ and the plane $\{x = 0\}$ intersect at the point

$$\mathbf{m}_+ = \left(0, \frac{1}{1+\lambda^2}, -\frac{\lambda}{1+\lambda^2}\right)^T.$$

The unstable two-dimensional manifold $W^u(\mathbf{p}_+)$ is locally contained in the half-plane

$$\mathcal{P}_+ = \left\{\lambda\left(1+\lambda^2\right)x - \lambda^2 y + \lambda z = 1 : x \ge 0\right\},$$

which is called the focal half-plane of $\mathbf{p}_+$. This half-plane is obtained from the eigenvectors associated to the complex eigenvalues of A^+. The half-plane $\mathcal{P}_+$ and the separation plane $\{x = 0\}$ intersect along the straight line

$$\mathcal{D}_+ = \left\{-\lambda^2 y + \lambda z = 1\right\}.$$

Let us emphasize that not every point in $\mathcal{D}_+$ belongs to the unstable manifold $W^u(\mathbf{p}_+)$; see Figure 1. The intersection point of $\mathcal{D}_+$ and the z-axis is $\mathbf{q}_+ = \left(0, 0, \frac{1}{\lambda}\right)^T$. Since $\mathbf{q}_+$ is a contact

point, the orbit $\gamma_{\mathbf{q}_+}$ is tangent to the separation plane $\{x = 0\}$ at $\mathbf{q}_+$. Thus, the segment $\mathcal{S}_+ \subset \mathcal{D}_+$ with endpoints $\mathbf{q}_+$ and $\Pi_+(\mathbf{q}_+)$ is contained in $W^u(\mathbf{p}_+)$.

The unstable manifold $W^u(\mathbf{p}_-)$ of $\mathbf{p}_-$ contains the half-line $\mathcal{L}_- = \{\mathbf{p}_- - \mu(1, \lambda, \lambda^2) : -\frac{1}{\lambda(1+\lambda^2)} \leq \mu < \infty\}$ generated by the eigenvector associated to the eigenvalue λ of the matrix A^-. The half-line and the plane $\{x = 0\}$ intersect at the point

$$\mathbf{m}_- = \left(0, \frac{1}{1+\lambda^2}, \frac{\lambda}{1+\lambda^2}\right)^T.$$

The stable two-dimensional manifold $W^s(\mathbf{p}_-)$ is locally contained in the half plane

$$\mathcal{P}_- = \left\{\lambda\left(1+\lambda^2\right)x + \lambda^2 y + \lambda z = -1 : x \leq 0\right\},$$

which is called the focal half-plane of $\mathbf{p}_-$. This half-plane is obtained from the eigenvectors associated to the complex eigenvalues of A^-. The half-plane $\mathcal{P}_-$ and the separation plane $\{x = 0\}$ intersect along the straight line

$$\mathcal{D}_- = \left\{\lambda^2 y + \lambda z = -1\right\}.$$

Let us emphasize that not every point in $\mathcal{D}_-$ belongs to the stable manifold $W^s(\mathbf{p}_-)$. The intersection point of $\mathcal{D}_-$ and the z-axis is $\mathbf{q}_- = \left(0, 0, -\frac{1}{\lambda}\right)^T$. Since $\mathbf{q}_-$ is a contact point, the orbit $\gamma_{\mathbf{q}_-}$ is tangent to the separation plane $\{x = 0\}$ at $\mathbf{q}_-$. Thus, the segment $\mathcal{S}_- \subset \mathcal{D}_-$ with endpoints $\mathbf{q}_-$ and $\Pi_-^{-1}(\mathbf{q}_-)$ is contained in $W^s(\mathbf{p}_-)$.

3. Existence of a two-dimensional heteroclinic orbit. In this section we prove the existence of a simple two-dimensional heteroclinic orbit $\rho_\pm$, that is, a heteroclinic orbit $\rho_\pm \subset W^u(\mathbf{p}_+) \cap W^s(\mathbf{p}_-)$ which intersects the plane $\{x = 0\}$ at exactly one point $\bar{\mathbf{q}}$.

A necessary and sufficient condition for the existence of the orbit $\rho_\pm$ is $\bar{\mathbf{q}} \in \mathcal{S}_- \cap \mathcal{S}_+$. This implies that $\bar{\mathbf{q}} = \left(0, -\lambda^{-2}, 0\right)^T$, because it is the intersection point of the straight lines $\mathcal{D}_+$ and $\mathcal{D}_-$. We now proceed to look for the values of the parameter λ for which the point $\bar{\mathbf{q}}$ belongs to $\mathcal{S}_-$ and, by reversibility, to $\mathcal{S}_+$.

By definition, the segment $\mathcal{S}_-$ is defined by the endpoints $\mathbf{q}_-$ and $\Pi_-^{-1}(\mathbf{q}_-)$. Since the third coordinate of $\mathbf{q}_-$ is negative and $\bar{\mathbf{q}}$ has the third coordinate equal to zero, then $\bar{\mathbf{q}} \in \mathcal{S}_-$ if and only if the third coordinate of $\Pi_-^{-1}(\mathbf{q}_-)$ is nonnegative. Therefore, system (1.5) has a simple two-dimensional heteroclinic orbit $\rho_\pm$ if and only if there exist $t_0 > 0$ and $\lambda_0 > 0$ such that $(t; \lambda) = (t_0; \lambda_0)$ is a solution of the system

$$\begin{cases} x_{\mathbf{q}_-}(-t; \lambda) = 0, \\ z_{\mathbf{q}_-}(-t; \lambda) \geq 0, \end{cases} \tag{3.1}$$

with

$$x_{\mathbf{q}_-}(-t; \lambda_0) < 0 \text{ for every } t \in (0, t_0). \tag{3.2}$$

We emphasize that condition (3.2) ensures that the point $\mathbf{x}_{\mathbf{q}_-}(-t_0; \lambda_0)$ is the preimage of $\mathbf{q}_-$ by the Poincaré map Π_-.

Taking into account condition (3.2), the expressions of $x_{\mathbf{q}_-}$ and $z_{\mathbf{q}_-}$ in (3.1) can be obtained by integrating the linear system in the half-space $\{x < 0\}$ in backward time with initial condition $\mathbf{x}(0;\lambda) = \mathbf{q}_-$. Thus, system (3.1) can be written as

(3.3)
$$\begin{cases} x_{\mathbf{q}_-}(-t;\lambda) = -\dfrac{1}{\lambda(1+\lambda^2)}\left[1 - e^{\frac{\lambda}{2}t}\left(\cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) - \dfrac{\lambda}{\sqrt{4+3\lambda^2}}\sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right)\right)\right] = 0, \\ z_{\mathbf{q}_-}(-t;\lambda) = -\dfrac{1}{\lambda}e^{\frac{\lambda}{2}t}\left[\cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) + \dfrac{\lambda}{\sqrt{4+3\lambda^2}}\sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right)\right] \geq 0. \end{cases}$$

Using the function $\varphi(\tau,\gamma) = 1 - e^{\gamma\tau}(\cos(\tau) - \gamma\sin(\tau))$ defined in [1], system (3.3) can be rewritten as

$$\begin{cases} -\dfrac{1}{\lambda(1+\lambda^2)}\varphi\left(-\frac{\sqrt{4+3\lambda^2}}{2}t, -\frac{\lambda}{\sqrt{4+3\lambda^2}}\right) = 0, \\ -\dfrac{1}{\lambda}e^{\frac{\lambda}{2}t}\left(1 - \varphi\left(\frac{\sqrt{4+3\lambda^2}}{2}t, -\frac{\lambda}{\sqrt{4+3\lambda^2}}\right)\right) \geq 0. \end{cases}$$

Hence, the existence of a solution $(t;\lambda) = (t_0;\lambda_0)$ of (3.1) satisfying inequality (3.2) is equivalent to the existence of a solution $(t;\lambda) = (t_0;\lambda_0)$ of system

$$\begin{cases} \varphi\left(-\frac{\sqrt{4+3\lambda^2}}{2}t, -\frac{\lambda}{\sqrt{4+3\lambda^2}}\right) = 0, \\ \varphi\left(\frac{\sqrt{4+3\lambda^2}}{2}t, -\frac{\lambda}{\sqrt{4+3\lambda^2}}\right) \geq 1 \end{cases} \tag{3.4}$$

with $t_0 > 0$, $\lambda_0 > 0$ and such that

$$\varphi\left(-\frac{\sqrt{4+3\lambda_0^2}}{2}t, -\frac{\lambda_0}{\sqrt{4+3\lambda_0^2}}\right) > 0 \tag{3.5}$$

for every t in $(0,t_0)$.

In the next result, whose proof is direct, we compile elementary properties of function $\varphi(\tau,\gamma)$, some of which can be found in [1].

Lemma 3.1. *Function $\varphi(\tau,\gamma) = 1 - e^{\gamma\tau}(\cos(\tau) - \gamma\sin(\tau))$ satisfies the following properties.*

(i) $\varphi(-\tau,-\gamma) = \varphi(\tau,\gamma)$.

(ii) $\frac{\partial\varphi}{\partial\tau} = (1+\gamma^2)e^{\gamma\tau}\sin(\tau)$ *and* $\frac{\partial^2\varphi}{\partial\tau^2} = (1+\gamma^2)e^{\gamma\tau}(\cos(\tau) + \gamma\sin(\tau))$.

(iii) *For every fixed $\gamma > 0$ the function $\varphi(\tau,\gamma)$ reaches its local maxima and minima values, respectively, at $\tau_{2k+1} = (2k+1)\pi$ and $\tau_{2k} = 2k\pi$ with $k \in \mathbb{Z}$. Moreover, $\varphi(\tau_{2k+1},\gamma) = 1 + e^{\gamma\tau_{2k+1}}$ and $\varphi(\tau_{2k},\gamma) = 1 - e^{\gamma\tau_{2k}}$.*

(iv) *There exists a unique function $\hat{\tau}_1 : (0,+\infty) \to (\pi,2\pi)$ such that $\varphi(\hat{\tau}_1(\gamma),\gamma) = 0$, $\varphi(\tau,\gamma) > 0$ for $\tau \in (0,\hat{\tau}_1(\gamma))$ and $\varphi(\tau,\gamma) < 0$ for $\tau \in (\hat{\tau}_1(\gamma),2\pi)$.*

(v) $\frac{\partial\varphi}{\partial\gamma} = e^{\gamma\tau}(-\tau\cos(\tau) + (1+\gamma\tau)\sin(\tau))$.

(vi) *There exists a unique function $\hat{\tau}_2 : (0,+\infty) \to (-2\pi,-\pi)$ such that $\varphi(\hat{\tau}_2(\gamma),\gamma) = 1$ and $\varphi(\tau,\gamma) > 1$ for $\tau \in (\hat{\tau}_2(\gamma),-\pi)$.*

The shape of functions $\hat{\tau}_1$ and $\hat{\tau}_2$ introduced in the previous lemma are described in the next result. Some of these properties are shown in Figure 2.

Lemma 3.2. *The following properties of functions $\hat{\tau}_1$ and $\hat{\tau}_2$, defined in Lemma 3.1, hold.*

(i) *The function $\hat{\tau}_1$ is differentiable, $\frac{d\hat{\tau}_1}{d\gamma} < 0$, $\lim_{\gamma\searrow 0}\hat{\tau}_1(\gamma) = 2\pi$, and $\lim_{\gamma\nearrow\infty}\hat{\tau}_1(\gamma) = \pi$.*

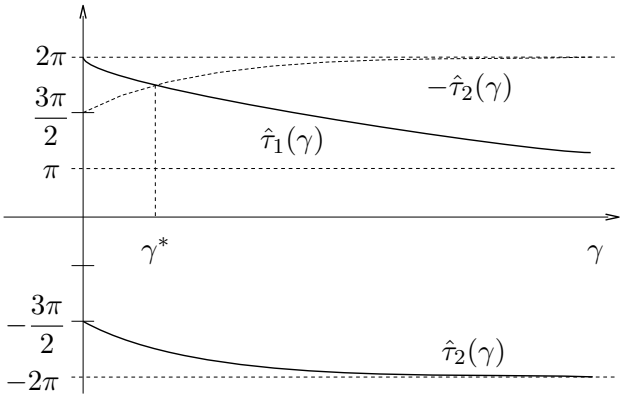


Figure 2. *Qualitative behavior of functions $\hat{\tau}_1$ and $\hat{\tau}_2$.*

(ii) *The function $\hat{\tau}_2$ is differentiable, $\frac{d\hat{\tau}_2}{d\gamma} < 0$, $\lim_{\gamma\searrow 0}\hat{\tau}_2(\gamma) = -\frac{3\pi}{2}$, and $\lim_{\gamma\nearrow\infty}\hat{\tau}_2(\gamma) = -2\pi$.*

(iii) *There exists a unique value $\gamma^* > 0$ such that $-\hat{\tau}_2(\gamma^*) = \hat{\tau}_1(\gamma^*)$. This value satisfies $\gamma^* < \frac{1}{\sqrt{19}}$. Moreover, $-\hat{\tau}_2(\gamma) < \hat{\tau}_1(\gamma)$ in $(0,\gamma^*)$ and $-\hat{\tau}_2(\gamma) > \hat{\tau}_1(\gamma)$ in $(\gamma^*,+\infty)$.*

(iv) *If $0 < \gamma < \gamma^*$, then $\varphi(-\hat{\tau}_1(\gamma),\gamma) < 1$, and $\varphi(-\hat{\tau}_1(\gamma),\gamma) > 1$ if $\gamma > \gamma^*$.*

Proof. (i) Since $\hat{\tau}_1(\gamma) \in (\pi, 2\pi)$, by the implicit function theorem it follows that $\hat{\tau}_1$ is a differentiable function and

$$\left.\frac{d\hat{\tau}_1}{d\gamma}\right|_\gamma = \frac{\hat{\tau}_1(\gamma)e^{-\gamma\hat{\tau}_1(\gamma)} - \sin(\hat{\tau}_1(\gamma))}{(1+\gamma^2)\sin(\hat{\tau}_1(\gamma))} < 0.$$

Taking the limit as γ tends to zero in the implicit expression $\varphi(\hat{\tau}_1(\gamma),\gamma) = 0$, it is easy to check that $\lim_{\gamma\searrow 0}\cos(\hat{\tau}_1(\gamma)) = 1$ and $\lim_{\gamma\nearrow\infty}\sin(\hat{\tau}_1(\gamma)) = 0$. Hence, $\lim_{\gamma\searrow 0}\hat{\tau}_1(\gamma) = 2\pi$ and $\lim_{\gamma\nearrow\infty}\hat{\tau}_1(\gamma) = \pi$.

(ii) This statement follows by the same arguments as in the previous item.

(iii) By the definition of φ and $\hat{\tau}_2$ it follows that, for every $\gamma > 0$, $\cos(\hat{\tau}_2(\gamma)) = \gamma\sin(\hat{\tau}_2(\gamma))$ and $\hat{\tau}_2(\gamma) \in \left(-2\pi, -\frac{3\pi}{2}\right)$. Fixing $\gamma = \frac{1}{\sqrt{19}}$ it is easy to obtain

$$\cos\left(\hat{\tau}_2\left(\frac{1}{\sqrt{19}}\right)\right) = \frac{1}{2\sqrt{5}} \quad \text{and} \quad \sin\left(\hat{\tau}_2\left(\frac{1}{\sqrt{19}}\right)\right) = \frac{\sqrt{19}}{2\sqrt{5}}.$$

This clearly gives $\varphi\left(-\hat{\tau}_2\left(\frac{1}{\sqrt{19}}\right), \frac{1}{\sqrt{19}}\right) = 1 - \frac{1}{\sqrt{5}}e^{-\frac{\hat{\tau}_2\left(\frac{1}{\sqrt{19}}\right)}{\sqrt{19}}}$, which is negative because of $\hat{\tau}_2(\gamma) < -\frac{3\pi}{2}$.

On the other hand, since $\varphi\left(\hat{\tau}_1\left(\frac{1}{\sqrt{19}}\right), \frac{1}{\sqrt{19}}\right) = 0$ and $\varphi\left(\tau, \frac{1}{\sqrt{19}}\right)$ is a monotone decreasing function for $\tau \in (\pi, 2\pi)$ (see Lemma 3.1(iii)), then $\hat{\tau}_1\left(\frac{1}{\sqrt{19}}\right) < -\hat{\tau}_2\left(\frac{1}{\sqrt{19}}\right)$.

Following statements (i) and (ii) in this lemma, the auxiliary function $h(\gamma) = \hat{\tau}_1(\gamma) + \hat{\tau}_2(\gamma)$ satisfies that $h'(\gamma) < 0$ for every $\gamma > 0$ and $\lim_{\gamma\searrow 0} h(\gamma) = \frac{\pi}{2} > 0$. Since $h\left(\frac{1}{\sqrt{19}}\right) < 0$, there

exists a value $\gamma^* \in \left(0, \frac{1}{\sqrt{19}}\right)$ such that $h(\gamma^*) = 0$, $h(\gamma) > 0$ for $0 < \gamma < \gamma^*$, and $h(\gamma) < 0$ for $\gamma > \gamma^*$.

(iv) If $0 < \gamma < \gamma^*$, then $-2\pi < -\hat{\tau}_1(\gamma) < \hat{\tau}_2(\gamma) < -\frac{3\pi}{2}$; see statement (iii). Therefore, by using that $\partial\varphi/\partial\tau > 0$ in $\tau \in (-2\pi, -\frac{3\pi}{2})$ (Lemma 3.1(ii)), it follows that $\varphi(-\hat{\tau}_1(\gamma), \gamma) < \varphi(\hat{\tau}_2(\gamma), \gamma) = 1$. The other inequality of the statement follows in a similar way. ■

The boundary value $\frac{1}{\sqrt{19}}$ appearing in statement (iii) of Lemma 3.2 has only computational meaning. In fact, it has been chosen to obtain the value $1/2$ in the statement of Theorem 1.1.

In the next result we prove the existence of a simple two-dimensional heteroclinic orbit in the piecewise linear differential system (1.5) for every value of the parameter λ greater than a certain value λ^*.

Proposition 3.3. *There exists a value λ^* in $(0, \frac{1}{2})$ such that*

(i) *if $\lambda \geq \lambda^*$, system (1.5) has a two-dimensional heteroclinic orbit with exactly one intersection point with the plane $\{x = 0\}$, and*

(ii) *if $\lambda < \lambda^*$, system (1.5) has no two-dimensional heteroclinic orbits with exactly one intersection point with the plane $\{x = 0\}$.*

Proof. For the sake of simplicity, let us consider the change of variables

$$\tau = \frac{\sqrt{4 + 3\lambda^2}}{2} t, \quad \gamma = \frac{\lambda}{\sqrt{4 + 3\lambda^2}}. \tag{3.6}$$

In these new variables, system (3.4) and condition (3.5) can be written as

$$\begin{cases} \varphi(-\tau, -\gamma) = 0, \\ \varphi(\tau, -\gamma) \geq 1, \end{cases} \quad \text{and} \quad \begin{cases} \varphi(-\bar{\tau}, -\gamma) > 0 \\ \text{for every } \bar{\tau} \text{ in } (0, \tau). \end{cases} \tag{3.7}$$

Hence, the existence of a solution $(\tau_0; \gamma_0)$ to (3.7) is equivalent to the existence of a two-dimensional heteroclinic orbit with exactly one intersection point with the plane $\{x = 0\}$. According to this, the proof of this proposition is reduced to the analysis of the existence of a solution to system (3.7).

Let γ^* be the value defined in Lemma 3.2(iii). Since $0 < \gamma^* < \frac{1}{\sqrt{19}}$, the corresponding value λ^* given by (3.6) is real and positive and satisfies $0 < \lambda^* < \frac{1}{2}$. Let us prove that this is the value λ^* for which both statements of the proposition hold.

(i) For $\lambda_0 \geq \lambda^*$, the corresponding value γ_0 given by (3.6) satisfies $\gamma_0 \geq \gamma^*$. Thus, using Lemmas 3.1(i) and 3.2(iv), we obtain $\varphi(\hat{\tau}_1(\gamma_0), -\gamma_0) \geq 1$. On the other hand, from Lemma 3.1(i) and (iv), we have $\varphi(-\hat{\tau}_1(\gamma_0), -\gamma_0) = 0$ and $\varphi(-\tau, -\gamma_0) > 0$ for every $\tau \in (0, \hat{\tau}_1(\gamma_0))$. That is, $(\hat{\tau}_1(\gamma_0); \gamma_0)$ is the desired solution of system (3.7).

(ii) For $\lambda_0 < \lambda^*$, the corresponding value γ_0 given by (3.6) satisfies $\gamma_0 < \gamma^*$. Note that the unique solution $\bar{\tau}$ of $\varphi(-\tau, -\gamma_0) = 0$ which satisfies $\varphi(-\tau, -\gamma_0) > 0$ for every $\tau \in (0, \bar{\tau})$ is $\bar{\tau} = \hat{\tau}_1(\gamma_0)$; see Lemma 3.1(i) and (iv). Because of $\gamma_0 < \gamma^*$ we have $\varphi(\hat{\tau}_1(\gamma_0), -\gamma_0) < 1$; see Lemma 3.2(iv). Consequently, we concluded that the system (3.7) does not have any solutions. ■

Note that a value λ satisfying the statements of Proposition 3.3 has to be unique in $\lambda > 0$. Moreover, the value $\lambda^* \approx 0.41527324$ can be numerically obtained.

4. Existence of a one-dimensional heteroclinic orbit. In this section we prove the existence of a simple one-dimensional heteroclinic orbit $\rho_{\mp}$, that is, a heteroclinic orbit $\rho_{\mp} = W^u(\mathbf{p}_-) = W^s(\mathbf{p}_+)$ which intersects the plane $\{x = 0\}$ at exactly three points. Two of these points are necessarily $\mathbf{m}_-$ and $\mathbf{m}_+$.

An equivalent condition for the existence of a simple one-dimensional heteroclinic orbit is $\Pi_+(\mathbf{m}_-) = \Pi_-^{-1}(\mathbf{m}_+)$. Due to reversibility, this occurs if and only if the point $\Pi_+(\mathbf{m}_-)$ belongs to the reversibility axis, that is, when the first and third coordinates of $\Pi_+(\mathbf{m}_-)$ are equal to zero. Therefore, system (1.5) has a simple one-dimensional heteroclinic orbit if and only if the system

$$\begin{cases} x_{\mathbf{m}_-}(t;\lambda) = 0, \\ z_{\mathbf{m}_-}(t;\lambda) = 0 \end{cases} \tag{4.1}$$

has a solution $(t_1;\lambda_1)$ such that $t_1 > 0$, $\lambda_1 > 0$, and

$$x_{\mathbf{m}_-}(t;\lambda_1) > 0 \text{ for every } t \in (0,t_1). \tag{4.2}$$

Note that condition (4.2) ensures that the point $\mathbf{x}_{\mathbf{m}_-}(t_1;\lambda_1)$ is the image of $\mathbf{m}_-$ by the Poincaré map Π_+.

Now, to prove the existence of solutions of system (4.1) we are going to simplify its equations. Integrating $\dot{\mathbf{x}} = A^+\mathbf{x} + \mathbf{e}_3$ in forward time with initial condition $\mathbf{x}(0;\lambda) = \mathbf{m}_-$, system (4.1) can be written as

$$\begin{cases} \dfrac{e^{-\lambda t}\left[(1+\lambda^2)\sqrt{4+3\lambda^2} + 2e^{\frac{3}{2}\lambda t}\lambda^2\sqrt{4+3\lambda^2}\cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) - 6e^{\frac{3}{2}\lambda t}\lambda^3\sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right)\right]}{\lambda\sqrt{4+3\lambda^2}(1+\lambda^2)(1+3\lambda^2)} = \dfrac{1}{\lambda(1+\lambda^2)}, \\ \dfrac{e^{-\lambda t}\lambda\left[-(1+\lambda^2)\sqrt{4+3\lambda^2} + 2e^{\frac{3}{2}\lambda t}(1+2\lambda^2)\sqrt{4+3\lambda^2}\cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) - 2\lambda e^{\frac{3}{2}\lambda t}\sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right)\right]}{\sqrt{4+3\lambda^2}(1+\lambda^2)(1+3\lambda^2)} = 0. \end{cases}$$

Adding $\frac{\lambda(1+\lambda^2)}{2e^{\frac{\lambda}{2}t}}$ times the first equation to $\frac{(1+\lambda^2)}{2\lambda e^{\frac{\lambda}{2}t}}$ times the second one gives

$$\cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) - \frac{\lambda}{\sqrt{4+3\lambda^2}}\sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) = \frac{1}{2e^{\frac{\lambda}{2}t}}. \tag{4.3}$$

Moreover, multiplying the second equation by $\frac{1}{\lambda}e^{\lambda t}(1+\lambda^2)(1+3\lambda^2)$, adding $1+\lambda^2$ to both sides of the obtained expression, and multiplying the result by $\frac{1}{2}e^{-\frac{3}{2}\lambda t}$, we get

$$(1+2\lambda^2)\cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) - \frac{\lambda}{\sqrt{4+3\lambda^2}}\sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) = \frac{1+\lambda^2}{2e^{\frac{3}{2}\lambda t}}. \tag{4.4}$$

Now, the trigonometric functions are determined by solving the system given by (4.3) and (4.4),

$$\begin{aligned} \cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) &= \frac{1+\lambda^2-e^{\lambda t}}{4\lambda^2 e^{\frac{3}{2}\lambda t}}, \\ \sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right) &= \frac{\sqrt{4+3\lambda^2}}{4\lambda^3 e^{\frac{3}{2}\lambda t}}(1+\lambda^2-e^{\lambda t}(1+2\lambda^2)), \end{aligned} \tag{4.5}$$

and thus

$$\cos\left(\tfrac{\sqrt{4+3\lambda^2}}{2}t\right)^2+\sin\left(\tfrac{\sqrt{4+3\lambda^2}}{2}t\right)^2$$
$$=\frac{4\lambda^6}{1+\lambda^2}e^{3\lambda t}-(1+4\lambda^2+3\lambda^4)e^{2\lambda t}+(2+6\lambda^2+3\lambda^4)e^{\lambda t}-(1+\lambda^2)^2+1$$

or, equivalently,

$$\frac{4\lambda^6}{1+\lambda^2}e^{3\lambda t}-(1+4\lambda^2+3\lambda^4)e^{2\lambda t}+(2+6\lambda^2+3\lambda^4)e^{\lambda t}-(1+\lambda^2)^2=0. \tag{4.6}$$

Let us consider the system given by (4.3) and (4.6). Note that every solution of system (4.1) is a solution of system (4.3) and (4.6), but the converse is not necessarily true. In the following lemma we establish the conditions on a solution of system (4.3) and (4.6) for being a solution of system (4.1).

Lemma 4.1. *Let $(t_1;\lambda_1)$ be a solution of system (4.3) and (4.6), with $t_1>0$ and $\lambda_1>0$. Then, $(t_1;\lambda_1)$ is a solution of system (4.1) if and only if $(2k-1)\pi<\frac{\sqrt{4+3\lambda_1^2}}{2}t_1<2k\pi$ with $k\in\mathbb{N}$.*

Proof. Suppose that $(t_1;\lambda_1)$ is a solution of system (4.1). Since $t_1>0$ and $\lambda_1>0$, from (4.5) it is clear that $\sin(\frac{\sqrt{4+3\lambda_1^2}}{2}t_1)<0$. Thus, the argument $\frac{\sqrt{4+3\lambda_1^2}}{2}t_1$ belongs to $((2k-1)\pi,2k\pi)$ with $k\in\mathbb{N}$.

For the other implication, let us consider the system

$$\begin{cases} X-\dfrac{\lambda_1}{\sqrt{4+3\lambda_1^2}}Y &= \dfrac{1}{2e^{\frac{\lambda_1}{2}t_1}},\\ X^2+Y^2 &= 1, \end{cases} \tag{4.7}$$

whose equations correspond to a straight line and a circle. Hence, it has at most two different solutions. Since the straight line defined by the first equation contains a point with $X=1$ and $Y>0$, at most one of such solutions has a negative second coordinate.

Note that

$$(X,Y)=\left(\frac{1+\lambda_1^2-e^{\lambda_1 t_1}}{4\lambda_1^2e^{\frac{3}{2}\lambda_1 t_1}},\frac{\sqrt{4+3\lambda_1^2}}{4\lambda_1^3e^{\frac{3}{2}\lambda_1 t_1}}(1+\lambda_1^2-e^{\lambda_1 t_1}(1+2\lambda_1^2))\right)$$

is a solution of system (4.7) with $Y<0$. On the other hand, since $(t_1;\lambda_1)$ is a solution of system (4.3) and (4.6), then

$$(\tilde{X},\tilde{Y})=\left(\cos\left(\frac{\sqrt{4+3\lambda_1^2}}{2}t_1\right),\sin\left(\frac{\sqrt{4+3\lambda_1^2}}{2}t_1\right)\right)$$

is also a solution of system (4.7) with $\tilde{Y}<0$. Therefore, we conclude that $(\tilde{X},\tilde{Y})=(X,Y)$, which means that $(t_1;\lambda_1)$ is a solution of (4.5) or, equivalently, a solution of (4.1). ∎

In the next result we prove that system (4.1) has at least a solution $(t_1;\lambda_1)$ which also satisfies some important conditions to verify inequality (4.2).

Lemma 4.2. *System* (4.1) *has a solution* $(t_1;\lambda_1)$ *satisfying*

$$\frac{1}{2}<\lambda_1<1 \quad \textit{and} \quad \pi<\frac{\sqrt{4+3\lambda_1^2}}{2}t_1<2\pi.$$

Proof. Let us define $\psi(s,\lambda)=\frac{4\lambda^6}{1+\lambda^2}s^3-(1+4\lambda^2+3\lambda^4)s^2+(2+6\lambda^2+3\lambda^4)s-(1+\lambda^2)^2$ and $\tau=\frac{\sqrt{4+3\lambda^2}}{2}t$. Thus, system (4.3) and (4.6) can be written as

$$\begin{cases} f_1(\tau,\lambda)=\varphi\left(\tau,\frac{\lambda}{\sqrt{4+3\lambda^2}}\right)-\dfrac{1}{2}=0,\\ f_2(\tau,\lambda)=\psi\left(e^{\frac{2\lambda}{\sqrt{4+3\lambda^2}}\tau},\lambda\right)=0, \end{cases} \tag{4.8}$$

where φ is the function defined in Lemma 3.1.

From Lemma 4.1, the proof is complete by showing the existence of a solution $(\tau_1;\lambda_1)$ of system (4.8) in $(\pi,2\pi)\times(\frac{1}{2},1)$. We will obtain such a solution by applying the Poincaré–Miranda theorem [17], which is an n-dimensional extension of Bolzano's theorem.

The definition of φ makes it obvious that $f_1(\pi,\lambda)>0$ and $f_1(2\pi,\lambda)<0$ for every $\lambda>0$. In particular, it is true for $\lambda\in[\frac{1}{2},1]$. That is, function f_1 takes different signs at the vertical sides of the rectangle $[\pi,2\pi]\times[\frac{1}{2},1]$.

Let us now analyze the sign of function $f_2(\tau,\lambda)$ at the horizontal sides of $[\pi,2\pi]\times[\frac{1}{2},1]$ by studying the cubic polynomials $\psi(s,1)$ and $\psi(s,\frac{1}{2})$. Since the derivative of $\psi(s,1)$ with respect to s is positive in $\mathbb{R}$ and $\psi(1,1)=1$, we have $\psi(s,1)>1$ for every $s\in(1,+\infty)$. Therefore, $f_2(\tau,1)>1$ for every $\tau>0$ and, in particular, $f_2(\tau,1)>0$ for every $\tau\in[\pi,2\pi]$.

For the last side of the rectangle, straightforward computations show that the derivative of $\psi(s,\frac{1}{2})$ with respect to s vanishes at two values $s_1<1<s_2$, where s_2 is a local minimum. Taking into account that $\psi(1,\frac{1}{2})=-\frac{1}{80}$ and $\psi(27,\frac{1}{2})=-\frac{41003}{80}$, it follows that $\psi(s,\frac{1}{2})<0$ for $s\in[1,27]$. Note that, for $\lambda=\frac{1}{2}$ and $\tau\in[\pi,2\pi]$, the exponential function $e^{\frac{2\lambda}{\sqrt{4+3\lambda^2}}\tau}$ takes values in $[1,27]$, as is easy to check from inequalities $4\pi<3\sqrt{19}$ and $e<3$. Therefore, $f_2(\tau,\frac{1}{2})<0$ for every $\tau\in[\pi,2\pi]$.

The lemma follows by the Poincaré–Miranda theorem. ∎

To conclude the proof of the existence of a simple one-dimensional heteroclinic orbit for system (1.5), it remains only to check that the solution $(t_1;\lambda_1)$ of system (4.1), given in Lemma 4.2, also satisfies condition (4.2). This is done in the following result.

Proposition 4.3. *For* $\lambda=\lambda_1$, *the system* (1.5) *has a simple one-dimensional heteroclinic orbit.*

Proof. From Lemma 4.2, there exists a solution $(t_1;\lambda_1)$ of system (4.1). Integrating $\dot{\mathbf{x}}=A^+\mathbf{x}+\mathbf{e}_3$ in forward time with initial condition $\mathbf{x}(0;\lambda)=\mathbf{m}_-$, we obtain the following expression for the second coordinate of the solution:

$$y_{\mathbf{m}_-}(t;\lambda)=\frac{2e^{\frac{1}{2}\lambda t}\lambda^2\sqrt{4+3\lambda^2}\cos\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right)+2e^{\frac{1}{2}\lambda t}\lambda(2+3\lambda^2)\sin\left(\frac{\sqrt{4+3\lambda^2}}{2}t\right)}{\sqrt{4+3\lambda^2}(1+\lambda^2)(1+3\lambda^2)}+\frac{e^{-\lambda t}}{(1+3\lambda^2)}. \tag{4.9}$$

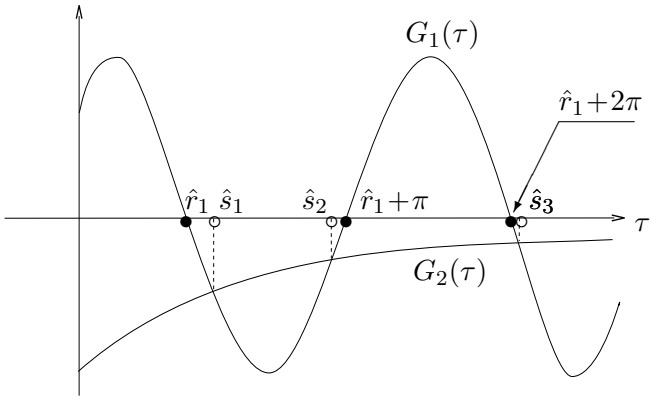


Figure 3. *Qualitative behavior of functions G_1 and G_2.*

Remember that, in our system, $\dot{x}_{\mathbf{m}_-}(t;\lambda) = y_{\mathbf{m}_-}(t;\lambda)$. Thus, using (4.5) to simplify $y_{\mathbf{m}_-}(t_1;\lambda_1)$, we obtain

$$\dot{x}_{\mathbf{m}_-}(t_1;\lambda_1) = \frac{e^{-\lambda_1 t_1} - 1}{\lambda_1^2} < 0.$$

On the other side, $\dot{x}_{\mathbf{m}_-}(0;\lambda_1) > 0$.

Assume that $(t_1;\lambda_1)$ does not satisfy condition (4.2). Therefore, there exists a value in $(0,t_1)$ where $x_{\mathbf{m}_-}(t;\lambda_1)$ vanishes. From $\dot{x}_{\mathbf{m}_-}(0;\lambda_1) > 0$ and $\dot{x}_{\mathbf{m}_-}(t_1;\lambda_1) < 0$ it may be concluded that the derivative $\dot{x}_{\mathbf{m}_-}(t;\lambda_1)$, that is, $y_{\mathbf{m}_-}(t;\lambda_1)$, has to vanish at least at three values s_1, s_2, s_3 such that $0 < s_1 < s_2 < s_3 < t_1$. For simplicity of notation, we consider the change $\tau = \frac{\sqrt{4+3\lambda_1^2}}{2}t$. Let $\hat{s}_1, \hat{s}_2, \hat{s}_3, \tau_1$ denote the respective values of s_1, s_2, s_3, t_1 by this change. Therefore, $0 < \hat{s}_1 < \hat{s}_2 < \hat{s}_3 < \tau_1$.

From (4.9), the equality $y_{\mathbf{m}_-}(t;\lambda_1) = 0$ is equivalent to $G_1(\tau) = G_2(\tau)$, where

$$\begin{aligned} G_1(\tau) &= \frac{\lambda_1^2}{1+\lambda_1^2}\cos(\tau) + \frac{\lambda_1(2+3\lambda_1^2)}{\sqrt{4+3\lambda_1^2}(1+\lambda_1^2)}\sin(\tau), \\ G_2(\tau) &= -\frac{1}{2}\exp\left(\frac{-3\lambda_1}{\sqrt{4+3\lambda_1^2}}\tau\right). \end{aligned} \tag{4.10}$$

The qualitative behavior of G_1 and G_2 is shown in Figure 3.

Since $G_2(\tau) < 0$ for every τ and $G_1(0) > 0$, there exists an $\hat{r}_1 \in (0, \hat{s}_1)$ where $G_1(\hat{r}_1) = 0$, $G_1(\tau) < 0$ for $\tau \in \bigcup_{k=0}^{\infty} I_k$, where $I_k = (\hat{r}_1 + 2k\pi, \hat{r}_1 + (2k+1)\pi)$. On the other hand, the second derivative G_2'' is always negative while the second derivative G_1'' is positive in every interval I_k with $k = 0, 1, 2, \ldots$. Therefore, in each interval I_k there may exist at most two values where G_1 and G_2 coincide.

We conclude that $\hat{s}_3 > \hat{r}_1 + 2\pi$. In particular, $\tau_1 > 2\pi$ and, equivalently, $\frac{\sqrt{4+3\lambda_1^2}}{2}t_1 > 2\pi$, which contradicts Lemma 4.2. Thus, the proposition follows. ■

From Propositions 3.3 and 4.3 we conclude that for $\lambda = \lambda_1$ system (1.5) has a simple two-dimensional heteroclinic orbit $\rho_\pm$ and a simple one-dimensional heteroclinic orbit $\rho_\mp$. That is, for $\lambda = \lambda_1$ system (1.5) has a $(3,1)$ heteroclinic cycle $\Gamma = \mathbf{p}_+ \cup \rho_\pm \cup \mathbf{p}_- \cup \rho_\mp$.

5. Existence of a reversible T-point heteroclinic cycle. In this section we finish the proof of Theorem 1.1. It remains only to verify that the heteroclinic cycle Γ, obtained in section 4 for $\lambda = \lambda_1$, satisfies the two following conditions, which are equivalent to those stated at the definition of a reversible T-point heteroclinic cycle:

(i) Manifolds $W^u(\mathbf{p}_+)$ and $W^s(\mathbf{p}_-)$ intersect transversally along the two-dimensional heteroclinic orbit $\rho_\pm$.

(ii) The difference between the third coordinates of $\Pi_+(\mathbf{m}_-)$ and $\Pi_-^{-1}(\mathbf{m}_+)$ is diffeomorphic to $\mu = \lambda - \lambda_1$ (note that the distance between two points in the plane $\{x = 0\}$ which are symmetric with respect to $\mathbf{R}$ is two times the absolute value of the third coordinate of any of them).

Statement (i) is a direct consequence of the transversality of manifolds $W^u(\mathbf{p}_+)$ and $W^s(\mathbf{p}_-)$ at $\{x = 0\}$ and the differentiability of the flow.

To verify condition (ii), it is necessary only to prove that the third coordinate of $\Pi_+(\mathbf{m}_-)$ is diffeomorphic to $\mu = \lambda - \lambda_1$.

Let $t^+_{\mathbf{m}_-}(\lambda)$ be the flying time of the solution through the point $\mathbf{m}_-$ for each λ in a neighborhood of λ_1. In particular, $t^+_{\mathbf{m}_-}(\lambda_1) = t_1$, where t_1 is defined in Lemma 4.2. The third coordinate of $\Pi_+(\mathbf{m}_-)$, as a function of λ, can be written as $h(\lambda) = z_{\mathbf{m}_-}(t^+_{\mathbf{m}_-}(\lambda); \lambda)$. Notice that $h(\lambda)$ is analytical because the solution depends only on the linear system in the half-space $\{x > 0\}$ and the orbit through $\Pi_+(\mathbf{m}_-)$ is transversal to the plane $\{x = 0\}$ for $\lambda = \lambda_1$. Thus, since $h(\lambda_1) = 0$, to conclude that $h(\lambda)$ is diffeomorphic to $\mu = \lambda - \lambda_1$ it is sufficient to prove that $h'(\lambda_1) \neq 0$.

Proposition 5.1. *Function $h(\lambda) = z_{\mathbf{m}_-}(t^+_{\mathbf{m}_-}(\lambda); \lambda)$ satisfies $h'(\lambda_1) < 0$.*

Proof. The first derivative $h'(\lambda_1)$ is given by

$$\left.\frac{dh}{d\lambda}\right|_{\lambda_1} = \dot{z}_{\mathbf{m}_-}(t_1; \lambda_1) \left.\frac{dt^+_{\mathbf{m}_-}}{d\lambda}\right|_{\lambda_1} + \left.\frac{\partial z_{\mathbf{m}_-}(t; \lambda)}{\partial \lambda}\right|_{(t_1;\lambda_1)}.$$

Since the derivative $\dot{x}_{\mathbf{m}_-}(t_1; \lambda_1) = y_{\mathbf{m}_-}(t_1; \lambda_1)$ does not vanish, the function $t^+_{\mathbf{m}_-}(\lambda)$ is implicitly defined by the equation $x_{\mathbf{m}_-}(t; \lambda) = 0$ in a neighborhood of $(t_1; \lambda_1)$. Moreover, $t^+_{\mathbf{m}_-}(\lambda)$ is analytical and

$$\left.\frac{dt^+_{\mathbf{m}_-}}{d\lambda}\right|_{\lambda_1} = -\frac{\left.\dfrac{\partial x_{\mathbf{m}_-}(t; \lambda)}{\partial \lambda}\right|_{(t_1;\lambda_1)}}{y_{\mathbf{m}_-}(t_1; \lambda_1)}.$$

On the other hand, since $z_{\mathbf{m}_-}(t; \lambda)$ is the third coordinate of the solution of system (1.5) with initial condition $\mathbf{m}_-$, we have

$$\dot{z}_{\mathbf{m}_-}(t_1; \lambda_1) = -\lambda_1(1 + \lambda_1^2)x_{\mathbf{m}_-}(t_1; \lambda_1) - y_{\mathbf{m}_-}(t_1; \lambda_1) + 1 = -y_{\mathbf{m}_-}(t_1; \lambda_1) + 1.$$

Therefore, substituting into the expression of $dh/d\lambda$, we obtain

$$\left.\frac{dh}{d\lambda}\right|_{\lambda_1} = \frac{y_{\mathbf{m}_-}(t_1; \lambda_1) - 1}{y_{\mathbf{m}_-}(t_1; \lambda_1)} \left.\frac{\partial x_{\mathbf{m}_-}(t; \lambda)}{\partial \lambda}\right|_{(t_1;\lambda_1)} + \left.\frac{\partial z_{\mathbf{m}_-}(t; \lambda)}{\partial \lambda}\right|_{(t_1;\lambda_1)}.$$

Integrating the linear system $\dot{\mathbf{x}} = A^+\mathbf{x} + \mathbf{e}_3$ with initial condition $\mathbf{m}_-$ and taking into account that $x_{\mathbf{m}_-}(t_1;\lambda_1) = 0$ and $z_{\mathbf{m}_-}(t_1;\lambda_1) = 0$, the following equalities hold:

$$\begin{aligned}
y_{\mathbf{m}_-}(t_1;\lambda_1) &= \frac{e^{-\lambda_1 t_1} - 1}{\lambda_1^2},\\
\left.\frac{\partial x_{\mathbf{m}_-}(t;\lambda)}{\partial\lambda}\right|_{(t_1;\lambda_1)} &= \frac{\lambda_1 t_1}{\lambda_1^2(1+3\lambda_1^2)(1+\lambda^2)(4+3\lambda^2)}\Big(9\lambda_1^4(e^{-\lambda_1 t_1}-1) + 18\lambda^2(e^{-\lambda_1 t_1}-1)\\
&\qquad + 9e^{-\lambda_1 t_1} - 5\Big)\\
&\quad + \frac{2}{\lambda_1^2(1+3\lambda_1^2)(1+\lambda^2)(4+3\lambda^2)}\Big(3\lambda_1^2(2e^{-\lambda_1 t_1}-3) + 7(e^{-\lambda_1 t_1}-1)\Big),\\
\left.\frac{\partial z_{\mathbf{m}_-}(t;\lambda)}{\partial\lambda}\right|_{(t_1;\lambda_1)} &= \frac{3\lambda_1 t_1}{\lambda_1^2(1+3\lambda_1^2)(1+\lambda^2)(4+3\lambda^2)}\Big(3\lambda_1^6 + \lambda_1^4(7-e^{-\lambda_1 t_1}) + \lambda_1^2(5-2e^{-\lambda_1 t_1})\\
&\qquad + 1 - e^{-\lambda_1 t_1}\Big)\\
&\quad + \frac{2}{\lambda_1^2(1+3\lambda_1^2)(1+\lambda^2)(4+3\lambda^2)}\Big(3\lambda_1^4(e^{-\lambda_1 t_1}-1) + \lambda_1^2(3e^{-\lambda_1 t_1}-2)\\
&\qquad + 1 - e^{-\lambda_1 t_1}\Big).
\end{aligned}$$

Straightforward computations show that

$$\left.\frac{dh}{d\lambda}\right|_{\lambda_1} = -\frac{2}{\lambda_1^2(1+3\lambda_1^2)(1+\lambda^2)(4+3\lambda^2)(e^{\lambda_1 t_1}-1)}P(t_1,\lambda_1),$$

where

$$\begin{aligned}
P(t_1,\lambda_1) &= 3(e^{\lambda_1 t_1} + e^{-\lambda_1 t_1} - 2)\lambda_1^5 t_1 + 3(4e^{\lambda_1 t_1} + e^{-\lambda_1 t^*} - 4)\lambda_1^4 + 2(2e^{\lambda_1 t_1} + 3e^{-\lambda_1 t_1} - 6)\lambda^3 t_1\\
&\quad + 9(2e^{\lambda_1 t_1} + e^{-\lambda_1 t_1} - 3)\lambda_1^2 + (e^{\lambda t_1} + 3e^{-\lambda_1 t_1} - 4\lambda_1 t_1) + 6(e^{\lambda_1 t_1} + e^{-\lambda_1 t_1} - 2).
\end{aligned}$$

From Lemma 4.2 we have $\frac{1}{2} < \lambda_1 < 1$ and $\frac{2\pi}{\sqrt{7}} < t_1 < \frac{8\pi}{\sqrt{19}}$. Therefore, $\frac{\pi}{\sqrt{7}} < \lambda_1 t_1 < \frac{8\pi}{\sqrt{19}}$ and $3 < e^{\lambda_1 t_1}$. This implies that the coefficients of λ^5, λ^4, λ^3, and λ^2 in the expression of $P(\lambda_1, t_1)$ are positive. Moreover, $e^{\lambda t_1} + 3e^{-\lambda_1 t_1} - 4\lambda_1 t_1 > -\lambda_1 t_1 > -\frac{8\pi}{\sqrt{19}}$, and $6(e^{\lambda_1 t_1} + e^{-\lambda_1 t_1} - 2) > 6$. Since $6\sqrt{19} > 8\pi$, it follows that $P(\lambda_1, t_1) > 0$, which proves the proposition. ■

From Proposition 5.1 we finish the proof of Theorem 1.1.

Acknowledgments. The authors would like to thank Dr. E. Freire, Dr. E. Ponce, and Dr. F. Torres for useful discussions.

REFERENCES

[1] A. A. Andronov, A. A. Vitt, and S. Khaikin, *Theory of Oscillators*, Dover, New York, 1987.
[2] A. Arneodo, P. Coullet, and C. Tresser, *Possible new strange attractors with spiral structure*, Comm. Math. Phys., 79 (1981), pp. 573–579.
[3] V. V. Bykov, *The bifurcations of separatrix contours and chaos*, Phys. D, 62 (1993), pp. 290–299.

[4] V. V. Bykov, *On systems with separatrix contour containing two saddle-foci*, J. Math. Sci. (New York), 95 (1999), pp. 2513–2522.
[5] V. V. Bykov, *Orbit structure in a neighborhood of a separatrix cycle containing two saddle-foci*, in Amer. Math. Soc. Transl. Ser. 2 200, AMS, Providence, RI, 2000, pp. 87–89.
[6] V. Carmona, E. Freire, E. Ponce, and F. Torres, *On simplifying and classifying piecewise linear systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 609–620.
[7] V. Carmona, E. Freire, E. Ponce, and F. Torres, *Invariant manifolds of periodic orbits for piecewise linear three-dimensional systems*, IMA J. Appl. Math., 69 (2004), pp. 71–91.
[8] L. O. Chua, M. Komuro, and T. Matsumoto, *The double scroll family.* II. *Rigorous analysis of bifurcation phenomena*, IEEE Trans. Circuits and Systems, 33 (1986), pp. 1097–1118.
[9] P. Coullet, C. Tresser, and A. Arneodo, *Transition to stochasticity for a class of forced oscillators*, Phys. Lett. A, 72 (1979), pp. 268–270.
[10] F. Dumortier, S. Ibañez, and H. Kokubu, *New aspects in the unfolding of the nilpotent singularity of codimension three*, Dyn. Syst., 16 (2001), pp. 63–95.
[11] F. Dumortier, S. Ibañez, and H. Kokubu, *Cocoon bifurcation in three-dimensional reversible vector fields*, Nonlinearity, 19 (2006), pp. 305–328.
[12] F. Fernández-Sánchez, E. Freire, and A. J. Rodríguez-Luis, *T-points in a $\mathbb{Z}_2$-symmetric electronic oscillator.* (I) *Analysis*, Nonlinear Dynam., 28 (2002), pp. 53–69.
[13] E. Freire, E. Gamero, A. J. Rodriguez-Luis, and A. Algaba, *A note on the triple-zero linear degeneracy: Normal forms, dynamical and bifurcation behaviors of an unfolding*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002), pp. 2799–2820.
[14] P. Glendinning and C. Sparrow, *T-points: A codimension two heteroclinic bifurcation*, J. Statist. Phys., 43 (1986), pp. 479–488.
[15] S. V. Gonchenko, D. V. Turaev, P. Gaspard, and G. Nicolis, *Complexity in the bifurcation structure of homoclinic loops to a saddle-focus*, Nonlinearity, 10 (1997), pp. 409–423.
[16] S. Ibañez and J. A. Rodriguez, *Shil'nikov configurations in any generic unfolding of the nilpotent singularity of codimension three on $\mathbb{R}^3$*, J. Differential Equations, 208 (2005), pp. 147–175.
[17] W. Kulpa, *The Poincaré–Miranda theorem*, Amer. Math. Monthly, 104 (1997), pp. 545–550.
[18] Y. Kuramoto and T. Tsuzuki, *Persistent propagation of concentration waves in dissipative media far for thermal equilibrium*, Progr. Theoret. Phys., 55 (1976), pp. 356–369.
[19] J. S. W. Lamb, M. A. Texeira, and K. N. Webster, *Heteroclinic bifurcations near Hopf-zero bifurcation in reversible vector fields in $\mathbb{R}^3$*, J. Differential Equations, 219 (2005), pp. 78–115.
[20] Y. T. Lau, *The "cocoon" bifurcation in three-dimensional systems with two fixed points*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 2 (1992), pp. 543–558.
[21] J. Llibre and A. E. Teruel, *Existence of Poincaré maps in piecewise linear differential systems in $\mathbb{R}^n$*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 8 (2004), pp. 2843–2851.
[22] J. Llibre, E. Ponce, and A. E. Teruel, *Horseshoes near homoclinic orbits for piecewise linear differential systems in $\mathbb{R}^3$*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 1171–1184.
[23] T. Matsumoto, L. O. Chua, and M. Komuro, *The double scroll*, IEEE Trans. Circuits and Systems, 32 (1985), pp. 797–818.
[24] T. Matsumoto, L. O. Chua, and K. Ayaki, *Reality of chaos in the double scroll circuit: A computer-assisted proof*, IEEE Trans. Circuits and Systems, 35 (1988), pp. 908–925.
[25] D. Michelson, *Steady solutions of the Kuramoto–Sivashinsky equation*, Phys. D, 19 (1986), pp. 89–111.
[26] L. P. Shil'nikov, *A case of the existence of a denumerable set of periodic motions*, Sov. Math. Dokl., 6 (1965), pp. 163–166.
[27] L. P. Shil'nikov, *A contribution to the problem of the structure of an extended neighbourhood of a rough equilibrium state of saddle-focus type*, Math. USSR Sbornik, 10 (1970), pp. 91–102.
[28] K. N. Webster and J. N. Elgin, *Asymptotic analysis of the Michelson system*, Nonlinearity, 16 (2003), pp. 2149–2162.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 1049–1100
© 2008 Society for Industrial and Applied Mathematics

Localized Hexagon Patterns of the Planar Swift–Hohenberg Equation*

David J. B. Lloyd†, Björn Sandstede†, Daniele Avitabile‡, and Alan R. Champneys‡

Abstract. We investigate stationary spatially localized hexagon patterns of the two-dimensional (2D) Swift–Hohenberg equation in the parameter region where the trivial state and regular hexagon patterns are both stable. Using numerical continuation techniques, we trace out the existence regions of fully localized hexagon patches and of planar pulses which consist of a strip filled with hexagons that is embedded in the trivial state. We find that these patterns exhibit snaking: for each parameter value in the snaking region, an infinite number of patterns exist that are connected in parameter space and whose width increases without bound. Our computations also indicate a relation between the limits of the snaking regions of planar hexagon pulses with different orientations and of the fully localized hexagon patches. To investigate which hexagons among the one-parameter family of hexagons are selected in a hexagon pulse or front, we derive a conserved quantity of the spatial dynamical system that describes planar patterns which are periodic in the transverse direction and use it to calculate the Maxwell curves along which the selected hexagons have the same energy as the trivial state. We find that the Maxwell curve lies within the snaking region, as expected from heuristic arguments.

Key words. localized patterns, hexagons, spots, Turing bifurcation, Swift–Hohenberg equation

AMS subject classifications. 35B32, 35B35, 35J60

DOI. 10.1137/070707622

1. Introduction.

Localized stationary structures play an important role in many biological, chemical, and physical processes (see, for instance, the textbooks [45, 72, 74]). Such structures have been observed in a variety of experiments ranging from vertically vibrated granular materials [38, 89], liquid crystals [15], binary-fluid convection [8, 65], autocatalytic chemical reactions such as the Belousov–Zhabotinsky system [31, 90], electrochemical systems [1, 6], and localized microstructures in solidification [48] to nonlinear optical devices [63, 77, 86]. Localized patterns have also been found in many nonlinear models such as those derived from magnetohydrodynamics [12], flame fronts [39], lasers [56], vibrated granular materials [36, 88], neural networks [54, 55], and cellular buckling [47] as well as in the Swift–Hohenberg equation [26, 28, 29, 44, 80], which often serves as a paradigm for general pattern-forming systems [30, 37].

In this paper, we consider stationary solutions of the Swift–Hohenberg equation [30, 85]

$$u_t = -(1+\Delta)^2 u - \mu u + \nu u^2 - u^3, \tag{1.1}$$

where $x \in \mathbb{R}$ for the one-dimensional (1D) version and $(x, y) \in \mathbb{R}^2$ in the planar case. We

*Received by the editors November 8, 2007; accepted for publication (in revised form) by M. Silber June 3, 2008; published electronically September 25, 2008.
http://www.siam.org/journals/siads/7-3/70762.html
†Department of Mathematics, University of Surrey, Guildford, GU2 7XH, UK (d.j.lloyd@surrey.ac.uk, b.sandstede@surrey.ac.uk).
‡Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (D.Avitabile@bristol.ac.uk, a.r.champneys@bristol.ac.uk).

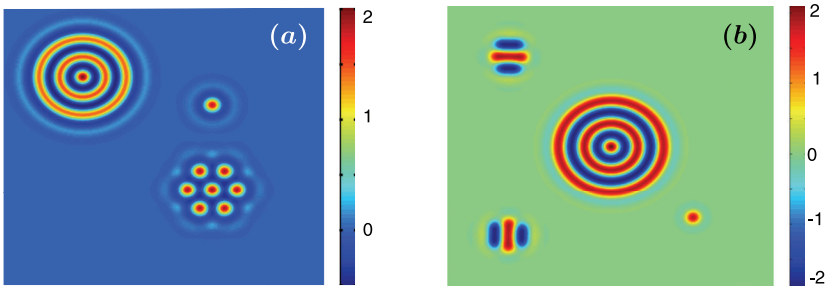


Figure 1. (a) *Localized stationary spots and hexagon patches of* (1.1) *for* $(\mu, \nu) = (0.5, 2.2)$. (b) *Localized stationary spots and stripes of* (1.5) *for* $(\mu, \nu) = (2.5, 4)$. *Both images are color plots of stationary solutions* $u(x, y)$, *with* x *plotted horizontally and* y *vertically, where the values of* $u(x, y)$ *are represented by colors as indicated in the color bars shown to the right of the color plots: The color plots in the remainder of this paper are produced in the same fashion.*

focus on the region $\nu \geq 0$ since the case $\nu < 0$ is then recovered upon replacing u by $-u$. The trivial state $u = 0$ is stable for $\mu > 0$ and destabilizes at $\mu = 0$ with respect to perturbations that have nonzero finite spatial wavelength. At $\mu = 0$, hexagons bifurcate in a transcritical bifurcation from $u = 0$ for each $\nu > 0$, while rolls bifurcate in a subcritical pitchfork bifurcation from $u = 0$ provided $\nu > \nu_r := \sqrt{27/38}$ [44]. While the bifurcating hexagons and rolls are initially unstable for $\mu > 0$, they stabilize in a subsequent saddle-node bifurcation, leading to bistability between the nontrivial patterns and the trivial state for $\mu > 0$. The bistability of trivial and patterned states opens up the possibility of finding fully localized stationary patches of hexagons or rolls such as those shown in Figure 1. It is patterns of this type that we shall focus on in this paper.

We first review briefly the situation in one dimension and refer the reader to section 2 for a more extensive discussion. In one space dimension, the Swift–Hohenberg equation exhibits localized structures, as shown in Figure 2. The patterns shown there are connected in parameter space, and their width increases as we move up on the bifurcation curve: this scenario is referred to as *snaking* [92]. There are several interesting questions one may ask about the patterns shown in Figure 2: can we predict for which values of μ these structures exist, and can we determine a priori which periodic pattern is selected to form the localized structure?

We begin with the second question: the steady-state equation

$$-(1 + \partial_x^2)^2 u - \mu u + \nu u^2 - u^3 = 0 \tag{1.2}$$

of the 1D Swift–Hohenberg equation exhibits, in the relevant parameter region, a one-parameter family of periodic patterns for each fixed (μ, ν). To decide which one of these makes up the core of the localized structure, we consider a front that connects the trivial state to a periodic pattern: this front corresponds to a heteroclinic orbit of the ordinary differential equation (1.2) that connects $u = 0$ to a periodic orbit. It turns out that the ODE (1.2) admits the first integral

$$\mathcal{H}(u) = u_{xxx} u_x - \frac{u_{xx}^2}{2} + u_x^2 + \frac{(1 + \mu)u^2}{2} - \frac{\nu u^3}{3} + \frac{u^4}{4} \tag{1.3}$$

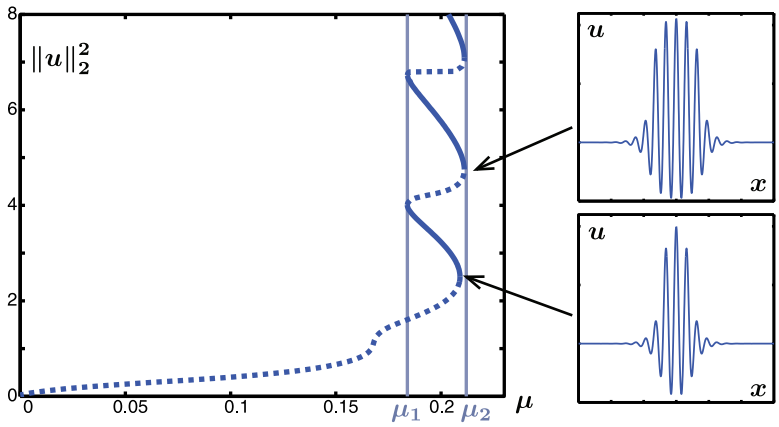


Figure 2. *A partial bifurcation diagram of localized stationary patterns in the 1D Swift–Hohenberg equation (1.1) is shown for $\nu = 1.6$ (other solution branches are given in Figure 8). Dotted blue lines indicate (temporally) unstable solutions, while solid blue lines denote stable solutions. At each fold along the snake, a pair of new rolls is formed in the pulse. The left and right fold bifurcations approach the vertical asymptotes $\mu_1 = 0.181$ and $\mu_2 = 0.211$, and the Maxwell point is $\mu_M = 0.2$.*

so that the value $[\mathcal{H}(u)](x)$ of $\mathcal{H}$ along a solution $u(x)$ of (1.2) does not change as a function of x. In particular, if we evaluate $\mathcal{H}$ along our front, we find that $\mathcal{H}$ must vanish along the limiting periodic pattern as it vanishes at $u = 0$. Generically, $\mathcal{H}$ will vanish only at finitely many periodic orbits in the family of periodic patterns and will therefore serve as a selection principle that involves only the periodic patterns.

We now address the first question, namely, for which parameter values stationary fronts exist. We recall that the Swift–Hohenberg equation (1.1) posed on $\mathbb{R}^d$ with $1 \le d \le 3$ is a gradient system,

$$u_t = -\nabla \mathcal{E}(u),$$

in $H^2(\mathbb{R}^d)$, where the energy functional $\mathcal{E}$ is given by

$$\mathcal{E}(u) = \int_{\mathbb{R}^d} \left[\frac{[(1+\Delta)u]^2}{2} + \frac{\mu u^2}{2} - \frac{\nu u^3}{3} + \frac{u^4}{4} \right] \mathrm{d}\mathbf{x}, \qquad \mathbf{x} \in \mathbb{R}^d, \tag{1.4}$$

and the gradient $\nabla \mathcal{E}(u) = \frac{\delta \mathcal{E}}{\delta u}(u)$ of $\mathcal{E}$ with respect to u is computed in $L^2(\mathbb{R}^d)$. In particular, $\mathcal{E}$ decreases strictly in time along solutions of (1.1) unless the solution is stationary. We record that the existence of the first integral $\mathcal{H}$ given above is actually a consequence of the translation invariance of the integrand of $\mathcal{E}$. While we cannot evaluate the energy functional along periodic patterns as they are not localized, whence the integral in (1.4) may not exist, we may, however, define a local energy by integrating over one spatial period of the underlying periodic pattern. We may then expect, on a heuristic level, that stationary interfaces between the trivial state and the periodic pattern can exist only when their local energies coincide; otherwise, one of the states would invade the other one to decrease energy, thus leading to moving fronts. This gives a heuristic criterion that allows us to determine for which values of μ stationary fronts can exist: for each μ, compute the periodic pattern $u_*(\mu)$ for which $\mathcal{H}(u_*(\mu)) = 0$, calculate its local energy $\mathcal{E}(u_*(\mu))$, and then find μ so that $\mathcal{E}(u_*(\mu)) = 0$. The

corresponding parameter value μ_{M} is referred to as the Maxwell point. It was first pointed out by Pomeau [73] that stationary fronts should exist not only at the Maxwell point $\mu = \mu_{\mathrm{M}}$ but in an entire interval that contains μ_{M} where fronts are pinned or locked. Consequently, we expect to find localized roll patches in an entire interval, and this is what happens in Figure 2. The specific pinning mechanism leading to Figure 2 was elucidated in [26, 52, 92] and will be discussed in section 2.

We now turn to the planar Swift–Hohenberg equation (1.1). Several numerical observations of localized spots and hexagon patches of (1.1) have been documented in the literature [26, 44, 78]. Localized square patterns have also been observed in the Swift–Hohenberg equation when an additional nonlinear gradient term is added [44, 79]. In addition, localized stripes and spots have been found in the cubic-quintic Swift–Hohenberg equation

$$u_t = -(1+\Delta)^2 u - \mu u + \nu u^3 - u^5. \tag{1.5}$$

Close to our approach is the paper [63], in which a complex Ginzburg–Landau equation with a saturable nonlinearity was studied in a cavity-soliton context. In [63], the steady-state equation was solved numerically as a boundary-value problem, and the bifurcation diagram was traced out for various localized patterns, including hexagon patches, using continuation techniques. However, the authors continued solutions only up to the first fold and not beyond.

The aim of this paper is to investigate hexagon fronts and fully localized hexagon patches in the two-dimensional (2D) Swift–Hohenberg equation (1.1). We have three main results. The first is the construction of a conserved quantity $\mathcal{H}$ for the 2D Swift–Hohenberg equation: the existence of $\mathcal{H}$ is a consequence of Noether's theorem since the integrand of the energy functional $\mathcal{E}$ is invariant under translations, i.e., does not depend explicitly on x.

Proposition 1 (conserved quantity for the 2D Swift–Hohenberg equation). *If $u(x,y)$ is a smooth solution of the planar Swift–Hohenberg equation (1.1) which is spatially periodic with period ℓ in the y-variable, then the quantity*

$$\mathcal{H}(u) = \int_0^\ell \left[u_{xxx}u_x - \frac{u_{xx}^2}{2} + u_x^2 + \frac{(1+\mu)u^2}{2} - \frac{\nu u^3}{3} + \frac{u^4}{4} - u_{xy}^2 - u_y^2 + \frac{u_{yy}^2}{2} \right] \mathrm{d}y \tag{1.6}$$

does not depend on x.

As in the 1D case, the first integral $\mathcal{H}$ provides a selection principle for hexagons: if we find a planar front that connects hexagons to the trivial state and is periodic in the transverse direction (see Figure 3), then $\mathcal{H}$ must vanish when evaluated along a single hexagon in the far field of the front. This selection principle together with the local energy $\mathcal{E}$ will allow us to compute Maxwell points for the planar Swift–Hohenberg equation (1.1). Our second result shows that, for each fixed $\nu > 0$, and all sufficiently small $\mu > 0$, there is a unique small-amplitude hexagon pattern along which $\mathcal{H}$ vanishes. We refer the reader to section 3.3 for a stronger result.

Proposition 2 (existence of hexagons with $\mathcal{H} = 0$). *For each fixed $\nu > 0$, there is a number $\mu_0 > 0$ so that the planar Swift–Hohenberg equation (1.1) admits a unique small-amplitude hexagon solution $u_*(\mu)$ that satisfies $\mathcal{H}(u_*(\mu)) = 0$ for each $\mu \in (0, \mu_0)$. These hexagons satisfy $u_*(0) = 0$, have wavenumber $\kappa_*(\mu)$ with $\kappa_*(0) = 1$, and depend smoothly on μ.*

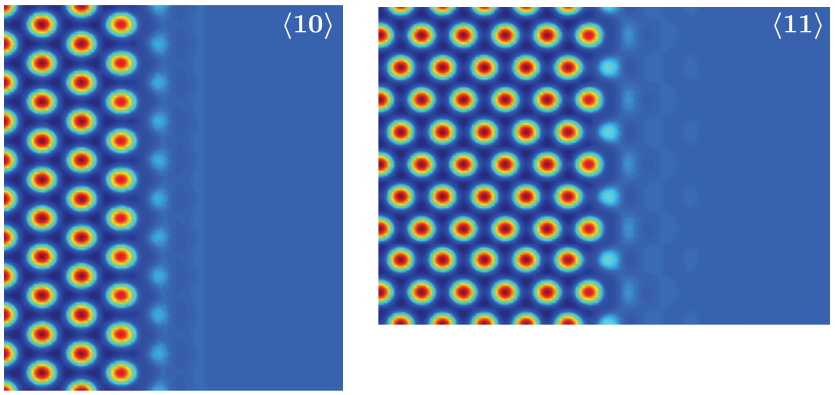


Figure 3. *Color plots of two stationary fronts are shown for the planar Swift–Hohenberg equation (1.1): Both fronts connect hexagons to the trivial state along the horizontal x-direction, are periodic in the vertical y-direction, and differ in the way in which their interfaces are aligned in a fixed hexagonal lattice.*

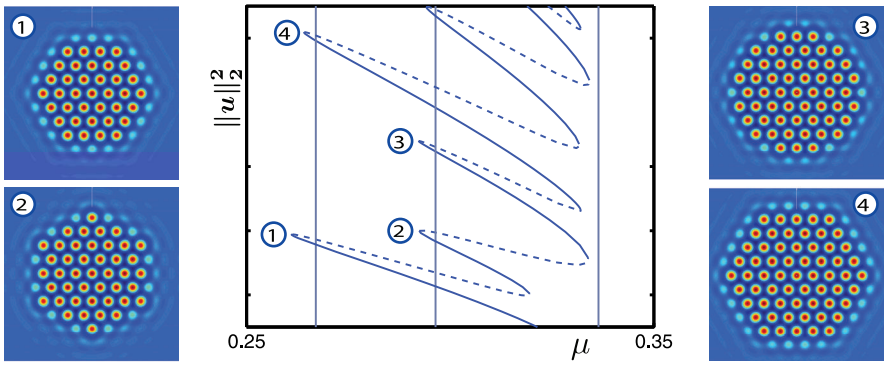


Figure 4. *Part of the bifurcation curve corresponding to localized hexagon patches of (1.1) with $\nu = 1.6$ is shown. Color plots of representative solutions are shown in panels 1–4. The entire snaking curve and color plots of the associated stationary solutions can be viewed in the accompanying animation (70762_01.mpg [10.8MB]).*

Our third result is a comprehensive numerical study of localized hexagon patches in the planar Swift–Hohenberg equation (1.1). Instead of giving a detailed list of these results, which can be found in section 5, we focus here on the observation that these localized structures snake.

Observation 1 (snaking of localized hexagon patches). *Localized hexagon patches of the Swift–Hohenberg equation exist and snake in a wedge-shaped region in the (μ, ν)-parameter plane. The shape of the hexagon patches changes along the snaking curve: their interfaces resemble planar hexagon fronts with different orientations with respect to a fixed hexagonal lattice; see Figures 3 and 4. The saddle-node bifurcations of the localized hexagon patches are aligned with saddle-nodes of planar hexagon fronts, which are shown as vertical lines in Figure 4.*

The remainder of this paper is organized as follows. We begin in section 2 with a review of snaking in one space dimension as this case motivated our paper to a large extent. We also review known results about regular hexagons and planar hexagon fronts. In section 3,

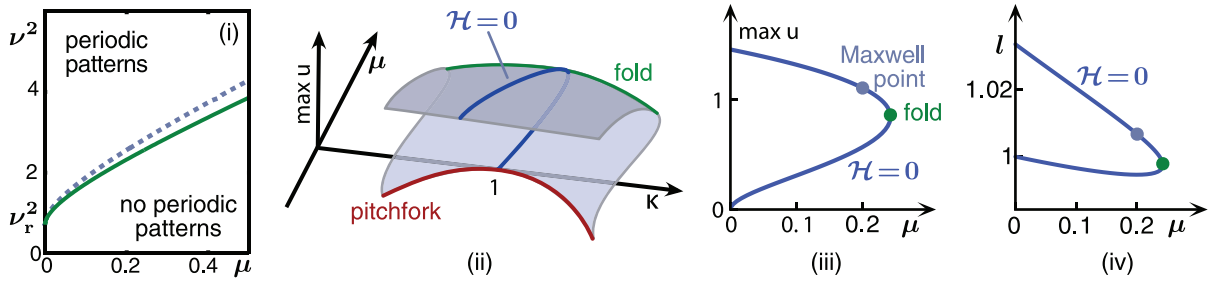


Figure 5. *Panel* (i) *contains the numerically computed fold bifurcation curve of* 1*D rolls of* (2.1) *with wavenumber* $\kappa = 1$ *(solid green) and the Maxwell curve (dotted grey) along which rolls with* $\mathcal{H} = 0$ *and* $\mathcal{E} = 0$ *exist. The schematic picture in panel* (ii) *indicates that, for fixed* $\nu > \nu_r$*, rolls exist for any wavenumber* κ *close to one. Panels* (iii) *and* (iv) *contain numerical bifurcation diagrams of the rolls that satisfy* $\mathcal{H} = 0$ *for* $\nu = 1.6$*: Shown are the amplitude and the wavelength* $l = 1/\kappa$ *(so that* $l = 1$ *corresponds to a period of* 2π*). The Maxwell point* $\mathcal{E} = 0$ *occurs on the upper branch, where rolls are stable.*

we discuss selection principles for hexagons and prove Propositions 1 and 2. We outline in section 4 the numerical algorithms that we used to compute planar hexagon fronts and localized hexagon patches and comment on their implementation. Our main results can be found in section 5, where we discuss fully localized hexagon and rhomboid patches. We end in section 6 with conclusions and a discussion.

Throughout this paper, we use color plots to illustrate the profiles of stationary planar patterns and refer the reader to the caption of Figure 1 for an explanation of what these plots represent. Two-parameter bifurcation diagrams are always drawn using (μ, ν^2) rather than (μ, ν), which makes the diagrams more legible.

2. Review of 1D snaking and planar hexagons.

2.1. Snaking in one space dimension.

Recall the steady-state equation

$$-(1+\partial_x^2)^2 u - \mu u + \nu u^2 - u^3 = 0, \qquad x \in \mathbb{R}, \tag{2.1}$$

of the 1D Swift–Hohenberg equation (1.1). Equation (2.1) has two important features that we shall exploit: it is invariant under the reflection $x \mapsto -x$ and admits the first integral $\mathcal{H}$ given in (1.3).

At $\mu = 0$, the trivial state $u = 0$ undergoes a pitchfork bifurcation to even spatially periodic patterns with period 2π or wavenumber $\kappa = 1$. These patterns bifurcate supercritically with $\mu < 0$ when $\nu < \nu_r := \sqrt{27/38}$ and subcritically with $\mu > 0$ when $\nu > \nu_r$. To accommodate the switch from super- to subcritical at $\nu = \nu_r$, a fold of periodic patterns with wavenumber $\kappa = 1$ emerges from $(\mu, \nu) = (0, \nu_r)$ into the positive half-plane, as shown in Figure 5(i). In fact, even periodic patterns bifurcate for any wavenumber κ close to one along a curve of pitchfork bifurcations, as indicated in Figure 5(ii) for a fixed $\nu > \nu_r$.

We are interested in standing localized structures such as those shown in Figure 2. As argued in section 1, the Maxwell curve predicts parameter regions where these structures may exist [11, 68, 70]: recall that Maxwell points are found by calculating, for each given (μ, ν), the periodic roll pattern $u_{\rm per}(\mu, \nu)$ that has vanishing first integral $\mathcal{H} = 0$ and subsequently adjusting the parameters so that the energy $\mathcal{E}(u_{\rm per}(\mu, \nu))$, computed over one period of $u_{\rm per}$,

vanishes as well. This can be done numerically, and the result [17, 19] is the Maxwell curve shown in Figure 5(i) which emerges from the codimension-two point $(0, \nu_r)$, where the bifurcation to rolls changes from super- to subcritical. We give two arguments that show that the Maxwell curve can emerge only from this point. First, small-amplitude rolls are unstable at subcritical bifurcations, while they bifurcate for $\mu < 0$ at supercritical bifurcations: in neither case can we find stable rolls for $\mu > 0$ that coexist with the stable trivial state. Alternatively, the normal form at pitchfork bifurcations is $\mu w - aw^3 = 0$, and the associated energy $\mathcal{E}(w) = \mu w^2/2 - aw^4/4$ does not vanish at the bifurcating state when $a \neq 0$.

Around the Maxwell curve, a pinning region exists where stable localized patterns can be found [73]. We shall now briefly review two different approaches that explain why this pinning region is present and why snaking occurs.

Asymptotics beyond all orders. Near the codimension-two point $(\mu, \nu) = (0, \nu_r)$ with $\nu_r = \sqrt{27/38}$, the snaking behavior in (1.1) can be explained by the interaction between the underlying periodic state and the slowly varying envelope which forms the localized structure. On the level of amplitude equations, the fast scale of the underlying periodic pattern is ignored, and these amplitude equations then predict that stationary pulses exist only at the Maxwell point. In the works [11, 13], the fast scale was formally reintroduced at lowest order into the amplitude equation, and the effect on locking and pinning was analyzed. This approach predicts an exponentially small width of the order $\exp(-\pi/|\nu - \nu_r|^2)$ of the locking region in the parameter μ when $\nu > \nu_r$ is fixed with $|\nu - \nu_r| \ll 1$ but does not capture the precise asymptotic behavior in ν. A consistent asymptotic analysis beyond all orders was recently carried out in [25, 52] which gives the precise asymptotic behavior of the width of the locking region in the parameter μ as a function of ν as ν approaches ν_r. Multiple scale expansions that capture the different scaling regimes near the bifurcation point and the Maxwell curve in one step, albeit without addressing terms beyond all orders, were introduced earlier in [16, 17]; in addition, the wavelength correction along the Maxwell curve was calculated in [17, (3.27)].

Dynamical-systems geometry. The second approach we shall discuss is of a more geometric nature and due to [92, 26]. We rewrite the fourth-order steady-state equation (2.1) as the first-order system

$$U_x = F(U; \mu, \nu), \qquad U = (u, u_x, u_{xx}, u_{xxx}) \in \mathbb{R}^4, \tag{2.2}$$

where we regard x as the time-like evolution variable. Recall that (2.2) has the first integral $\mathcal{H}$ from (1.3). The reflection invariance $x \mapsto -x$ of (2.1) means that (2.2) is reversible with reverser $\mathcal{R}U := (u, -u_x, u_{xx}, -u_{xxx})$: if $U(x)$ is a solution, so is $\mathcal{R}U(-x)$. Reversible solutions $U(x)$ of (2.2), which by definition satisfy $U(0) \in \operatorname{Fix}\mathcal{R}$, correspond to even solutions $u(x)$ of (2.1).

The trivial state $u = 0$ and the even periodic patterns $u_{per}(x)$ of (2.1) correspond to the equilibrium $U = 0$ and reversible periodic orbits $U_{per}(x)$, respectively, of (2.2). If the trivial state $u = 0$ and the rolls $u_{per}(x)$ are temporally stable with respect to the Swift–Hohenberg equation, then the corresponding solutions $U = 0$ and $U_{per}(x)$ of (2.2) are hyperbolic. If, for instance, $U = 0$ were not hyperbolic, then the matrix $F_U(0; \mu, \nu)$ would have a purely imaginary eigenvalue $\mathrm{i}\omega$, and $u(x, t) = \mathrm{e}^{\mathrm{i}\omega x}$ would satisfy the linearization of (1.1) about $u = 0$, which contradicts temporal stability. A similar argument applies to purely imaginary Floquet exponents of $U_{per}(x)$, which are related to the dispersion relation of $u_{per}(x)$: if the

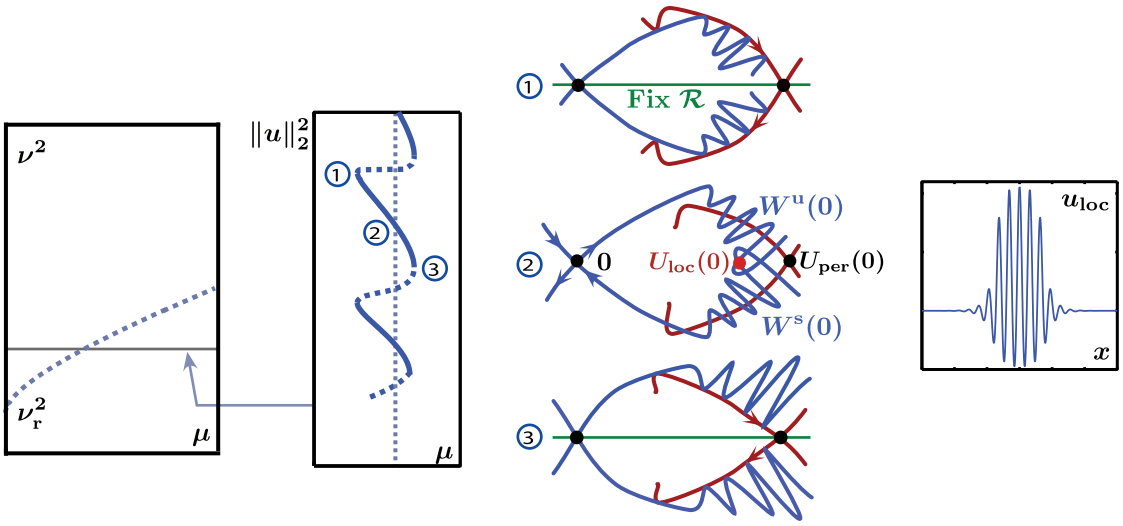


Figure 6. *The leftmost panel shows the Maxwell curve in the (μ, ν^2)-parameter space, while the center-left panel shows the bifurcation diagram for a fixed value of ν. The three pictures in the center-right panel illustrate the geometry of the stable and unstable manifolds of the equilibrium $U = 0$ and the periodic orbit $U_{\rm per}$ in a two-dimensional Poincaré section (see text for further details): The tangles of the unstable and stable manifolds of $U = 0$ are caused by the expansion and contraction near the periodic orbit and are therefore more pronounced near the periodic orbit. The localized pattern shown in the rightmost panel corresponds to the intersection $U_{\rm loc}$ of the stable and unstable manifolds of $U = 0$ near the periodic orbit $U_{\rm per}$.*

rolls $u_{\rm per}(x)$ are spectrally stable, then the periodic orbit $U_{\rm per}(x)$ will have only two Floquet exponents at zero; see [66, 67, 82, 83] and [81, sect. 3.4.2] for further details. In summary, we can identify the localized patterns $u_{\rm loc}(x)$ shown in Figure 2 with homoclinic orbits of (2.2) that lie in the intersections of the stable and unstable manifolds of $U = 0$ which come close to the hyperbolic periodic orbit $U_{\rm per}(x)$.

To visualize this situation better, we restrict (2.2) to the three-dimensional invariant zero level set $\mathcal{H}^{-1}(0)$ of the first integral $\mathcal{H}$. Next, we choose a two-dimensional Poincaré section Σ in the three-dimensional set $\mathcal{H}^{-1}(0)$ at the point $U_{\rm per}(0)$ on the periodic orbit $U_{\rm per}(x)$. The fixed-point space $\mathrm{Fix}\,\mathcal{R}$ of the reverser $\mathcal{R}$ becomes a line in the section Σ which can be used to identify symmetric orbits: note that the phase diagrams will be symmetric under the reverser $\mathcal{R}$. We now make the assumption that the unstable manifold of $U = 0$ intersects the stable manifold of the periodic orbit $U_{\rm per}(x)$ transversely in the section Σ and that the parameter μ moves these manifolds transversely through each other,[1] as shown in Figure 6. Numerical computations in [42] confirmed this assumption in the snaking regime of a reversible system of two coupled second-order equations. Figure 6 illustrates the resulting geometry which explains why the existence region of localized structures is an interval in parameter space. The end points of the intervals correspond to parameter values where fronts that connect the trivial state to the patterned state disappear in saddle-node bifurcations. Figure 7 explains in more detail why the localized structures get broader as we move up along the snaking curve.

[1]Due to reversibility, the same is then true for the stable manifold of $U = 0$ and the unstable manifold of $U_{\rm per}(x)$.

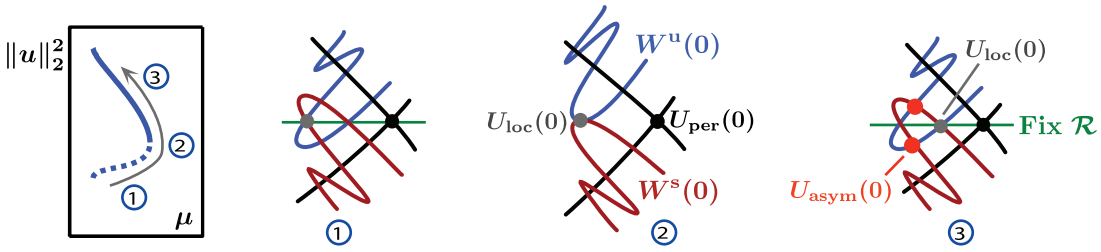


Figure 7. *As we move upward on the snaking curve, μ first increases and then decreases. Geometrically, the localized structure moves from one branch of the invariant manifolds through a saddle-node bifurcation to the adjacent branch. In doing so, it moves closer to the periodic orbit: Thus, it spends more "time" x near U_{per} and therefore broadens in space. As a result, its norm increases, which explains the structure of the snaking curve. Besides saddle-node bifurcations, the localized structures also undergo pitchfork bifurcations, which result in two asymmetric structures explained further in the text and in Figure 8.*

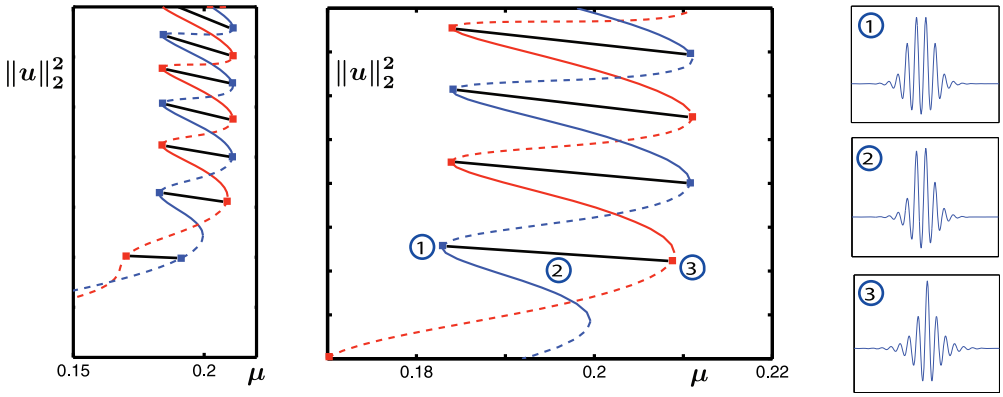


Figure 8. *The left and center panels contain the bifurcation diagrams of asymmetric localized structures that bifurcate at pitchfork bifurcations from even localized structures [20, 22]: Solutions along the curved branches are even with minima (blue) or maxima (red) at $x = 0$, while solutions along the horizontal ladders (black) are asymmetric. Panels 1–3 contain the graphs of selected solution profiles $u(x)$.*

As can be seen from Figure 7, the reflection symmetry $x \mapsto -x$ of the Swift–Hohenberg equation has another interesting consequence: each saddle-node bifurcation of an even localized structure is accompanied by a pitchfork bifurcation at which two asymmetric structures bifurcate. In the Swift–Hohenberg equation, these asymmetric states and the associated bifurcation diagrams were recently computed numerically in [20, 22], and we reproduce their numerical computations in Figure 8. As shown in Figure 8, the asymmetric structures connect two different families of even localized structures in parameter space. We give now a brief heuristic explanation of this phenomenon and refer the reader to [10] for a rigorous approach. Here and in [10], we assume that the parameter μ unfolds the intersection of $W^u(0)$ and $W^s(U_{per})$ as indicated in Figure 6 and that the resulting bifurcation diagram of even localized structures is as shown in Figure 8. Each symmetric periodic orbit $U_{per}(x)$ intersects the fixed-point space $\mathrm{Fix}\,\mathcal{R}$ precisely twice, and the two intersection points correspond to maxima and minima of u_{per}. Even localized structures arise as intersections of the unstable manifold $W^u(0)$ with $\mathrm{Fix}\,\mathcal{R}$ near either the maximum or the minimum of U_{per}. In the Swift–Hohenberg

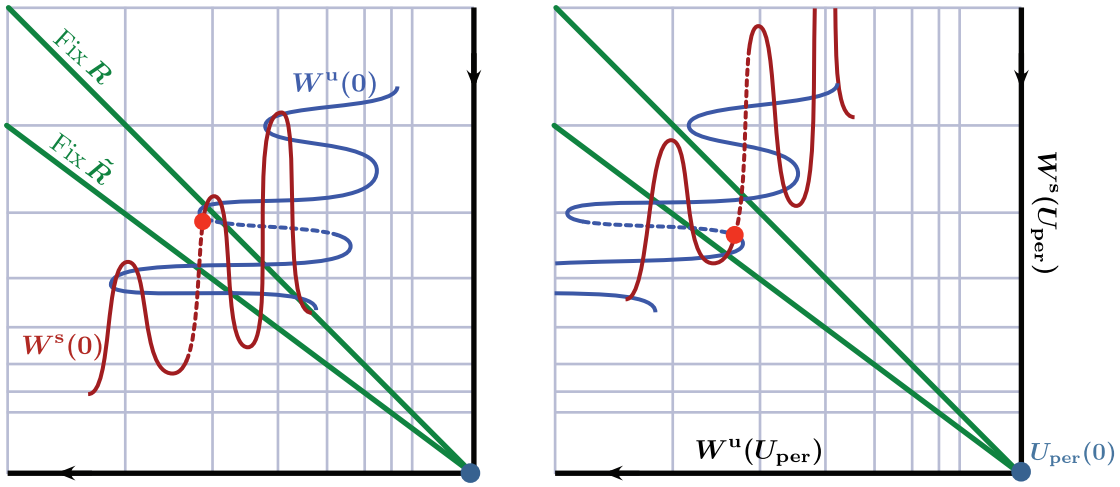


Figure 9. *Shown is the dynamics of the linearized Poincaré map near the periodic orbit U_{per}. On the left, the parameter μ is close to the pitchfork bifurcation in $\mathrm{Fix}\,\mathcal{R}$ that creates an asymmetric localized structure $U_{\mathrm{asym}}(0)$ (red bullet). As μ is increased, we assume that $W^{\mathrm{u}}(0)$ is moved to the left and, correspondingly, by reversibility, $W^{\mathrm{s}}(0)$ is moved upward. Following their intersection $U_{\mathrm{asym}}(0)$, which occurs along the branches plotted with dashes, we observe that it ends at a second pitchfork bifurcation in $\mathrm{Fix}\,\tilde{\mathcal{R}}$. Thus, the asymmetric localized structures connect, in parameter space, even solutions with a minimum at $x = 0$ to even solutions with a maximum at $x = 0$. The points $U_{\mathrm{asym}}(0)$ stay on the same horizontal and vertical segments of the invariant manifolds, and their distance from $U_{\mathrm{per}}(0)$ therefore does not change much: In particular, their L^2-norm cannot change by too much, which explains why the ladders are approximately horizontal. See also the accompanying animation* (70762_02.mov [533KB]).

equation, numerical evidence suggests that both these structures exist; see Figure 8. To capture them, we can work with two different Poincaré sections $\Sigma_{\max}$ and $\Sigma_{\min}$, placed at the maximum and minimum, or else work with a second reverser $\tilde{\mathcal{R}} := \Pi\mathcal{R}\Pi^{-1}$ near the section $\Sigma := \Sigma_{\min}$, where $\Pi : \Sigma_{\max} \to \Sigma_{\min}$ is the first-return map induced by the flow of (2.2). We choose the latter approach as it allows us to visualize the entire dynamics in one section. We straighten out the invariant manifolds of U_{per} and use the linearized Poincaré map near the symmetric orbit U_{per}. Using the assumptions made above on the unfolding of the intersections with respect to the parameter μ, we obtain the diagrams shown in Figure 9 which reproduce the numerically observed ladder structure geometrically.

We now turn to a discussion of the shape of the snaking region in the (μ, ν)-parameter space. Figure 6 indicates that snaking occurs in an interval in μ that is bounded by fold bifurcations of heteroclinic orbits which connect $U = 0$ to the rolls $U_{\mathrm{per}}(x)$. To demarcate the snaking region in (μ, ν)-space, we should therefore continue these fold bifurcations in parameter space, which is a difficult numerical task as we would need to find simultaneously periodic solutions, their Floquet eigenfunctions, and the heteroclinic orbits at a structurally unstable saddle-node bifurcation (see [53] for a recent numerical approach to this problem). Instead, we continue fold bifurcations of localized rolls in the parameters (μ, ν). As shown in Figure 10, these fold curves do not reach the codimension-two point $(\mu, \nu) = (0, \nu_{\mathrm{r}})$ but instead collide pairwise in cusp bifurcations. Continuing fold curves of localized rolls therefore gives a good approximation of the snaking region which fails, however, near the codimension-two

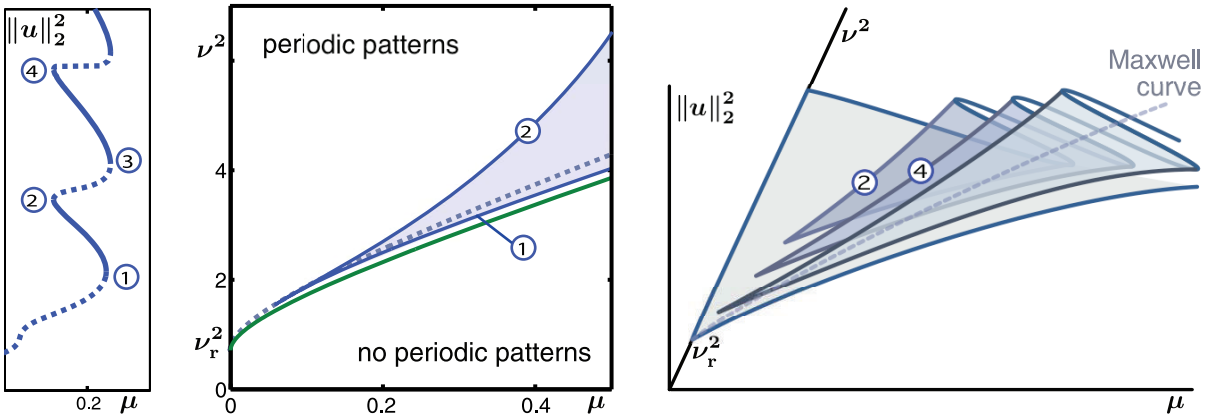


Figure 10. *The middle panel contains the numerically computed fold curves of localized rolls of (2.1) associated with the folds labeled (1) and (2) in the left panel: The two fold curves collide in a cusp bifurcation. The right panel contains a schematic illustration of the sheet of localized rolls: Fold curves of localized rolls will collide pairwise in cusps, and we believe that the sequence of cusps approaches the codimension-two point $(0, \nu_r)$ along the Maxwell curve.*

point.

2.2. Regular hexagonal patterns. We briefly review known results on the existence and stability of regular hexagons [33, 40, 41, 49] (see also [35, sect. 2] for a review), which can be tiled to cover the entire plane.

The planar Swift–Hohenberg equation

$$u_t = -(1+\Delta)^2 u - \mu u + \nu u^2 - u^3, \qquad \mathbf{x} = (x, y) \in \mathbb{R}^2 \tag{2.3}$$

is equivariant under the action of the Euclidean symmetry group $\mathbb{E}(2)$ which consists of rotations, translations, and reflections of the plane. Thus, we may seek stationary solutions to this equation that are invariant under a given fixed subgroup of $\mathbb{E}(2)$. We focus on hexagons with wavenumber $\kappa = 1$ and consider therefore the planar hexagonal lattice $\mathcal{L}$,

$$\mathcal{L} = \{n_1 l_1 + n_2 l_2 + n_3 l_3 \in \mathbb{R}^2;\, n_1, n_2, n_3 \in \mathbb{Z}\},$$

where

$$l_1 = 2\pi\left(1, \frac{1}{\sqrt{3}}\right), \qquad l_2 = 2\pi\left(0, -\frac{2}{\sqrt{3}}\right), \qquad l_3 = 2\pi\left(-1, \frac{1}{\sqrt{3}}\right);$$

see Figure 11. It is convenient to retain the lattice vector l_3 even though it is redundant. Hexagons are time-independent solutions $u(\mathbf{x})$ of (2.3) which are invariant under the subgroup $\mathbb{D}_6$ of $\mathbb{E}(2)$ and which are $\mathcal{L}$-periodic so that

$$u(\mathbf{x} + l) = u(\mathbf{x}) \qquad \forall l \in \mathcal{L},\ \forall \mathbf{x} \in \mathbb{R}^2.$$

To find hexagons, it is convenient to use the dual lattice $\mathcal{L}^*$ defined via

$$\mathcal{L}^* = \{n_1 k_1 + n_2 k_2 + n_3 k_3 \in \mathbb{R}^2;\, n_1, n_2, n_3 \in \mathbb{Z}\}$$

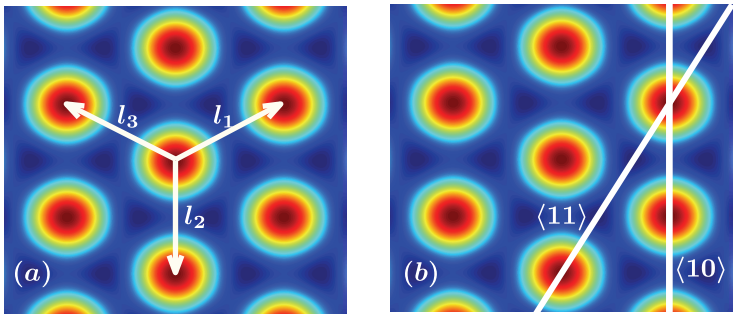


Figure 11. *Shown are color plots of regular hexagons on the lattice $\mathcal{L}$ together with the lattice vectors l_1, l_2, and l_3 in panel* (a) *and with two lines with Bravais–Miller indices* $\langle 10\rangle$ *and* $\langle 11\rangle$ *in panel* (b).

with

$$k_1 = (-1, 0), \qquad k_2 = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right), \qquad k_3 = \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right),$$

which allows us to represent $\mathcal{L}$-periodic solutions of (2.3) by the Fourier series

$$u(\mathbf{x}, t) = \sum_{k\in\mathcal{L}^*} \hat{u}_k(t) \mathrm{e}^{\mathrm{i}k\cdot\mathbf{x}}.$$

We now restrict our attention to hexagons that bifurcate from $u = 0$. Linearizing (2.3) about $u = 0$ gives the linear operator $-(1+\Delta)^2-\mu$. Posed on appropriate spaces of $\mathcal{L}$-periodic functions, the first instability occurs at $\mu = 0$, where an eigenvalue of algebraic multiplicity six, with eigenfunctions $\exp(\pm \mathrm{i}k_j \cdot \mathbf{x})$ for $j = 1, 2, 3$, crosses into the right half-plane. On the resulting six-dimensional center manifold, we can parametrize solutions as

(2.4)
$$u(\mathbf{x}, t) = A_1(t)\mathrm{e}^{-\mathrm{i}x} + A_2(t)\mathrm{e}^{\mathrm{i}(x+\sqrt{3}y)/2} + A_3(t)\mathrm{e}^{\mathrm{i}(x-\sqrt{3}y)/2} + h(A_1, A_2, A_3, \overline{A}_1, \overline{A}_2, \overline{A}_3; \mu, \nu) + \text{c.c.},$$

where the function h represents the higher-order contributions to the center manifold. The flow on the center manifold can be calculated for $|\mu| \ll 1$ and $|\nu| \ll 1$ by evaluating the Swift–Hohenberg equation on the elements of the center manifold and projecting the resulting expression back onto the center eigenspace.[2] The result is

(2.5)
$$\begin{aligned}
\dot{A}_1 &= -\mu A_1 + \alpha_1 \overline{A}_2\overline{A}_3 + \alpha_2 A_1|A_1|^2 + \alpha_3 A_1(|A_2|^2 + |A_3|^2) + r(A_1, A_2, A_3),\\
\dot{A}_2 &= -\mu A_2 + \alpha_1 \overline{A}_1\overline{A}_3 + \alpha_2 A_2|A_2|^2 + \alpha_3 A_2(|A_1|^2 + |A_3|^2) + r(A_2, A_3, A_1),\\
\dot{A}_3 &= -\mu A_3 + \alpha_1 \overline{A}_2\overline{A}_1 + \alpha_2 A_3|A_3|^2 + \alpha_3 A_3(|A_2|^2 + |A_1|^2) + r(A_3, A_1, A_2),
\end{aligned}$$

plus the complex conjugated equations of six ODEs for the complex amplitudes A_j, where the coefficients α_j are real and are given by

$$\alpha_1 = 2\nu + \mathrm{O}(|\mu|(|\mu| + |\nu|)), \qquad \alpha_2 = -3 + \mathrm{O}(|\mu| + |\nu|), \qquad \alpha_3 = -6 + \mathrm{O}(|\mu| + |\nu|),$$

[2] If ν is small, then the function h is not needed for the calculation of the cubic terms of the reduced vector field: The coefficients in (2.5) for arbitrary ν can be found in [44].

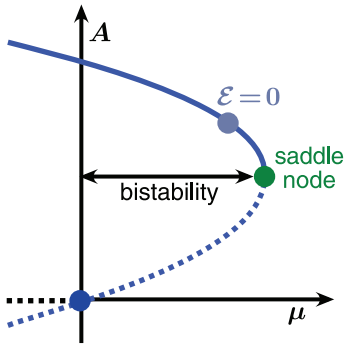


Figure 12. *The bifurcation diagram of hexagons with critical wavenumber $\kappa = 1$ in the Swift–Hohenberg equation is shown for fixed $0 < \nu \ll 1$. Solid lines correspond to stable patterns, while dotted lines correspond to unstable ones. The hexagons undergo a saddle-node bifurcation at amplitude $A = \nu/15 + \mathrm{O}(\nu^2)$ when $\mu = \nu^2/15 + \mathrm{O}(\nu^3)$. The energy $\mathcal{E}$ vanishes only for $\mu = \mu_{\mathcal{E}}$ given in (2.9) for the hexagons with amplitude $A = 4\nu/45 + \mathrm{O}(\nu^2)$ on the upper branch.*

and the remainder term $r(z) = \mathrm{O}(|z|^4)$ is of higher order. Equivariance of the Swift–Hohenberg equation and invariance of the hexagonal lattice with respect to rotations by $\pi/3$, reflections, and translations manifest themselves in the equivariance of the reduced system with respect to the transformations

$$\begin{aligned}
\sigma &: (A_1, A_2, A_3) \mapsto (\overline{A}_3, \overline{A}_1, \overline{A}_2),\\
\rho &: (A_1, A_2, A_3) \mapsto (A_1, A_3, A_2),\\
\tau_a &: (A_1, A_2, A_3) \mapsto (\mathrm{e}^{\mathrm{i}a_1} A_1, \mathrm{e}^{-\mathrm{i}(a_1+\sqrt{3}a_2)/2} A_2, \mathrm{e}^{\mathrm{i}(-a_1+\sqrt{3}a_2)/2} A_3),
\end{aligned}$$

respectively. Hexagons are invariant under $\mathbb{D}_6$ and therefore lie, in particular, in the one-dimensional intersection of the fixed-point spaces of σ^2 and σ^3, which is given by $A := A_1 = A_2 = A_3 \in \mathbb{R}$. As an intersection of fixed-point spaces, the line $A_1 = A_2 = A_3 \in \mathbb{R}$ is invariant under (2.5), and hexagons can therefore be found as nontrivial equilibria of the differential equation

$$\dot{A} = -\mu A + 2\nu A^2 - 15A^3 + \mathrm{O}((|\mu| + |\nu|)(|\mu| + |A|)|A|^2 + |A|^4) \tag{2.6}$$

which exist for

$$\mu = 2\nu A - (15 + \mathrm{O}(|\nu| + |A|))A^2. \tag{2.7}$$

The stability of these hexagons with respect to $\mathcal{L}$-periodic perturbations is calculated by considering the linearization of (2.5) about the hexagons [23, 40]. The resulting bifurcation diagram of hexagons is plotted for $\nu > 0$ in Figure 12; other solution branches corresponding to mixed modes, which bifurcate in secondary bifurcations, and rolls exist but are not shown in Figure 12. The hexagons with $A > 0$ shown in Figure 12 are up-hexagons: our focus is on $\mu, \nu > 0$, and we shall therefore encounter only up-hexagons in the rest of this paper. When $\nu < 0$, the bifurcation diagram in Figure 12 does not change except that A is reflected via $A \mapsto -A$: in this case, localized hexagon patterns consist of down-hexagons, though

the bifurcation diagrams shown in the remainder of the paper would not change due to the symmetry $(u,\nu)\mapsto(-u,-\nu)$.

We now calculate the energy $\mathcal{E}(u)$ given in (1.4) along the hexagon branch (2.7). Substituting

$$u(x,y)=2A\left[\cos(x)+\cos((x+\sqrt{3}y)/2)+\cos((x-\sqrt{3}y)/2)\right]$$

into (1.4) and integrating over the fundamental domain $[0,4\pi]\times[0,4\pi/\sqrt{3}]$, we find that the energy of the hexagons with wavenumber $\kappa=1$ is given by

$$\mathcal{E}=\frac{8\pi^2A^3}{\sqrt{3}}(4\nu-45A)+\mathrm{O}((|\nu|+|A|)A^4), \tag{2.8}$$

which vanishes precisely when $A=4\nu/45+\mathrm{O}(\nu^2)$, which corresponds to the parameter curve

$$\mu_{\mathcal{E}}=\frac{8\nu^2}{135}+\mathrm{O}(\nu^3). \tag{2.9}$$

For fixed $\nu>0$, the energy (2.8) along the bifurcating hexagons is strictly larger than zero for $0<|A|\ll 1$.

Throughout this section, we considered only hexagons with wavenumber $\kappa=1$. Hexagons with wavenumbers κ close to one bifurcate for $\mu=-(1-\kappa^2)^2$, and these can be captured by an analogous analysis upon using the arguments $(\kappa x,\kappa y)$ in place of (x,y) in the right-hand side of (2.4).

2.3. Planar hexagon fronts. We now discuss planar stationary hexagon fronts that connect hexagons to the trivial state. As illustrated in Figure 3, these fronts can have different orientations with respect to the hexagonal lattice $\mathcal{L}$, which can be classified using the Bravais–Miller index [5].

Definition 1 (Bravais–Miller index). *Fix the hexagonal lattice $\mathcal{L}$. The Bravais–Miller index $\langle n_1\,n_2\,n_3\rangle$ of a line in the plane is given by the reciprocals n_j of the intercepts of the line with the lines $\mathbb{R}l_j$ generated by the lattice vectors l_j (assigning the reciprocal $n_j=0$ if the line does not intersect $\mathbb{R}l_j$). Negative indices $-n$ with $n>0$ are conventionally written as $\overline{n}:=-n$. Since $n_1+n_2+n_3=0$, we may write the index using only two indices: our choice is $\langle n_1\,n_2\rangle$, and we refer the reader to Figure 11(b) for examples.*

If a hexagon front has a straight interface, we assign the Bravais–Miller index of its interface to it. Two examples of Bravais–Miller indices are given in Figure 11(b), and the corresponding hexagon fronts are shown in Figure 3. We shall present more detailed numerical results for the existence of fronts with these orientations later in the paper.

For μ close to zero, stationary fronts between hexagons and the trivial state can be found using a formal multiscale expansion as carried out in [60], though we note that this approach cannot capture the expected pinning and locking of these fronts. To construct stationary $\langle 10\rangle$ fronts, we substitute the ansatz

$$u(x,y)=\epsilon\left(A_1(\epsilon x)\mathrm{e}^{-\mathrm{i}x}+A_2(\epsilon x)\left[\mathrm{e}^{\mathrm{i}(x+\sqrt{3}y)/2}+\mathrm{e}^{\mathrm{i}(x-\sqrt{3}y)/2}\right]+\text{c.c.}\right)$$

into (2.3), set $\mu = \epsilon^2\tilde{\mu}$ and $\nu = \epsilon\tilde{\nu}$, and expand in powers of ϵ. To leading order, we obtain the system

$$
\begin{aligned}
0 &= 4\partial_X^2 A_1 - \tilde{\mu}A_1 + 2\tilde{\nu}\overline{A}_2^2 - 3A_1|A_1|^2 - 12A_1|A_2|^2,\\
0 &= \partial_X^2 A_2 - \tilde{\mu}A_2 + 2\tilde{\nu}\overline{A}_1\overline{A}_2 - 9A_2|A_2|^2 - 6A_2|A_1|^2,
\end{aligned}
\tag{2.10}
$$

where the nonlinearity on the right-hand side comes from (2.5) upon setting $A_2 = A_3$. A similar ODE can be derived for $\langle 11\rangle$ fronts and, in fact, for fronts of any given orientation [60].

Stationary fronts with index $\langle 10\rangle$ that connect the trivial state to hexagons with wavenumber $\kappa = 1$ correspond to heteroclinic orbits of (2.10) that connect $(A_1, A_2) = (0, 0)$ to the hexagon solutions $(A_1, A_2) = (\tilde{A}, \tilde{A})$, where $A = \epsilon\tilde{A}$ is an equilibrium of (2.6). Equation (2.10) admits the first integral

$$
\begin{aligned}
\mathcal{H}(A_1, A_2) = {}& 2|\partial_X A_1|^2 + |\partial_X A_2|^2 - \tilde{\mu}\left(\frac{|A_1|^2}{2} + |A_2|^2\right)\\
&+ \tilde{\nu}(\overline{A}_1\overline{A}_2^2 + A_1A_2^2) - \frac{3}{4}|A_1|^4 - 6|A_1|^2|A_2|^2 - \frac{9}{2}|A_2|^4,
\end{aligned}
\tag{2.11}
$$

which is constant along solutions of (2.10). In particular, heteroclinic orbits between $(0, 0)$ and $(\tilde{A}, \tilde{A})$ can exist only when $\mathcal{H}(\tilde{A}, \tilde{A}) = \mathcal{H}(0, 0) = 0$, which gives the condition

$$
\mathcal{H}(\tilde{A}, \tilde{A}) = -\frac{3\tilde{\mu}}{2}\tilde{A}^2 + 2\tilde{\nu}\tilde{A}^3 - \frac{45}{4}\tilde{A}^4 = 0
$$

that we need to solve for $\tilde{A}$. Using (2.7) and interpreting the results in the original unscaled parameters, we arrive at the condition

$$
\mu_{\mathrm{M}} = \frac{8\nu^2}{135} + \mathrm{O}(\nu^3),
\tag{2.12}
$$

along which $\mathcal{H}$ vanishes at the bifurcating hexagons with wavenumber $\kappa = 1$. Equation (2.9) shows that the energy $\mathcal{E}$ vanishes at the same hexagons for the same parameter values, to the order in which we computed them. Thus, the curve defined by (2.12) gives the Maxwell curve, which provides a heuristic criterion for the existence of planar fronts but does not account for pinning and locking phenomena.

The analysis reviewed here is valid only in the limit $(\mu, \nu) \to 0$ and does not address the wavenumber selection as we fixed $\kappa = 1$. In section 3, we will construct a conserved quantity $\mathcal{H}$ which defines the Maxwell curve for general parameter values and prove Proposition 3 on the selection of hexagons that satisfy $\mathcal{H} = 0$. In section 5.2, we shall compare the predictions made by the Maxwell curve (2.12) and its extension from section 3 with numerical computations for the full 2D Swift–Hohenberg equation.

3. Spatial dynamics, and selection principles for hexagons. In this section, we prove Propositions 1 and 2.

3.1. Proof of Proposition 1. Proposition 1 states that if $u(x,y)$ is a bounded solution of the planar stationary Swift–Hohenberg equation

$$-(1+\Delta)^2 u - \mu u + \nu u^2 - u^3 = 0, \qquad (x,y) \in \mathbb{R}^2, \tag{3.1}$$

which is spatially periodic with period ℓ in the y-variable, then the quantity

$$\mathcal{H}(u) = \int_0^\ell \left[u_{xxx}u_x - \frac{u_{xx}^2}{2} + u_x^2 + \frac{(1+\mu)u^2}{2} - \frac{\nu u^3}{3} + \frac{u^4}{4} - u_{xy}^2 - u_y^2 + \frac{u_{yy}^2}{2} \right] \mathrm{d}y \tag{3.2}$$

does not depend on x. This can, of course, be verified directly by computing the derivative of (3.2) with respect to x and using (3.1), and we omit this straightforward calculation.

Instead, we outline how the first integral $\mathcal{H}$ can be derived in the first place from the translation invariance of the Lagrangian associated with (3.1) and refer the reader to [2, 3] and [24, Ch. 15] for a general abstract approach. We start with a general energy functional[3]

$$\mathcal{E}(u) = \int_{\mathbb{R}^d} \mathcal{L}(u(\mathbf{x}), \nabla u(\mathbf{x}), \Delta u(\mathbf{x}))\, \mathrm{d}\mathbf{x}, \qquad \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

where $\mathcal{L}(q,p,r) : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is a given smooth function that does not depend explicitly upon $\mathbf{x}$. The Euler–Lagrange equation associated with the energy functional $\mathcal{E}$ reflects the extremum condition $\nabla\mathcal{E}(u) = \frac{\delta\mathcal{E}}{\delta u}(u) = 0$, where the gradient with respect to u is calculated in the L^2-scalar product, and is given by

$$\mathcal{L}_q(u, \nabla u, \Delta u) - \nabla \cdot \mathcal{L}_p(u, \nabla u, \Delta u) + \Delta \mathcal{L}_r(u, \nabla u, \Delta u) = 0, \qquad \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d, \tag{3.3}$$

where the partial derivatives $\mathcal{L}_q$, $\mathcal{L}_p$, and $\mathcal{L}_r$ of the function $\mathcal{L}(q,p,r)$ are evaluated at $(u(\mathbf{x}), \nabla u(\mathbf{x}), \Delta u(\mathbf{x}))$. Assume now that $u(\mathbf{x})$ is a smooth solution of the Euler–Lagrange equation (3.3). We compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x_1}\left[\mathcal{L}(u, \nabla u, \Delta u)\right] &= \mathcal{L}_q u_{x_1} + \mathcal{L}_p \cdot \nabla u_{x_1} + \mathcal{L}_r \Delta u_{x_1} \\
&\overset{(3.3)}{=} (\nabla \cdot \mathcal{L}_p - \Delta \mathcal{L}_r) u_{x_1} + \mathcal{L}_p \cdot \nabla u_{x_1} + \mathcal{L}_r \Delta u_{x_1} \\
&= u_{x_1} \nabla \cdot \mathcal{L}_p + \nabla u_{x_1} \cdot \mathcal{L}_p + \mathcal{L}_r \Delta u_{x_1} - (\Delta \mathcal{L}_r) u_{x_1} \\
&= \nabla \cdot (u_{x_1} \mathcal{L}_p) + \nabla \cdot (\mathcal{L}_r \nabla u_{x_1} - u_{x_1} \nabla \mathcal{L}_r).
\end{aligned}$$

Thus, we have established the existence of a conservation law for the Euler–Lagrange equation.

Lemma 1. *Assume that $u(\mathbf{x})$ is a smooth solution of the Euler–Lagrange equation (3.3) associated with the Lagrangian $\mathcal{L}(q,p,r)$; then the conservation law*

$$\partial_{x_1}\mathcal{L}(u, \nabla u, \Delta u) - \nabla \cdot \left[u_{x_1} \mathcal{L}_p(u, \nabla u, \Delta u) + \mathcal{L}_r(u, \nabla u, \Delta u) \nabla u_{x_1} - u_{x_1} \nabla \mathcal{L}_r(u, \nabla u, \Delta u) \right] = 0 \tag{3.4}$$

is satisfied for all $\mathbf{x} \in \mathbb{R}^d$.

[3] In section 3.1, we use $\mathcal{L}$ exclusively for the Lagrangian; in all other sections, this letter refers to the hexagonal lattice.

We now return to the planar Swift–Hohenberg equation (3.1), which is the Euler–Lagrange equation associated with the energy functional

$$\mathcal{E}(u) = \int_{\mathbb{R}^2} \mathcal{L}(u, \nabla u, \Delta u)\, \mathrm{d}x\, \mathrm{d}y$$

for the Lagrangian

$$\mathcal{L}(q, p, r) := \frac{(q+r)^2}{2} + \frac{\mu q^2}{2} - \frac{\nu q^3}{3} + \frac{q^4}{4}, \tag{3.5}$$

where $(q, p, r) \in \mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}$. Lemma 1 asserts that (3.4) is met for any smooth solution $u(x, y)$ of (3.1). If we further assume that $u(x, y)$ is periodic with period ℓ in y, then we can integrate (3.4) in y over $[0, \ell]$ and use periodicity in y to find that the equation

$$\partial_x \int_0^\ell [\mathcal{L}(u, \nabla u, \Delta u) - u_x \mathcal{L}_{p_1}(u, \nabla u, \Delta u) - \mathcal{L}_r(u, \nabla u, \Delta u) u_{xx} + u_x \partial_x \mathcal{L}_r(u, \nabla u, \Delta u)]\, \mathrm{d}y = 0$$

is met for all x, where we write $p = (p_1, p_2)$. In particular, along such solutions,

$$\int_0^\ell [\mathcal{L}(u, \nabla u, \Delta u) - u_x \mathcal{L}_{p_1}(u, \nabla u, \Delta u) - \mathcal{L}_r(u, \nabla u, \Delta u) u_{xx} + u_x \partial_x \mathcal{L}_r(u, \nabla u, \Delta u)]\, \mathrm{d}y$$

does not depend on x. Substituting the expression for $\mathcal{L}$ from (3.5), we find that the integral above coincides with the expression (3.2) for $\mathcal{H}$, as claimed.

3.2. Spatial dynamics. As in the 1D situation, the quantity $\mathcal{H}(u)$ given in (3.2) determines which hexagons can be connected by a stationary planar front to the trivial state. We use spatial dynamics to gain further insight into why snaking should occur for stationary planar fronts. We focus on fronts that connect hexagons to the trivial state and that are spatially periodic along the interface.

Thus, assume that $u(x, y)$ is a smooth solution of (3.1) which is periodic in y with positive minimal period ℓ. We define the wavenumber

$$\kappa = \frac{2\pi}{\ell}$$

and introduce the rescaling $\phi = \kappa y$. With this rescaling, $\phi \in [0, 2\pi]$ corresponds to $y \in [0, \ell]$, and the planar Swift–Hohenberg equation (3.1) becomes

$$\partial_x^4 u + 2\kappa^2 \partial_x^2 \partial_\phi^2 u + \kappa^4 \partial_\phi^4 u + 2(\partial_x^2 u + \kappa^2 \partial_\phi^2 u) + (1+\mu)u - \nu u^2 + u^3 = 0. \tag{3.6}$$

Exploiting that we are now interested in solutions with period 2π in ϕ, we write this equation as a first-order system in x and obtain

$$U_x = \mathcal{A}(\mu, \kappa) U + \mathcal{N}(U; \nu), \qquad U \in \mathcal{U}, \tag{3.7}$$

where

$$\mathcal{A}(\mu, \kappa) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\kappa^4 \partial_\phi^4 - 2\kappa^2 \partial_\phi^2 - (1+\mu) & 0 & -2 - 2\kappa^2 \partial_\phi^2 & 0 \end{pmatrix}, \qquad \mathcal{N}(U; \nu) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \nu U_1^2 - U_1^3 \end{pmatrix},$$

and $U(x) = U(x)(\varphi)$ is, for each fixed x, a function of φ that lies in $\mathcal{U} := H^3(S^1) \times H^2(S^1) \times H^1(S^1) \times L^2(S^1)$, where S^1 is the interval $[0, 2\pi]$ with the end points identified. Equation (3.7) is reversible with respect to the reverser

$$\mathcal{R}U = \mathcal{R}(U_1, U_2, U_3, U_4)^t = (U_1, -U_2, U_3, -U_4)^t,$$

which corresponds to reflections in x, and admits the first integral

$$\mathcal{H}(U; \mu, \nu, \kappa) = \int_0^{2\pi} \Big[U_2 U_4 - \frac{U_3^2}{2} + U_2^2 + \frac{(1+\mu)U_1^2}{2} - \frac{\nu U_1^3}{3} + \frac{U_1^4}{4} - \kappa^2(\partial_\phi U_2)^2 - \kappa^2(\partial_\phi U_1)^2 + \frac{\kappa^4(\partial_\phi^2 U_1)^2}{2} \Big] \mathrm{d}\phi, \tag{3.8}$$

which is simply (3.2) with y rescaled and $(u, u_x, u_{xx}, u_{xxx})$ replaced by U. Even though the system (3.7) is ill-posed in the sense that (3.7) may not have a solution for a given initial condition, we can still apply the theory developed in [71, 82, 83] to show that stable and unstable manifolds of equilibria and periodic orbits of (3.7) exist. These manifolds are infinite-dimensional, but the results in [10, 71, 83] imply that the geometric situation for (3.7) is analogous to the 1D situation.

The existence of the first integral $\mathcal{H}$ implies that if there is a heteroclinic orbit of (3.7) that connects $U = 0$ to a periodic orbit, then $\mathcal{H}$ must vanish along the heteroclinic orbit and on the periodic orbit. In particular, if we seek stationary fronts between the trivial state and regular hexagons, then for each fixed (μ, ν) there will typically be a unique regular hexagon, with a uniquely selected wavenumber κ, that satisfies the condition $\mathcal{H} = 0$. Under appropriate existence and transversality assumptions on the heteroclinic orbits that correspond to such fronts, we can use spatial dynamics to prove the existence of transverse homoclinic orbits, corresponding to planar hexagon pulses with the same selected hexagonal wavenumber, and of complex snaking bifurcation diagrams, and we refer the reader to [10] for details.

For fully localized hexagon patches such as those shown in Figures 1(a) and 4, spatial dynamics may not work: while we can view the radial variable as the evolution variable, it is not clear how appropriate function spaces can be set up that allow for the increasingly finer hexagon structure in the angular variable. Nevertheless, we may formally move along the radial direction from the center of a localized hexagon patch toward infinity and consider the interface between regular hexagons and the trivial state at the patch boundary: if this boundary becomes approximately planar as the patch grows, as seems to be the case for the hexagon patches that we present later in this paper, then we should expect, on a formal level, that the selection principle for hexagon fronts applies to localized patches, too: the existence region of localized hexagon patches should therefore be centered around the Maxwell curve of the planar hexagon fronts between trivial state and regular hexagons. Furthermore, the hexagons inside the hexagon patch should satisfy $\mathcal{H} = 0$, which selects their wavenumber.

So far, we have focused on planar hexagon fronts between the trivial state and regular hexagons as these are relevant for localized hexagon patches. However, other stationary planar hexagon fronts exist, and we outline now how they arise and what their spatial profiles look like. Assume that we found a transversely constructed heteroclinic orbit of (3.7) between

regular hexagons and the trivial state for the parameter values (μ_0, ν_0) and the wavenumber κ_0. If we vary the wavenumber κ near κ_0 while keeping (μ_0, ν_0) fixed, then the assumed transversality implies that the heteroclinic orbit will persist. However, this orbit will now connect the trivial state to frustrated hexagons that have minimal period $\ell = 2\pi/\kappa$ in the y-direction but are slightly compressed or expanded in the x-direction to accommodate the condition $\mathcal{H} = 0$ on the new cross-section $[0, \ell]$. The planar hexagon fronts constructed in this fashion live on the domain $\mathbb{R} \times [0, \ell]$ with periodic boundary conditions in the y-variable. Periodic boundary conditions allow us to view these fronts as planar hexagon fronts on the domains $\mathbb{R} \times [-n\ell, n\ell]$ for arbitrary positive integers n on the plane: the resulting fronts connect the trivial state to frustrated hexagon patterns that are periodic with minimal period ℓ in the transverse y-direction. The corresponding Maxwell curves will depend on ℓ (or κ), and we therefore obtain an entire family of Maxwell curves.

3.3. Regular hexagons. We prove Proposition 2, which states that, for each fixed ν, there is a unique branch of regular hexagons which bifurcate from $u = 0$ at $\mu = 0$ along which the first integral $\mathcal{H}$ from (3.2) vanishes. The wavelength of these selected hexagons may vary along the branch and therefore needs to be treated as an unknown which will adjust itself to satisfy the constraint $\mathcal{H} = 0$. This is a special case of the more general problem of finding solutions to Hamiltonian systems with a prescribed value of the Hamiltonian [9, 14, 18, 69, 75, 76]. Indeed, we prove Proposition 2 by applying the following general bifurcation theorem.

Theorem 3.1 (see [9, Theorem 2.2]). *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be Banach spaces. Let $\psi \in \mathcal{Z}^*$, $\mathcal{F} \in \mathcal{C}^\omega(\mathcal{X} \times \mathbb{R}^2, \mathcal{Y})$, and $\mathcal{H} \in \mathcal{C}^\omega(\mathcal{X} \times \mathbb{R}^2, \mathcal{Z})$, and consider the equation*

$$\begin{pmatrix} \mathcal{F}(u, \mu, \kappa) \\ \psi(\mathcal{H}(u, \mu, \kappa)) \end{pmatrix} = 0, \tag{3.9}$$

where $u \in \mathcal{X}$ and $(\mu, \kappa) \in \mathbb{R}^2$. We assume that $u = 0$ is a solution of (3.9) for all (μ, κ) with $\mathcal{H}_u(0, \mu, \kappa) \equiv 0$. Furthermore, assume that $\mathcal{F}_u(0, \mu_0, \kappa_0) \in L(\mathcal{X}, \mathcal{Y})$ is Fredholm[4] with index zero with $\mathrm{N}(\mathcal{F}_u(0, \mu_0, \kappa_0)) = \mathbb{R}\hat{u} \neq \{0\}$. We write $\mathcal{X} = \mathbb{R}\hat{u} \oplus \hat{\mathcal{X}}$, $\hat{\mathcal{Y}} := \mathrm{R}(\mathcal{F}_u(0, \mu_0, \kappa_0))$, and $\mathcal{Y} = \mathbb{R}\hat{v} \oplus \hat{\mathcal{Y}}$ and denote by $P : \mathcal{Y} \to \hat{\mathcal{Y}}$ the projection along $\hat{v}$. We assume now that $\psi(\mathcal{H}_{uu}(0, \mu_0, \kappa_0)[\hat{u}, \hat{u}]) = 0$ and that the operator $D \in L(\hat{\mathcal{X}} \times \mathbb{R}^2, \hat{\mathcal{Y}} \times \mathbb{R}^2)$ given by

$$D := \begin{pmatrix} P\mathcal{F}_u(0, \mu_0, \kappa_0) & P\mathcal{F}_{u\mu}(0, \mu_0, \kappa_0)[\hat{u}, 1] & P\mathcal{F}_{u\kappa}(0, \mu_0, \kappa_0)[\hat{u}, 1] \\ 0 & (1-P)\mathcal{F}_{u\mu}(0, \mu_0, \kappa_0)[\hat{u}, 1] & (1-P)\mathcal{F}_{u\kappa}(0, \mu_0, \kappa_0)[\hat{u}, 1] \\ \psi(\mathcal{H}_{uu}[\hat{u}, \cdot]) & \psi(\mathcal{H}_{uu\mu}[\hat{u}, \hat{u}, 1]) & \psi(\mathcal{H}_{uu\kappa}[\hat{u}, \hat{u}, 1]) \end{pmatrix} \tag{3.10}$$

is an isomorphism. Under these assumptions, $(0, \mu_0, \kappa_0)$ is a bifurcation point for (3.9), and there is an interval I containing 0 and a unique analytic branch $(u, \mu, \kappa)(s)$ of solutions of (3.9), defined for $s \in I$, which satisfies $(u, \mu, \kappa)(0) = (0, \mu_0, \kappa_0)$ and $\|u(s) - s\hat{u}\|_{\mathcal{X}} = \mathrm{O}(s^2)$ as $s \to 0$.

We now set up an appropriate framework that allows us to appeal to the preceding theorem to prove Proposition 2. We first rescale (x, y) by setting $x = X/\kappa$ and $y = Y/\kappa$. In the rescaled

[4] A linear operator L is called Fredholm if its null space $\mathrm{N}(L)$ is finite-dimensional, its range $\mathrm{R}(L)$ is closed, and the range $\mathrm{R}(L)$ has finite codimension. In this case, the Fredholm index is defined to be the difference $\dim \mathrm{N}(L) - \operatorname{codim} \mathrm{R}(L)$.

variables, the Swift–Hohenberg equation is given by

$$(1+\kappa^2\Delta)^2 u + \mu u - \nu u^2 + u^3 = 0.$$

We are interested in regular hexagons and therefore seek, as in section 2.2, solutions with $\mathbb{D}_6$-symmetry. In addition, we require that solutions be centered at the origin to reduce the multiplicity of solutions. Thus, we set

$$\begin{aligned}
\mathcal{X} = \Big\{ & u \in \mathcal{C}^4(\mathbb{R}^2,\mathbb{R});\ u(X,Y) = u(X,Y+4\pi/\sqrt{3}) = u(X+2\pi,Y+2\pi/\sqrt{3}),\\
& u(X,Y) = u((X+\sqrt{3}Y)/2,(-\sqrt{3}X+Y)/2) \quad \forall (X,Y) \Big\},\\
\mathcal{Y} = \Big\{ & u \in \mathcal{C}^0(\mathbb{R}^2,\mathbb{R});\ u(X,Y) = u(X,Y+4\pi/\sqrt{3}) = u(X+2\pi,Y+2\pi/\sqrt{3}),\\
& u(X,Y) = u((X+\sqrt{3}Y)/2,(-\sqrt{3}X+Y)/2) \quad \forall (X,Y) \Big\},\\
\mathcal{Z} = {} & \mathcal{C}^0(\mathbb{R},\mathbb{R})
\end{aligned}$$

and define

$$\begin{aligned}
\mathcal{F}(u,\mu,\kappa) &:= (1+\kappa^2\Delta)^2 u + \mu u - \nu u^2 + u^3,\\
\mathcal{H}(u,\mu,\kappa) &:= \int_0^{\frac{4\pi}{\sqrt{3}}} \left[\kappa^4 \left(u_{XXX} u_X - \frac{u_{XX}^2}{2} \right) + \kappa^2 u_X^2 + \frac{(1+\mu)u^2}{2} - \frac{\nu u^3}{3} + \frac{u^4}{4} \right.\\
&\qquad \left. - \kappa^4 u_{XY}^2 - \kappa^2 u_Y^2 + \frac{\kappa^4 u_{YY}^2}{2} \right] \mathrm{d}Y,\\
\psi\mathcal{H}(u,\mu,\kappa) &:= \mathcal{H}(u,\mu,\kappa)\big|_{X=0}.
\end{aligned}$$

The required regularity assumptions are then met, and it is clear that $u=0$ is a solution for all (μ,κ) with $\mathcal{H}_u(0,\mu,\kappa) \equiv 0$. We calculate the derivatives

$$\begin{aligned}
\mathcal{F}_u(0,\mu,\kappa)[v] &= (1+\kappa^2\Delta)^2 v + \mu v,\\
\mathcal{F}_{u\mu}(0,\mu,\kappa)[v] &= v,\\
\mathcal{F}_{u\kappa}(0,\mu,\kappa)[v] &= 4\kappa^3\Delta^2 v + 2\kappa\Delta v,\\
\mathcal{H}_{uu}(0,\mu,\kappa)[v,w] &= \int_0^{\frac{4\pi}{\sqrt{3}}} [\kappa^4(v_X w_{XXX} + v_{XXX} w_X - v_{XX} w_{XX}) + 2\kappa^2 v_X w_X + (1+\mu) vw\\
&\qquad - 2\kappa^4 v_{XY} w_{XY} - 2\kappa^2 v_Y w_Y + \kappa^4 v_{YY} w_{YY}]\,\mathrm{d}Y,\\
\mathcal{H}_{uu\mu}(0,\mu,\kappa)[v,w] &= \int_0^{\frac{4\pi}{\sqrt{3}}} vw\,\mathrm{d}Y,\\
\mathcal{H}_{uu\kappa}(0,\mu,\kappa)[v,w] &= \int_0^{\frac{4\pi}{\sqrt{3}}} [4\kappa^3(v_X w_{XXX} + v_{XXX} w_X - v_{XX} w_{XX}) + 4\kappa v_X w_X\\
&\qquad - 8\kappa^3 v_{XY} w_{XY} - 4\kappa v_Y w_Y + 4\kappa^3 v_{YY} w_{YY}]\,\mathrm{d}Y.
\end{aligned}$$

The analysis reviewed in section 2.2 implies that we should pick $(\mu_0,\kappa_0) = (0,1)$, for which $\mathcal{F}_u(0,0,1)$ has a one-dimensional null space spanned by

$$\hat{u}(X,Y) = \cos(X) + \cos((X+\sqrt{3}Y)/2) + \cos((X-\sqrt{3}Y)/2).$$

We may set $\hat{v} := \hat{u}$ and use the L^2-inner product on the fundamental periodicity domain $[0, 4\pi] \times [0, 4\pi/\sqrt{3}]$ to define complements $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$ of $\mathbb{R}\hat{u}$ in $\mathcal{X}$ and $\mathcal{Y}$, as $\mathcal{F}_u(0, 0, 1)$ is symmetric with respect to this inner product.

Substituting these expressions, we find that $\mathcal{H}_{uu}(0, 0, 1)[\hat{u}, \hat{u}] = 0$. Furthermore, we find that the operator D defined in (3.10) is given by

$$D = \begin{pmatrix} P\mathcal{F}_u(0,0,1) & 0 & 0 \\ 0 & 1 & 2 \\ \star & 4\pi\sqrt{3} & -8\pi\sqrt{3} \end{pmatrix} \in L(\hat{\mathcal{X}} \times \mathbb{R}^2, \hat{\mathcal{Y}} \times \mathbb{R}^2).$$

In particular, this operator is invertible, and we obtain the following result, which is formulated again in the original spatial variables.

Proposition 3 ($\mathcal{L}$-periodic hexagons with $\mathcal{H} = 0$). *Fix any ν; then there exist an interval $I \subset \mathbb{R}$ and a unique branch $s \mapsto (u, \mu, \kappa)(s)$ of nontrivial $\mathcal{L}$-periodic hexagons of the Swift–Hohenberg equation (3.1) with zero constraint (3.8) and aspect ratio $\kappa(s)$, which are defined and smooth for $s \in I$. Moreover, $(u, \mu, \kappa)(0) = (0, 0, 1)$ and*

(3.11)
$$\Big\| [u(s)](x,y) - s\Big[\cos(\kappa(s)x) + \cos(\kappa(s)(x + \sqrt{3}y)/2) + \cos(\kappa(s)(x - \sqrt{3}y)/2)\Big] \Big\|_{\mathcal{C}^4} = \mathrm{O}(s^2)$$

as $s \to 0$.

For $\nu > 0$, (2.7) and (3.11) imply that $\mu(s) > 0$ for all sufficiently small $s > 0$, as claimed in Proposition 2.

4. Numerical algorithms.

In this section, we describe the numerical algorithms, and their implementation, that we used to compute regular hexagons, planar hexagon pulses, and localized hexagon patches.

Though we will rely on continuation methods for most of our computations, we shall also occasionally employ an initial value problem solver, which we discuss first. Afterward, we outline the computation of regular hexagons and the associated hexagon Maxwell curve. This information will guide us as to where we may find hexagon pulses and localized hexagon structures in the Swift–Hohenberg equation. We then move on to the computation of planar hexagon pulses that are periodic in the transverse direction. Last, we present the numerical methods for the computation of localized hexagon patches: these methods are designed to take advantage of the $\mathbb{D}_6$-symmetry of localized structures and allow us to compute localized structures that extend over large spatial regions. The computations of localized hexagon patches were also repeated with other methods to check the reliability of the numerical results.

The actual computations were carried out on FINCH, a dual core 2.7 GHz PowerPC G5 with 4GB of RAM, and PHOENIX, a server with two 3GHz dual core Xeon processors with 8GB of RAM, both running Mac OS 10.4.

4.1. The initial value problem solver.

To quickly find solutions of the Swift–Hohenberg equation, investigate the stability of patterns with respect to small symmetry-breaking perturbations, and confirm the solutions obtained from our other numerical solvers, we employ an initial value problem solver for the Swift–Hohenberg equation (1.1), which we shall now discuss briefly. First, we use the 2D Fourier transform to reduce the initial value problem

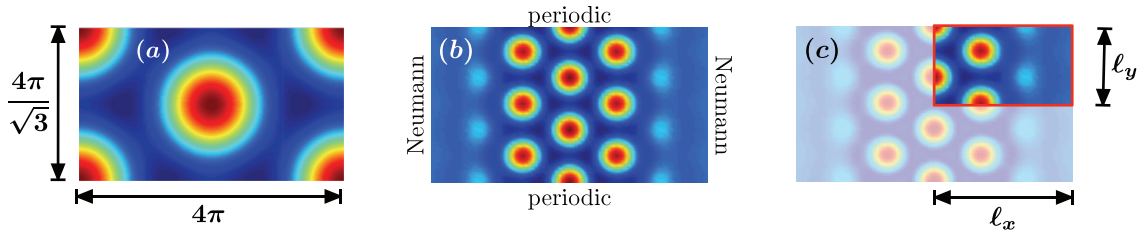


Figure 13. *Regular hexagons are computed with Neumann boundary conditions on the domain* $\Omega_{\rm hex}$ *shown in panel* (a). *Reflection-symmetric planar hexagon pulses are defined on the domain shown in panel* (b) *with Neumann conditions in the horizontal* x*-direction and periodic boundary conditions in the vertical* y*-direction. Panel* (c) *illustrates the computational domain* $\Omega = (0, \ell_x) \times (0, \ell_y)$ *with Neumann conditions.*

on a rectangular box with periodic boundary conditions to a system of ODEs. The resulting ODE system is truncated at a sufficiently large Fourier mode and solved in time using the first-order exponential time-stepping algorithm developed in [27]. We implemented this solver in MATLAB. Computations are done on domains of size 60×60 with 256×256 and 512×512 Fourier modes. Typical time steps are 0.01 and 0.001.

4.2. Regular hexagons and Maxwell curves. To find regular hexagons, we proceed initially as in section 3.3. It has been shown in [62, 61] that regular hexagons can be computed in a rectangular box with Neumann boundary conditions provided the ratio of the lengths of the sides of the rectangle is an integer multiple of $\sqrt{3}$. Thus, we introduce new independent coordinates $X = \kappa x$ and $Y = \kappa y$ and use the rescaled Swift–Hohenberg equation

$$\mathcal{F}(u; \mu, \nu, \kappa) = (1 + \kappa^2 \Delta)^2 u + \mu u - \nu u^2 + u^3 = 0, \qquad (X, Y) \in \Omega_{\rm hex} \tag{4.1}$$

on the computational domain $\Omega_{\rm hex} = (0, 4\pi) \times (0, 4\pi/\sqrt{3})$ with Neumann boundary conditions; see Figure 13(a).

For the computation of Maxwell curves, we add the constraints

$$\mathcal{H}(u; \mu, \nu, \kappa) = \int_0^{\frac{4\pi}{\sqrt{3}}} \left[-\frac{\kappa^4 (u_{XX})^2}{2} + \frac{(1+\mu)u^2}{2} - \frac{\nu u^3}{3} + \frac{u^4}{4} - \kappa^2 (u_Y)^2 + \frac{\kappa^4 (u_{YY})^2}{2} \right]_{X=0} \mathrm{d}Y = 0, \tag{4.2}$$

which ensures that the first integral $\mathcal{H}$ vanishes, and

$$\mathcal{E}(u; \mu, \nu, \kappa) = \int_{\Omega_{\rm hex}} \left[\frac{[(1 + \kappa^2 \Delta)u]^2}{2} + \frac{\mu u^2}{2} - \frac{\nu u^3}{3} + \frac{u^4}{4} \right] \mathrm{d}X \, \mathrm{d}Y = 0, \tag{4.3}$$

which enforces zero energy. Note that several terms in the original expression (3.2) for $\mathcal{H}$ vanish on account of the Neumann conditions $u_X(0, Y) = u_{XXX}(0, Y) = 0$. The choice of our computational domain means that we accurately compute the energy of two full hexagons in (4.3); see Figure 13(a).

We expect that the equation $\mathcal{F}(u; \mu, \nu, \kappa) = 0$ has a locally unique regular zero u for each fixed (μ, ν, κ) in appropriate regions[5] in parameter space. We can also use (4.1) together with

[5]For instance, for μ close to zero, due to the results in section 2.2; see also Proposition 3.

the constraints (4.2)–(4.3) in a numerical continuation framework where we expect to find a curve of solutions (u, μ, ν, κ) that depend on an arclength parameter s. Since the resulting Maxwell curves do not exhibit any folds in the parameter μ, we can, in fact, compute this curve by stepping in the parameter μ.

For the actual computations, we evaluate derivatives using spectral differentiation as in [87, sect. 3] and compute the integrals in the constraints (4.2)–(4.3) using the periodic trapezoid rule [87, sect. 12]. We now briefly outline how this is done in our context. We choose the mesh $X_i = 4\pi i/N$ and $Y_j = 4\pi j/N\sqrt{3}$ for $i, j = 1, \dots N$ and write $u_{ij} = u(X_i, Y_j)$ on this mesh. A convenient way of evaluating first-order derivatives and the Laplacian is via Kronecker products: if A is an $m \times n$ matrix and B is a $p \times q$ matrix, then the Kronecker product $A \otimes B$ is an $mp \times nq$ matrix which consists of $m \times n$ blocks, where each block is a $p \times q$ matrix. The (i, j)th block is given by $a_{ij}B$. Introducing the step size $h = 2\pi/N$, the spectral differentiation matrices for functions of one variable are given by

$$D_N = \text{toeplitz}\left[0, \left(\frac{(-1)^j}{2\tan(jh/2)}\right)_{j=1,\dots,N-1}\right]$$

for the first derivative and by

$$D_N^{(2)} = \text{toeplitz}\left[-\frac{\pi^2}{3h^2} - \frac{1}{6}, \left(-\frac{(-1)^j}{2\sin^2(jh/2)}\right)_{j=1,\dots,N-1}\right]$$

for the second derivative [87, sect. 3], where toeplitz$[v]$ denotes the symmetric Toeplitz matrix (a matrix whose entries are constant along each diagonal) formed by the row vector $v \in \mathbb{R}^N$. Using Kronecker products, we can now set up the $N^2 \times N^2$ differentiation matrices for $u(X, Y)$ which are given by

$$D_{X,N} = I \otimes \left(\frac{1}{4}\right) D_N, \quad D_{Y,N} = \left(\frac{\sqrt{3}}{4}\right) D_N \otimes I, \quad \Delta_N = I \otimes \left(\frac{1}{4}\right)^2 D_N^{(2)} + \left(\frac{\sqrt{3}}{4}\right)^2 D_N^{(2)} \otimes I,$$

corresponding, respectively, to the first-order derivatives ∂_X and ∂_Y and to the Laplacian Δ.

The above procedure results in a finite-dimensional system. We solve this system in MATLAB using the nonlinear Newton trust-region solver FSOLVE. For (μ, ν) close to zero, we choose

$$u(X, Y) = A[\cos(X) + \cos((X + \sqrt{3}Y)/2) + \cos((X - \sqrt{3}Y)/2)]$$

as an initial guess, where a good approximation for the amplitude A can be obtained from normal-form theory by solving (2.7). MATLAB's Newton trust-region solver has the advantage of often achieving global convergence even when starting from poor initial guesses. We have frequently obtained better convergence by solving initially only the Swift–Hohenberg equation, without the integral constraints (4.2)–(4.3), with $\kappa = 1$ and ν fixed. Afterward, using this solution as initial data, we solve the Swift–Hohenberg equation together with one or both of the integral constraints by including one or two of the parameters κ and ν as unknowns. Once we have a solution to (4.1)–(4.3), we continue it in μ by stepping in the parameter μ and solving (4.1)–(4.3) for (u, ν, κ) for each fixed μ. We use $18 \cdot 18 = 324$ interpolation points in the box $\Omega_{\text{hex}} = (0, 4\pi) \times (0, \frac{4\pi}{\sqrt{3}})$ and compute solutions within an absolute tolerance of 10^{-4}. The entire Maxwell curve was computed in a couple of minutes.

4.3. Numerical continuation of planar hexagon pulses. In this section, we discuss the computation of stationary planar hexagon pulses with Bravais–Miller indices $\langle 10\rangle$ and $\langle 11\rangle$ such as those shown in Figure 3. The interfaces in these solutions are vertical, and the overall patterns are periodic in the transverse y-direction and reflection symmetric.

Hence, we will focus on computing stationary solutions $u(x,y)$ of the planar Swift–Hohenberg equation that are periodic in the y-direction and are symmetric under reflections in x and y so that $u(-x,y) = u(x,y) = u(x,-y)$ for all (x,y). These solutions therefore satisfy

$$(1+\Delta)^2 u + \mu u - \nu u^2 + u^3 = 0, \qquad (x,y) \in \Omega, \tag{4.4}$$

on $\Omega = (0,\ell_x)\times(0,\ell_y)$ with Neumann boundary conditions

$$u_x|_{\{x=0,\ell_x\}} = u_{xxx}|_{\{x=0,\ell_x\}} = u_y|_{\{y=0,\ell_y\}} = u_{yyy}|_{\{y=0,\ell_y\}} = 0 \tag{4.5}$$

on $\partial\Omega$; see Figure 13. We need to choose ℓ_x large enough to avoid boundary effects (we used $\ell_x = 50$), while ℓ_y is chosen in such a way as to accommodate hexagon interfaces with $\langle 10\rangle$ or $\langle 11\rangle$ orientation: we pick $\ell_y = 4\pi n/\sqrt{3}$ for interfaces with index $\langle 10\rangle$ and $\ell_y = 2\pi n$ for interfaces with index $\langle 11\rangle$. The choice of $n \in \mathbb{N}$ allows us to compute patterns for several wavelengths in the vertical y-direction.

Fixing the length ℓ_y of the domain in the y-direction may frustrate the hexagons: since regular hexagons can no longer choose their wavenumber freely to satisfy the constraint $\mathcal{H} = 0$, the patterned state will typically consist of frustrated hexagons that have a fixed period ℓ_y/N in the y-variable for some integer N to accommodate the fixed length in the y-direction, while their wavelength in the x-direction adjusts itself to satisfy the constraint $\mathcal{H} = 0$; the resulting frustrated hexagons are therefore slightly compressed or elongated in the x-direction and no longer $\mathbb{D}_6$-symmetric. We could add the constraint (4.2) and allow ℓ_y to vary so that regular hexagons always fit into the domain. Since we have found that the regular hexagons for which $\mathcal{H} = 0$ have wavenumbers κ very close to $\kappa = 1$, we believe that the effect of fixing ℓ_y on the selected patterns is negligible. However, the snaking limits of the planar hexagon pulses may coincide better with the snaking limits of the localized hexagon patches had we elected to allow ℓ_y to vary.

To solve (4.4)–(4.5) numerically, we used a 13-point finite difference stencil for the spatial discretization. We implemented the resulting system in the continuation framework PARACONT [7], a module built on top of the continuation module LOCA of the package Trilinos, which is written and maintained by Sandia Laboratories [43]. Since Trilinos does not currently offer a direct solver for LOCA that works on parallel processors, we employed a multilevel preconditioner on a coarse level so that an exact sparse linear solve is done. The computations were carried out on the domain $\Omega = (0,\ell_x)\times(0,\ell_y)$ with $\ell_x = 50$. We used $\ell_y = 20\pi/\sqrt{3}$ for $\langle 10\rangle$ pulses and $\ell_y = 10\pi$ for $\langle 11\rangle$ pulses and worked with both 128×256 and 256×256 mesh points for both computations.

We remark that localized pulses on long cylinders $(0,\ell_x)\times S^1$ have been computed previously in the von Karman–Donnell equations, a coupled system of elliptic PDEs that describe equilibria of axially compressed cylindrical shells [46, 59]. The approach adopted there was to carry out a Fourier decomposition in the angular direction leading to a large system of ODEs that were solved with the boundary-value solver AUTO97.

4.4. Numerical continuation of localized hexagon patches. We now turn to the computation and continuation of planar localized hexagon patches such as the ones presented in Figures 1(a) and 4. We focus on the computation of patterns with $\mathbb{D}_6$-symmetry. Since we found that sparse Cartesian meshes give a preference to $\mathbb{D}_4$-symmetric square patterns, we discretize the planar Swift–Hohenberg equation in polar coordinates. In particular, we found that a spectral Fourier discretization in the angular coordinate combined with an adaptive collocation mesh in the radial coordinate appears to be a very efficient method for computing localized hexagon patches.

In the following, we will outline our approach for computing localized patterns with an arbitrary $\mathbb{D}_{2k}$-symmetry.[6] Restricting ourselves to solutions with $\mathbb{D}_{2k}$-symmetry allows us to compute them on the first quadrant $\Omega = \{x, y > 0\}$ with Neumann boundary conditions, which is advantageous for various reasons. First, it factors out, in a natural fashion and without the need for introducing additional constraints, the continuous translation and rotation symmetries in $\mathbb{E}(2)$ of the Swift–Hohenberg equation, whose presence would otherwise yield a singular Jacobian, which is problematic for Newton solvers. Second, computing solely on the first quadrant greatly reduces the size of the discretized system. Third, as already mentioned, we can center localized solutions at the origin and compute efficiently in polar coordinates. The main disadvantage of computing on the first quadrant is that temporal stability cannot be deduced and bifurcations to $\mathbb{D}_{2k+1}$ patterns cannot be detected. Overall, we believe that the advantages outweigh the disadvantage of potentially failing to detect instabilities during continuation as these can often be identified a posteriori by direct numerical simulations.

We therefore consider the stationary planar Swift–Hohenberg equation

$$(1 + \Delta_{r,\theta})^2 u + \mu u - \nu u^2 + u^3 = 0 \tag{4.6}$$

written in polar coordinates $(r, \theta) \in (0, \infty) \times [0, 2\pi)$, where

$$\Delta_{r,\theta} u = u_{rr} + \frac{u_r}{r} + \frac{u_{\theta\theta}}{r^2}.$$

Polar coordinates are singular at $r = 0$, and we need to find appropriate boundary conditions at the origin to remove this singularity. To do this, we follow [84]. Assuming that u is a sufficiently localized solution, we multiply (4.6) by another localized function v and subsequently integrate over (r, θ) to arrive at the weak formulation

$$\begin{aligned} &\int_0^{2\pi}\int_0^{\infty} \left[\Delta u \Delta v - 2\nabla u \nabla v + (1+\mu)uv - \nu u^2 v + u^3 v\right] r \,\mathrm{d}r \,\mathrm{d}\theta \\ &= \int_0^{2\pi}\int_0^{\infty} \left[\left(\frac{(ru_r)_r}{r} + \frac{u_{\theta\theta}}{r^2}\right)\left(\frac{(rv_r)_r}{r} + \frac{v_{\theta\theta}}{r^2}\right) - 2\left(u_r v_r + \frac{u_\theta v_\theta}{r^2}\right)\right. \\ &\qquad \left. + (1+\mu)uv - \nu u^2 v + u^3 v\right] r \,\mathrm{d}r \,\mathrm{d}\theta = 0 \end{aligned} \tag{4.7}$$

of (4.6). The boundary conditions at $r = 0$ which make the bilinear form (4.7) meaningful are

$$u_r|_{(0,\theta)} = (ru_r)_r|_{(0,\theta)} = u_\theta|_{(0,\theta)} = u_{\theta r}|_{(0,\theta)} = u_{\theta\theta}|_{(0,\theta)} = (u_{\theta\theta})_r|_{(0,\theta)} = 0 \qquad \forall \theta \in [0, 2\pi).$$

[6] This method can be extended to localized $\mathbb{D}_{2k+1}$-symmetric patterns, but we do not go into the details here.

Since $(ru_r)_r = ru_{rr} + u_r$, the conditions above reduce to

$$u_r(0,\theta) = u_{\theta\theta}(0,\theta) = u_{\theta\theta r}(0,\theta) = 0 \qquad \forall \theta \in [0, 2\pi). \tag{4.8}$$

We now expand $u(r,\theta)$ in a Fourier series which we truncate at order $N \in \mathbb{N}$ to get a finite-dimensional system. Thus, we set

$$u(r,\theta) = \sum_{n=-N}^{N} a_n(r) \mathrm{e}^{-\mathrm{i}n\theta}, \tag{4.9}$$

where N is the truncation order, and $a_n(r)$ is complex-valued for each n. The Laplacian becomes

$$\Delta_{r,\theta} u = \sum_{n=-N}^{N} \left[\partial_r^2 a_n + \frac{\partial_r a_n}{r} - \frac{n^2 a_n}{r^2} \right] \mathrm{e}^{-\mathrm{i}n\theta}.$$

Substituting these expressions, we find that the truncated planar Swift–Hohenberg equation (4.6) can be written as

$$\begin{aligned} &\partial_r^4 a_n + \frac{2\partial_r^3 a_n}{r} - \frac{\partial_r^2 a_n}{r^2} + \frac{\partial_r a_n}{r^3} - \frac{2n^2 \partial_r^2 a_n}{r^2} + \frac{2n^2 \partial_r a_n}{r^3} - \frac{4n^2 a_n}{r^4} + \frac{n^4 a_n}{r^4} \\ &+ 2\left(\partial_r^2 a_n + \frac{\partial_r a_n}{r} - \frac{n^2 a_n}{r^2} \right) + (1+\mu) a_n - \nu \sum_{p+q=n} a_p a_q + \sum_{p+q+s=n} a_p a_q a_s = 0, \end{aligned} \tag{4.10}$$

while the boundary conditions (4.8) at the origin reduce to

$$\begin{aligned} \partial_r a_0\big|_{r=0} = \partial_r^3 a_0\big|_{r=0} = 0, \\ a_n\big|_{r=0} = \partial_r a_n\big|_{r=0} = 0 \qquad \forall n \neq 0. \end{aligned}$$

Solutions $u(r,\theta)$ with $\mathbb{D}_{2k}$-symmetry are invariant under the reflection $\theta \mapsto -\theta$ and the rotation $\theta \mapsto \theta - \pi$ so that

$$u(r,\theta) = u(r,-\theta) \quad \text{and} \quad u(r,\theta) = u(r,\theta-\pi)$$

for all (r,θ). These identities imply that

$$a_n = a_{-n} \quad \text{and} \quad a_n = (-1)^n a_n \quad \forall n,$$

for the coefficients $a_n(r)$ of the Fourier representation (4.9) of u. The first of these two conditions implies that we need to compute the coefficients a_n only for $n \geq 0$, while the second condition implies that all odd Fourier coefficients a_{2n+1} must vanish identically. The two summations in (4.10) can then be simplified by noting that

$$\begin{aligned} \sum_{p+q=n} a_p a_q &= \sum_{p=-N}^{N} a_{|p|} a_{|n-p|}, \\ \sum_{p+q+s=n} a_p a_q a_s &= \sum_{p=-N}^{N} \sum_{q=-N}^{N} a_{|p|} a_{|q|} a_{|n-p-q|}. \end{aligned}$$

So far, we have considered solutions with an arbitrary $\mathbb{D}_{2k}$-symmetry. From now on, we restrict ourselves to solutions with $\mathbb{D}_6$-symmetry, such as hexagons. Such solutions are in addition invariant under rotations by an angle of $\pi/3$, which is equivalent to requiring that

$$a_n = \mathrm{e}^{\mathrm{i}n\pi/3} a_n \qquad \forall n.$$

The only nonzero modes that can satisfy this constraint are those for which $n = 6m$ for some $m \in \mathbb{Z}$. Hence, for the purpose of computing localized $\mathbb{D}_6$-symmetric solutions, we need to consider only the Fourier modes $a_{6m}(r)$: note that the space spanned by these modes is the fixed-point space under the $\mathbb{D}_6$-action and therefore invariant under the evolution of (4.10). Thus, we set

$$A_m(r) := a_{6m}(r), \qquad |m| \le M,$$

with $N = 6M$ and write (4.10) as the first-order system

$$\begin{aligned}
\partial_r A_m &= B_m,\\
\partial_r B_m &= C_m,\\
\partial_r C_m &= D_m,\\
\partial_r D_m &= -\frac{2D_m}{r} + \frac{C_m}{r^2} - \frac{B_m}{r^3} + \frac{2(6m)^2 C_m}{r^2} - \frac{2(6m)^2 B_m}{r^3}\\
&\quad + \frac{4(6m)^2 A_m}{r^4} - \frac{(6m)^4 A_m}{r^4}\\
&\quad - 2\left(C_m + \frac{B_m}{r} - \frac{(6m)^2 A_m}{r^2}\right) - (1+\mu) A_m\\
&\quad + \nu \sum_{p=-M}^{M} A_{|p|} A_{|m-p|} - \sum_{p=-M}^{M} \sum_{q=-M}^{M} A_{|p|} A_{|q|} A_{|m-p-q|}
\end{aligned} \tag{4.11}$$

with boundary conditions

$$\begin{aligned}
\partial_r A_0\big|_{r=0} = D_0\big|_{r=0} &= 0,\\
A_m\big|_{r=0} = C_m\big|_{r=0} &= 0 \qquad \forall m \neq 0,\\
A_m\big|_{r=R} = B_m\big|_{r=R} &= 0 \qquad \forall m,
\end{aligned} \tag{4.12}$$

where R indicates the radial domain $(0, R)$ on which we compute solutions.

Both the domain truncation parameter R and the Fourier truncation parameter M must be set to suitably large values to ensure that the Neumann boundary conditions at $r = R$ do not influence the localized patterns and to make sure that the Fourier modes can resolve the angular dependence of the computed patterns. Close to $(\mu, \nu) = 0$, the localized patterns are small in amplitude but are also well spread out: this requires both R and M to be large. Specifically, we expect that the number of Fourier modes required to resolve a localized hexagon patch corresponds roughly to the number of hexagon rings one wishes to interpolate: if we wish to compute a hexagon patch of radius R, then it will have approximately R hexagons, or more, located on its interface. To resolve these R hexagons, we need at least $M \approx R$ Fourier modes in the angular variable, and this is indeed what we find in our numerical computations.

A major problem is acquiring good initial data for continuation. We use two different methods for preparing good initial guesses for the boundary-value problem (4.11)–(4.12). The first method is to use an initial value problem solver and find a stable stationary localized hexagon patch by direct numerical simulations: as we can first compute the Maxwell curve, we do know where to look for stable hexagon patches. Our experience was, however, that convergence to stationary solutions tends to be very slow. Instead, we have found that it is better to discretize the boundary-value problem (4.11)–(4.12) in the radial variable using a Chebyshev decomposition with an infinite mapping in the radial coordinate r that bunches the collocation points near the origin, as in [57]. We then use MATLAB's Newton trust-region solver FSOLVE starting from initial data of the form

$$u(x,y) = a\,\mathrm{sech}(b\sqrt{x^2+y^2})\left[\cos(x)+\cos((x+\sqrt{3}y)/2)+\cos((x-\sqrt{3}y)/2)\right],$$

where a is chosen to be close to or greater than the maximum height of a single hexagon cell and b determines the size of the localized hexagon patch. This procedure gives excellent convergence results and allows us to obtain accurate starting data for subsequent continuation in parameters.

As already mentioned, initial data for hexagon patches were computed using MATLAB's Newton trust-region solver. To continue these solutions, we implemented the boundary-value problem (4.11)–(4.12) in AUTO-07P [34]. Within AUTO-07P, we computed the L^2-norm of solutions by appending an additional equation together with another parameter that corresponds to the value of the L^2-norm of a solution; we exclude this additional equation from the calculation of the pseudoarclength. The Jacobian of the right-hand side of (4.11) was supplied in analytic form to speed up the computation. We use standard AUTO-07P tolerances and choose the collocation mesh size NTST between 200 and 400. The radial domain truncation parameter R was set to $R = 80, 100, 200$, while the number of angular Fourier modes was taken to be $M = 20, 30, 40$. The computation of the full hexagon snake took up to one day on PHOENIX.

Hexagon patches have been computed previously in [63] in the context of nonlinear optics. In fact, the authors there computed and continued several different localized states and traced out the beginning of the snaking diagram. They discretized the underlying PDE on an equidistant mesh, used the fast Fourier transform for evaluating the spatial derivatives, and solved the resulting large system of algebraic equations using Newton's method. This method tends to be computationally expensive (their computations required the use of 300 servers with 500MHz processors) since the mesh requires a large number of modes even in the tails of the localized pattern.

5. Localized hexagon and rhomboid patches: Numerical results.

In this section, we present our numerical results. We emphasize that the computational domains of numerical solutions are typically much larger than the domains visible in the figures presented below as we frequently cropped images to highlight the features of the localized patterns.

5.1. Regular hexagons and Maxwell curves.

We first compute regular hexagons of the planar Swift–Hohenberg equation that satisfy $\mathcal{H} = 0$ as solutions to (4.1)–(4.2) with $\nu = 1.6$. As discussed in section 3, only these hexagons can be connected to the trivial state by a

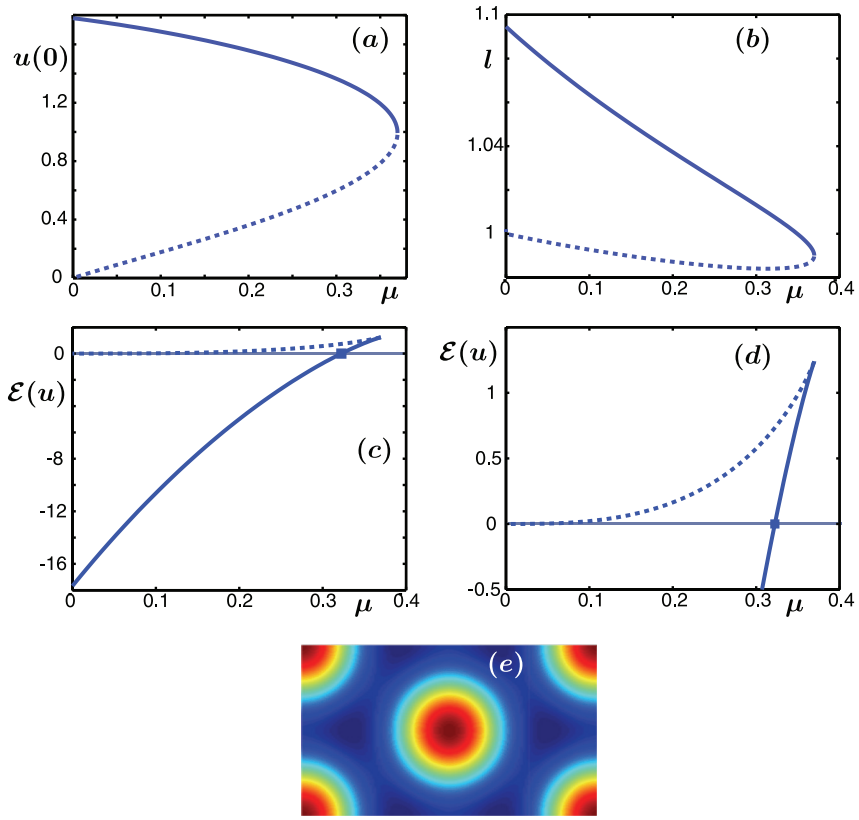


Figure 14. *We computed regular hexagons with $\mathcal{H} = 0$ as solutions of (4.1)–(4.2) with $\nu = 1.6$. Stable hexagons are plotted in solid lines, and unstable ones are plotted in dashed lines. We plot the amplitude $u(0,0)$ at the origin in panel* (a)*, the wavelength $l = 1/\kappa$ of the hexagons in* (b)*, and the energy $\mathcal{E}(u)$ along the branch in* (c)–(d)*. The Maxwell point $\mathcal{E}(u) = 0$ occurs at $\mu = \mu_{\rm M} = 0.3224$ on the stable branch. Compared with the trivial state, stable regular hexagons have less energy to the left and higher energy to the right of the Maxwell point. Panel* (e) *contains a color plot of the regular hexagons $u(x, y)$ on the domain $[0, 4\pi l] \times [0, 4\pi l/\sqrt{3}]$ at the Maxwell point.*

stationary planar front. The bifurcation diagram shown in Figure 14(a) is qualitatively similar to that found in section 2.2 in the normal-form analysis for $|\nu| \ll 1$: regular hexagons bifurcate off the trivial solution at $\mu = 0$ and are initially unstable but regain stability in a saddle-node bifurcation. Figure 14(b) contains a plot of the wavelength $l := 1/\kappa$ of the hexagons with $\mathcal{H} = 0$ as μ is varied. The energy $\mathcal{E}(u)$ of these hexagons, computed over two hexagons, is shown in Figure 14(c)–(d) as a function of μ. In particular, the Maxwell point, where $\mathcal{E} = 0$, occurs at $\mu = \mu_{\rm M} = 0.3224$, and we plot the computed hexagon at the Maxwell point in Figure 14(e). We remark that the dependence of the wavelength and the energy on the parameter μ is qualitatively similar for hexagons and 1D rolls; see Figure 5(iii)–(iv) or [19, Figure 2] for results on rolls.

Next, we solve (4.1)–(4.3) for (u, μ, ν, κ), which gives the hexagon Maxwell curve along which hexagons with zero energy $\mathcal{E} = 0$ and zero first integral $\mathcal{H} = 0$ exist. As discussed in section 3, this curve serves as a guide to where hexagon fronts and pulses as well as fully

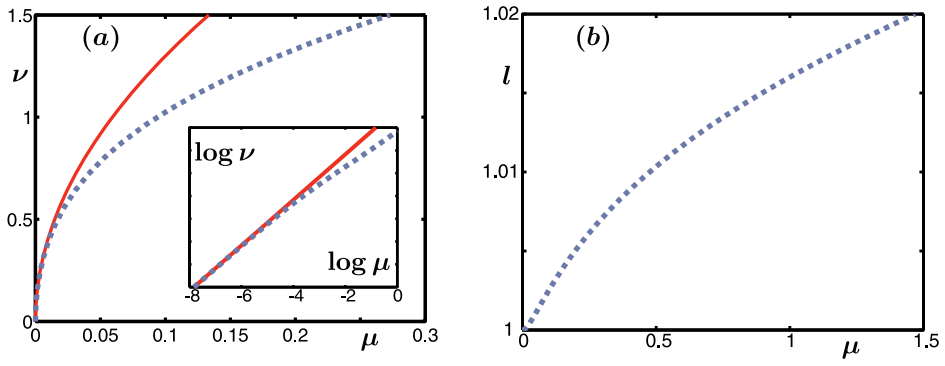


Figure 15. *In panel* (a)*, we plot the Maxwell curve along which the regular hexagons of* (4.1) *satisfy the constraints* $\mathcal{H} = 0$ *and* $\mathcal{E} = 0$ *from* (4.2) *and* (4.3)*, respectively: The numerical result is dashed, while the analytic prediction* $\mu = \mu_{\mathrm{M}} = 8\nu^2/135$*, valid in the limit* $\nu \to 0$*, is plotted as a solid line. Panel* (b) *gives the wavelength* $l = 1/\kappa$ *of the corresponding hexagons.*

localized hexagon patches can be expected. In Figure 15, we plot both the hexagon Maxwell curve in (μ, ν)-parameter space and the dependence of the wavelength $l = 1/\kappa$ of the selected hexagons on the parameter μ. As predicted by the theory outlined in section 2.3, the Maxwell curve emerges from the codimension-two point $(\mu, \nu) = (0, 0)$ and agrees well with the analytic prediction $\mu_{\mathrm{M}} = 8\nu^2/135$ given in (2.12). The wavelength of the hexagons increases along the Maxwell curve. We remark that, for $0 \leq \mu \leq 0.6$, the Maxwell curve agrees well with the curve obtained from setting $\mathcal{E} = 0$ and allowing arbitrary values for $\mathcal{H}$ while keeping $\kappa = 1$ fixed (we do not show a comparison of these curves though).

As shown in Figure 14(d), stable regular hexagons have less energy than the trivial state to the left of the Maxwell curve and higher energy to its right. Thus, we expect that hexagons will invade the trivial state for μ sufficiently far to the left of the Maxwell curve, while the trivial state will invade hexagons for μ sufficiently far to its right.

5.2. Planar hexagon pulses: Bifurcation diagram for $\nu = 1.6$. Throughout this section, we fix $\nu = 1.6$ and recall that the hexagon Maxwell point is given by $\mu = 0.3224$.

We compute planar hexagon pulses of the Swift–Hohenberg equation with Bravais–Miller indices $\langle 10 \rangle$ and $\langle 11 \rangle$. Example plots of these solutions for $\mu = 0.31$ are given in Figure 3. Since we fixed the computational domain in these computations, all hexagons are slightly compressed by the same fraction in the y-direction instead of being fully $\mathbb{D}_6$-symmetric: each vertical slice $u(x, \cdot)$ of the planar hexagon pulse $u(x, y)$ must satisfy $\mathcal{H} = 0$ for each x, and since regular hexagons cannot adjust their wavelength in the y-direction to accommodate this condition due to the fixed domain dimension, the selected patterns are slightly frustrated hexagons. The frustrated hexagons are still periodic in both the x- and the y-directions, but their wavelengths in the x- and y-directions are no longer in a $\sqrt{3} : 1$ ratio as those of regular hexagons.

Upon varying μ, we find that planar $\langle 10 \rangle$ hexagon pulses snake as shown in Figure 16. Upon passing through a pair of fold bifurcations, the pulses acquire an additional full column of hexagons and thereby widen in the horizontal x-direction. As outlined in section 3.2, we can consider the planar Swift–Hohenberg equation as a dynamical system in an unbounded

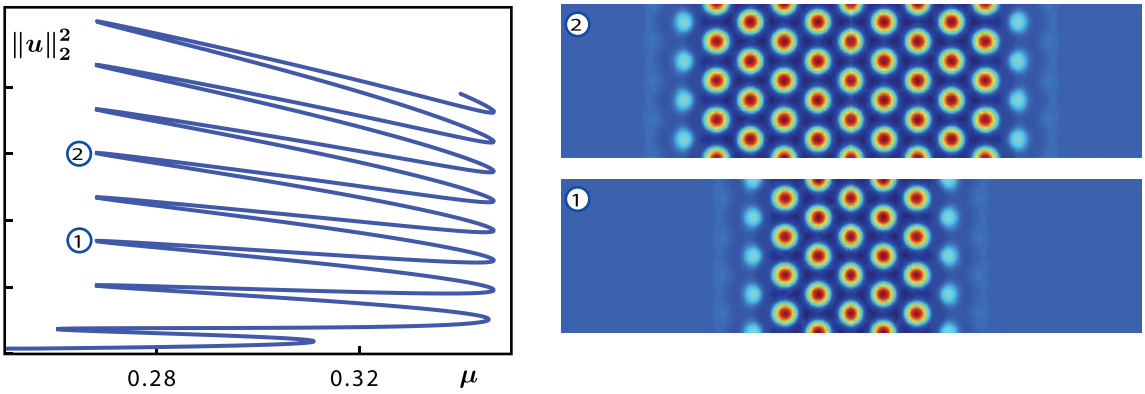


Figure 16. *The left panel contains the bifurcation diagram of planar* $\langle 10 \rangle$ *hexagon pulses, while two selected profiles at the labeled parameter values are shown in the right two panels.*

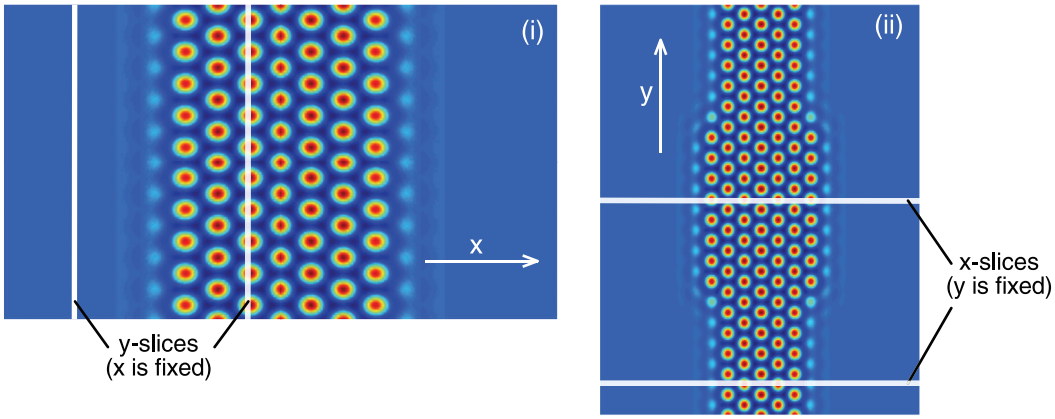


Figure 17. *Panel* (i) *contains a color plot of a planar* $\langle 10 \rangle$ *hexagon pulse. We view* x *as the evolution variable which is used to evolve* y*-slices forward and backward: The left* y*-slice corresponds to an equilibrium of the resulting spatial dynamical system in* x *(the profile in the* y*-slice does not change when* x *is varied nearby), while the* y*-slice in the hexagon region corresponds to a periodic orbit (the profile changes periodically in* x *when the* y*-slice is moved to the left and right). Thus, we can interpret a planar* $\langle 10 \rangle$ *hexagon pulse as a homoclinic orbit that passes close to a periodic orbit which is formed of hexagons, and we therefore expect snaking on account of the results in sections 2.1 and 3.2. Panel* (ii) *contains a color plot of an almost-planar hexagon pulse. Here, we view* y *as the evolution variable which propagates* x*-slices up- and downward. The two indicated* x*-slices correspond to two different periodic orbits: Their profiles change periodically when* y *is varied, but the horizontal extent of the hexagon regions is different for the two profiles. We can therefore interpret an almost-planar hexagon pulse as a homoclinic orbit in the* y*-dynamics which connects the periodic orbit at the top and bottom to itself and which passes near a second periodic orbit. The results in sections 2.1 and 3.2 imply again that snaking should occur.*

direction provided we restrict ourselves to a bounded cross-section in the remaining spatial variable. To explain the snaking of planar $\langle 10 \rangle$ pulses, we treat the x-variable as our time-like variable and restrict y to a bounded interval with Neumann boundary conditions. As can be seen from Figure 17(i), a planar hexagon pulse corresponds to a homoclinic orbit of the trivial state $U = 0$ which passes close to a periodic orbit in the x-dynamics that is formed of

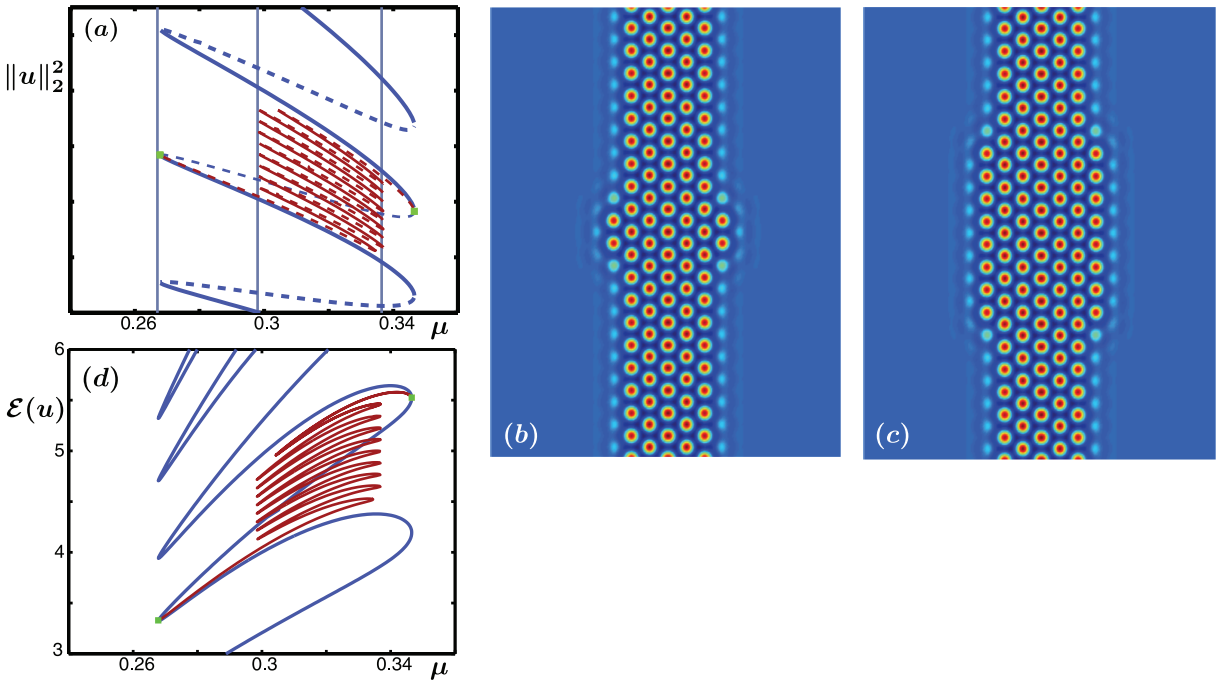


Figure 18. *In panel* (a)*, the bifurcation diagrams of planar* $\langle 10 \rangle$ *hexagon pulses (blue) and the bifurcating almost-planar pulses (red) are shown, where stable pulses are indicated by solid lines and unstable ones by dashed lines. Panels* (b)–(c) *are color plots of two different almost-planar hexagon pulses for* $\mu = 0.3$ *along the smaller red snaking diagram. In panel* (d)*, we plot the energy* $\mathcal{E}(u)$*, computed as an integral over the entire computational domain, along the branches.*

hexagons. Thus, we are in the situation discussed in sections 2.1 and 3.2 and expect snaking to set in [10]. As already alluded to in section 3.2, the scenario we just described persists if we change the height of the y-interval: the selected hexagons will become slightly frustrated, and, accordingly, the Maxwell point may change slightly, but the resulting planar hexagon pulses will continue to snake.

As shown in Figure 18(a), planar $\langle 10 \rangle$ pulses undergo additional pitchfork bifurcations near each fold. Figure 18(b)–(c) shows that the patterns bifurcating at the pitchfork bifurcations are almost-planar hexagon pulses: at onset, either one or two hexagon cells appear in new columns to the left and right at the center of the $\langle 10 \rangle$ pulse. As we move along the bifurcating branch, the almost-planar hexagon pulses begin to snake, and, at each fold, additional pairs of hexagon cells are added symmetrically above and below the already added hexagon cells until the entire column is filled. At this point, the branch of almost-planar pulses terminates in a second pitchfork bifurcation at the planar $\langle 10 \rangle$ hexagon pulses. Almost-planar hexagon pulses undergo only a finite number of folds due to the finite height of the computational domain. Similar to the case of planar pulses, the left and right fold bifurcations of almost-planar hexagon pulses line up. To explain the snaking of almost-planar pulses, we consider the y-variable as our time-like variable and restrict x to a large bounded interval with Neumann boundary conditions; see Figure 17(ii). In this spatial-dynamics interpretation, an almost-planar hexagon pulse corresponds to a homoclinic orbit

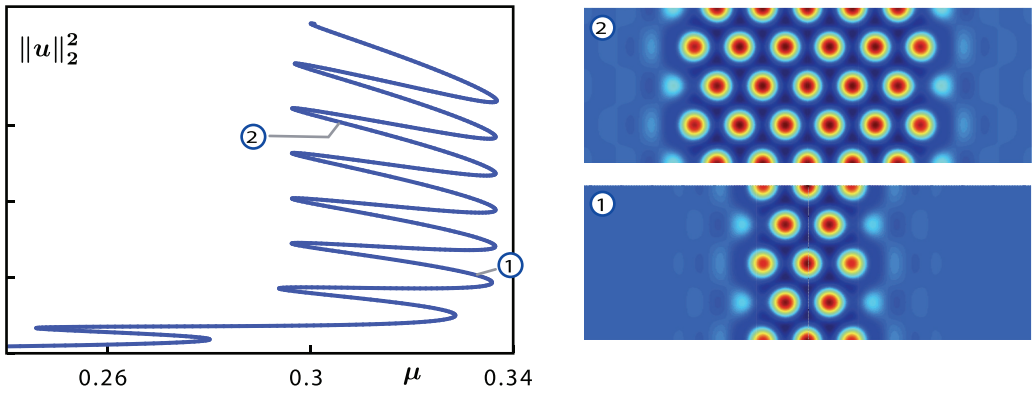


Figure 19. *The left panel contains the bifurcation diagram of planar* $\langle 11 \rangle$ *hexagon pulses. Two selected profiles are shown on the right.*

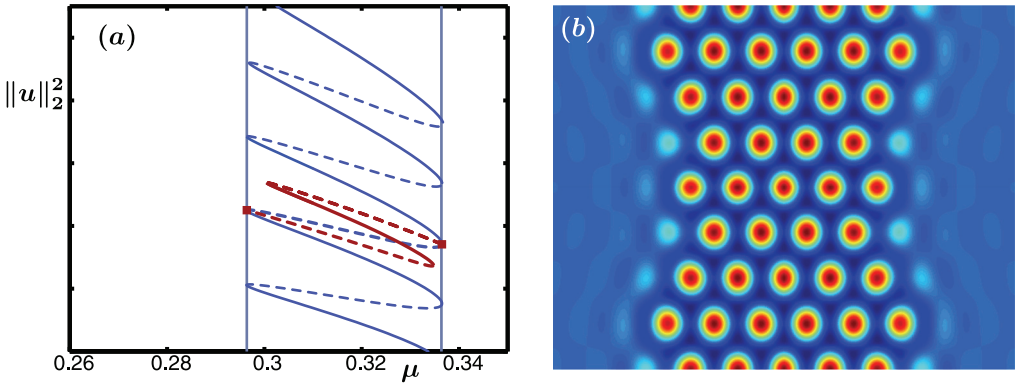


Figure 20. *The bifurcation diagrams of planar* $\langle 11 \rangle$ *hexagon pulses (blue) and the bifurcating almost-planar pulses (red) are plotted in panel* (a). *Panel* (b) *is a color plot of an almost planar* $\langle 11 \rangle$ *pulse for* $\mu = 0.309$.

of a periodic orbit that passes close to a second periodic orbit as y increases. Each periodic orbit in the y-dynamics consists of a localized hexagon pulse in the x-variable with a different number of hexagons in its center. Homoclinic orbits between periodic orbits will snake in the same fashion as the homoclinic orbits between equilibria that we discussed in section 2.1.

We now turn to a discussion of planar $\langle 11 \rangle$ hexagon pulses which also snake (see Figure 19) and exhibit pitchfork bifurcations to almost-planar $\langle 11 \rangle$ pulses, as shown in Figure 20. The almost-planar $\langle 11 \rangle$ hexagon pulses undergo only two folds since the computational domain allows only eight hexagons in the y-direction: there would be more folds if the height ℓ_y of the computational domain used in Figure 13 were larger.

Figure 21 contains the bifurcation diagrams of both planar $\langle 10 \rangle$ and $\langle 11 \rangle$ hexagon pulses. This figure shows that the orientation of the hexagon pulse has a significant effect on the width of the snaking region. The different vertical lines along which the folds line up will play an important role later when we discuss fully localized hexagon patches. We believe that there are many other hexagon pulses with orientations different from $\langle 10 \rangle$ and $\langle 11 \rangle$. The other

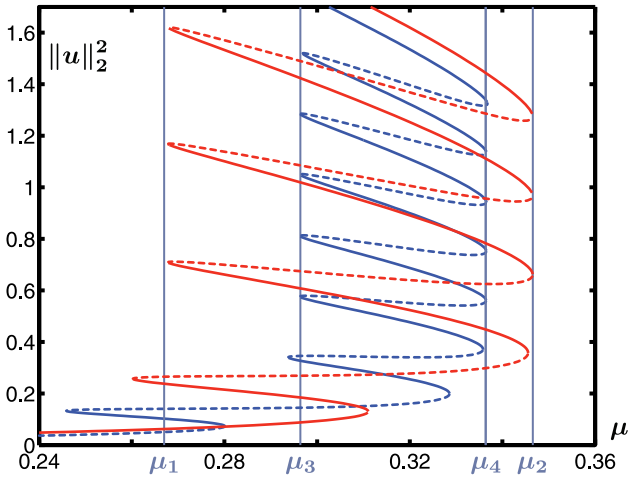


Figure 21. *The bifurcation diagrams of planar hexagon pulses with Bravais–Miller indices* $\langle 10\rangle$ *(in red) and* $\langle 11\rangle$ *(in blue) are shown. The* $\langle 10\rangle$ *pulse snakes between the limits* $\mu_1 = 0.267$ *and* $\mu_2 = 0.3454$*, while the* $\langle 11\rangle$ *pulse snakes between* $\mu_3 = 0.2964$ *and* $\mu_4 = 0.3364$*. The stability of the branches alternates between unstable (dashed) and stable (solid) at each fold.*

planar hexagon pulses will have larger Bravais–Miller indices, and we expect heuristically that these interfaces have higher energy. These pulses can be computed in exactly the same fashion as the $\langle 10\rangle$ and $\langle 11\rangle$ pulses, but we have not carried out these computations at present. The existence of infinitely many planar hexagon pulses with different orientations can also be inferred from an energy argument. At the Maxwell point, a single hexagon cell has zero energy, and so any combination of the hexagon cells that lie on the hexagon lattice can be used to create differently oriented pulses. We expect the resulting pulses to pin or lock to produce snaking regions similar to those found for $\langle 10\rangle$ and $\langle 11\rangle$ pulses.

Finally, we briefly address the nature of the pitchfork bifurcations to almost-planar hexagon pulses that occur near each fold bifurcation. We begin with the fold bifurcation: the eigenfunction $v_0(x, y)$ associated with the fold eigenvalue $\lambda = 0$ is periodic in the y-variable with minimal period $\ell = 4\pi/\sqrt{3}$ for $\langle 10\rangle$ and $\ell = 4\pi$ for $\langle 11\rangle$ pulses. Now consider the planar pulse on the entire plane and apply Floquet–Bloch theory (see, for instance, [67, Theorem 2.1]): we find a one-parameter family $\lambda(\gamma)$ of eigenvalues, defined for all γ sufficiently close to zero, whose eigenfunctions are of the form $v(x, y; \gamma)e^{i\gamma y}$, where $v(x, y; \gamma)$ has period ℓ in the y-variable. For $\gamma = 0$, we recover the fold eigenvalue $\lambda(0) = 0$ with eigenfunction $v(x, y; 0) = v_0(x, y)$. Next, we consider domains of height $N\ell$ in the y-direction for large integers $N \gg 1$ with periodic boundary conditions in y: this is the situation shown in Figure 16. The eigenfunctions we found on the plane fit into this domain provided $\gamma = \gamma_n := 2\pi n/N\ell$ for integers $n \geq 0$. The smallest nonzero value of γ is $\gamma_1 = 2\pi/N\ell$: the associated eigenvalue $\lambda(\gamma_1) = O(1/N)$ is close to zero for $N \gg 1$, and its eigenfunction

$$v(x, y; \gamma_1)e^{i\gamma_1 y} = [v_0(x, y) + O(1/N)]e^{2\pi i y/N\ell}$$

is a harmonic modulation of the fold eigenfunction in the y-direction. Figure 22 shows a $\langle 10\rangle$ hexagon pulse and the associated fold and pitchfork eigenfunctions on a domain with

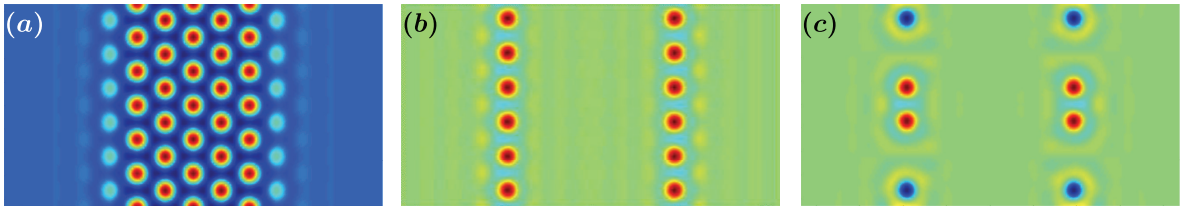


Figure 22. *Shown are color plots of a planar* $\langle 10 \rangle$ *hexagon pulse at a fold bifurcation in* (a)*, the associated fold eigenfunction in* (b)*, and the pitchfork eigenmode that leads to almost-planar hexagon pulses in* (c)*.*

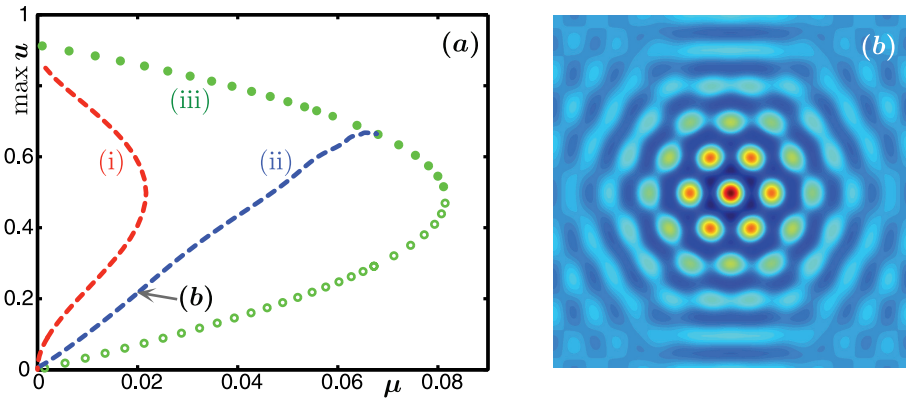


Figure 23. *Localized radial pulses* (i)*, localized hexagon patches* (ii)*, and regular hexagons* (iii) *with wavenumber* $\kappa = 1$ *of the planar Swift–Hohenberg equation* (1.1) *with* $\nu = 0.9$ *are shown in panel* (a)*. Dashed solutions and regular hexagons with open bullets are unstable, while regular hexagons with filled bullets are stable. Note that neither branch* (i) *nor* (ii) *terminates at the branch* (iii) *of regular hexagons (see text for details). Panel* (b) *contains a color plot of a localized hexagon patch for* $\mu = 0.02$*.*

$N = 6$: the pitchfork eigenfunction is indeed a cosine modulation of the fold eigenfunction with maximal period in y, as claimed.

5.3. Localized hexagon patches. In this section, we discuss our results for localized hexagon patches of the planar Swift–Hohenberg equation (1.1). We shall also report on numerical results for localized radial pulses.[7] We focus first on two different representative slices $\nu = 0.9$ and $\nu = 1.6$ of the bifurcation diagram in (μ, ν)-parameter space before we consider the full diagram in (μ, ν)-space and comment on the special value $\nu = 1.049$ that separates regions of qualitatively similar behavior.

5.3.1. Bifurcation diagram for $0 < \nu < 1.049$. We first consider the region where $0 < \nu < 1.049$ and illustrate our results in Figure 23 for $\nu = 0.9$. We find localized radial pulses, localized hexagon patches, and regular hexagons. All these solutions bifurcate from $u = 0$ at $\mu = 0$ and are initially unstable. Regular hexagons stabilize in a fold bifurcation and later cross into the left half-plane $\mu < 0$. Localized radial pulses gain stability with respect to radial perturbations at the first fold bifurcation but continue to be unstable with respect to hexagonal perturbations. They cross with nonzero amplitude into the left half-plane, where

[7] Existence results for these localized radial solutions are proved in [58].

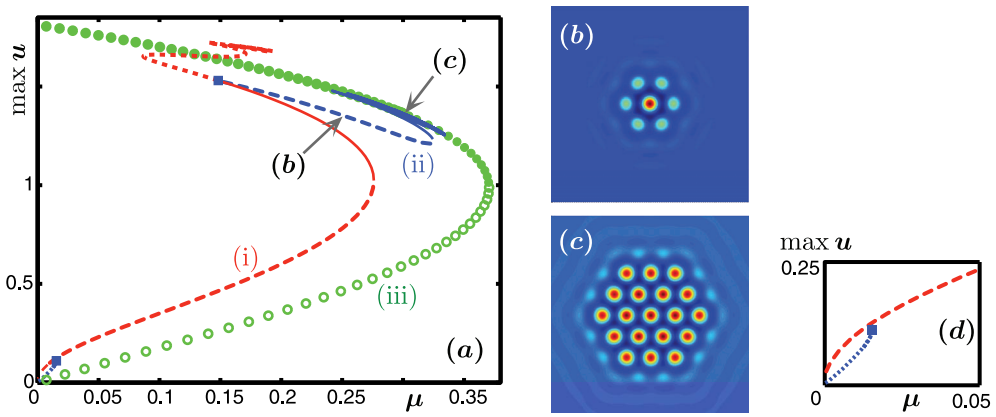


Figure 24. *Localized radial pulses* (i)*, localized hexagon patches* (ii)*, and regular hexagons* (iii) *with wavenumber* $\kappa = 1$ *of the planar Swift–Hohenberg equation* (1.1) *with* $\nu = 1.6$ *are shown in panel* (a)*; the area near the origin is enlarged in panel* (d)*. Dashed solutions and regular hexagons with open bullets are unstable, while solutions along solid lines and regular hexagons with filled bullets are stable. Panels* (b) *and* (c) *are color plots of localized hexagon patches for* $\mu = 0.25$ *near the bifurcation off the radial pulse* (b) *and for* $\mu = 0.3$ *on the snaking curve* (c)*.*

they turn into nonlocalized target patterns. The unstable localized hexagon patches appear to begin to snake for $\mu \approx 0.065$, but we were only able to continue through the first fold. The Maxwell point of regular hexagons for $\nu = 0.9$ is $\mu = \mu_{\mathrm{M}} \approx 0.07$.

5.3.2. Bifurcation diagram for $1.049 < \nu < 2.23$. In Figure 24, we summarize the bifurcation diagram for $\nu = 1.6$, where localized hexagon patches have previously been found in [78] via direct numerical simulations. The localized hexagon patches that arise in this parameter region tend to be highly localized, and the numerical methods described in section 4 should therefore work particularly well. Other previous direct numerical simulations have shown that temporally stable localized radial pulses exist in this region of parameter space: our numerical continuation methods will corroborate these findings and establish a strong link between localized radial and hexagonal structures.

As shown in Figure 24, radial pulses bifurcate off the trivial solution at $\mu = 0$ and are initially unstable with respect to radial and hexagonal perturbations. Also bifurcating at $\mu = 0$ are unstable localized hexagon patches which disappear in a subcritical pitchfork bifurcation of the radial pulse at $\mu \approx 0.015$: from this point onward, the radial pulses are unstable only with respect to radial perturbations. Subsequently, the radial pulses stabilize in a saddle-node bifurcation at $\mu \approx 0.276$ and later on, for $\mu \approx 0.15$, undergo a second subcritical bifurcation to unstable localized hexagon patches. The radial pulses continue on and begin to snake. The unstable localized hexagon patches that bifurcate at the second pitchfork bifurcation of the radial pulses gain stability in a fold bifurcation at $\mu \approx 0.325$ and begin to snake around the hexagon Maxwell point $\mu_{\mathrm{M}} \approx 0.3222$. While snaking, the localized hexagon patches become wider until they fill the entire domain. In addition to these localized patterns, unstable regular hexagons also bifurcate from the trivial solution at $\mu = 0$ and stabilize in a fold bifurcation at $\mu \approx 0.37$. As can be seen from Figure 24(a), the localized hexagon patterns snake close to the regular hexagons.

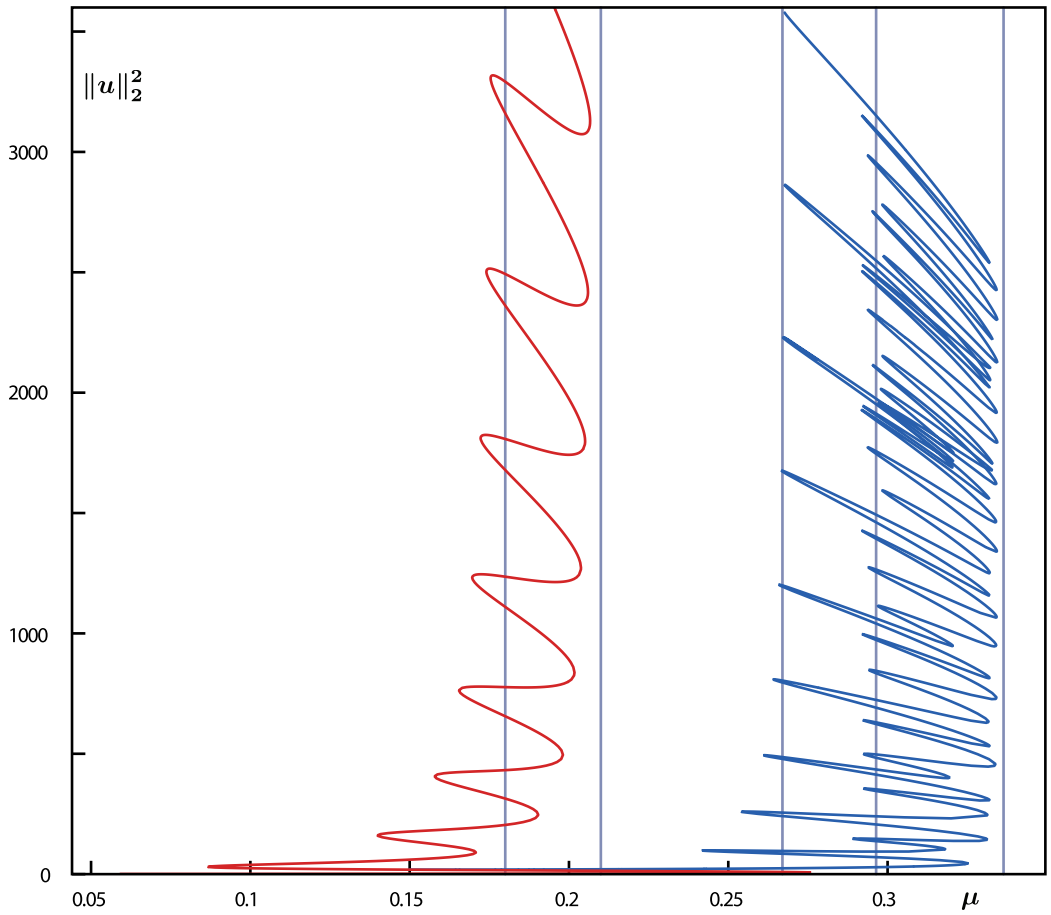


Figure 25. *The snaking curves of radial pulses (left, in red) and fully localized hexagon patches (right, in blue) are plotted for $\nu = 1.6$. Radial solutions are computed on a disk of radius $R = 100$: The fold asymptotes of $1D$ rolls are indicated by vertical grey lines, and the Maxwell point of $1D$ rolls occurs at $\mu_{\mathrm{M}} = 0.2$. The localized hexagon patches are computed with $M = 19$ angular Fourier modes on a domain of radius $R = 80$: The fold asymptotes of planar $\langle 10 \rangle$ and $\langle 11 \rangle$ hexagon pulses are plotted as vertical grey lines, and the hexagon Maxwell point occurs for $\mu_{\mathrm{M}} \approx 0.3222$. The hexagon snaking curve and the associated solution profiles can be viewed in the accompanying animation* (70762_01.mpg [10.8MB]).

We focus now on the snaking behavior of localized hexagon patches. The bifurcation curves of localized radial and hexagon patterns are shown in Figure 25. In particular, we see that the snaking of localized hexagon patches is qualitatively very different from the snaking of the radial pulses and, in fact, also from the snaking of planar hexagon fronts, whose diagram is shown in Figure 21. Indeed, the fold bifurcations of radial pulses and planar hexagon pulses occur near two well-defined limiting values, while the folds of localized hexagon patches align themselves along at least three distinct vertical asymptotes.

The spatial shapes of the fully localized hexagon patches along the snaking curve can be viewed in the accompanying animation (70762_01.mpg [10.8MB]). It is clear from the movie that the localized hexagon structures change in a complicated fashion as the parameter μ is

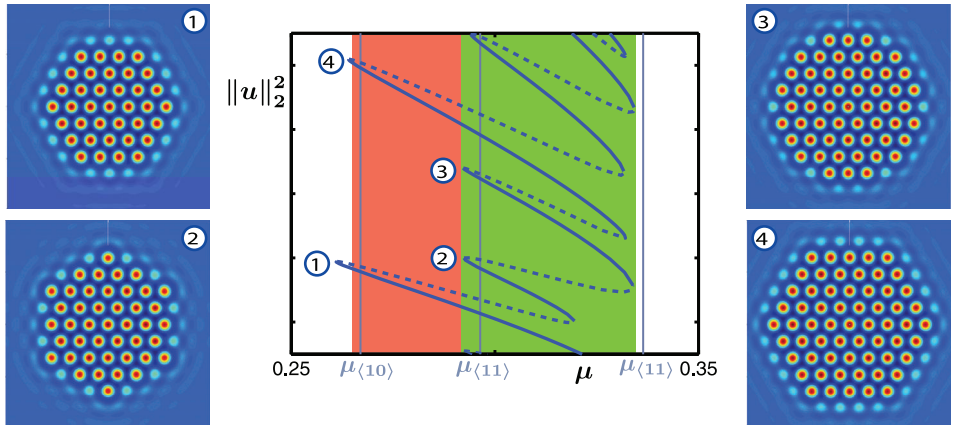


Figure 26. *The center panel contains part of the bifurcation diagram of localized hexagon patches for* $\nu = 1.6$. *Stable and unstable solutions are plotted in solid and dashed lines, respectively. The vertical lines in grey correspond to the fold limits of planar* $\langle 10\rangle$ *and* $\langle 11\rangle$ *hexagon pulses. The red and green regions indicate where temporal self-completion, as shown in Figure* 31(b) *and* (c), *does or does not occur, respectively. Panels* 1–4 *contain color plots of the hexagon patches at the inner and outer left folds.*

varied, and we shall now discuss some of the features visible in the movie in more detail and attempt to identify the underlying mechanisms.

In Figure 26, we plot a few selected spatial profiles along a small segment of the bifurcation curve. We note that localized structures alternately lose and regain stability at consecutive folds. Moving along the bifurcation branch from panel (1) to panel (4) in Figure 26, we see that the localized structure acquired an additional ring of hexagons and thus grew from four to five rings. The process of adding rings is much more complicated than simply adding a new full ring at each fold. First, as shown in panel (2), a single hexagon is added at the center of each side: note that the faces of the resulting localized structure resemble the planar $\langle 11\rangle$ hexagon pulses encountered in section 5.2 and that the addition of the single hexagon occurs at a fold that aligns itself with the snaking curve of $\langle 11\rangle$ pulses. Next, in panel (3), hexagons are added symmetrically to either side of the centered hexagon cell created previously in panel (2): this happens again near an inner fold. The final step is to add an additional pair of hexagons symmetrically to either side of the previously created hexagons to complete the row: this occurs near the fold corresponding to a $\langle 10\rangle$ pulse. The faces of the "superhexagon" structure in panels (1) and (4) resemble planar hexagon pulses with $\langle 10\rangle$ orientation. In summary, as the snake in Figure 26 is traversed, new hexagon cells emerge symmetrically on each face, starting at the center of each face. This cellular growth is reminiscent of the bifurcation diagram of almost-planar $\langle 10\rangle$ pulses shown in Figure 18.

Figure 27 contains a sequence of pictures of localized hexagon structures further up on the snaking curve to illustrate the transition from a localized structure with nine hexagon rings to a pattern with ten rings. Comparing panels (1) and (2), we find that two new hexagon cells appear in the center of each face: there are two new hexagon cells, rather than just one, as the number of hexagons in the outermost row in Figure 27(1) is odd, rather than even, as for the four-ring structure in Figure 26. While cells are added in the center, the corners of the superhexagon structure recede and disappear, which did not happen for four-ring structures.

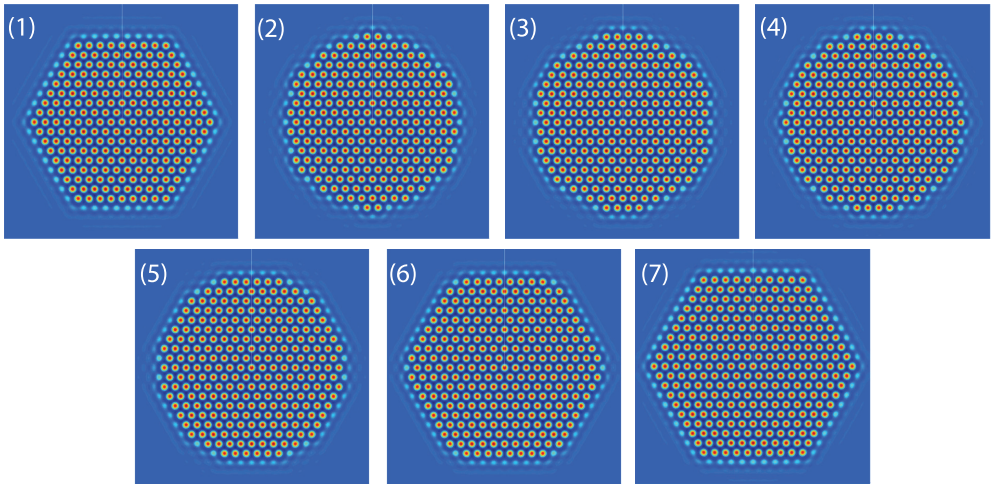


Figure 27. *Color plots of localized hexagon patches at the leftmost folds are shown to illustrate the growth from a* 9*-ring hexagon patch to a* 10*-ring patch. The computations are done on a domain of radius* $R = 100$ *with* $M = 40$ *angular Fourier modes.*

In panel (3), the diagonal faces resemble planar $\langle 11 \rangle$ pulses which move forward to complete the superhexagon visible in panel (7) as we move along the bifurcation curve. The removal of the six cells in the corners of the superhexagon in panels (1)–(2) leads to a reduction of the L^2-norm of the localized structures, which creates the apparent self-intersections of the bifurcation curve visible in Figure 25.

The results discussed above indicate that localized hexagon patches expand initially by adding one or two new hexagon cells at the center of each face. In Figure 28, we compare the bifurcation curves of almost-planar $\langle 10 \rangle$ hexagon pulses with one and two cells to the relevant segments of the snaking curve of localized hexagon patches. The parameter values for the first two folds along the bifurcation curves agree well, but the bifurcation curves of almost-planar pulses and localized hexagon patches separate soon after. On a heuristic level, we believe that hexagon patches do not grow by adding full rows because it costs too much energy to grow hexagons at the corners where two adjacent faces join up. This belief is supported by the observation that the corner hexagons actually recede, as is visible in panels (1)–(2) of Figure 27. Thus, even though cells initially emerge at the centers of each face, the overall growth mechanism is clearly more global, which is why the bifurcation curves of the almost-planar $\langle 10 \rangle$ hexagon pulses agree with the snaking curve of hexagon patches only initially.

We believe that infinitely many planar hexagon fronts with different Bravais–Miller indices play a role in forming the bifurcation diagram shown in Figure 25, though we were not able to go up far enough on the bifurcation curve to identify additional vertical asymptotes that may belong to planar hexagons with different indices. It is remarkable, though, that all rightmost folds seem to line up near the asymptote coming from the planar $\langle 11 \rangle$ hexagon pulse. Most of these folds seem to involve structures that resemble either planar $\langle 11 \rangle$ or almost-planar $\langle 10 \rangle$ pulses. On a heuristic level, it appears that growing these structures involves the same mechanism, and we illustrate this further in Figure 29. However, hexagons emerge along the

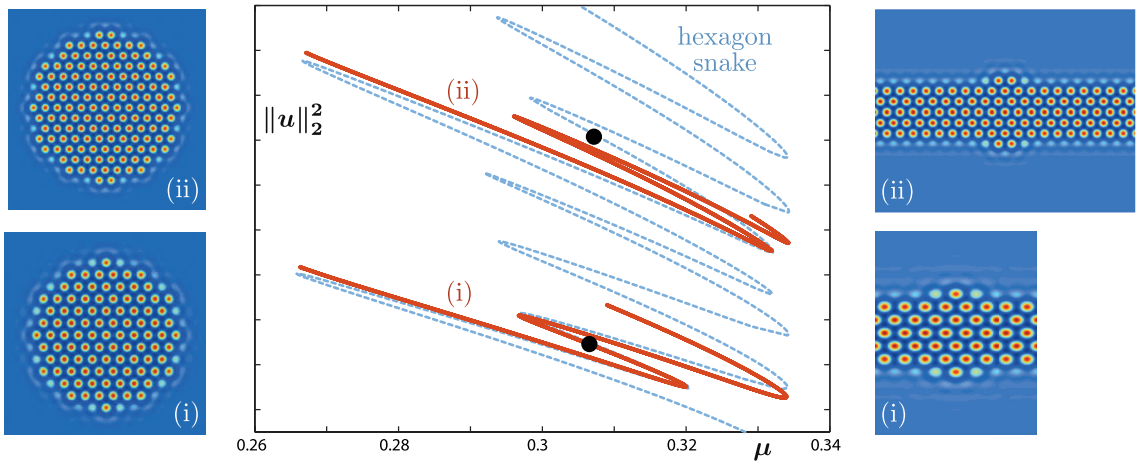


Figure 28. *In the middle panel, we plot part of the hexagon snake from Figure 25 and overlay the snaking curves of almost-planar* $\langle 10 \rangle$ *hexagon pulses for which initially one cell* (i) *or two cells* (ii) *emerge at the center. We rescaled the vertical coordinate of the snaking curves of the almost-planar pulses linearly to allow a comparison with the hexagon data. The remaining panels show color plots of localized hexagon patches (left) for parameter values indicated by bullets in the middle panel and representative almost-planar pulses (right) on the curves* (i) *and* (ii).

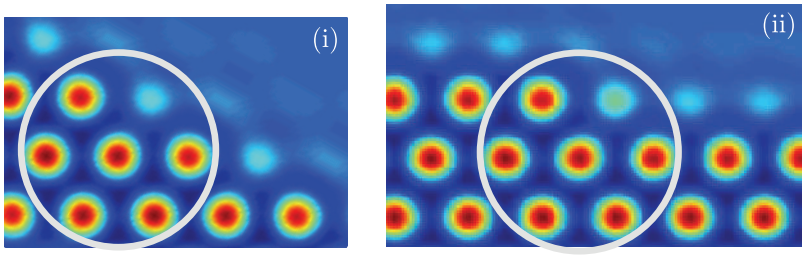


Figure 29. *We plot a rotated planar* $\langle 11 \rangle$ *hexagon pulse in panel* (i) *and an almost-planar* $\langle 10 \rangle$ *hexagon pulse in panel* (ii). *The circles enclose incomplete hexagon structures: Along the next fold in the snaking curve, the missing hexagon in the circle will be filled in.*

entire interface of a planar $\langle 11 \rangle$ pulse, while only a single hexagon is added to an almost-planar $\langle 10 \rangle$ structure.

In section 3.2, we showed that regular hexagon cells in a stationary planar hexagon front satisfy $\mathcal{H} = 0$ for the conserved quantity $\mathcal{H}$ that we defined in (3.8). The condition $\mathcal{H} = 0$ selects the wavenumber of these hexagons. In section 3.2, we stated our belief that this selection criterion for the wavenumber should also apply to localized hexagon patches. In Figure 30, we compare the wavenumbers of regular hexagons for which $\mathcal{H} = 0$ with the wavenumbers of the hexagon cells at the center of localized hexagon patches. We find that the wavenumbers of the center hexagon cells in localized hexagon patches get closer to the predicted wavenumbers as we move up on the snaking curve.

The different vertical asymptotes visible in Figure 26 have interesting consequences for

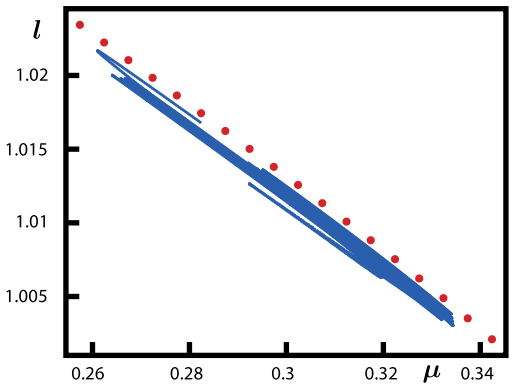


Figure 30. *Plotted are the wavelengths $l = 1/\kappa$ of the regular hexagons for which $\mathcal{H} = 0$ (red bullets) and of the hexagon cells located at the center of the localized hexagon patches (blue dots) as functions of μ for $\nu = 1.6$. The wavelength $l = 1$ corresponds to a spatial period of 2π.*

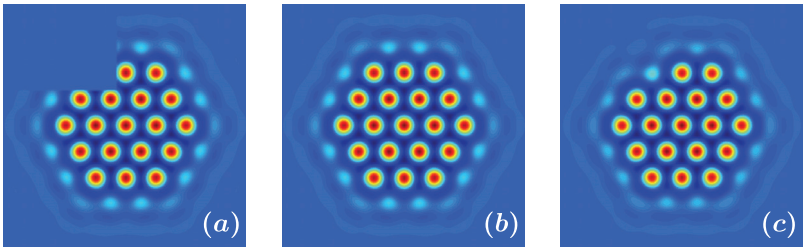


Figure 31. *We illustrate the temporal evolution of localized hexagon structures in the Swift–Hohenberg equation (1.1) with $\nu = 1.6$. The initial condition is shown in panel* (a)*. The solutions at time $t = 100$ are shown in panel* (b) *for $\mu = 0.27$ inside the red region of Figure 26 and in panel* (c) *for $\mu = 0.3$ inside the green region of Figure 26.*

the temporal dynamics of the Swift–Hohenberg equation. As indicated in Figure 26, we divide the μ-parameter space into two intervals, shown in red and green, depending on whether we are to the left or right of the folds that are aligned with the leftmost asymptote of the planar $\langle 11 \rangle$ hexagon pulse. These regions seem to be intimately linked with different self-completion behaviors of localized hexagon patches. We choose the pattern shown in Figure 31(a) as our initial condition and solve the planar Swift–Hohenberg equation. For μ in the red region, to the left of the leftmost $\langle 11 \rangle$ fold, the solution evolves in time toward the completed superhexagon shown in Figure 31(b): since $\langle 11 \rangle$ pulses do not exist in this parameter region, the pattern evolves in time so that all interfaces are $\langle 10 \rangle$ pulses. In contrast, for μ in the green region to the right of the leftmost $\langle 11 \rangle$ fold, the solution converges in time to a localized hexagon patch that is not $\mathbb{D}_6$-symmetric: stable $\langle 11 \rangle$ pulses exist in this region, and the interface of the pattern finds it easier to evolve toward an $\langle 11 \rangle$ pulse rather than a $\langle 10 \rangle$ pulse. Phrased in terms of the bifurcation diagram of Figure 26, the solution moves upward to the stable $\mathbb{D}_6$-symmetric pattern on the hexagon bifurcation curve for parameters in the red region. In contrast, the solution evolves to an asymmetric hexagon patch for parameters in the green region: it appears as if the asymmetric patterns block the evolution toward symmetric patches. We conjecture

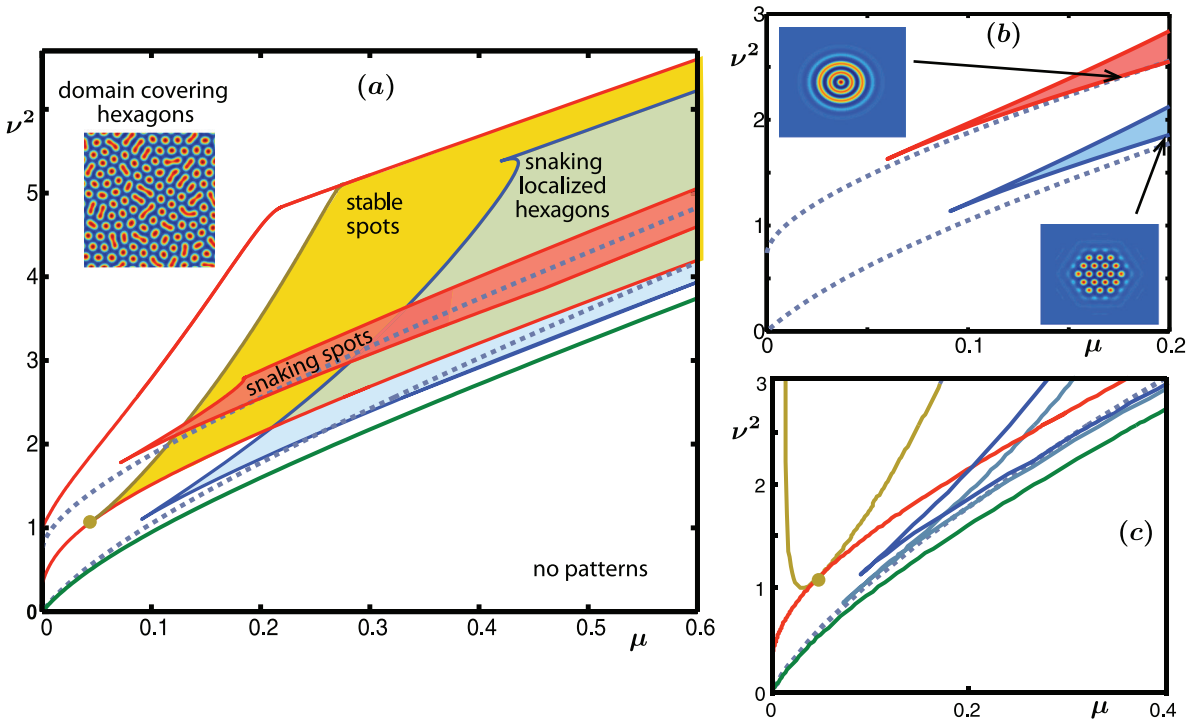


Figure 32. *This figure summarizes our numerical results for radial and hexagonal patterns in the planar Swift–Hohenberg equation* (1.1). *Dashed grey curves correspond to* 1*D roll and* 2*D hexagonal Maxwell curves, the green curve corresponds to the disappearance of regular hexagons with wavenumber* $\kappa = 1$ *in a fold bifurcation, and red curves correspond to fold bifurcations of localized radial spots. The existence region of stable spots is shown in yellow in panel* (a). *Panels* (a)–(c) *also indicate the snaking regions of spots and localized hexagon patches that are delimited by the first pair of fold bifurcations of these patterns: As discussed in the main text, we expect that there is a sequence of fold curves which disappear in a sequence of cusps that accumulate at the origin, so that the full snaking regions are expected to extend along the two Maxwell curves all the way to the* ν*-axis. Panel* (c) *contains four such fold bifurcation curves of localized hexagon patches that are aligned along the hexagon Maxwell curve and which disappear at two cusp bifurcations. Panel* (c) *also contains the pitchfork bifurcation curves of localized hexagon patches from spots in yellow and the fold bifurcation curve of spots in red: These curves meet at the mode interaction point* $(\mu, \nu^2) = (0.048, 1.1)$ $(\nu = 1.049)$.

that these asymmetric hexagon patches may bifurcate from $\mathbb{D}_6$-symmetric hexagon patches in a planar version of ladders, similar to those observed in one space dimension in [19, 20, 22] which we reproduced in Figure 8.

5.3.3. Bifurcation diagram in (μ, ν)-parameter space. Using numerical continuation, we have also traced out partial bifurcation diagrams for localized patterns of the planar Swift–Hohenberg equation in the parameters (μ, ν). These results are summarized in Figure 32 and presented in schematic form in Figure 33.

Regular hexagons exist above the green curve in Figure 32(a). We find that localized hexagon patches seem to bifurcate from the trivial state $u = 0$ along the entire positive ν-axis into the positive quadrant $\mu > 0$. Hexagon patches also bifurcate from localized radial spots along a pitchfork bifurcation curve, where symmetry is broken from O(2) to $\mathbb{D}_6$. Overall, we

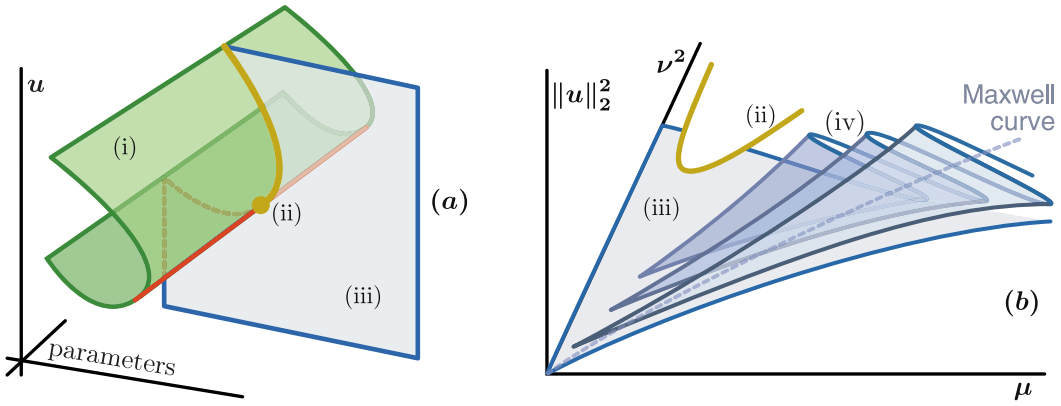


Figure 33. *Schematic illustrations of the bifurcation diagram of localized hexagon patches are shown. Panel* (a) *shows the sheet of localized radial spots* (i) *which undergo pitchfork bifurcations to localized hexagon patches along the yellow curve* (ii). *In panel* (b), *localized hexagon patches* (iii) *bifurcate from* $u = 0$ *at* $\mu = 0$ *and from spots along the pitchfork bifurcation curve* (ii). *The hexagon patches begin to snake* (iv) *near the hexagonal Maxwell curve (shown in dotted grey) in an infinite sequence of fold bifurcations that disappear closer to the origin in a sequence of cusps. See Figure* 32(c) *for the corresponding numerical results.*

obtain a connected surface along which localized hexagon patches exist; see Figure 33 for a schematic picture and Figures 23 and 24 for numerical computations. Localized spots are stable in a wedge delimited by the fold and pitchfork bifurcation curves that emerge from the mode interaction point $(\mu, \nu) = (0.048, 1.049)$ $(\nu^2 = 1.1)$ in Figure 32(a).

Localized hexagon patches snake in a wedge-like region which is aligned with the hexagon Maxwell curve and appears to extend all the way to the origin, where the bifurcation to regular hexagons changes from super- to subcritical, as outlined in Figure 33(b). In Figure 32(c), we show four numerically continued fold bifurcation curves of localized hexagon patches which disappear in two cusp bifurcations. For $\nu \leq 0.9$, localized hexagon patches are spread out so far and fold bifurcations occur so close to each other that we were not able to continue beyond the first fold: these numerical difficulties prevented us from further probing the sequence of fold and cusp bifurcations. Spots also snake, and we show in Figure 32(a) the region delimited by the first two fold bifurcation curves along their snaking curve together with the cusp at which the fold curves collide and disappear. Again, we expect that the snaking region of spots extends along the Maxwell curve associated with 1D rolls to the codimension-two point $(\mu, \nu) = (0, \sqrt{27/38})$, where the bifurcations to rolls change from super- to subcritical.

In Figure 26, we observed that the leftmost snaking limit of the hexagon patches coincides with the leftmost fold of the planar $\langle 10 \rangle$ hexagon pulses. To illuminate this feature further, we show in Figure 34 the snaking region of fully localized hexagon patches and, for comparison, the snaking regions of planar $\langle 10 \rangle$ and $\langle 11 \rangle$ hexagon pulses in panels (a) and (b), respectively. For $\mu < 0.35$, the leftmost fold curve of the $\langle 10 \rangle$ pulses aligns itself with the leftmost boundary of the hexagon patches, while the rightmost fold curve of the $\langle 11 \rangle$ pulses aligns itself with the rightmost boundary of the hexagon patches.

We see in Figure 32 that a cusp forms at $(\mu, \nu^2) \approx (0.4205, 5.4173)$ on the upper fold curve on the wedge belonging to localized hexagon patches. This upper cusp is similar to that found in [19, 20, 22] for 1D structures and indicates that a new Maxwell curve crosses into

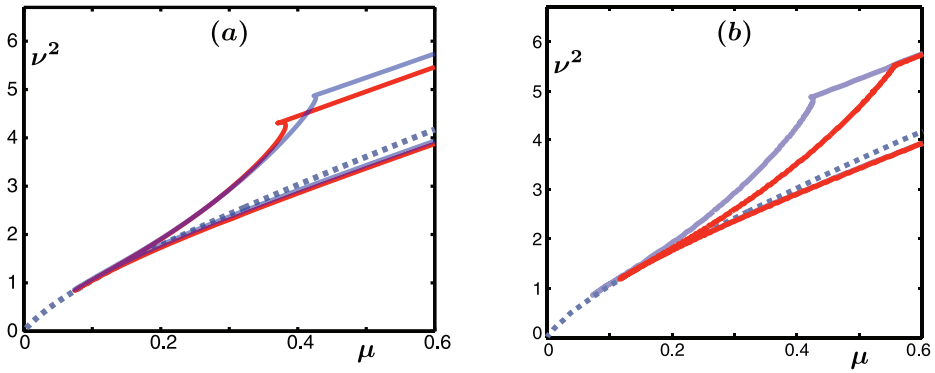


Figure 34. *Comparison of the snaking region of localized hexagon patches (blue) and the snaking regions of planar* $\langle 10 \rangle$ *hexagon pulses (red) in panel* (a) *and* $\langle 11 \rangle$ *hexagon pulses (red) in panel* (b)*. The hexagon Maxwell curve is also shown (dashed grey).*

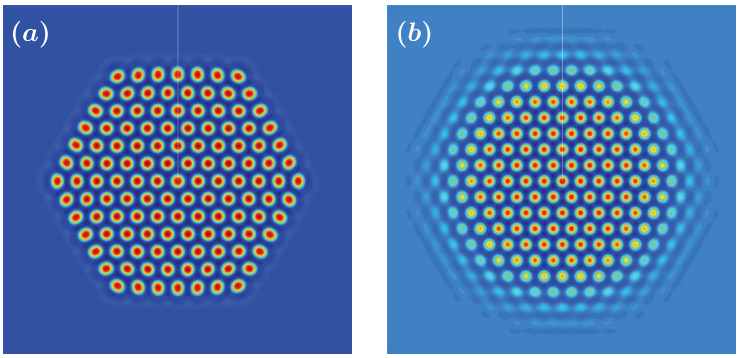


Figure 35. *Panels* (a) *and* (b) *contain color plots of localized hexagon patches for* $(\mu, \nu^2) = (0.7, 6.2355)$ *and* $(\mu, \nu^2) = (0.0738, 0.8612)$*, respectively.*

the snaking region. We plot a localized hexagon patch further up on the upper fold curve for $(\mu, \nu^2) = (0.7, 6.2355)$ in Figure 35(a): note that the individual hexagon cells on the outer ring are elongated. For comparison, we plot in Figure 35(b) the localized hexagon patch at the lower cusp at $(\mu, \nu^2) = (0.0738, 0.8612)$.

5.4. Localized rhomboid patches. We also investigated fully localized rhomboid patches. These solutions have an interior cellular hexagonal structure, but, as shown in Figure 36(b), the overall patch does not possess hexagonal symmetry: the underlying cellular pattern is shifted by half a spatial period so that the center of the localized patch does not coincide with the center of an interior hexagon. Unlike localized hexagon patches, localized rhomboids bifurcate from a pair of localized hexagons, which in turn bifurcate from the trivial state. Hexagon pairs, triplets, and rhomboids consisting of four hexagons have previously been computed in [63]. The snaking diagram of the localized rhomboid patches, shown in Figure 36(a), appears to be qualitatively similar to the snaking diagram of localized hexagon patches shown in Figure 25. There seem to be roughly three snaking limits: the leftmost limit corresponds to the completion of a superrhomboid, shown in Figure 36(b), while the other two snaking limits

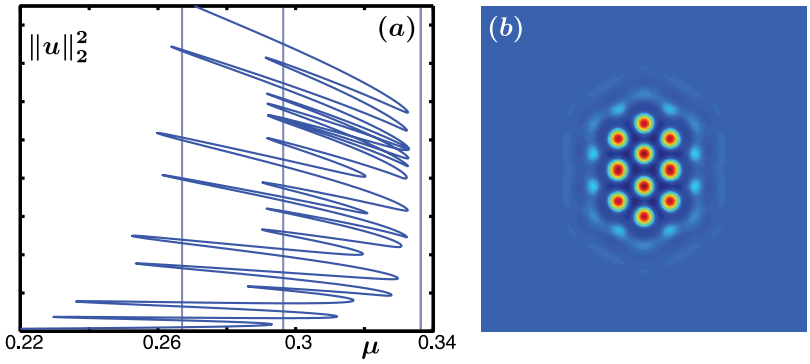


Figure 36. *Panel* (a) *contains the bifurcation diagram of localized rhomboid patches for* $\nu = 1.6$*, while the rhomboidal structure itself is shown as a color plot in panel* (b) *for* $\mu = 0.2817$*. The vertical lines are the asymptotes of the fold bifurcations of* $\langle 10 \rangle$ *and* $\langle 11 \rangle$ *hexagon pulses shown in Figure 26. The computations were carried out on a square domain with dimensions* 80×80*.*

correspond to the emergence of individual hexagons along the sides of the localized structure, as in the case of hexagon patches. Further up the snake, the leftmost folds get closer to the folds of the planar $\langle 10 \rangle$ hexagon pulses. Similar to the case of hexagon patches, the bifurcation curves of localized rhomboids intersect as individual corner cells are suppressed, leading to a decrease of the overall L^2-norm. We computed the localized rhomboids on a large square box with Neumann boundary conditions rather than with polar coordinates as we did for localized hexagon patches: in particular, we are confident that the damping of the corner cells is not due to boundary conditions or numerical errors.

We believe that the localized hexagon and rhomboid patches are connected in parameter space by bifurcation curves of asymmetric patterns similar to those found in [19, 21] for 1D structures. These ladders will effectively shift the cellular pattern between the localized hexagon and rhomboid patches.

6. Conclusions and discussion.

Summary. We briefly summarize our main findings. First, we provided a selection principle for hexagons that can appear as asymptotic states in planar fronts that connect to the trivial state $u = 0$. Any such hexagon must satisfy $\mathcal{H} = 0$, where the function $\mathcal{H}$ is a first integral of the spatial dynamical system that describes solutions $u(x, y)$ of the Swift–Hohenberg equation that are periodic in the transverse y-direction. The expression (1.6) of $\mathcal{H}$ was derived from a conservation law that arises, via Noether's theorem, due to the translation symmetry of the Lagrangian of the energy (1.4) of the Swift–Hohenberg equation. Using a theorem proved in [9], we also showed that a unique branch of regular hexagons along which $\mathcal{H} = 0$ bifurcates from the trivial state at $\mu = 0$: these hexagons have a uniquely selected wavelength.

In section 3.2, we gave a spatial-dynamics formulation of solutions of the Swift–Hohenberg equation that are periodic in one of the two spatial variables. This formulation implies that planar hexagon pulses, such as those shown in Figures 16 and 19, will exist in open regions of parameter space provided they are transversely constructed, a condition we expect to hold generically. This indicates that snaking should occur for planar hexagon pulses, and we found numerically in section 5.2 that snaking does indeed occur for hexagons pulses with two differ-

ent orientations, namely, with Bravais–Miller indices $\langle 10 \rangle$ and $\langle 11 \rangle$. We also computed and continued almost-planar hexagon pulses which bifurcate from the planar hexagon pulses in pitchfork bifurcations near each fold bifurcation: these almost-planar pulses appear prominently in the snaking diagram of localized hexagon patches.

Heuristically, we expect that the snaking regions are centered around the hexagon Maxwell curve which corresponds to the curve in (μ, ν)-parameter space along which hexagons exist that satisfy $\mathcal{H} = 0$ and that have the same energy, $\mathcal{E}(0) = 0$, as the trivial state $u = 0$. Indeed, only when the trivial state and the hexagons have roughly the same energy can we expect that stationary interfaces between them exist; otherwise, one of the states will invade the other one to reduce the overall energy. Our numerical computations confirmed this heuristic picture and showed furthermore that the Maxwell curve emerges from the codimension-two point where the bifurcation to hexagons changes from super- to subcritical, as it is there that we can expect regions of bistability to exist.

Our main numerical findings consist of the continuation results of localized hexagon and rhomboid patches in the Swift–Hohenberg equation. Our computations suggest that infinitely many hexagon and rhomboid patches coexist in open parameter regions. The localized hexagon patches lie on the same solution branch and increase in width as we move along the branch. Strikingly, the hexagon patches do not grow by adding a full ring of hexagons at each fold but instead seem to follow, at least initially, the almost-planar hexagon pulses with indices $\langle 10 \rangle$ and $\langle 11 \rangle$ that we computed in section 5.2. Overall, we found a rich snaking structure with several, possibly infinitely many, vertical asymptotes for fold bifurcations of localized structures, compared with only two asymptotes for planar hexagon pulses and 1D structures. We identified three asymptotes as arising from fold bifurcations of planar hexagon pulses with indices $\langle 10 \rangle$ and $\langle 11 \rangle$. However, we did not identify an overarching mechanism that predicts how hexagon patches might grow as we move up further along the branch.

We also investigated self-completion of asymmetric hexagon patches and found evidence that self-completion occurs only to the left of the snaking region of the $\langle 11 \rangle$ fronts. The self-completion study was motivated by results in [4, 6], where this process was addressed by using interaction theory for localized spots. The alternative explanation put forward here is based on the existence regions of planar hexagon fronts with different orientations. Though we do not have any conclusive evidence, we do not believe that the hexagon structures found in our paper can be viewed as bound states of localized spots: Figure 24, for instance, shows that localized hexagon patches can exist well outside the existence region of localized spots.

Finally, we mention that Figure 32 contains various results on localized radial structures. In particular, the branch of localized hexagon patches that bifurcates from the trivial state at $(\mu, \nu) = 0$ and later begins to snake splits, for larger values of ν, into two branches which begin or end at pitchfork bifurcations of localized radial structures. We refer the reader to [58] for a more detailed analytical and numerical study of these radial spots.

Open problems. We now outline what we believe to be interesting questions for further research on multidimensional localized patterns and refer the reader to [50] for another recent list of open problems in this area.

A major goal is to uncover the mechanism that underlies the snaking behavior of localized hexagon patches and to prove that it does occur in the Swift–Hohenberg equation. Currently, there do not seem to be any methods available that can be used to carry out such a compre-

hensive analysis. Thus, we discuss first a number of more modest open problems that may give better insight into certain aspects of hexagon snaking.

Snaking of planar hexagon pulses seems more amenable to an analytic approach. On a formal level, asymptotics beyond all orders has recently been used in [25, 52] to predict the snaking width for the 1D structures shown in Figure 2 near the codimension-two point $(\mu, \nu) = 0$. The idea behind this approach is to look more closely into the derivation of the amplitude equations (2.10) which govern the existence of 1D pulses. In the standard derivation, anisotropic terms that depend on the small scale x, rather than the large scale $X = \epsilon x$, are neglected. In [25, 52], these terms and their effect on the remaining modes through the nonlinearity are taken into account, and an analysis of the resulting exponentially small coupling terms between rolls and the 1D pulse gave an extremely accurate prediction for the snaking region in the 1D setting. The same approach may perhaps work in the planar case to capture the interaction terms between small-scale hexagons and large-scale hexagon pulses with different Bravais–Miller indices.

The energy functional of the Swift–Hohenberg equation may also help to illuminate snaking of hexagon pulses. Our numerical results indicate that the widths of the snaking regions of planar hexagon pulses depend on the orientation of their interfaces, i.e., on their Bravais–Miller indices, and it may be possible to capture this effect through an appropriate interfacial energy. Along the same lines, the growth of cells along an interface for almost-planar hexagon pulses appears qualitatively similar to the growth of interface boundaries in polycrystalline structures. Numerical studies of polycrystalline structures in [64] via two-dimensional Ising models have shown that the orientation of interfaces in hexagonal lattices has a significant effect on the propagation speed of these interfaces. In the context of the Swift–Hohenberg equation, the speeds of planar $\langle 10 \rangle$ and $\langle 11 \rangle$ fronts outside the pinning region may be similar to those seen in Ising models.

Another approach to understanding snaking of planar hexagon pulses is to assume that there exists a generic planar hexagon front for a certain value of μ which disappears, as μ decreases or increases beyond a certain threshold, via a saddle-node bifurcation as indicated in Figure 6. Instead of using geometric methods to prove that this results in snaking of planar hexagon pulses, an analytic result via Lin's method could be used to generalize the intuitive picture given in Figure 6 for two-dimensional Poincaré sections to the infinite-dimensional spatial-dynamics setting of section 3.2. This analysis has recently been carried out in [10], where it was also shown that it captures asymmetric ladder structures.

While an analysis of snaking of hexagon patches seems currently out of reach, it may be possible to say more about the underlying mechanisms by carrying out more comprehensive numerical studies. The numerical methods we have used significantly reduce the computational cost required to compute fully localized 2D patterns by using an adaptive mesh and by taking into account the symmetry of localized hexagon patches. However, the computation of larger patterns requires a less expensive way of computing hexagon patches. We believe that implementing triangular finite elements in Trilinos should result in a significant speed-up. In addition, we do not expect that the core region of hexagon patches changes much as the patterns grow. Thus, it might be possible to work with an annular region as the computational domain where the boundary conditions on the inner boundary are chosen to ensure compatibility with previously computed hexagon patches. This should lead to a further reduction of

the size of the system.

Other localized 2D patterns. Last, we comment on other localized planar structures.

In [78], fully localized stripe patches were observed in the cubic-quintic Swift–Hohenberg equation

$$u_t = -(1+\Delta)^2 u - \mu u + \nu u^3 - u^5. \tag{6.1}$$

These patterns, reproduced in Figure 1(b), are clearly anisotropic. To describe them, one could use, as in [32], the Newell–Whitehead–Segel equation

$$4\left(\partial_X - \frac{\mathrm{i}}{2}\partial_{YY}^2\right)^2 A = \epsilon A - |A|^2 A + |A|^4 A$$

for the envelope function $A(X,Y)$ of stripes that are parallel to the y-direction, where $(X,Y) = (\epsilon^{1/2}x, \epsilon^{1/4}y)$. This equation admits localized fronts $A_1(X)$ and $A_2(Y)$ with different spatial widths. The front $A_1(X)$ occurs at the 1D Maxwell point found in the normal-form analysis of the 1D cubic-quintic Swift–Hohenberg equation (6.1): this front corresponds to an "equilibrium to periodic orbit" connection for the associated spatial dynamical system, and we therefore expect snaking and the growth of additional rolls along the x-direction as an equation parameter is varied. The front $A_2(Y)$ in the Y-direction, on the other hand, corresponds to an "equilibrium to equilibrium" connection, and we do not expect snaking to occur; instead, we expect that the bifurcation curve converges, in an oscillatory fashion, to a single vertical asymptote [51]. This latter behavior is precisely what was observed in [78] in numerical simulations of fully localized stripe patterns.

Another interesting Swift–Hohenberg model is

$$u_t = -(1+\Delta)^2 u - \mu u + \nu u^3 - u^5 + \alpha\nabla\cdot[|\nabla u|^2\nabla u].$$

The last term in the above equation gives preference to patterns with square symmetry, and localized patches of squares have indeed been found in direct numerical simulations [44, 79]. We expect that these patches exhibit snaking and predict that localized square patches of square shape grow by adding new cells starting from the middle of each face.

Localized pentagonal structures have also been observed numerically in [91, Figure 10(a)–(b)] in a model of driven optical cavities. Our numerical methods could be extended easily to compute and continue these structures by expanding u as a Fourier series

$$u(r,\theta) = \sum_{n\in\mathbb{Z}} a_n(r)\mathrm{e}^{5ni\theta}$$

using five-fold symmetric terms. Pentagons do not tile the plane, so the question of snaking for localized pentagons would be interesting.

Acknowledgments. We are grateful to John Burke, Edgar Knobloch, Paul Matthews, Arnd Scheel, and Thomas Wagenknecht for many helpful comments and discussions. We also thank José Antonio Medina Hernández, Gregory Kozyreff, and the anonymous referees for many comments that helped us improve the presentation of this paper. David Lloyd and Björn Sandstede thank the Newton Institute for its hospitality during the theme programme "Pattern Formation in Large Domains" in Autumn 2005, and Björn Sandstede gratefully acknowledges a Royal Society Wolfson Research Merit Award.

REFERENCES

[1] E. Ammelt, Y. A. Astrov, and H.-G. Purwins, *Hexagon structures in a two-dimensional dc-driven gas discharge system*, Phys. Rev. E, 58 (1998), pp. 7109–7117.

[2] S. C. Anco and G. Bluman, *Direct construction method for conservation laws of partial differential equations* I: *Examples of conservation law classifications*, European J. Appl. Math., 13 (2002), pp. 545–566.

[3] S. C. Anco and G. Bluman, *Direct construction method for conservation laws of partial differential equations* II: *General treatment*, European J. Appl. Math., 13 (2002), pp. 567–585.

[4] I. S. Aranson, K. A. Gorshkov, A. S. Lomov, and M. I. Rabinovich, *Stable particle-like solutions of multidimensional nonlinear fields*, Phys. D, 43 (1990), pp. 435–453.

[5] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Harcourt, New York, 1976.

[6] Y. Astrov and Y. Logvin, *Formation of clusters of localized states in a gas discharge system via a self-completion scenario*, Phys. Rev. Lett., 79 (1997), pp. 2983–2986.

[7] D. Avitabile, *Computation of Planar Patterns and Their Stability*, Ph.D. thesis, University of Surrey, Guildford, UK, 2008.

[8] O. Batiste, E. Knobloch, A. Alonso, and I. Mercader, *Spatially localized binary-fluid convection*, J. Fluid Mech., 560 (2006), pp. 149–158.

[9] R. E. Beardmore, M. A. Peletier, C. J. Budd, and M. Ahmer Wadee, *Bifurcations of periodic solutions satisfying the zero-Hamiltonian constraint in reversible differential equations*, SIAM J. Math. Anal., 36 (2005), pp. 1461–1488.

[10] M. Beck, J. Knobloch, D. J. B. Lloyd, B. Sandstede, and T. Wagenknecht, *Snakes, Ladders, and Isolas of Localised Patterns*, preprint, 2008.

[11] C. Bensimon, B. Shraiman, and V. Croquette, *Nonadiabatic effects in convection*, Phys. Rev. A, 38 (1988), pp. 5461–5464.

[12] S. Blanchflower, *Magnetohydrodynamic convectons*, Phys. A, 261 (1999), pp. 74–81.

[13] U. Bortolozzo, M. G. Clerc, C. Falcon, S. Residori, and R. Rojas, *Localized states in bistable pattern-forming systems*, Phys. Rev. Lett., 96 (2006), 214501.

[14] M. Bougchariou, *Closed orbits of Hamiltonian systems on non-compact prescribed energy surfaces*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 603–616.

[15] H. R. Brand, C. Fradin, P. Finn, W. Pesch, and P. Cladis, *Electroconvection in nematic liquid crystals: Comparison between experimental results and the hydrodynamic model*, Phys. Lett. A, 235 (1997), pp. 508–514.

[16] C. J. Budd, G. W. Hunt, and R. Kuske, *Asymptotics of cellular buckling close to the Maxwell load*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 2935–2964.

[17] C. J. Budd and R. Kuske, *Localized periodic patterns for the non-symmetric generalized Swift-Hohenberg equation*, Phys. D, 208 (2005), pp. 73–95.

[18] B. Buffoni and J. F. Toland, *Global existence of homoclinic and periodic orbits for a class of autonomous Hamiltonian systems*, J. Differential Equations, 118 (1995), pp. 104–120.

[19] J. Burke and E. Knobloch, *Localized states in the generalized Swift-Hohenberg equation*, Phys. Rev. E (3), 73 (2006), 056211.

[20] J. Burke and E. Knobloch, *Homoclinic snaking: Structure and stability*, Chaos, 17 (2007), 037102.

[21] J. Burke and E. Knobloch, *Normal form for spatial dynamics in the Swift-Hohenberg equation*, Discrete Contin. Dyn. Syst. Suppl., (September) (2007), pp. 170–180.

[22] J. Burke and E. Knobloch, *Snakes and ladders: Localized states in the Swift-Hohenberg equation*, Phys. Lett. A, 360 (2007), pp. 681–688.

[23] E. Buzano and M. Golubitsky, *Bifurcation on the hexagonal lattice and the planar Bénard problem*, Philos. Trans. Roy. Soc. London Ser. A, 308 (1983), pp. 617–667.

[24] B. J. Cantwell, *Introduction to Symmetry Analysis*, Cambridge University Press, Cambridge, UK, 2002.

[25] S. J. Chapman and G. Kozyreff, *Exponential Asymptotics of Localised Patterns and Snaking Bifurcation Diagrams*, preprint, 2008.

[26] P. Coullet, C. Riera, and C. Tresser, *Stable static localised structures in one dimension*, Phys. Rev. Lett., 84 (2000), pp. 3069–3072.

[27] S. M. Cox and P. C. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.
[28] S. M. Cox and P. C. Matthews, *Instability and localisation of patterns due to a conserved quantity*, Phys. D, 175 (2003), pp. 196–219.
[29] C. Crawford and H. Riecke, *Oscillon-type structures and their interaction in a Swift-Hohenberg model*, Phys. D, 129 (1999), pp. 83–92.
[30] M. Cross and P. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.
[31] P. Davies, P. Blanchedeau, E. Dulos, and P. D. Kepper, *Dividing blobs, chemical flowers and patterned islands in a reaction-diffusion system*, J. Phys. Chem. A, 102 (1998), pp. 8236–8244.
[32] R. J. Deissler and H. R. Brand, *Two-dimensional localized solutions for subcritical bifurcations in systems with broken rotational symmetry*, Phys. Rev. E, 51 (1995), pp. R852–R855.
[33] B. Dionne, M. Silber, and A. C. Skeldon, *Stability results for steady, spatially periodic planforms*, Nonlinearity, 10 (1997), pp. 321–353.
[34] E. J. Doedel, *AUTO-07P: Continuation and Bifurcation Software for Ordinary Differential Equations*, Tech. rep., Concordia University, Montreal, Canada, 2007.
[35] A. Doelman, B. Sandstede, A. Scheel, and G. Schneider, *Propagation of hexagonal patterns near onset*, European J. Appl. Math., 14 (2003), pp. 85–110.
[36] J. Eggers and H. Riecke, *Continuum description of vibrated sand*, Phys. Rev. E, 59 (1999), pp. 4476–4483.
[37] P. C. Fife, *Pattern formation in gradient systems*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 677–722.
[38] J. Fineberg, *Physics in a jumping sandbox*, Nature, 382 (1996), pp. 793–764.
[39] A. A. Golovin, B. J. Matkowsky, and A. A. Nepomnyashchy, *A complex Swift-Hohenberg equation coupled to the Goldstone mode in the nonlinear dynamics of flames*, Phys. D, 179 (2003), pp. 183–210.
[40] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory* II, Springer-Verlag, New York, 1988.
[41] M. Golubitsky, J. W. Swift, and E. Knobloch, *Symmetries and pattern selection in Rayleigh-Bénard convection*, Phys. D, 10 (1984), pp. 249–276.
[42] D. Gomila, A. J. Scroggie, and W. J. Firth, *Bifurcation structure of dissipative solitons*, Phys. D, 227 (2007), pp. 70–77.
[43] M. A. Heroux, R. A. Bartlett, V. E. Howle, R. J. Hoekstra, J. J. Hu, T. G. Kolda, R. B. Lehoucq, K. R. Long, R. P. Pawlowski, E. T. Phipps, A. G. Salinger, H. K. Thornquist, R. S. Tuminaro, J. M. Willenbring, A. Williams, and K. S. Stanley, *An overview of the Trilinos project*, ACM Trans. Math. Software, 31 (2005), pp. 397–423.
[44] M. F. Hilali, S. Metens, P. Borckmans, and G. Dewel, *Pattern selection in the generalised Swift-Hohenberg model*, Phys. Rev. E, 51 (1995), pp. 2046–2052.
[45] R. B. Hoyle, *Pattern Formation*, Cambridge University Press, Cambridge, UK, 2006.
[46] G. W. Hunt, G. J. Lord, and A. R. Champneys, *Homoclinic and heteroclinic orbits underlying the post-buckling of axially-compressed cylindrical shells*, in Localization and Solitary Waves in Solid Mechanics, World Scientific, River Edge, NJ, 1999, pp. 285–297.
[47] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. A. Wadee, C. J. Budd, and G. J. Lord, *Cellular buckling in long structures*, Nonlinear Dynam., 21 (2000), pp. 3–29.
[48] H. Jamgotchian, N. Bergeon, D. Benielle, P. Voge, B. Billia, and R. Guerin, *Localised microstructures induced by fluid flow in directional solidification*, Phys. Rev. Lett., 87 (2001), 166105.
[49] E. Knobloch, *Pattern selection in long-wavelength convection*, Phys. D, 41 (1990), pp. 450–479.
[50] E. Knobloch, *Spatially localized structures in dissipative systems: Open problems*, Nonlinearity, 21 (2008), pp. T45–T60.
[51] J. Knobloch and T. Wagenknecht, *Homoclinic snaking near a heteroclinic cycle in reversible systems*, Phys. D, 206 (2005), pp. 82–93.
[52] G. Kozyreff and S. J. Chapman, *Asymptotics of large bound states of localized structures*, Phys. Rev. Lett., 97 (2006), 044502.
[53] B. Krauskopf and T. Riess, *A Lin's method approach to finding and continuing heteroclinic connections involving periodic orbits*, Nonlinearity, 21 (2008), pp. 1655–1690.

[54] C. R. Laing and W. C. Troy, *PDE methods for nonlocal models*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 487–516.
[55] C. R. Laing, W. C. Troy, B. Gutkin, and G. B. Ermentrout, *Multiple bumps in a neuronal model of working memory*, SIAM J. Appl. Math., 63 (2002), pp. 62–97.
[56] J. Lega, J. V. Moloney, and A. C. Newell, *Swift-Hohenberg equation for lasers*, Phys. Rev. Lett., 73 (1994), pp. 2978–2981.
[57] D. J. B. Lloyd and A. R. Champneys, *Efficient numerical continuation and stability analysis of spatiotemporal quadratic optical solitons*, SIAM J. Sci. Comput., 27 (2005), pp. 759–773.
[58] D. J. B. Lloyd and B. Sandstede, *Localized radial solutions of the Swift–Hohenberg equation*, preprint, 2008.
[59] G. J. Lord, A. R. Champneys, and G. W. Hunt, *Computation of homoclinic orbits in partial differential equations: An application to cylindrical shell buckling*, SIAM J. Sci. Comput., 21 (1999), pp. 591–619.
[60] B. A. Malomed, A. A. Nepomnyashchy, and M. I. Tribelsky, *Domain boundaries in convection patterns*, Phys. Rev. A, 42 (1990), pp. 7244–7263.
[61] P. C. Matthews, *Hexagonal patterns in finite domains*, Phys. D, 116 (1998), pp. 81–94.
[62] P. C. Matthews, M. R. E. Proctor, and N. O. Weiss, *Compressible magnetoconvection in three dimensions: Planforms and nonlinear behaviour*, J. Fluid Mech., 305 (1995), pp. 281–305.
[63] J. M. McSloy, W. J. Firth, G. K. Harkness, and G.-L. Oppo, *Computationally determined existence and stability of transverse structures* II: *Multipeaked cavity solitons*, Phys. Rev. E, 66 (2002), 046606.
[64] M. Medelev, D. J. Srolovitz, L. Shvindlerman, and G. Gottstein, *Interface mobility under different driving forces*, J. Mater. Res., 17 (2002), pp. 234–245.
[65] I. Mercader, A. Alonso, and O. Batiste, *Spatiotemporal dynamics near the onset of convection for binary mixtures in cylindrical containers*, Phys. Rev. E, 77 (2008), 036313.
[66] A. Mielke, *A spatial center manifold approach to steady state bifurcations from spatially periodic patterns*, in Dynamics of Nonlinear Waves in Dissipative Systems: Reduction, Bifurcation and Stability, G. Dangelmayr, B. Fiedler, K. Kirchgässner, and A. Mielke, eds., Longman, Harlow, UK, 1996, pp. 209–262.
[67] A. Mielke, *Instability and stability of rolls in the Swift-Hohenberg equation*, Comm. Math. Phys., 189 (1997), pp. 829–853.
[68] A. A. Nepomnyaschy, M. I. Tribelsky, and M. G. Velarde, *Wave number selection in convection and related problems*, Phys. Rev. E, 50 (1994), pp. 1194–1197.
[69] L. A. Peletier and W. C. Troy, *Spatial Patterns*, Birkhäuser Boston, Boston, 2001.
[70] M. A. Peletier, *Sequential buckling: A variational analysis*, SIAM J. Math. Anal., 32 (2001), pp. 1142–1168.
[71] D. Peterhof, B. Sandstede, and A. Scheel, *Exponential dichotomies for solitary-wave solutions of semilinear elliptic equations on infinite cylinders*, J. Differential Equations, 140 (1997), pp. 266–308.
[72] L. M. Pismen, *Patterns and Interfaces in Dissipative Dynamics*, Springer-Verlag, Berlin, 2006.
[73] Y. Pomeau, *Front motion, metastability, and subcritical bifurcations in hydrodynamics*, Phys. D, 23 (1986), pp. 3–11.
[74] M. I. Rabinovich, A. B. Ezersky, and P. D. Weidman, *The Dynamics of Patterns*, World Scientific, River Edge, NJ, 2000.
[75] P. H. Rabinowitz, *Periodic solutions of a Hamiltonian system on a prescribed energy surface*, J. Differential Equations, 33 (1979), pp. 336–352.
[76] P. H. Rabinowitz, *The prescribed energy problem for periodic solutions of Hamiltonian systems*, in Hamiltonian Dynamical Systems, AMS, Providence, RI, 1988, pp. 183–191.
[77] S. Residori, *Patterns, fronts and structures in a liquid-crystal-light-valve with optical feedback*, Phys. Rep., 416 (2005), pp. 201–272.
[78] H. Sakaguchi and H. R. Brand, *Stable localised solutions of arbitrary length for the quintic Swift-Hohenberg equation*, Phys. D, 97 (1996), pp. 274–285.
[79] H. Sakaguchi and H. R. Brand, *Stable localised squares in pattern-forming nonequilibrium systems*, Europhys. Lett., 38 (1997), pp. 341–346.
[80] H. Sakaguchi and H. R. Brand, *Localised patterns for the quintic complex Swift-Hohenberg equation*, Phys. D, 117 (1998), pp. 95–105.

[81] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 983–1055.
[82] B. Sandstede and A. Scheel, *Defects in oscillatory media: Toward a classification*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 1–68.
[83] B. Sandstede and A. Scheel, *Relative Morse indices, Fredholm indices, and group velocities*, Discrete Contin. Dyn. Syst., 20 (2008), pp. 139–158.
[84] J. Shen, *Efficient spectral-Galerkin methods* III: *Polar and cylindrical geometries*, SIAM J. Sci. Comput., 18 (1997), pp. 1583–1604.
[85] J. Swift and P. C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. A, 15 (1977), pp. 319–328.
[86] M. Tlidi, A. G. Vladimirov, and P. Mandel, *Interaction and stability of periodic and localised structures in optical bistable systems*, IEEE J. Quant. Electr., 39 (2003), pp. 216–226.
[87] L. N. Trefethen, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.
[88] L. S. Tsimring and I. S. Aranson, *Localised and cellular patterns in vibrated granular layer*, Phys. Rev. Lett., 79 (1997), pp. 213–216.
[89] P. B. Umbanhowar, F. Melo, and H. L. Swinney, *Localised excitations in a vertically vibrated granular layer*, Nature, 382 (1996), pp. 793–796.
[90] V. K. Vanag and I. R. Epstein, *Stationary and oscillatory localised patterns, and subcritical bifurcations*, Phys. Rev. Lett., 92 (2004), 128301.
[91] A. G. Vladimirov, J. M. McSloy, D. V. Skryabin, and W. J. Firth, *Two-dimensional clusters of solitary structures in driven optical cavities*, Phys. Rev. E, 65 (2002), 046606.
[92] P. D. Woods and A. R. Champneys, *Heteroclinic tangles and homoclinic snaking in the unfolding of a degenerate reversible Hamiltonian-Hopf bifurcation*, Phys. D, 129 (1999), pp. 147–170.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 3, pp. 1101–1129
© 2008 Society for Industrial and Applied Mathematics

Neuronal Networks with Gap Junctions: A Study of Piecewise Linear Planar Neuron Models*

S. Coombes†

Abstract. The presence of gap junction coupling among neurons of the central nervous systems has been appreciated for some time now. In recent years there has been an upsurge of interest from the mathematical community in understanding the contribution of these direct electrical connections between cells to large-scale brain rhythms. Here we analyze a class of exactly soluble single neuron models, capable of producing realistic action potential shapes, that can be used as the basis for understanding dynamics at the network level. This work focuses on planar piecewise linear models that can mimic the firing response of several different cell types. Under constant current injection the periodic response and phase response curve (PRC) are calculated in closed form. A simple formula for the stability of a periodic orbit is found using Floquet theory. From the calculated PRC and the periodic orbit a phase interaction function is constructed that allows the investigation of phase-locked network states using the theory of weakly coupled oscillators. For large networks with global gap junction connectivity we develop a theory of strong coupling instabilities of the homogeneous, synchronous, and splay states. For a piecewise linear caricature of the Morris–Lecar model, with oscillations arising from a homoclinic bifurcation, we show that large amplitude oscillations in the mean membrane potential are organized around such unstable orbits.

Key words. piecewise linear models, gap junctions, Floquet theory, coupled-oscillator theory, phase-density function

AMS subject classification. 92C20

DOI. 10.1137/070707579

1. Introduction.

Gap junctions allow for direct communication between cells. They are typically formed from the juxtaposition of two hemichannels (connexin proteins) and allow the free movement of ions or molecules across the intercellular space separating the plasma membrane of one cell from another. Gap junction coupling is known to occur between many cell types, including, for example, pancreatic-β cells [21], heart cells [23], and astrocytes [9]. It is no understatement to say that they are now believed to be ubiquitous throughout the central nervous system [16]. Indeed it has been appreciated for some time that they exist between inhibitory neurons of the neocortex [35]. As well as being found in the neocortex [36, 4, 39, 34], they occur in many other brain regions, including the hippocampus [34], inferior olivary nucleus in the brain stem [75], the spinal cord [71], and the thalamus [47], and have recently been shown to form axo-axonic connections between *excitatory* cells in the hippocampus (on mossy fibers) [41]. Without the need for receptors to recognize chemical messengers, gap junctions are much faster than chemical synapses at relaying signals. The

*Received by the editors November 7, 2007; accepted for publication (in revised form) by B. Ermentrout June 3, 2008; published electronically September 25, 2008.
http://www.siam.org/journals/siads/7-3/70757.html
†Department of Mathematical Sciences, University of Nottingham, Nottingham, NG7 2RD, UK (stephen.coombes@nottingham.ac.uk).

synaptic delay for a chemical synapse is typically in the range 1–100 ms, while the synaptic delay for an electrical synapse may be only about 0.2 ms. There is now little doubt that gap junctions play a substantial role in the generation of neural rhythms, both functional [3, 46, 8] and pathological [82, 24], and that they may subserve system level computations [62].

The presence of gap junctional coupling in a neuronal network necessarily means that neurons directly "feel" the shape of action potentials (APs) from other neurons to which they are connected. From a modeling perspective one must therefore be careful to work with single neuron models that have an accurate representation of an AP shape. To date there is now a zoo of single neuron models that can accurately reflect these shapes for different neuronal cell types (see, for example, [49]). Typically such models, being based around that of Hodgkin–Huxley [44], are high dimensional and can often only be analyzed using perturbative techniques, such as geometric singular perturbation theory (see [72] for a review). When combined with an initial reduction of the model, say, using the techniques in [54], this has proven a remarkably powerful approach for gaining insight into single neuron behavior. However, it does not necessarily pave the way for tractable network studies. In this case, starting from a one-dimensional integrate-and-fire (IF) type model is often advocated [11]. However, since the IF model does not generate an AP shape, it must be augmented in some way as in [15, 37], leading one naturally to consider the spike-response model [38]. However, in this case the AP is considered to have a universal shape, triggered as the voltage reaches a constant voltage threshold. This does not quite capture the dynamics of a truly excitable system (with gating variables), where instead one would expect a state-dependent threshold and a variable AP shape. Thus we are naturally led to a search for planar models possessing one voltage and one gating variable that can mimic the behavior of high dimensional conductance-based models. Perhaps the most famous example of such a model is the FitzHugh–Nagumo model [31], which has many of the same characteristics as the Hodgkin–Huxley model. In this case analytical progress has been possible with one further step, namely, the introduction of piecewise linear (PWL) nullclines. This gives rise to the so-called McKean model [64], for which a number of results about the existence and stability of periodic orbits are now known [83, 84]. In this paper we introduce a broader class of PWL models that can mimic the behavior of many common cell types and describe how to analyze periodic orbits explicitly. Importantly we show that the study of such models does indeed allow for mathematical studies of the rich dynamical behavior seen in large networks with strong gap junction coupling. In this sense our work is complementary to many other theoretical studies that focus on *weak coupling* [74, 69, 7, 61, 56, 26] as well as *computational* studies with strong coupling [55, 60, 3, 76].

One of the main motivations for pursuing the work in this paper is that it may underpin the development of a tractable firing rate model of neural tissue possessing gap junctions. Necessarily this must require an understanding of strong coupling if gap induced variations in firing rate are of interest. With the exception of work by van Vreeswijk [81] (for synaptic interactions), results for strong coupling are rare. Hence, although we focus on the special case of PWL neuron models, this is useful as it allows us to gain some specific insight into dynamics in the strong coupling regime. Moreover, some of the techniques we develop here, notably for determining the stability of the asynchronous state in a strongly gap junction coupled global network, are valid not just for PWL systems but also for more general limit cycle oscillator networks. The more detailed structure of this paper is as follows. In section 2

we introduce the class of PWL models that we study throughout the paper. In particular, we focus on two distinct examples, one of which is the McKean model and the other a new PWL model that caricatures the conductance-based Morris–Lecar model with oscillations generated via a homoclinic bifurcation [68]. Next, in section 3, we show how to analyze periodic orbits that arise in such models under constant current injection. This includes the construction of orbits, the determination of their stability, and the calculation of the phase response curve (PRC) for the orbit. Stability is analyzed using Floquet theory and shown to lead to a simple formula for the nonzero Floquet exponent. Network studies are pursued in section 4 for two important cases: (i) weak coupling, and (ii) strong coupling. In the former case we show how to calculate the phase interaction function for a network in closed form using a Fourier representation. This is used to investigate phase-locked states in both small and large networks. Focusing on synchronous and splay states in globally coupled networks, we further show how to treat the case of strong coupling. Our results for existence and stability recover those of the weak coupling theory in the appropriate limit. In section 5 we use this strong coupling theory to understand large amplitude oscillations seen in the mean field signal of networks of Morris–Lecar neurons with gap junction coupling [42]. Finally, in section 6 we discuss natural extensions of the work in this paper.

2. Piecewise linear neuron models. The excitable properties of neural cells can often be summarized simply by determining their firing rate response to constant current injection. Broadly speaking one then classifies a neuron as being either Type I or Type II. Type I is obtained when repetitive APs are generated with an arbitrarily low frequency, whereas in Type II APs emerge at a nonzero frequency. In this latter case it is natural to think of oscillations as arising through a Hopf bifurcation. Indeed, from the seminal experimental and modeling work of Hodgkin and Huxley, this is known to be the case for the squid giant axon. Thus, although the original Hodgkin–Huxley model consists of four nonlinear ordinary differential equations (ODEs), it is not surprising that alternative planar models can also be invoked to fit at least the firing rate response. The classic example is the FitzHugh–Nagumo model [31], though others such as those obtained by a systematic reduction of the Hodgkin–Huxley equations are known [1]. These planar models are described by two coupled nonlinear ODEs—one for voltage and the other for a single effective gating variable. The nullcline for the voltage variable has a cubic shape typical of many excitable systems. Although powerful geometric techniques may be brought to bear on such planar models, their analysis in closed form is precluded by the presence of the cubic nonlinearity. This has motivated the introduction and study of PWL caricatures, such as the McKean model [64, 79]. The equations for a single two-dimensional McKean neuron take the form

$$C\dot{v} = f(v) - w + I, \tag{2.1}$$

$$\dot{w} = g(v, w), \tag{2.2}$$

where the functions $f(v)$ and $g(v, w)$ are given by

$$f(v) = \begin{cases} -v, & v < a/2, \\ v - a, & a/2 \le v \le (1+a)/2, \\ 1 - v, & v > (1+a)/2, \end{cases} \tag{2.3}$$

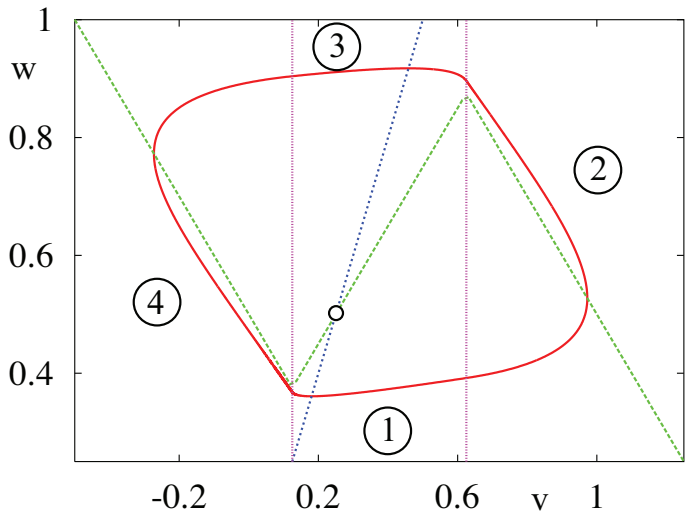


Figure 1. *The phase plane for the McKean model has a nullcline with a piecewise linear cubic shape (dashed green line) corresponding to $\dot{v}=0$ and a linear one associated with $\dot{w}=0$ (dotted blue line). Parameters are $C=0.1$, $I=0.5$, $\gamma=0.5$, and $a=0.25$. The red line corresponds to a stable periodic orbit.*

$$g(v,w)=v-\gamma w. \tag{2.4}$$

Here, $C>0$, $\gamma>0$, I is a constant drive, and $f(v)$ is a PWL caricature of the cubic FitzHugh–Nagumo nonlinearity $f(v)=v(1-v)(v-a)$, while $g(v,w)$ describes the linear dynamics of the gating variable. Another popular choice for $f(v)$ is the function $f(v)=-v+\Theta(v-a)$, where Θ is the Heaviside step function. The analysis of this latter nonlinearity has been pursued in detail by Tonnelier [78, 80]. A phase-plane plot of the McKean model is shown in Figure 1. Generating Type I behavior, often associated with either a homoclinic bifurcation or a saddle-node on an invariant cycle (SNIC) [27], necessarily requires the introduction of a nonlinear dynamics for the gating variable, as in the Morris–Lecar model or the cortical neuron model of Wilson [85]. A PWL idealization of the Morris–Lecar model has already been introduced by Tonnelier and Gerstner [80], and since the nullcline of the gating variable in the Wilson model has a quadratic shape, it too is easy to caricature. Indeed, many of the shapes for $g(v,w)$ underlying a Type I response appear to be described with the simple continuous choice

$$g(v,w)=\begin{cases}(v-\gamma_1 w+b^*\gamma_1-b)/\gamma_1, & v<b,\\ (v-\gamma_2 w+b^*\gamma_2-b)/\gamma_2, & v\geq b,\end{cases} \tag{2.5}$$

with $-a/2<b^*<(1-a)/2$ and $a/2<b<(1+a)/2$. Here we take $\gamma_2>0$, though we allow γ_1 to take both positive and negative values. Another natural choice, though this time discontinuous, is $g(v,w)=v-\gamma w+\Theta(v-b)$, which has been used to caricature the Morris–Lecar model in particular [80]. Note that (up to a constant shift) we recover the PWL McKean model with the choice $\gamma_1=\gamma=\gamma_2$ in (2.5). An example with dynamics that is bistable between a fixed point and a limit cycle is shown in Figure 2. Here the emergence of low frequency oscillations is associated with a homoclinic bifurcation, whereby the amplitude

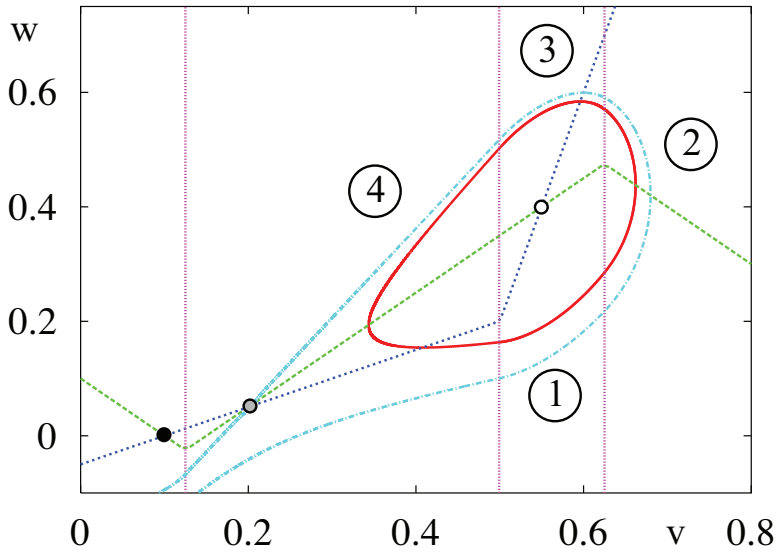


Figure 2. *The phase plane for the piecewise linear Morris–Lecar (PML) model with* $\gamma_1 = 2$, $\gamma_2 = 0.25$, $C = 0.825$, $I = 0.1$, $a = 0.25$, $b = 0.5$, *and* $b^* = 0.2$. *The pale blue line passing through the saddle (gray filled circle) is the separatrix between the stable fixed point (black filled circle) and the stable limit cycle (in red).*

of the periodic orbit grows with a decrease in I and collides with a saddle point. We regard this model as a PWL caricature of the Morris–Lecar neuron, with oscillations arising from a homoclinic bifurcation, and as such shall refer to it as the PML model. On a technical point it is important to note that it is not possible to have a smooth SNIC with a PWL model, since it would not contain any quadratic parts (necessary to define a saddle-node bifurcation). One such example would be the nonlinear IF neuron, described by Karbowski and Kopell [52] with subthreshold dynamics $\dot{v} = |v| + I$. Throughout the rest of this paper we shall work with the PWL model defined by (2.3) and (2.5), though we stress here that the techniques we develop work for all of the PWL choices for $f(v)$ and $g(v, w)$ that we have discussed above.

3. Periodic orbits. Much can be said about the dynamics of models of PWL planar systems defined by (2.3) and (2.5). For the special case that $f(v) = -v + \Theta(v - a)$ and $g(v, w) = v$ Tonnelier [78] has shown how to use the method of harmonic balance [2] to obtain information about periodic orbits. Here we present an alternative approach that can tackle more general choices for f and g. In essence we solve the system in each of its linear regimes and demand continuity of solutions to construct orbits of the full nonlinear flow. To see how we do this it is first convenient to consider a two-dimensional linear system of the form

$$\dot{z} = Az + b, \qquad z = \begin{bmatrix} v \\ w \end{bmatrix}, \tag{3.1}$$

where the 2×2 matrix A has components a_{ij}, $i, j = 1, 2$, and b is a constant 2×1 input vector. The solution to (3.1) may be written in the form

$$z(t) = G(t)z(0) + K(t)b, \qquad G(t) = \mathrm{e}^{At}, \qquad K(t) = \int_0^t G(s)\mathrm{d}s. \tag{3.2}$$

If A has real eigenvalues $\lambda_\pm$, such that $Aq_\pm = \lambda_\pm q_\pm$ with $q_\pm \in \mathbb{R}^2$, given by

$$\lambda_\pm = \frac{\mathrm{Tr}\, A \pm \sqrt{(\mathrm{Tr}\, A)^2 - 4 \det A}}{2}, \tag{3.3}$$

then we may "diagonalize" and write $G(t)$ in the computationally useful form $G(t) = P\mathrm{e}^{\Lambda t}P^{-1}$, where $\Lambda = \mathrm{diag}(\lambda_+, \lambda_-)$, $P = [q_+, q_-]$, and $q_\pm = [(\lambda_\pm - a_{22})/a_{21}, 1]^T$. If A has complex eigenvalues $\rho \pm i\omega$, then the associated complex eigenvector is q such that $Aq = (\rho + i\omega)q$, $q \in \mathbb{C}^2$. In this case $G(t) = \mathrm{e}^{\rho t}P\mathcal{R}_{\omega t}P^{-1}$, where

$$\mathcal{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad P = [\mathrm{Im}(q), \mathrm{Re}(q)] = \begin{bmatrix} 0 & 1 \\ \widehat{\omega} & \widehat{\rho} \end{bmatrix}, \tag{3.4}$$

with $\widehat{\omega} = \omega/a_{12}$ and $\widehat{\rho} = (\rho - a_{11})/a_{12}$. Note that ρ and ω may be written using the invariance of Tr and det as $\rho = (a_{11} + a_{22})/2$, $\omega^2 = a_{11}a_{22} - a_{12}a_{21} - \rho^2 > 0$. The explicit form for $G(t)$, necessary for carrying out computations, is given in Appendix A.

To specify a periodic orbit of the PWL model of choice it is convenient to break the solution into pieces such that on each piece the dynamics is governed by a linear dynamical system. As a concrete example we will focus on the type of periodic orbits shown in Figures 1 and 2. In both these examples we need only consider four distinct pieces, labeled by $\mu = 1, \ldots, 4$. We denote the time spent in each of these four states as T_μ. For each piece we write $z_\mu(t) = G_\mu(t)z_\mu(0) + K_\mu(t)b_\mu$ with the forms for G_μ and K_μ given by (3.2) under the replacement of A by A_μ. For the McKean model we have that $A_1 = A_3$, $A_2 = A_4$, where

$$A_1 = \begin{bmatrix} 1/C & -1/C \\ 1 & -\gamma \end{bmatrix}, \qquad A_2 = \begin{bmatrix} -1/C & -1/C \\ 1 & -\gamma \end{bmatrix}, \tag{3.5}$$

with

$$b_1 = \begin{bmatrix} (I-a)/C \\ 0 \end{bmatrix}, \quad b_2 = \begin{bmatrix} (1+I)/C \\ 0 \end{bmatrix}, \quad b_4 = \begin{bmatrix} I/C \\ 0 \end{bmatrix}, \tag{3.6}$$

and $b_3 = b_1$. For the PML model defined by (2.5)

$$A_1 = \begin{bmatrix} 1/C & -1/C \\ 1/\gamma_2 & -1 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} -1/C & -1/C \\ 1/\gamma_2 & -1 \end{bmatrix}, \qquad A_4 = \begin{bmatrix} 1/C & -1/C \\ 1/\gamma_1 & -1 \end{bmatrix}, \tag{3.7}$$

with

$$b_1 = \begin{bmatrix} (I-a)/C \\ b^* - b/\gamma_2 \end{bmatrix}, \quad b_2 = \begin{bmatrix} (1+I)/C \\ b^* - b/\gamma_2 \end{bmatrix}, \quad b_4 = \begin{bmatrix} (I-a)/C \\ b^* - b/\gamma_1 \end{bmatrix}, \tag{3.8}$$

$A_3 = A_1$, and $b_3 = b_1$. Introducing two voltage thresholds v_{th}^1 and v_{th}^2, where $(v_{\mathrm{th}}^1, v_{\mathrm{th}}^2) = (a/2, (1+a)/2)$ for the McKean model and $(v_{\mathrm{th}}^1, v_{\mathrm{th}}^2) = (b, (1+a)/2)$ for the PML model, means that we can parameterize a periodic orbit by choosing initial data such that $z_1(0) = (v_{\mathrm{th}}^1, w^*)$ (with w^* as yet undetermined) and $z_{\mu+1}(0) = G_\mu(T_\mu)z_\mu(0) + K_\mu(T_\mu)b_\mu$ for $\mu = 1, 2, 3$. The "times-of-flight" T_μ are determined by solving the threshold crossing conditions $v_1(T_1) = v_{\mathrm{th}}^2$, $v_2(T_2) = v_{\mathrm{th}}^2$, $v_3(T_3) = v_{\mathrm{th}}^1$, and $v_4(T_4) = v_{\mathrm{th}}^1$. A periodic solution can then be found by solving

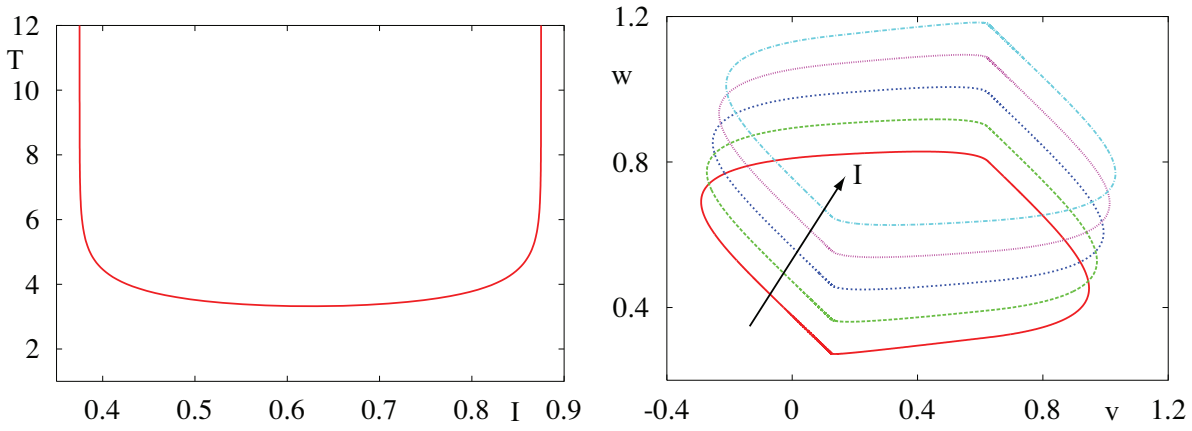


Figure 3. *McKean model. Left: Period of solution as a function of background drive I. Right: Shape of orbits for $I = 0.4, 0.5, 0.6, 0.7, 0.8$. Other parameters are as in Figure 1.*

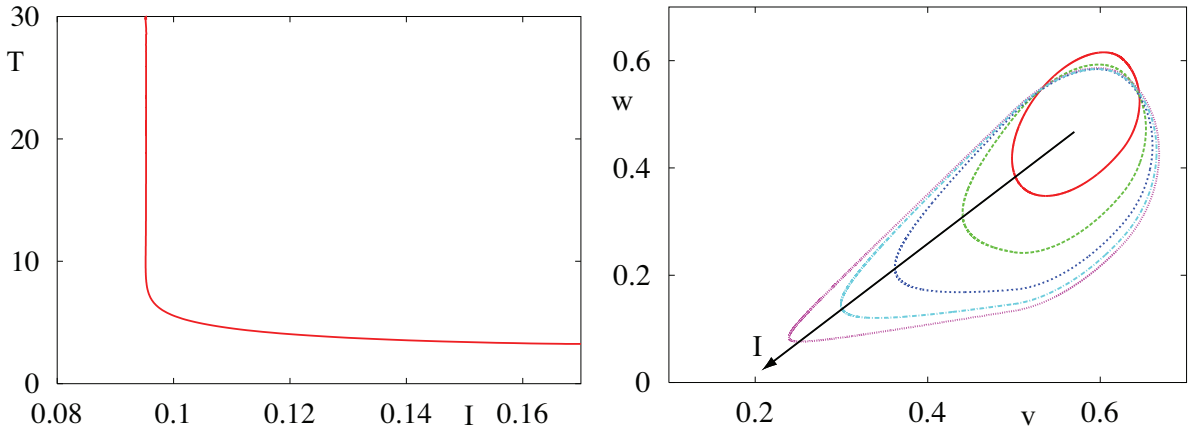


Figure 4. *PML model. Left: Period of solution as a function of background drive I. Right: Shape of orbits for I ranging from 0.17 to 0.09. Other parameters are as in Figure 2.*

$w_4(T_4) = w_1(0)$, thus yielding w^* and the period $T = \sum_{\mu=1}^{4} T_\mu$. In Figure 3 we plot the period and orbit shape as a function of the external drive I obtained using the prescription above. A similar plot for the PML model is shown in Figure 4. In contrast to the McKean model, the firing rate of the PML model at the onset of repetitive behavior increases from zero, as expected for a system with a homoclinic bifurcation.

Three other types of periodic solution are also possible. Two of these involve only a single threshold crossing—namely, one which crosses through the section $v = v_{\rm th}^1$ but not $v = v_{\rm th}^2$ and another which crosses through $v = v_{\rm th}^2$ but not $v = v_{\rm th}^1$. Calling these the *sub-* and *supra*threshold periodic orbits, respectively, we may solve each using the approach (and notation) above. The subthreshold orbit is specified by the restriction $\mu = \{1, 4\}$ with $z_1(0) = (v_{\rm th}^1, w^*)$, subject to $v_1(T_1) = v_{\rm th}^1 = v_4(T_4)$ and $w_4(T_4) = w_1(0)$, so that $T = T_1 + T_4$. The suprathreshold orbit is specified by the restriction $\mu = \{2, 3\}$ with $z_2(0) = (v_{\rm th}^2, w^*)$, subject to $v_2(T_2) = v_{\rm th}^2 = v_3(T_3)$ and $w_3(T_3) = w_2(0)$, so that $T = T_2 + T_3$. Examples of

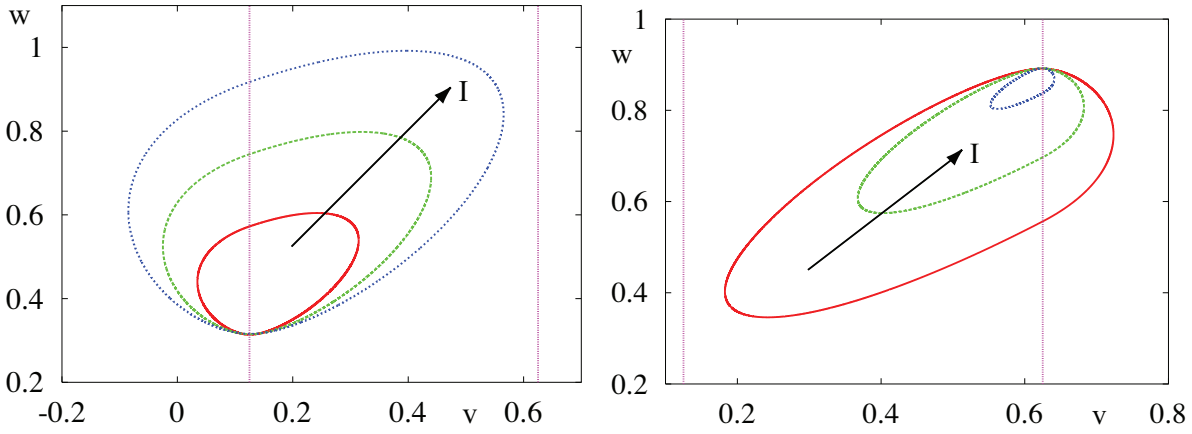


Figure 5. *McKean model. Left:* Subthreshold *orbits with* $C = 1$, $\gamma = 0.4$, *and* $I = 0.475, 0.5, 0.525$. *All these subthreshold orbits have a common period. Right:* Suprathreshold *orbits with* $C = 1$, $\gamma = 0.7$, *and* $I = 0.47, 0.49, 0.51$. *All these suprathreshold orbits have a common period. Other parameters are as in Figure* 1.

such orbits are shown in Figure 5. The final type of orbit does not cross any thresholds and is defined simply by $v(T) = v_{\rm th}$ and $v(T) = v(0)$ for some section $v_{\rm th}$ through the orbit. We shall call such an orbit *harmonic*, because its shape will be determined by a linear system of ODEs. Note, however, that it will exist only at an isolated point in parameter space, namely, where the coefficient matrix A has purely complex eigenvalues ($\mathrm{Tr}\, A = 0$, $\det A > 0$).

3.1. Phase response curve. It is common practice in neuroscience to characterize a neuronal oscillator in terms of its phase response to a perturbation. This gives rise to the notion of a so-called phase response curve (PRC). For a detailed discussion of PRCs we refer the reader to [29, 30, 45]. It suffices to say that there are three main ways to calculate PRCs, attributed to Winfree, Kuramoto, and Malkin. A nice comparison of these three approaches can be found in [50]. For concreteness we shall follow the exposition in [13] for the Malkin adjoint method. Consider a dynamical system $\dot{z} = F(z)$ with a T-periodic solution $Z(t) = Z(t+T)$ and introduce an infinitesimal perturbation Δz_0 to the trajectory $Z(t)$ at time $t = 0$. This perturbation evolves according to the linearized equation of motion:

$$\frac{\mathrm{d}\Delta z}{\mathrm{d}t} = DF(Z(t))\Delta z, \qquad \Delta z(0) = \Delta z_0. \tag{3.9}$$

Here $DF(Z)$ denotes the Jacobian of F evaluated along Z. Introducing a time-independent phase shift $\Delta\theta$ as $\theta(Z(t) + \Delta z(t)) - \theta(Z(t))$, we have to first order in Δz that

$$\Delta\theta = \langle Q(t), \Delta z(t)\rangle, \tag{3.10}$$

where $\langle\cdot,\cdot\rangle$ defines the standard inner product, and $Q = \nabla_Z\theta$ is the gradient of θ evaluated at $Z(t)$. Taking the time-derivative of (3.10) gives

$$\left\langle \frac{\mathrm{d}Q}{\mathrm{d}t}, \Delta z \right\rangle = -\left\langle Q, \frac{\mathrm{d}\Delta z}{\mathrm{d}t} \right\rangle = -\langle Q, DF(Z)\Delta z\rangle = -\left\langle DF^T(Z)Q, \Delta z\right\rangle. \tag{3.11}$$

Since the above equation must hold for arbitrary perturbations, we see that the gradient $Q = \nabla_Z \theta$ satisfies the linear equation

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = D(t)Q, \qquad D(t) = -DF^T(Z(t)), \tag{3.12}$$

subject to the conditions $\nabla_{Z(0)}\theta \cdot F(Z(0)) = 1/T$ and $Q(t) = Q(t+T)$. The first condition simply guarantees that $\dot{\theta} = 1/T$ (at any point on the periodic orbit), and the second enforces periodicity. The (vector) PRC, R, is related to Q according to the simple scaling $R = QT$. In general (3.12) must be solved numerically to obtain the PRC, say, using the *adjoint* routine in XPP [25]. However, for PWL models $DF(Z)$ is piecewise constant, and we can obtain a solution in closed form. Introducing a labeling as for the periodic orbit in section 3, we rewrite (3.12) in the form $\dot{Q}_\mu = D_\mu Q_\mu$, where $D_\mu = -A_\mu^T$. The solution of each subsystem is given by $Q_\mu(t) = G_\mu^T(T_\mu - t)Q_\mu(T_\mu)$ with $Q_\mu(T_\mu) = Q_{\mu+1}(0)$ for $\mu = 1, 2, 3$. Denoting $Q_4(T_4) = (q_1, q_2)$, we have the relation

$$\frac{q_1}{\mu}\left[f(v_{\mathrm{th}}^1) - w^* + I\right] + q_2 g(v_{\mathrm{th}}^1, w^*) = \frac{1}{T}. \tag{3.13}$$

Periodicity is ensured by choosing $Q_1(0) = Q_4(T_4)$. After introducing the 2×2 matrix $\Gamma = G_1^T(T_1)G_2^T(T_2)G_3^T(T_3)G_4^T(T_4)$, this periodicity condition takes the form

$$(\Gamma_{11} - 1)q_1 + \Gamma_{12}q_2 = 0. \tag{3.14}$$

Hence (3.13) and (3.14) define a pair of linear equations for (q_1, q_2) that we may write in the form

$$\Psi \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \begin{bmatrix} 1/T \\ 0 \end{bmatrix}, \qquad \Psi = \begin{bmatrix} (f(v_{\mathrm{th}}^1) - w^* + I)/\mu & g(v_{\mathrm{th}}^1, w^*) \\ \Gamma_{11} - 1 & \Gamma_{12} \end{bmatrix}. \tag{3.15}$$

This is easily solved, with, say, Cramer's rule, giving $q_i = \det(\Psi_i)/\det(\Psi)$, where

$$\Psi_1 = \begin{bmatrix} 1/T & g(v_{\mathrm{th}}^1, w^*) \\ 0 & \Gamma_{12} \end{bmatrix}, \qquad \Psi_2 = \begin{bmatrix} (f(v_{\mathrm{th}}^1) - w^* + I)/\mu & 1/T \\ \Gamma_{11} - 1 & 0 \end{bmatrix}. \tag{3.16}$$

Similarly we may also construct the PRCs for the sub- and suprathreshold orbits (though we omit the details here). Note that the discussion above assumes that the underlying dynamical system is described by a continuous vector field, so that we are free to choose any point on the orbit to fix the condition $\dot{\theta} = 1/T$. For discontinuous systems such as would arise in the singular limit $C = 0$ or with a discontinuous choice of $g(v, w)$, then conditions (3.13) and (3.14) are not sufficient. Techniques for tackling relaxation style oscillations that arise in the former case have been developed in [48, 17], while the latter case can easily be treated by writing down the matching conditions to fix $\dot{\theta} = 1/T$ at any jump discontinuities in $g(v, w)$. Plots of two example PRCs constructed using the above approach are shown in Figure 6.

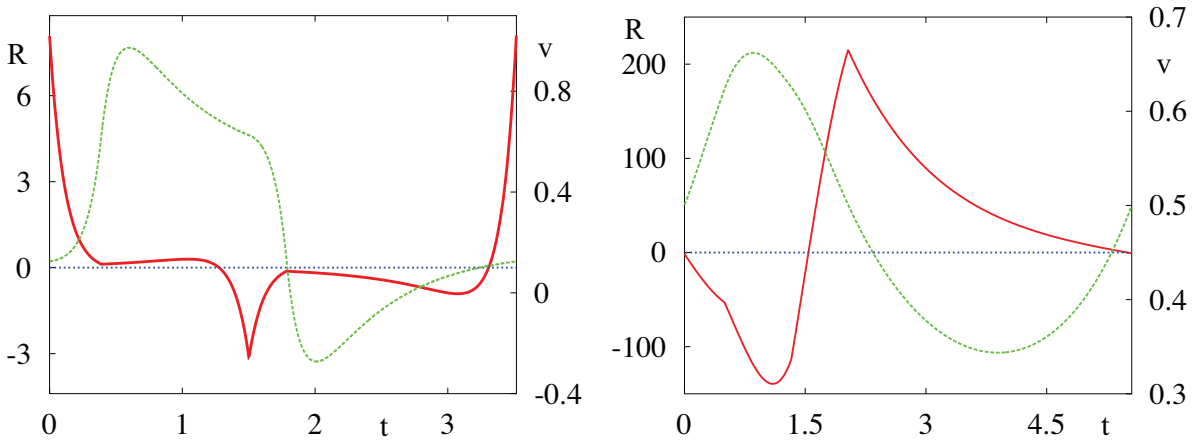


Figure 6. *PRC (first component of $Q(t)$ scaled by T). The dashed line shows the underlying shape of the periodic voltage solution. Left: McKean model PRC with parameters as in Figure 1. Right: PML model with parameters as in Figure 2.*

3.2. Stability: Floquet theory. The natural way to determine the stability of a periodic orbit is to use Floquet theory (see, for example, [14]). The linearized equations describing the evolution of perturbations around the periodic orbit are given by (3.9). Note that with the use of a time-ordering operator $\mathcal{T}$ we may write the fundamental matrix solution of this T-periodic system as

$$G(t) = \mathcal{T}\left\{\exp\left[\int_0^t DF(Z(s))\mathrm{d}s\right]\right\}, \tag{3.17}$$

where $\mathcal{T}D(t)D(s) = \Theta(t-s)D(t)D(s)+\Theta(s-t)D(s)D(t)$. Let μ_k be the (distinct) eigenvalues of $G(T)$, and write $\sigma_k = \ln(\mu_k)/T \bmod 2\pi i$. It follows that the periodic orbit will be stable if all the Floquet exponents have negative real part, namely, $\mathrm{Re}\,\sigma_k < 0$ for all $k = 1, 2$. Note that one of the Floquet exponents is always zero since it corresponds to perturbations along the periodic orbit (i.e., $\dot{Z}$ is a solution of (3.9) with a Floquet multiplier equal to unity). For PWL models time-ordering is not an issue (since DF is piecewise constant), and we have that $G(T) = G_4(T_4)G_3(T_3)G_2(T_2)G_1(T_1) = \Gamma^T$.

We now make use of the well-known result $\mu_1\mu_2 = \exp(\int_0^T \mathrm{Tr}\, DF(s)\mathrm{d}s)$ to obtain $(\sigma_1, \sigma_2) = (0, \sigma)$, where

$$\sigma = \frac{1}{T}\sum_{\mu=1} T_\mu \,\mathrm{Tr}\, A_\mu. \tag{3.18}$$

The nonzero exponent for the sub- and suprathreshold orbit is given by (3.18) with $T_2 = T_3 = 0$ and $T_1 = T_4 = 0$, respectively. For a harmonic splay state we have simply that $\sigma = \mathrm{Tr}\, A_1$. Periodic solutions are stable if $\sigma < 0$. For example, a periodic solution of the McKean model has a nonzero Floquet exponent $\sigma = (T_1 - T_2 + T_3 - T_4)/(CT) - \gamma$. Note that in the singular limit $C \to 0$ we expect $T_{1,3} \to 0$, so that $\sigma \leq 0$. Hence, any periodic orbits that persist in this limit will be stable. For the PML model $\sigma = (T_1 - T_2 + T_3 + T_4)/(CT) - 1$. Some example plots of the nonzero Floquet exponent as a function of the external drive I are shown in Figure 7.

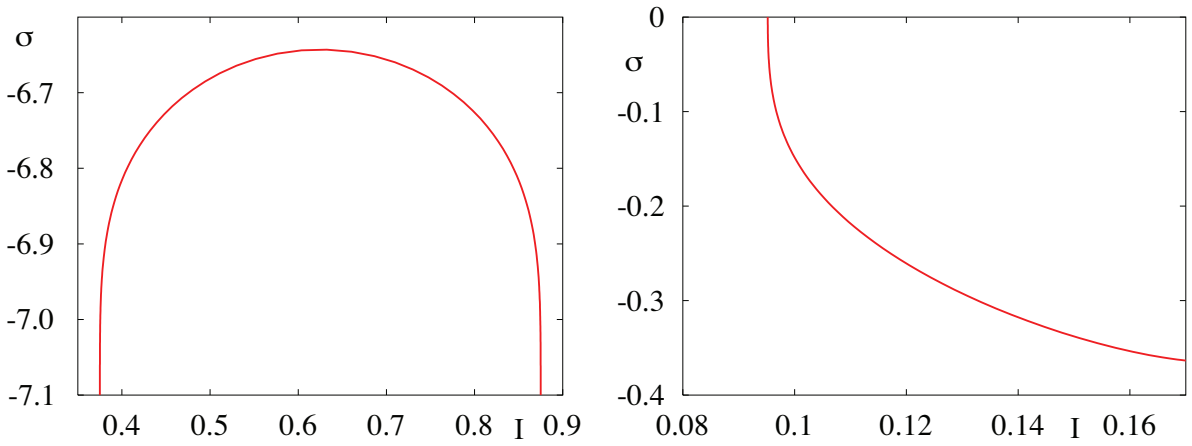


Figure 7. *Plot of the Floquet exponent σ. Left: Result for the McKean model using the solution branch of Figure 3 (left). Right: Result for the PML model using the solution branch of Figure 4 (left). Since $\sigma < 0$, the solution branches in these two examples are stable.*

4. Gap junction coupling.

To model the direct gap junction coupling between two cells, one labeled *post* and the other *pre*, we introduce an extra current to the right-hand side of (2.1) of the form

$$I_{\text{gap}} = g_{\text{gap}}(v_{\text{pre}} - v_{\text{post}}), \tag{4.1}$$

where g_{gap} is the conductance of the gap junction. Indexing neurons in a network with the label $i = 1, \ldots, N$ and defining a gap junction conductance strength between neurons i and j as g_{ij} means that neuron i experiences a drive of the form $N^{-1}\sum_{j=1}^{N} g_{ij}(v_j - v_i)$. For a phase-locked state then $z_i(t) = z(t - \phi_i T)$, $z(t) = z(t + T)$ (for some constant phases $\phi_i \in [0, 1)$), and we have N equations distinguished by the driving terms $N^{-1}\sum_{j=1}^{N} g_{ij}(v(t + (\phi_i - \phi_j)T) - v(t))$. In this section we pursue two approaches for studying networks of identical PWL neurons with such coupling terms. The first is the more familiar coupled-oscillator approach, valid for weak coupling. The second approach exploits a Fourier representation to obtain closed form solutions for splay states with arbitrary coupling strength.

4.1. Weak coupling.

The theory of weakly coupled oscillators [57, 28] is now a standard tool of dynamical systems theory and has been invoked by several authors to study networks with gap junctions [69, 61, 70, 26, 63, 53]. It has also previously been used to study networks of McKean neurons in the singular limit $C \to 0$ [17, 22]. We introduce a time-dependent phase along the T-periodic orbit of an uncoupled neuron such that $\dot{\theta}_i(t) = 1/T$ for $i = 1, \ldots, N$, with $\theta_i \in [0, 1)$. In the presence of weak coupling (small g_{ij}), the dynamics for a gap junction coupled network then takes the form

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \frac{1}{T} + \frac{1}{N}\sum_{j=1}^{N} g_{ij} H(\theta_j - \theta_i), \qquad i = 1, \ldots, N, \tag{4.2}$$

where $H(\theta)$ is the so-called phase interaction function. For gap junction coupling this is given by

$$H(\theta) = \frac{1}{T}\int_0^T Q^T(t)(v(t+\theta T) - v(t), 0)\mathrm{d}t, \tag{4.3}$$

where $v(t)$ is a periodic solution of (2.1) and (2.2), and $Q(t)$ is the associated adjoint. It is convenient to introduce Fourier series for the 2×1 vectors z and Q and write

$$z(t) = \sum_n z_n \mathrm{e}^{2\pi i n t/T}, \qquad Q(t) = \sum_n Q_n \mathrm{e}^{2\pi i n t/T}. \tag{4.4}$$

The phase interaction function then has the series representation

$$H(\theta) = \sum_n R_n v_{-n}[\mathrm{e}^{-2\pi i n\theta} - 1], \tag{4.5}$$

where v_n denotes the first component of z_n and R_n is the first component of Q_n. The Fourier coefficients z_n and Q_n may be obtained in closed form by taking Fourier transforms of the solutions for $z(t)$ and $Q(t)$. A straightforward calculation, using the forms of $z(t)$ and $Q(t)$ derived in sections 3 and 3.1, gives

$$z_n = \frac{1}{T}\sum_{\mu=1}^4 \left[\alpha_\mu^n z_\mu(0) + \gamma_\mu^n b_\mu\right] \mathrm{e}^{-2\pi i n\nu_\mu}, \qquad Q_n = \frac{1}{T}\sum_{\mu=1}^4 \beta_\mu^n Q_\mu(T_\mu)\mathrm{e}^{-2\pi i n\nu_\mu}, \tag{4.6}$$

where $(\nu_1, \nu_2, \nu_3, \nu_4) = (0, T_1, T_1+T_2, T_1+T_2+T_3)/T$ and the coefficients α_μ^n, β_μ^n, and γ_μ^n are given explicitly by

$$\alpha_\mu^n = \int_0^{T_\mu} G_\mu(t)\mathrm{e}^{-2\pi i n t/T}\mathrm{d}t, \quad \beta_\mu^n = \int_0^{T_\mu} G_\mu^T(T_\mu - t)\mathrm{e}^{-2\pi i n t/T}\mathrm{d}t, \quad \gamma_\mu^n = \int_0^{T_\mu} K_\mu(t)\mathrm{e}^{-2\pi i n t/T}\mathrm{d}t. \tag{4.7}$$

Computationally useful forms for these matrix coefficients are given in Appendix B. Writing $H(\theta)$ as the Fourier series $\sum_n H_n \mathrm{e}^{2\pi i n\theta}$, we have that $H_n = R_{-n}v_n$ for $n \neq 0$ and $H_0 = -\sum_{n\neq 0} H_n$. From the structure of (4.7) given in Appendix B we see that the Fourier coefficients for the orbit and the response function decay as $1/n$, and hence those of the phase interaction function decay as $1/n^2$. Examples of phase interaction functions constructed using the above prescription are shown in Figure 8.

We define a phase-locked solution to be of the form $\theta_i(t) = \phi_i + \Omega t$, where ϕ_i is a constant phase and Ω is the collective frequency of the coupled oscillators. Substitution into the averaged system (4.2) gives

$$\Omega = \frac{1}{T} + \frac{1}{N}\sum_{j=1}^N g_{ij} H(\phi_j - \phi_i), \qquad i = 1, \ldots, N. \tag{4.8}$$

After choosing some reference oscillator, these N equations determine the collective frequency Ω and $N-1$ relative phases. In order to analyze the local stability of a phase-locked solution $\Phi = (\phi_1, \ldots, \phi_N)$, we linearize the system by setting $\theta_i(t) = \phi_i + \Omega t + \widetilde{\theta}_i(t)$ and expand to first order in $\widetilde{\theta}_i$ to obtain

$$\frac{\mathrm{d}\widetilde{\theta}_i}{\mathrm{d}t} = \frac{1}{N}\sum_{j=1}^N \widehat{\mathcal{H}}_{ij}(\Phi)\widetilde{\theta}_j, \qquad \widehat{\mathcal{H}}_{ij}(\Phi) = g_{ij}H'(\phi_j - \phi_i) - \delta_{i,j}\sum_{k=1}^N g_{ik}H'(\phi_k - \phi_i), \tag{4.9}$$

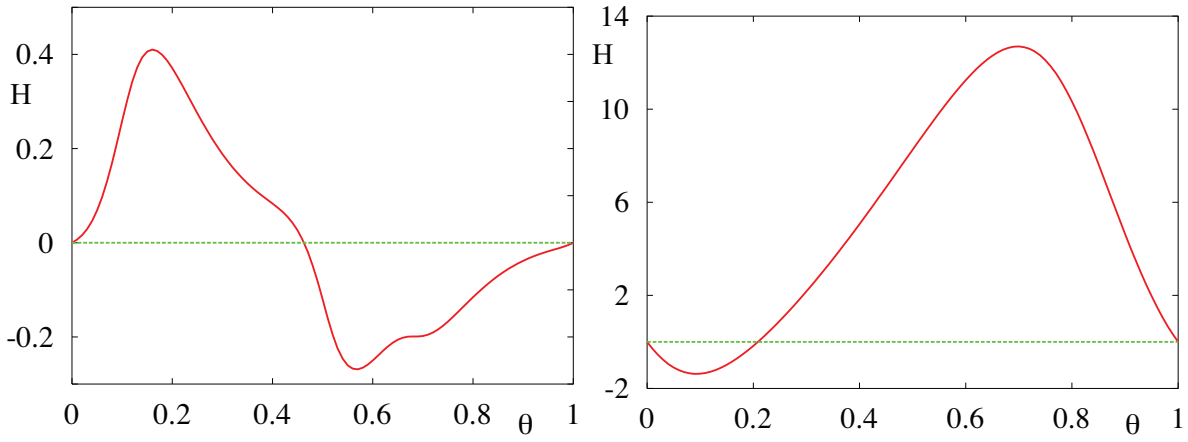


Figure 8. *Phase interaction functions corresponding to Figure 6. Left: McKean model. Right: PML model.*

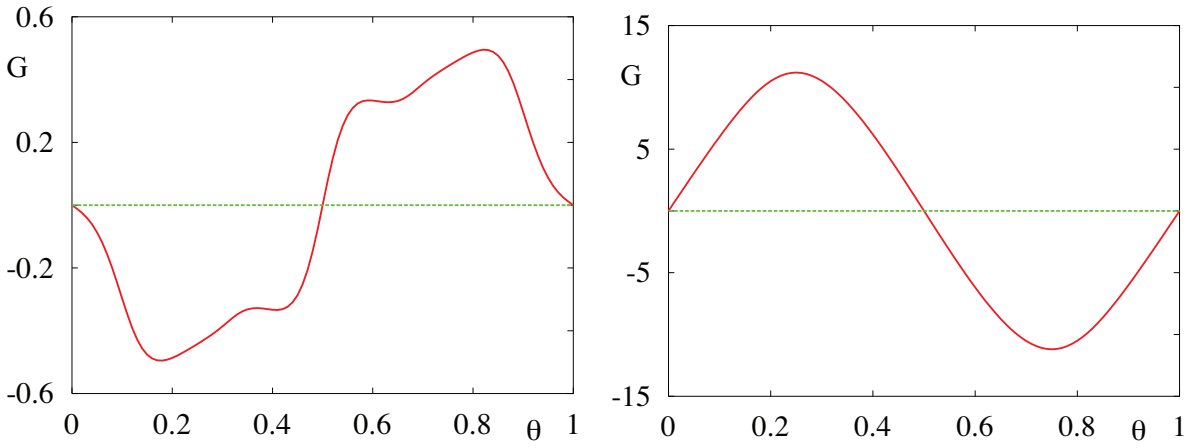


Figure 9. *Phase interaction function between two neurons; $G(\theta) = H(-\theta) - H(\theta)$ corresponding to Figure 8. Left: McKean model. Right: PML model.*

where $H'(\phi) = \mathrm{d}H(\phi)/\mathrm{d}\phi$. One of the eigenvalues of the Jacobian $\widehat{\mathcal{H}}$ is always zero, and the corresponding eigenvector points in the direction of the flow, that is, $(1, 1, \ldots, 1)$. The phase-locked solution will be stable provided that all other eigenvalues have a negative real part. For two neurons, with $g_{ij} = g$, a phase-locked state is defined by $G(\phi) = 0$, where $G(\phi) = g[H(-\phi) - H(\phi)]$ and ϕ is the relative phase between the two. The condition for stability is simply $G'(\phi) < 0$. By symmetry the phase-locked state ($\phi = 0$) and the antisynchronous state ($\phi = 1/2$) are guaranteed to exist. In Figure 9 we plot $G(\phi)$ for the phase interaction functions of Figure 8. In this example we see that the McKean model admits a stable synchronous solution, while the PML model admits a stable antisynchronous solution.

For globally coupled networks with $g_{ij} = g$ the system (4.2) is $\mathbf{S}_N \times \mathbf{T}^1$ equivariant. By the equivariant branching lemma maximally symmetric solutions describing synchronous, splay, and cluster states are expected to be generic [5]. For the synchronous state, defined by $\phi_i(t) = 0$, the collective frequency is given simply as $\Omega = 1/T$, and $\widehat{\mathcal{H}}_{ij}(\Phi) = gH'(0)[1 - N\delta_{ij}]$.

Hence, there is a single zero eigenvalue and an eigenvalue $\lambda = -gH'(0)$ of multiplicity $N-1$. For the examples in Figure 8 we see that the McKean model has a stable synchronous solution, while the PML model, with oscillations generated by a homoclinic bifurcation, does not. If the underlying single neuron model has an oscillation generated by a SNIC bifurcation (for which the PRC is well known), then synchrony is stable [26, 50]. For a splay state of the form $\phi_i = i/N$ the eigenvalues of $\widehat{\mathcal{H}}$ are given by $\lambda_n = g\sum_j H'(j/N)(\mathrm{e}^{2\pi inj/N}-1)/N$ for $n = 0,\dots,N-1$. Such solutions are often called merry-go-round states, since all oscillators in the network pass through some fixed phase at regularly spaced time intervals of T/N. For a recent review of the stability of cluster states (in which subsets of the oscillator population synchronize, with oscillators belonging to different clusters behaving differently) we refer the reader to [40, 12]. We shall not focus on them further here.

In the limit $N \to \infty$ we have the useful result that (for global coupling) network averages may be replaced by time averages:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} F(jT/N) = \frac{1}{T}\int_0^T F(t)\mathrm{d}t = F_0 \tag{4.10}$$

for some T-periodic function $F(t) = F(t+T)$. Hence in the large N limit the collective frequency of a splay state (global coupling) is given by $\Omega = 1/T + gH_0$, with eigenvalues

$$\lambda_n = \frac{g}{T}\int_0^T H'(t/T)\mathrm{e}^{2\pi int/T}\mathrm{d}t = -2\pi ingH_{-n}. \tag{4.11}$$

Hence a splay state is stable if $-ng\,\mathrm{Im}\,H_n < 0$, where we have used the fact that since $H(\theta)$ is real, then $\mathrm{Im}\,H_{-n} = -\mathrm{Im}\,H_n$. A numerical examination of the eigenvalues (4.11) (using the analytical expressions for H_n obtained via (4.6)) for the phase interaction functions shown in Figure 8 shows that the splay state is unstable for both these examples. One natural way to stabilize the splay state is to include some synaptic coupling as in the work of [37, 26]. Another mechanism is to include noise, as originally noted by Kuramoto [57]. If we consider the addition of zero mean white noise with variance σ^2 to the voltage dynamics, then the phase-reduced system also feels an additive zero mean white noise source, though with variance σ_θ^2 given by $\sigma_\theta^2 = \sigma^2\int_0^T [R(t)]^2\,\mathrm{d}t/T$. For a globally coupled network the asynchronous state is stable if $-ng\,\mathrm{Im}\,H_n < \sigma_\theta^2 n^2$ for all $n \neq 0$ [58]. This nicely shows us that if the eigenvalues associated with the deterministic model stray slightly into the right-hand complex plane, then a small amount of noise can be used to compensate and restabilize the splay state. However, since this is an argument that relies upon weak coupling, then it can necessarily work only if the unstable eigenvalues are sufficiently close to the imaginary axis.

4.2. Beyond weak coupling. Here we develop techniques for the study of the synchronous and splay states in the strong coupling regime for global coupling ($g_{ij} = g$ for all i,j). First, we show how to construct such solutions, extending techniques used in our weak coupling analysis, and use this to explore the effect of gap junction strength on network firing rates. Second, we show how to analyze the stability of the synchronous solution using Floquet theory and that of the splay state using a phase-density formalism.

4.2.1. Existence and stability of a synchronous state. A synchronous network solution exists whenever a periodic orbit for an isolated oscillator exists ($g = 0$) and has the period of the uncoupled isolated oscillator. However, stability will depend on the value of coupling g. For convenience we define a matrix $A_\mu(g) = A_\mu - B(g)$ with

$$B(g) = \frac{g}{C}J, \qquad J = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.12}$$

Following reasoning similar to that in [40] the $2N$ Floquet multipliers are given as the eigenvalues of a $2N \times 2N$ matrix $G(T)$, where $G(T)$ has the form of (3.17), with DF given by

$$DF = \begin{bmatrix} A(g) + \frac{1}{N}B(g) & \frac{1}{N}B(g) & \frac{1}{N}B(g) & \dots & \frac{1}{N}B(g) \\ \frac{1}{N}B(g) & A(g) + \frac{1}{N}B(g) & \frac{1}{N}B(g) & \dots & \frac{1}{N}B(g) \\ \vdots & & \ddots & & \\ \frac{1}{N}B(g) & \frac{1}{N}B(g) & \dots & A(g) + \frac{1}{N}B(g) & \frac{1}{N}B(g) \\ \frac{1}{N}B(g) & \frac{1}{N}B(g) & \dots & \frac{1}{N}B(g) & A(g) + \frac{1}{N}B(g) \end{bmatrix}, \tag{4.13}$$

where $A(g) = A_\mu(g)$. Here μ is chosen according to $\mu = \mu_1$ if $t \in [0, T_1)$, $\mu = \mu_2$ if $t \in [T_1, T_2)$, $\mu = \mu_3$ if $t \in [T_2, T_3)$, and $\mu = \mu_4$ if $t \in [T_3, T_4)$, defining four distinct *phases* of the orbit. On each of these four phases $DF = DF_\mu$ is independent of time. The matrix DF_μ is block circulant, with a generating row given by $[A_\mu(g) + B(g)/N \; B(g)/N \; \dots \; B(g)/N]$, and can be diagonalized by Fourier transform. Introducing the 2-component vector q_n, $n = 0, \dots, N-1$, the components of the eigenvectors of DF_μ can be listed as a set of N 2-component vectors with entries

$$q_n e^{2\pi i n m/N}, \qquad n, m = 0, \dots, N-1, \tag{4.14}$$

where q_0 is an eigenvector of A_μ and $q_{n\neq 0}$ is an eigenvector of $A_\mu(g)$. Hence, we may calculate $G(T) = G_4(T_4)G_3(T_3)G_2(T_2)G_1(T_1)$ using the representation $G_\mu(T_\mu) = P_\mu \exp(\Lambda_\mu T_\mu) P_\mu^{-1}$, where P_μ is the matrix of eigenvectors of DF_μ and Λ_μ is the corresponding diagonal matrix of eigenvalues, comprising the two eigenvalues of A_μ and $N-1$ copies of the two eigenvalues of $A_\mu(g)$. However, motivated by these observations, there is a much simpler way of calculating the Floquet multipliers that avoids the computation of $G(T)$ (and its eigenstructure).

We write the N-dimensional linearized system of equations in the form $\dot{Z} = DF\, Z$. First consider the two-dimensional system $\dot{z} = A(g)z$. By direct inspection we see that

$$Z = \begin{bmatrix} z \\ -z \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \begin{bmatrix} z \\ 0 \\ -z \\ \vdots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} z \\ 0 \\ 0 \\ \vdots \\ -z \end{bmatrix} \tag{4.15}$$

are linearly independent solutions of $\dot{Z} = DF\, Z$. Hence the two Floquet multipliers of $\dot{z} = A(g)z$ are also Floquet multipliers of $\dot{Z} = DF\, Z$, with $N-1$ degeneracy. Now con-

sider the two-dimensional system $\dot{z} = Az$. Again by direct inspection we see that

$$Z = \begin{bmatrix} z \\ z \\ z \\ \vdots \\ z \end{bmatrix} \tag{4.16}$$

is also a solution of $\dot{Z} = DF\,Z$, with two Floquet multipliers. Hence we can account for all the $2N$ Floquet multipliers (including the one which is unity). Using the analysis of section 3.2 the three relevant Floquet exponents are given by (3.18) and the pair (σ_1, σ_2), where $\sigma_k = \ln(\mu_k)/T \bmod 2\pi i$. Here the μ_k are the two (distinct) eigenvalues of $G(T) = G_4(T_4)G_3(T_3)G_2(T_2)G_1(T_1)$, where $G_\mu(T_\mu) = \exp(A_\mu(g)T_\mu)$ is a 2×2 matrix. Hence the synchronous network state is stable if an uncoupled isolated oscillator has a stable periodic orbit ($\sigma < 0$, and see (3.18)) and if the absolute values of the eigenvalues of $G(T)$ are less than unity. The condition for eigenvalues to cross the unit circle along the real axis is $\det[G(T) \pm I] = 0$, and off of the real axis we have the condition $\det G(T) = 1$. This latter condition is equivalent to $\sum_\mu T_\mu \operatorname{Tr} A_\mu(g) = 0$. For the examples in section 4.1 we find that the McKean model supports a stable synchronous state for weak coupling and that this stability persists with increasing g. For the PML model the synchronous state is unstable for weak g and can restabilize with increasing g when $\det[G(T) - I] = 0$. For the parameters of Figure 2 this occurs at $g \sim 0.45$.

4.2.2. Existence and stability of a splay state. Here we will focus on a globally coupled network in the large N limit. We first rewrite the coupling term for a splay state, $(v_i(t), w_i(t)) = (v(t - iT/N),\, w(t - iT/N))$ with $(v(t), w(t)) = (v(t + T),\, w(t + T))$, as

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} v(t + jT/N) = \frac{1}{T} \int_0^T v(t)\mathrm{d}t, \tag{4.17}$$

which is independent of both i and t. Hence, for a splay state every neuron in the network is described by the same dynamical system, namely,

$$C\dot{v} = f(v) - gv - w + I + gv_0, \qquad \dot{w} = g(v, w), \tag{4.18}$$

where $v_0 = T^{-1} \int_0^T v(t)\mathrm{d}t$. We note that because of the dependence on v_0 (4.18) is an advanced-retarded differential delay equation (see Appendix D for a general numerical method of solution). In the notation of section 3 we write $\dot{z}_\mu = A_\mu(g)z_\mu + b_\mu(g)$, where $b_\mu(g) = b_\mu + b$, with

$$b = \frac{g}{C} J z_0, \qquad z_0 = \begin{bmatrix} v_0 \\ w_0 \end{bmatrix}. \tag{4.19}$$

The same techniques as deployed in section 3 can be invoked to obtain a formal solution describing a T-periodic orbit. In doing so we see that the generalization of (4.6) expresses the

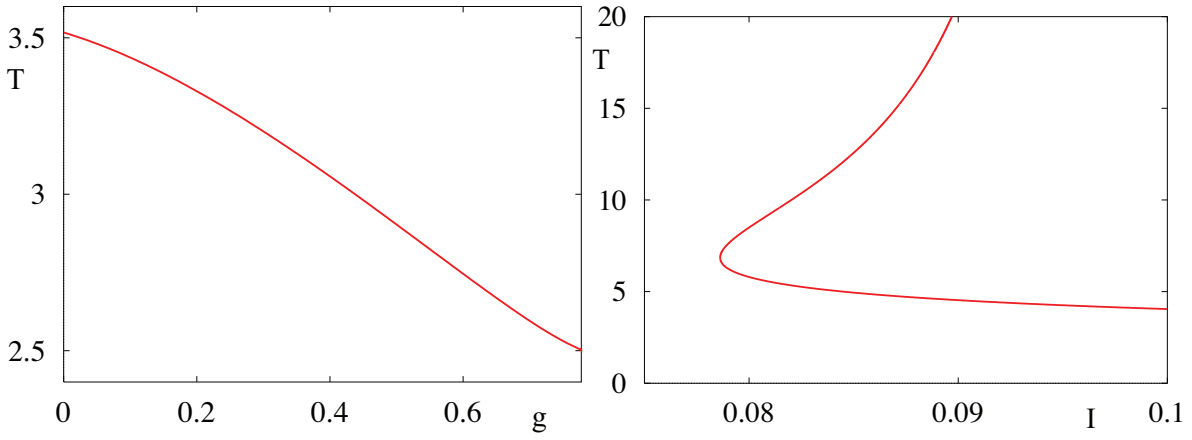


Figure 10. *Period of the splay state as a function of the coupling strength g. Left: McKean model with parameters as in Figure 1. Right: Period of PML model with g = 0.1 as a function of drive I and other parameters as in Figure 2. Note the coexistence of long and short period splay states.*

solution in terms of itself via the dependence of $b_\mu(g)$ on the Fourier component z_0. Setting $n = 0$ in this equation gives a self-consistent expression for z_0 given by

$$z_0 = \frac{1}{T}\sum_{\nu=1}^{4}\left\{\alpha_\nu^0(g)z_\nu(0) + \gamma_\nu^0(g)b_\nu(g)\right\}, \tag{4.20}$$

where $\alpha_\mu^n(g)$ and $\gamma_\mu^n(g)$ are the natural generalizations of α_μ^n and γ_μ^n (obtained under the replacement of $G_\mu(t) = \exp(A_\mu t)$ by $\exp(A_\mu(g)t)$ in (4.7)). Equation (4.20) may be rearranged to obtain an explicit equation for z_0 in the form $z_0 = Mz_1(0)$, where the 2×2 matrix M is a function of system parameters and the unknowns w^* and T_μ. The threshold crossing conditions may then be solved for as before to determine w^* and T_μ. The elements of α_μ^0 are given explicitly by $K_\mu(T_\mu)$, and those of γ_μ^0 are given in Appendix C. The dependence of the period on the strength of coupling g is shown in Figure 10. Typically we find that if a splay state exists for $g = 0$, then with increasing g its period decreases. However, in some parameter regimes it can also begin to increase again, as originally noted in [26]. Interestingly for the PML model it is easy to find parameter regimes where there is a coexistence of solutions, as in Figure 10, right. Note that in this example with $I < 0.09$ (where there are two solutions) the splay state does not exist at $g = 0$ so that weak coupling theory cannot tell us anything about either existence or stability.

In general the stability of a phase-locked state can be determined by determining the $2N$ Floquet exponents of the linearized system. Indeed, pursuing this approach for a splay state, we would find a similar coefficient matrix as in (4.13) with diagonal entries not equal to each other, but rather phase shifted, making analytical progress more cumbersome. However, for large N we may pursue an alternative phase reduction technique for networks of limit cycle oscillators with synaptic coupling developed by van Vreeswijk [81] and later used to study resonate-and-fire networks [66]. To do this we first write the coupling term $N^{-1}\sum_{j=1}^{N} v_j(t)$ in a more convenient form for studying perturbations of the mean field; namely, we write

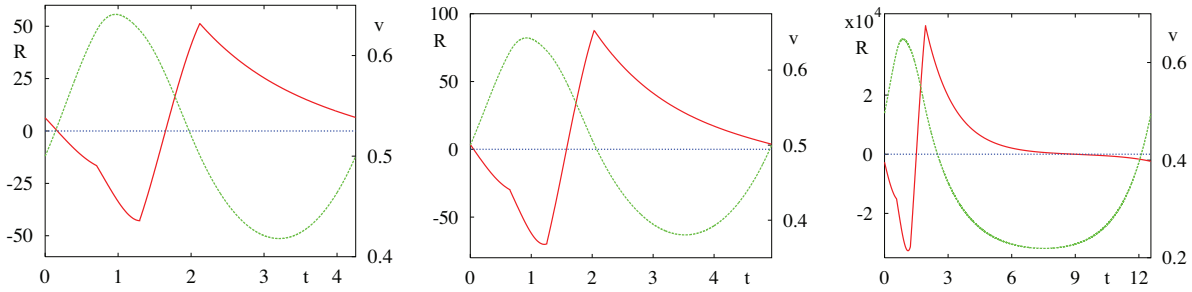


Figure 11. *The PRC of the splay state for the PML model for three different points on the solution branch shown in Figure 10, right. Left:* $I = 0.095$. *Middle:* $I = 0.085$, *lower branch. Right:* $I = 0.085$, *upper branch.*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}v_j(t)=\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}\sum_{m\in\mathbb{Z}}u(t-T_j^m), \tag{4.21}$$

where $T_j^m = mT + jT/N$. Here $u(t) = 0$ for $t < 0$ and is chosen such that $v(t) = \sum_{m\in\mathbb{Z}} u(t - mT)$, ensuring that $v(t) = v(t + T)$. For arbitrary values of T_j^m the coupling term (4.21) is time-dependent, and we may write it in the form

$$E(t)=\int_0^\infty f(t-s)u(s)\mathrm{d}s,\qquad f(t)=\lim_{N\to\infty}\frac{1}{N}\sum_{j,m}\delta(t-T_j^m), \tag{4.22}$$

where we recognize $f(t)$ as a firing rate. We now consider perturbations of the mean field such that $E(t)$ (the average membrane voltage) is split into a stationary part (arising from the splay state) and an infinitesimal perturbation. Namely, we write $E(t) = v_0 + \epsilon(t)$, with small $\epsilon(t)$. Since this perturbation to the oscillator defined by (4.18), the *splay* oscillator, is small, we may use phase reduction techniques to study the stability of the splay state.

In terms of a phase $\theta \in [0, 1)$ along the asynchronous state, we may write the evolution of this phase variable in response to a perturbation in the mean field as

$$\frac{\mathrm{d}\theta}{\mathrm{d}t}=\frac{1}{T}+g\Gamma(\theta)\epsilon(t), \tag{4.23}$$

where $\Gamma(\theta)$ is the (g-dependent) voltage component of the adjoint for the splay oscillator. This can again be calculated in closed form using the techniques developed in section 3.1. Some examples of splay state PRCs are shown in Figure 11. In fact, we need to treat N phase variables θ_i, each described by an equation of the form (4.23), which are coupled by the dependence of $\epsilon(t)$ on these variables. To make this more explicit we write

$$\epsilon(t)=\int_0^\infty \delta f(t-s)u(s)\mathrm{d}s \tag{4.24}$$

and use a phase-density description to calculate the dependence of the perturbed firing rate δf on the phases. We define a phase-density function as the fraction of neurons in the interval $[\theta, \theta + \mathrm{d}\theta]$, namely, $\rho(\theta, t) = N^{-1}\sum_j \delta(\theta_j(t) - \theta)$. Introducing the flux $J(\theta, t) = \rho(\theta, t)\dot{\theta}$, we

have the continuity equation

$$\frac{\partial \rho}{\partial t} = -\frac{\partial J}{\partial \theta}, \tag{4.25}$$

with boundary condition $J(1,t) = J(0,t)$. The firing rate is the flux through $\theta = 1$, so that $f(t) = J(1,t)$. Considering perturbations around the splay state, $(\rho, J) = (1, T^{-1})$, means writing $\rho(\theta,t) = 1 + \delta\rho(\theta,t)$, with a corresponding perturbation of the flux that takes the form $\delta J(\theta,t) = \delta\rho(\theta,t)/T + g\Gamma(\theta)\epsilon(t)$. In fact, the analysis that follows applies to the *asynchronous* state and not just the splay state. The distinction between the splay and asynchronous states is subtle; in the splay state, the phases are distributed along a cycle with phase differences of $1/N$ between two adjacent phases. In the asynchronous state, the definition is simply $\rho(\theta,t) = \rho_0(\theta)$, namely, that the phase density function is independent of time.

Differentiation of $\delta J(\theta,t)$ gives the partial differential equation

$$\partial_t \delta J(\theta,t) = -\frac{1}{T}\partial_\theta \delta J(\theta,t) + g\Gamma(\theta)\epsilon'(t), \tag{4.26}$$

where

$$\epsilon(t) = \int_0^\infty u(s)\delta J(1, t-s)\mathrm{d}s. \tag{4.27}$$

Assuming a solution of the form $\delta J(\theta,t) = \mathrm{e}^{\lambda t}\delta J(\theta)$ gives

$$\epsilon(t) = \delta J(1)\mathrm{e}^{\lambda t}\widetilde{u}(\lambda), \tag{4.28}$$

where $\widetilde{u}(\lambda) = \int_0^\infty u(t)\mathrm{e}^{-\lambda t}\mathrm{d}t$ is the Laplace transform of $u(t)$. In this case $\epsilon'(t) = \lambda\epsilon(t)$. Equation (4.26) then reduces to the ODE

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\delta J(\theta)\mathrm{e}^{\lambda T\theta} = g\lambda T\Gamma(\theta)\delta J(1)\widetilde{u}(\lambda)\mathrm{e}^{\lambda T\theta}. \tag{4.29}$$

Integrating (4.29) from $\theta = 0$ to $\theta = 1$ and using the fact that $\delta J(1) = \delta J(0)$ yields an implicit equation for λ as

$$\frac{1}{\widetilde{u}(\lambda)} - \frac{g\lambda T}{\mathrm{e}^{\lambda T} - 1}\int_0^1 \Gamma(\theta)\mathrm{e}^{\lambda\theta T}\mathrm{d}\theta = 0. \tag{4.30}$$

By taking the Laplace transform of $v(t) = \sum_m u(t - mT)$, we have that

$$\widetilde{u}(\lambda) = (1 - \mathrm{e}^{-\lambda T})\widetilde{v}(\lambda). \tag{4.31}$$

Hence, we may write λ as the solution to $\mathcal{E}(\lambda) = 0$, where

$$\mathcal{E}(\lambda) = \frac{\mathrm{e}^{\lambda T}}{\widetilde{v}(\lambda)} - g\lambda T\int_0^1 \Gamma(\theta)\mathrm{e}^{\lambda\theta T}\mathrm{d}\theta. \tag{4.32}$$

Since $1/\widetilde{v}(0) = 0$, we see that $\mathcal{E}(0) = 0$, as expected. Writing $\lambda = \nu + i\omega$, we may find the pair (ν,ω) by the simultaneous solution of $\mathcal{E}_{\mathrm{R}}(\nu,\omega) = 0$ and $\mathcal{E}_{\mathrm{I}}(\nu,\omega) = 0$, where $\mathcal{E}_{\mathrm{R}}(\nu,\omega) =$

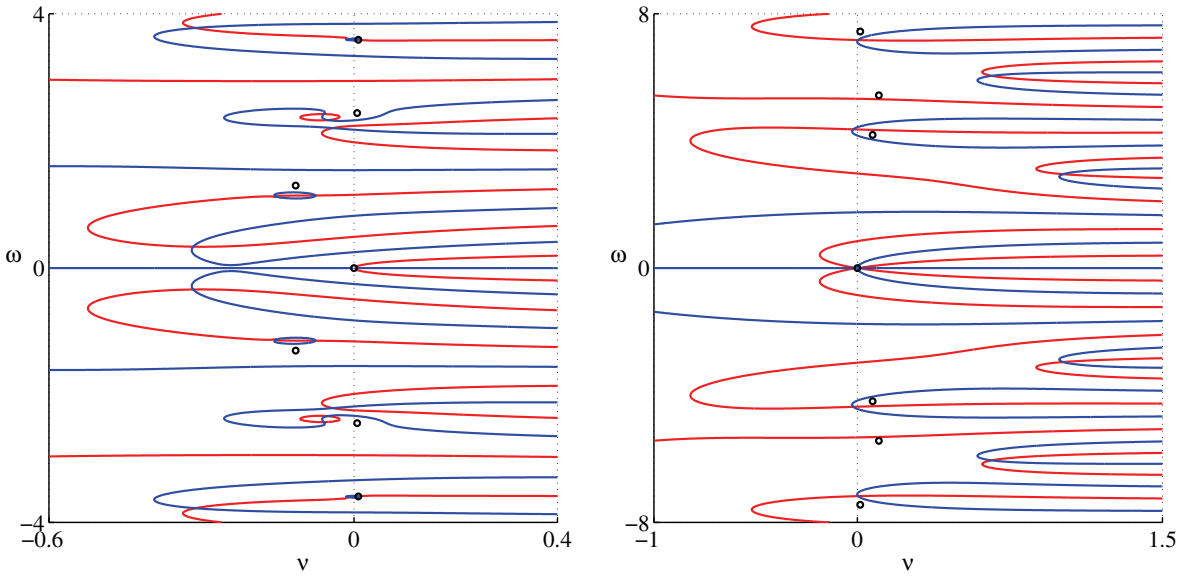


Figure 12. *Spectrum for the splay state in the PML model. Eigenvalues are at the positions where the red and blue curves intersect. The small circles denote the predictions from weak coupling theory. Parameters are as in Figure 2. Left: $I = 0.1$, $g = 0.01$. Note the unstable mode with $\omega \sim \pm 3.6$. As expected, eigenvalues from the weak coupling theory are close to the zeros of the full stability function. Right: Spectrum for the splay state with $g = 0.1$, $C = 0.9$, $I = 0.085$. In this case predictions from weak coupling theory break down. Note the occurrence of a double-zero eigenvalue signaling a bifurcation to a branch of solutions with $T_2 = 0$ (i.e., orbits tangential to $v = v_{th}^2$).*

$\mathrm{Re}\,\mathcal{E}(\nu + i\omega)$ and $\mathcal{E}_{\mathrm{I}}(\nu, \omega) = \mathrm{Im}\,\mathcal{E}(\nu + i\omega)$. In terms of the Fourier coefficients for $\Gamma(\theta)$ and $v(t)$, we may obtain a useful representation for (4.32) using

$$\int_0^1 \Gamma(\theta) \mathrm{e}^{\lambda\theta T} \mathrm{d}\theta = (\mathrm{e}^{\lambda T} - 1) \sum_n \frac{R_n}{2\pi i n + \lambda T}, \tag{4.33}$$

$$\widetilde{v}(\lambda) = T \sum_n \frac{v_{-n}}{2\pi i n + \lambda T}. \tag{4.34}$$

Examples of the spectrum obtained from the zeros of $\mathcal{E}(\lambda)$, for the PML model, are shown in Figure 12. In all cases we find the splay state is unstable.

For small g we expect to recover the stability result obtained using weakly coupled oscillator theory (see section 4.1). To check this we consider solutions of the form $2\pi i n + \lambda T = 2\pi i n g R_n v_{-n} T$ for $n \neq 0$ and $g \ll 1$. In this case we have that

$$\frac{\mathcal{E}(\lambda)}{g} = \frac{1}{\sum_n 1/(2\pi i n R_n)} - (\lambda T)^2 \sum_n \frac{R_n}{2\pi i n + \lambda T}. \tag{4.35}$$

Using the facts that R_n decays as $1/n$ (and so is an odd function of n) and λ scales with g, we may write

$$\sum_{n,m} \frac{1}{2\pi i n R_n} \frac{R_m}{2\pi i m + \lambda T} \approx \sum_n \frac{1}{2\pi i n (2\pi i n + \lambda T)} = \frac{1}{\mathrm{e}^{\lambda T} - 1} \int_0^1 S(\theta) \mathrm{e}^{\lambda\theta T} d\theta, \tag{4.36}$$

where we introduce the function $S(\theta) = \sum_n S_n e^{2\pi i n\theta}$, with $S_n = 1/(2\pi i n)$. Recognizing $S(\theta)$ as the Fourier series for the sawtooth function $S(\theta) = S(\theta+1)$ with $S(\theta) = -\theta$ for $\theta \in [0,1)$, we may evaluate (4.36) as $1/(\lambda T)^2$. Using (4.36) in (4.35) shows that $\mathcal{E}(\lambda_n) = 0$, where $\lambda_n = 2\pi i n - 2\pi i n g H_{-n}$, and we recover the stability condition for weak coupling, namely, $-ng \operatorname{Im} H_n < 0$ for $n \neq 0$.

5. Mean field rhythms. We are now in an ideal position to explore observations of Han, Kurrer, and Kuramoto [42, 43] concerning large amplitude oscillations seen in the average membrane potential of globally gap junction coupled Morris–Lecar networks. These novel rhythms were interpreted as cyclic transitions between coherent and incoherent network states and described as "bursting." However, to distinguish this from the type of behavior commonly associated with fast-slow systems [18], we shall not use this terminology here. For the rest of this section we focus on the PML model. We begin our discussion by analyzing the homogeneous fixed point behavior of the network. Using arguments similar to those in section 4.2.1, we may easily construct conditions for the stability of the fixed point (v^{ss}, w^{ss}). Considering the case $b < v^{ss} < (1+a)/2$, we have that

$$v^{ss} = \frac{a - I + b_s - b/\gamma_2}{1 - g - 1/\gamma_2}, \qquad w^{ss} = \frac{v^{ss}}{\gamma_2} + b_s - \frac{b}{\gamma_2}. \tag{5.1}$$

If it exists, this homogeneous steady state is independent of C. The conditions for network stability of this homogeneous state are $\operatorname{Tr} A_1 = 1/C - 1 < 0$ and $\operatorname{Tr} A_1(g) = (1-g)/C - 1 < 0$. Thus a homogeneous network state (if it exists) is stable only for $C > 1$ and unstable otherwise. Note, however, that when $\operatorname{Tr} A_1(g) = 0$, namely, $C = C_H = 1 - g$, we expect the existence of a harmonic splay state (since the dynamics is governed by a purely linear system with imaginary eigenvalues). Generically the results in sections 4.2.1 and 4.2.2 show that both the synchronous and splay states will be unstable for the PML model. However, knowledge of these states and the stability of the network steady state can be used to understand the original observations in [42, 43] regarding oscillations in the mean membrane potential. These authors suggested that such states could be viewed as being pushed and pulled between the unstable synchronous state and the unstable fixed point. However, in light of the work presented here, we now see that such oscillations may also occur as oscillations around an unstable orbit that can be either a fixed point or a splay state. We illustrate this idea with the aid of Figure 13. In each of the upper panels we show plots of the unstable orbits that exist for $C < C_H$, $C = C_H$, and $C > C_H$, with $C < 1$. For $C < C_H$ there is an unstable splay state that "sits" between the unstable synchronous state and the unstable homogeneous steady state. Direct numerical simulations show that the network fluctuates around the splay state, cycling between the other two unstable states. A similar behavior occurs at $C = C_H$, though the network can fluctuate around and between three coexisting unstable splay states. For $C > C_H$ the network dynamics fluctuates around the unstable homogeneous steady state. In Figure 14 we show an example of large amplitude oscillations in the mean membrane potential with a value of C just less than $C = 1$, beyond which point the homogeneous steady state is stable.

6. Discussion. Motivated by the desire to understand the dynamics of neuronal networks with gap junction coupling, we have developed a number of results for planar PWL neuron models. We focus on these as they are minimal models capable of generating AP shapes.

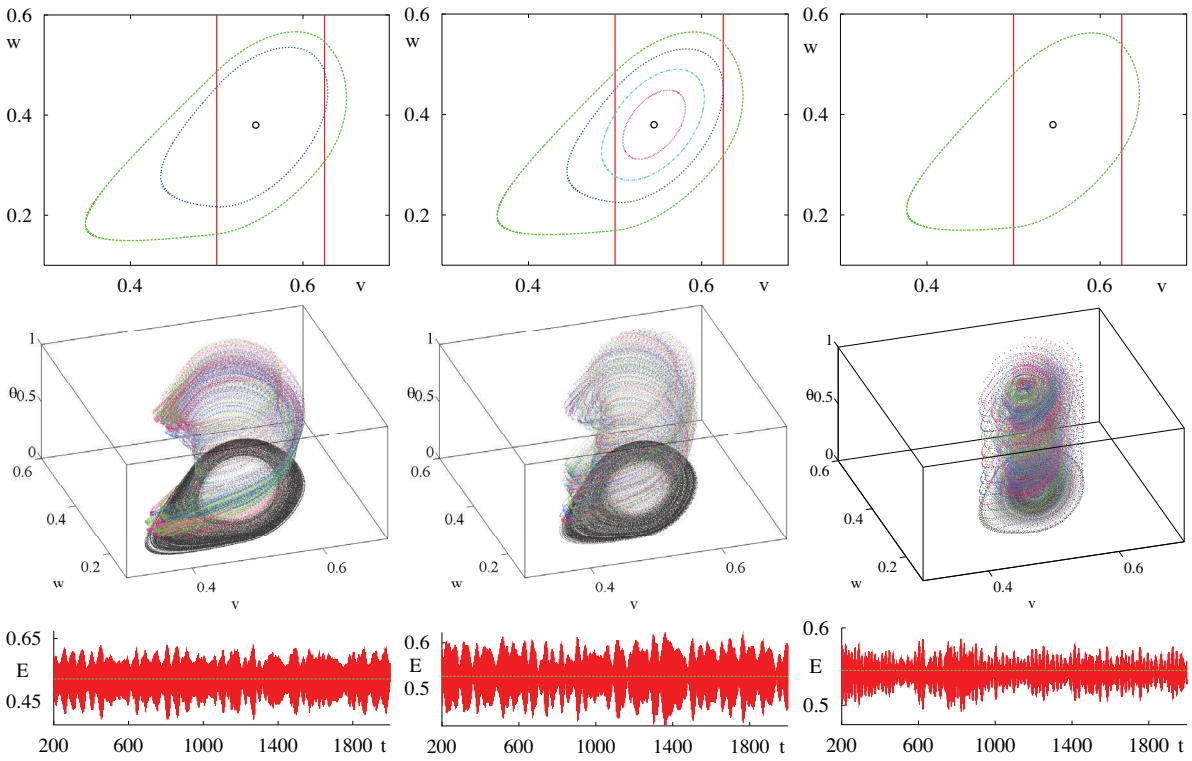


Figure 13. *Top: A family of coexisting unstable orbits in the PML model; synchronous (green), splay (blue), subthreshold splay (light blue), and harmonic splay (red). Here* $g = 0.1$, $I = 0.085$, *and other parameters are as in Figure 2. Left:* $C < 1 - g$ $(C = 0.89)$. *Middle:* $C = 1 - g$ $(C = 0.9)$. *Right:* $C > 1 - g$ $(C = 0.91)$. *Middle: Numerical simulation (after dropping transients) with* $N = 100$ *neurons showing a pseudo–color plot of the triple* (θ, v_i, w_i), *where* $\theta = t/\Delta \bmod 1$ *for some fixed* Δ. *Initial data is chosen to lie between the splay and synchronous state. Left: The network cycles between the unstable synchronous state and the unstable splay state.* Δ *is chosen as the mean of the synchronous and splay period. Middle: The network cycles between the unstable synchronous state and the unstable harmonic splay state.* Δ *is chosen as the mean of the synchronous and harmonic splay period. Right: The network cycles between the unstable synchronous state and the unstable fixed point.* Δ *is chosen as the period of the synchronous state. Bottom: Mean field signal* $E(t)$ *showing large amplitude fluctuations. Left: Fluctuations around the splay state (with* $v_0 = 0.52107$*). Middle: Fluctuations around the splay state (with* $v_0 = 0.52583$*). Right: Fluctuations around the fixed point (with* $v_0 = 0.545$*).*

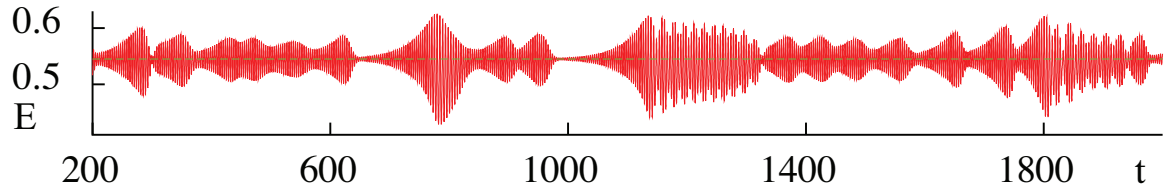


Figure 14. *Time varying mean field behavior organized around the unstable fixed point. Parameters are as in Figure 13, right, with* $C = 0.95$.

Unlike synaptically coupled networks, the shape of an AP is all-important in a gap junction coupled network as it is communicated directly between cells. For any PWL planar single

neuron model we have shown how to build periodic orbits in a self-consistent way, by piecing together trajectories from neighboring regions of phase space. Moreover, this procedure naturally lends itself to the construction of the associated PRC. The stability of periodic orbits has been established using Floquet theory, which in this case generates closed form expressions for the nonzero exponent of the orbit. As well as paving the way for the more obvious weakly coupled network analysis, we have found that the simplicity of the PWL model can allow for studies in the strong coupling regime, albeit for global coupling. In illustration of the utility of studying PWL networks, we have further shown how this can underpin a systematic explanation of the original observations of Han, Kurrer, and Kuramoto [42] on the generation of exotic mean field signals in networks of Morris–Lecar neurons with gap junction coupling.

Looking forward, it is worth mentioning here a number of possible extensions of the work in this paper that will lead to a deeper understanding of the role of gap junctions in shaping brain rhythms. As we have stressed, the techniques in this paper are general and are applicable to many PWL systems. In particular, it would be valuable to study Type I models which rely on a SNIC to generate their firing rate response [33, 32]. The cortical neuron model of Wilson [85] is a classic example of this, and its quadratic recovery variable is easily caricatured by choosing the parameter γ_1 in (2.5) to be negative. For systems with local chain-like coupling it may also prove possible to adapt techniques in [65] to study both synchronization and transient dynamics. Another natural step is to endow the purely gap junction coupled networks that we have described here with synaptic interactions. At the level of weak coupling the coupled oscillator theory that we have described here is naturally generalized along the lines of [26, 53]. Another mechanism available to neurons for the initiation of a firing event is that of anode break excitation, whereby a neuron can fire on release from a hyperpolarized state. The planar models that we have considered here are capable of such behavior, and thus when connected by inhibitory synapses emergent network periodic orbits might also be analyzed in a PWL fashion. In the strong coupling regime the challenge of studying phase-locked states that are neither synchronous nor splay effectively reduces to the problem of studying ODEs with delays. The techniques for doing this for PWL systems are relatively well developed in the engineering community, and one may therefore revisit the work in this paper, making explicit use of Lambert functions to define trajectories [6, 86]. In light of the recent interest in the analysis of gap junctions between dendritic trees [73], it would be interesting to explore the possibility of moving away from point neuron models, as studied here, to models with a spatially extended character [10]. For dendrites without active processes, the tools for doing this have been partially developed in [19]. However, of all the possible next steps, we regard the development of a tissue level firing rate model that can properly treat gap junction coupling as the major challenge facing the mathematical neuroscience community. Although at the level of fast voltage variables it is natural to think of gap junction coupling between nearest neighbors as generating a diffusive coupling, it is not clear that it is appropriate to simply add diffusive terms to existing rate models [77], such as the Wilson–Cowan, Amari, or Liley models (recently reviewed in [20]), since the state variables in these examples have no direct interpretation as fast voltage variables. In future work we hope to combine ideas from mean field dynamics, particularly those in [51, 67], with "equation-free" modeling [59] to tackle this challenge.

Appendix A. For real eigenvalues of A, we have the explicit form for $G(t)$ in section 3:

$$
\begin{aligned}
G_{11}(t) &= \frac{1}{\lambda_+ - \lambda_-}\left\{\lambda_+ \mathrm{e}^{\lambda_+ t} - \lambda_- \mathrm{e}^{\lambda_- t} - a_{22}\left[\mathrm{e}^{\lambda_+ t} - \mathrm{e}^{\lambda_- t}\right]\right\},\\
G_{12}(t) &= -\frac{\lambda_+ - a_{22}}{\lambda_+ - \lambda_-}\frac{\lambda_- - a_{22}}{a_{21}}\left[\mathrm{e}^{\lambda_+ t} - \mathrm{e}^{\lambda_- t}\right],\\
G_{21}(t) &= \frac{a_{21}}{\lambda_+ - \lambda_-}\left[\mathrm{e}^{\lambda_+ t} - \mathrm{e}^{\lambda_- t}\right],\\
G_{22}(t) &= \frac{1}{\lambda_+ - \lambda_-}\left\{\lambda_+ \mathrm{e}^{\lambda_- t} - \lambda_- \mathrm{e}^{\lambda_+ t} + a_{22}\left[\mathrm{e}^{\lambda_+ t} - \mathrm{e}^{\lambda_- t}\right]\right\}.
\end{aligned} \tag{A.1}
$$

The matrix $K(t)$ may then be calculated as

$$
\begin{aligned}
K_{11}(t) &= \frac{1}{\lambda_+ - \lambda_-}\left\{\mathrm{e}^{\lambda_+ t} - \mathrm{e}^{\lambda_- t} - a_{22}\left[\frac{\mathrm{e}^{\lambda_+ t}-1}{\lambda_+} - \frac{\mathrm{e}^{\lambda_- t}-1}{\lambda_-}\right]\right\},\\
K_{12}(t) &= -\frac{\lambda_+ - a_{22}}{\lambda_+ - \lambda_-}\frac{\lambda_- - a_{22}}{a_{21}}\left[\frac{\mathrm{e}^{\lambda_+ t}-1}{\lambda_+} - \frac{\mathrm{e}^{\lambda_- t}-1}{\lambda_-}\right],\\
K_{21}(t) &= \frac{a_{21}}{\lambda_+ - \lambda_-}\left[\frac{\mathrm{e}^{\lambda_+ t}-1}{\lambda_+} - \frac{\mathrm{e}^{\lambda_- t}-1}{\lambda_-}\right],\\
K_{22}(t) &= \frac{1}{\lambda_+ - \lambda_-}\left\{\frac{\lambda_+}{\lambda_-}\left[\mathrm{e}^{\lambda_- t}-1\right] - \frac{\lambda_-}{\lambda_+}\left[\mathrm{e}^{\lambda_+ t}-1\right] + a_{22}\left[\frac{\mathrm{e}^{\lambda_+ t}-1}{\lambda_+} - \frac{\mathrm{e}^{\lambda_- t}-1}{\lambda_-}\right]\right\}.
\end{aligned} \tag{A.2}
$$

For complex eigenvalues of A, we have the explicit form for $G(t)$:

$$
G(t) = \frac{\mathrm{e}^{\rho t}}{\widehat{\omega}}\begin{bmatrix} \widehat{\omega}\cos\omega t - \widehat{\rho}\sin\omega t & \sin\omega t \\ -(\widehat{\rho}^2+\widehat{\omega}^2)\sin\omega t & \widehat{\omega}\cos\omega t + \widehat{\rho}\sin\omega t \end{bmatrix}. \tag{A.3}
$$

The matrix $K(t)$ may then be calculated as

$$
K(t) = \frac{1}{\widehat{\omega}}\begin{bmatrix} \widehat{\omega}K_R(t) - \widehat{\rho}K_I(t) & K_I(t) \\ -(\widehat{\rho}^2+\widehat{\omega}^2)K_I(t) & \widehat{\omega}K_R(t) + \widehat{\rho}K_I(t) \end{bmatrix}, \tag{A.4}
$$

where

$$
K_R(t) = \frac{1}{\rho^2+\omega^2}\left\{\rho\left[\mathrm{e}^{\rho t}\cos(\omega t) - 1\right] + \omega \mathrm{e}^{\rho t}\sin(\omega t)\right\}, \tag{A.5}
$$

$$
K_I(t) = \frac{1}{\rho^2+\omega^2}\left\{\omega\left[1 - \mathrm{e}^{\rho t}\cos(\omega t)\right] + \rho \mathrm{e}^{\rho t}\sin(\omega t)\right\}. \tag{A.6}
$$

Appendix B. Computationally useful forms for the matrix elements in (4.7) are as follows:

$$
\alpha_\mu^n = P_\mu \Psi_\mu^n P_\mu^{-1}, \tag{B.1}
$$

$$
\beta_\mu^n = \widetilde{P}_\mu \Psi_\mu^{-n} \widetilde{P}_\mu^{-1} \mathrm{e}^{-2\pi i n T_\mu/T}, \tag{B.2}
$$

$$
\gamma_\mu^n = P_\mu \widetilde{\Psi}_\mu^n P_\mu^{-1}, \tag{B.3}
$$

where

$$\Psi_\mu^n = \operatorname{diag}\left(\frac{e^{(\lambda_+^\mu - 2\pi i n/T)T_\mu}-1}{\lambda_+^\mu - 2\pi i n/T}, \frac{e^{(\lambda_-^\mu - 2\pi i n/T)T_\mu}-1}{\lambda_-^\mu - 2\pi i n/T}\right), \tag{B.4}$$

$$\widetilde{\Psi}_\mu^n = \operatorname{diag}(1/\lambda_+^\mu, 1/\lambda_-^\mu)\left[\Psi_\mu^n + \frac{e^{-2\pi i n T_\mu/T}-1}{2\pi i n/T} I\right], \qquad n \neq 0, \tag{B.5}$$

$$\widetilde{\Psi}_\mu^0 = \operatorname{diag}(1/\lambda_+^\mu, 1/\lambda_-^\mu)\operatorname{diag}\left(\frac{e^{\lambda_+^\mu T_\mu}-1}{\lambda_+^\mu} - T_\mu, \frac{e^{\lambda_-^\mu T_\mu}-1}{\lambda_-^\mu} - T_\mu\right), \tag{B.6}$$

with P_μ and $\widetilde{P}_\mu$ as the matrix of eigenvectors of A_μ and A_μ^T, respectively, with associated eigenvalues $\lambda_\pm^\mu$. From the structure of Ψ_μ^n and $\widetilde{\Psi}_\mu^n$ above we see that α_μ^n, β_μ^n, and γ_μ^n all decrease as $1/n$.

Appendix C. In section 4.2.2 the elements of γ_μ^0 are calculated explicitly by noting that $\gamma_\mu^0 = \int_0^{T_\mu} K_\mu(t)\mathrm{d}t$. The structure of $K_\mu(t)$ for real and imaginary eigenvalues of the associated matrix A_μ are given by (A.2) and (A.4), respectively. Denoting $\int_0^t K(s)\mathrm{d}s$ by $F(t)$ gives

$$\begin{aligned}
F_{11}(t) &= \frac{1}{\lambda_+ - \lambda_-}\left\{\frac{e^{\lambda_+ t}-1}{\lambda_+} - \frac{e^{\lambda_- t}-1}{\lambda_-} - a_{22}\left[\frac{1}{\lambda_+}\left[\frac{e^{\lambda_+ t}-1}{\lambda_+} - t\right] - \frac{1}{\lambda_-}\left[\frac{e^{\lambda_- t}-1}{\lambda_-} - t\right]\right]\right\},\\
F_{12}(t) &= -\frac{\lambda_+ - a_{22}}{\lambda_+ - \lambda_-}\frac{\lambda_- - a_{22}}{a_{21}}\left[\frac{1}{\lambda_+}\left[\frac{e^{\lambda_+ t}-1}{\lambda_+} - t\right] - \frac{1}{\lambda_-}\left[\frac{e^{\lambda_- t}-1}{\lambda_-} - t\right]\right],\\
F_{21}(t) &= \frac{a_{21}}{\lambda_+ - \lambda_-}\left[\frac{1}{\lambda_+}\left[\frac{e^{\lambda_+ t}-1}{\lambda_+} - t\right] - \frac{1}{\lambda_-}\left[\frac{e^{\lambda_- t}-1}{\lambda_-} - t\right]\right],\\
F_{22}(t) &= \frac{1}{\lambda_+ - \lambda_-}\left\{\frac{\lambda_+}{\lambda_-}\left[\frac{e^{\lambda_- t}-1}{\lambda_-} - t\right] - \frac{\lambda_-}{\lambda_+}\left[\frac{e^{\lambda_+ t}-1}{\lambda_+} - t\right]\right.\\
&\quad \left. + a_{22}\left[\frac{1}{\lambda_+}\left[\frac{e^{\lambda_+ t}-1}{\lambda_+} - t\right] - \frac{1}{\lambda_-}\left[\frac{e^{\lambda_- t}-1}{\lambda_-} - t\right]\right]\right\}
\end{aligned} \tag{C.1}$$

for real $\lambda_\pm$ and

$$F(t) = \frac{1}{\widehat{\omega}}\begin{bmatrix}\widehat{\omega}F_R(t) - \widehat{\rho}F_I(t) & F_I(t)\\ -(\widehat{\rho}^2 + \widehat{\omega}^2)F_I(t) & \widehat{\omega}F_R(t) + \widehat{\rho}F_I(t)\end{bmatrix}, \tag{C.2}$$

where

$$F_R(t) = \frac{1}{\rho^2 + \omega^2}\left\{\rho\left[K_R(t) - t\right] + \omega K_I(t)\right\}, \tag{C.3}$$

$$F_I(t) = \frac{1}{\rho^2 + \omega^2}\left\{\omega\left[t - K_R(t)\right] + \rho K_I(t)\right\} \tag{C.4}$$

for complex $\lambda_\pm$. The matrices $\gamma_\mu^0 = F_\mu(T_\mu)$ may then be calculated using the above forms for $F(t)$ (under the replacement of $\lambda_\pm$ by $\lambda_\pm^\mu$).

Appendix D. For a splay state we may rewrite the advanced-retarded system of equations (4.18) as a set of ODEs by introducing

$$X_-(t) = \frac{1}{T}\int_0^t v(t)\mathrm{d}t, \qquad X_+(t) = \frac{1}{T}\int_t^T v(t)\mathrm{d}t. \tag{D.1}$$

After rescaling time as $\tau = t/T$ we may write

$$\frac{C}{T}\frac{\mathrm{d}v}{\mathrm{d}\tau} = f(v) - gv - w + I + g(X_- + X_+), \tag{D.2}$$

$$\frac{1}{T}\frac{\mathrm{d}w}{\mathrm{d}\tau} = g(v,w), \tag{D.3}$$

$$\frac{\mathrm{d}X_-}{\mathrm{d}\tau} = v, \tag{D.4}$$

$$\frac{\mathrm{d}X_+}{\mathrm{d}\tau} = -v, \tag{D.5}$$

subject to the boundary conditions $v(0) = v_{\rm th} = v(1)$, $w(0) = w^* = w(1)$, $X_-(0) = 0$, $X_-(1) = v_0$, $X_+(0) = v_0$, and $X_+(1) = 0$ (for some voltage section $v_{\rm th}$). We have four ODEs with seven boundary conditions which we may treat as a boundary value problem for the free parameters (v_0, w^*, T). For general choices of f and g it is natural to use numerical shooting for the solution of this problem. Alternatively, for the PWL models discussed in this paper we may analytically construct solutions according to the prescription described in section 4.2.2.

Acknowledgments. I would like to thank Bard Ermentrout and three anonymous referees for useful comments that have improved the presentation of the work in this paper.

REFERENCES

[1] L. F. Abbott and T. B. Kepler, *Model neurons: From Hodgkin–Huxley to Hopfield*, in Statistical Mechanics of Neural Networks, Lecture Notes in Phys. 368, L. Garrido, ed., Springer-Verlag, Berlin, 1990, pp. 5–18.

[2] D. J. Allwright, *Harmonic balance and the Hopf bifurcation*, Math. Proc. Cambridge Philos. Soc., 82 (1977), pp. 453–467.

[3] V. A. Alvarez, C. C. Chow, E. J. Van Bockstaele, and J. T. Williams, *Frequency-dependent synchrony in locus ceruleus: Role of electrotonic coupling*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 4032–4036.

[4] Y. Amitai, J. R. Gibson, M. B. S. L. Patrick, A. M. Ho, B. W. Connors, and D. Golomb, *The spatial dimensions of electrically coupled networks of interneurons in the neocortex*, J. Neurosci., 22 (2002), pp. 4142–4152.

[5] P. Ashwin and J. W. Swift, *The dynamics of n weakly coupled identical oscillators*, J. Nonlinear Sci., 2 (1992), pp. 69–108.

[6] F. A. Asl and A. G. Ulsoy, *Analysis of a system of linear delay differential equations*, J. Dynamic Systems, Measurement, and Control, 125 (2003), pp. 215–223.

[7] T. Bem and J. Rinzel, *Short duty cycle destabilizes a half-center oscillator, but gap junctions can restabilize*, J. Neurophysiology, 91 (2003), pp. 693–703.

[8] M. V. L. Bennet and R. S. Zukin, *Electrical coupling and neuronal synchronization in the mammalian brain*, Neuron, 41 (2004), pp. 495–511.

[9] M. Bennett, J. Contreras, F. Bukauskas, and J. Sáez, *New roles for astrocytes: Gap junction hemichannels have something to communicate*, Trends in Neurosciences, 26 (2003), pp. 610–617.

[10] P. C. Bressloff and S. Coombes, *Physics of the extended neuron*, Internat. J. Modern Phys. B, 11 (1997), pp. 2343–2392.

[11] P. C. Bressloff and S. Coombes, *Dynamics of strongly coupled spiking neurons*, Neural Computation, 12 (2000), pp. 91–129.

[12] E. Brown, P. Holmes, and J. Moehlis, *Globally coupled oscillator networks*, in Perspectives and Problems in Nonlinear Science: A Celebratory Volume in Honor of Larry Sirovich, Springer-Verlag, New York, 2003, pp. 183–215.
[13] E. Brown, J. Moehlis, and P. Holmes, *On the phase reduction and response dynamics of neural oscillator populations*, Neural Computation, 16 (2004), pp. 673–715.
[14] C. Chicone, *Ordinary Differential Equations with Applications*, 2nd ed., Texts in Applied Mathematics 34, Springer-Verlag, New York, 2006.
[15] C. C. Chow and N. Kopell, *Dynamics of spiking neurons with electrical coupling*, Neural Computation, 12 (2000), pp. 1643–78.
[16] B. W. Connors and M. A. Long, *Electrical synapses in the mammalian brain*, Ann. Rev. Neurosci., 27 (2004), pp. 393–418.
[17] S. Coombes, *Phase locking in networks of synaptically coupled McKean relaxation oscillators*, Phys. D, 160 (2001), pp. 173–188.
[18] S. Coombes and P. C. Bressloff, eds., *Bursting: The Genesis of Rhythm in the Nervous System*, World Scientific, New York, 2005.
[19] S. Coombes, Y. Timofeeva, C.-M. Svensson, G. J. Lord, K. Josic, S. J. Cox, and C. M. Colbert, *Branching dendrites with resonant membrane: A "sum-over-trips" approach*, Biol. Cybernet., 97 (2007), pp. 137–149.
[20] S. Coombes, N. A. Venkov, L. Shiau, I. Bojak, D. T. J. Liley, and C. R. Laing, *Modeling electrocortical activity through improved local approximations of integral neural field equations*, Phys. Rev. E (3), 76 (2007), 051901.
[21] G. de Vries and A. Sherman, *Beyond synchronization: Modulatory and emergent effects of coupling in square-wave bursting*, in Bursting: The Genesis of Rhythm in the Nervous System, World Scientific, New York, 2005, pp. 243–272.
[22] M. Denman-Johnson and S. Coombes, *A continuum of weakly coupled McKean neurons*, Phys. Rev. E (3), 67 (2003), 051903.
[23] S. Dhein and J. S. Borer, eds., *Cardiovascular Gap Junctions (Advances in Cardiology)*, Karger, Basel, 2006.
[24] F. E. Dudek, *Gap junctions and fast oscillations: A role in seizures and epileptogenesis?*, Epilepsy Currents, 2 (2002), pp. 133–136.
[25] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.
[26] B. Ermentrout, *Gap junctions destroy persistent states in excitatory networks*, Phys. Rev. E (3), 74 (2006), 031918.
[27] G. B. Ermentrout, *Neural nets as spatio-temporal pattern forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.
[28] G. B. Ermentrout and N. Kopell, *Frequency plateaus in a chain of weakly coupled oscillators,* I, SIAM J. Math. Anal., 15 (1984), pp. 215–237.
[29] G. B. Ermentrout and N. Kopell, *Oscillator death in systems of coupled neural oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 125–146.
[30] G. B. Ermentrout and N. Kopell, *Multiple pulse interactions and averaging in systems of coupled neural oscillators*, J. Math. Biol., 29 (1991), pp. 195–217.
[31] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membranes*, Biophys. J., 1182 (1961), pp. 445–466.
[32] H. Fujii and I. Tsuda, *Itinerant dynamics of class I^* neurons coupled by gap junctions*, in Computational Neuroscience: Cortical Dynamics, Lecture Notes in Comput. Sci. 3416, P. Érdi, A. Esposito, M. Marinaro, and S. Scarpetta, eds., Springer-Verlag, New York, 2004, pp. 140–160.
[33] H. Fujii and I. Tsuda, *Neocortical gap junction-coupled interneuron systems may induce chaotic behavior itinerant among quasi-attractors exhibiting transient synchrony*, Neurocomputing, 58–60 (2004), pp. 151–157.
[34] T. Fukuda, T. Kosaka, W. Singer, and R. A. W. Galuske, *Gap junctions among dendrites of cortical GABAergic neurons establish a dense and widespread intercolumnar network*, J. Neurosci., 26 (2006), pp. 3434–3443.
[35] E. J. Furshpan and D. D. Potter, *Mechanism of nerve-impulse transmission at a crayfish synapse*, Nature, 180 (1957), pp. 342–343.

[36] M. Galarreta and S. Hestrin, *A network of fast-spiking cells in the neocortex connected by electrical synapses*, Nature, 402 (1999), pp. 72–75.
[37] J. Gao and P. Holmes, *On the dynamics of electrically coupled neurons with inhibitory synapses*, J. Computational Neuroscience, 22 (2007), pp. 39–61.
[38] W. Gerstner and J. L. van Hemmen, *Associative memory in a network of 'spiking' neurons*, Network, 3 (1992), pp. 139–164.
[39] J. R. Gibson, M. Beierlein, and B. W. Connors, *Functional properties of electrical synapses between inhibitory interneurons of neocortical layer* 4, J. Neurophysiology, 93 (2005), pp. 467–480.
[40] D. Golomb, X. J. Wang, and J. Rinzel, *Synchronization properties of spindle oscillations in a thalamic reticular nucleus model*, J. Neurophysiology, 72 (1994), pp. 1109–1126.
[41] F. Hamzei-Sichani, N. Kamasawa, W. G. M. Janssen, T. Yasumura, K. G. V. Davidson, P. R. Hof, S. L. Wearne, M. G. Stewart, S. R. Young, M. A. Whittington, J. E. Rash, and R. D. Traub, *Gap junctions on hippocampal mossy fiber axons demonstrated by thin-section electron microscopy and freeze-fracture replica immunogold labeling*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 12548–12553.
[42] S. K. Han, C. Kurrer, and Y. Kuramoto, *Dephasing and bursting in coupled neural oscillators*, Phys. Rev. Lett., 75 (1995), pp. 3190–3193.
[43] S. K. Han, C. Kurrer, and Y. Kuramoto, *Diffusive interaction leading to dephasing of coupled neural oscillators*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 7 (1997), pp. 869–876.
[44] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve tissue*, J. Physiology (London), 116 (1952), pp. 449–472.
[45] F. C. Hoppensteadt and E. M. Izhikevich, *Weakly Connected Neural Networks*, Appl. Math. Sci. 126, Springer-Verlag, New York, 1997.
[46] S. G. Hormuzdi, M. A. Filippov, G. Mitropoulou, H. Monyer, and R. Bruzzone, *Electrical synapses: A dynamic signaling system that shapes the activity of neuronal networks*, Biochimica et Biophysica Acta, 1662 (2004), pp. 113–137.
[47] S. W. Hughes and V. Crunelli, *Just a phase they're going through: The complex interaction of intrinsic high-threshod bursting and gap junctions in the generation of thalamic α and θ rhythms*, Internat. J. Psychophysiology, 74 (2007), pp. 3–17.
[48] E. M. Izhikevich, *Phase equations for relaxation oscillators*, SIAM J. Appl. Math., 60 (2000), pp. 1789–1804.
[49] E. M. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press, Cambridge, MA, 2007.
[50] E. M. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press, Cambridge, MA, 2007, Ch. 10; also available online at http://www.izhikevich.com.
[51] C. G. A. V. K. Jirsa and J. A. S. Kelso, *Synchrony and clustering in heterogeneous networks with global coupling and parameter dispersion*, Phys. Rev. Lett., 94 (2005), 018106.
[52] J. Karbowski and N. Kopell, *Multispikes and synchronization in a large-scale neural network with delays*, Neural Computation, 12 (2000), pp. 1573–1606.
[53] F. G. Kazanci and B. Ermentrout, *Pattern formation in an array of oscillators with electrical and chemical coupling*, SIAM J. Appl. Math., 67 (2007), pp. 512–529.
[54] T. B. Kepler, L. F. Abbott, and E. Marder, *Reduction of conductance-based neuron models*, Biol. Cybernet., 66 (1992), pp. 381–387.
[55] T. B. Kepler, E. Marder, and L. F. Abbott, *The effect of electrical coupling on the frequency of model neuronal oscillators*, Science, 248 (1990), pp. 83–85.
[56] N. Kopell and B. Ermentrout, *Chemical and electrical synapses perform complementary roles in the synchronization of interneuronal networks*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 15482–15487.
[57] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, New York, 1984.
[58] Y. Kuramoto, *Collective synchronization of pulse-coupled oscillators and excitable units*, Phys. D, 50 (1991), pp. 15–30.
[59] C. R. Laing, *On the application of "equation-free modelling" to neural systems*, J. Computational Neuroscience, 20 (2006), pp. 5–23.
[60] T. J. Lewis and J. Rinzel, *Self-organized synchronous oscillations in a network of excitable cells coupled by gap junctions*, Network, 11 (2000), pp. 299–320.

[61] T. J. Lewis and J. Rinzel, *Dynamics of spiking neurons connected by both inhibitory and electrical coupling*, J. Computational Neuroscience, 14 (2003), pp. 283–309.
[62] Y. Loewenstein, Y. Yarom, and H. Sompolinsky, *The generation of oscillations in networks of electrically coupled cells*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 8095–8100.
[63] J. G. Mancilla, T. J. Lewis, D. J. Pinto, J. Rinzel, and B. W. Connors, *Synchronization of electrically coupled pairs of inhibitory interneurons in neocortex*, J. Neurosci., 27 (2007), pp. 2058–2073.
[64] H. P. McKean, *Nagumo's equation*, Adv. Math., 4 (1970), pp. 209–223.
[65] G. S. Medvedev and N. Kopell, *Synchronization and transient dynamics in the chains of electrically coupled FitzHugh–Nagumo oscillators*, SIAM J. Appl. Math., 61 (2001), pp. 1762–1801.
[66] K. Miura and M. Okada, *Globally coupled resonate-and-fire models*, Progr. Theoret. Phys. Suppl., 161 (2006), pp. 255–259.
[67] S. D. Monte, F. d'Ovidio, and E. Mosekilde, *Coherent regimes of globally coupled dynamical systems*, Phys. Rev. Lett., 90 (2003), 054102.
[68] C. Morris and H. Lecar, *Voltage oscillations in the barnacle giant muscle fiber*, Biophys. J., 35 (1981), pp. 193–213.
[69] B. Pfeuty, G. Mato, D. Golomb, and D. Hansel, *Electrical synapses and synchrony: The role of intrinsic currents*, J. Neurosci., 23 (2003), pp. 6280–6294.
[70] B. Pfeuty, G. Mato, D. Golomb, and D. Hansel, *The combined effects of inhibitory and electrical synapses in synchrony*, Neural Computation, 17 (2005), pp. 633–670.
[71] J. E. Rash, R. K. Dillman, B. L. Bilhartz, H. S. Duffy, L. R. Whalen, and T. Yasumura, *Mixed synapses discovered and mapped throughout mammalian spinal cord*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 4235–4239.
[72] J. Rubin and D. Terman, *Geometric singular perturbation analysis of neuronal dynamics*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, G. Ioos, and N. Kopell, eds., North–Holland, Amsterdam, 2002, pp. 93–146.
[73] F. Saraga, L. Ng, and F. K. Skinner, *Distal gap junctions and active dendrites can tune network dynamics*, J. Neurophysiology, 95 (2006), pp. 1669–1682.
[74] A. Sherman and J. Rinzel, *Rhythmogenic effects of weak electrotonic coupling in neuronal models*, Proc. Natl. Acad. Sci. USA, 89 (1992), pp. 2471–2474.
[75] C. Sotelo, R. Llinas, and R. Baker, *Structural study of inferior olivary nucleus of the cat: Morphological correlates of electrotonic coupling*, J. Neurophysiology, 37 (1974), pp. 541–559.
[76] C. Soto-Treviño, P. Rabbah, E. Marder, and F. Nadim, *Computational model of electrically coupled, intrinsically distinct pacemaker neurons*, J. Neurophysiology, 94 (2005), pp. 590–604.
[77] M. L. Steyn-Ross, D. A. Steyn-Ross, M. T. Wilson, and J. W. Sleigh, *Gap junctions mediate large-scale Turing structures in a mean-field cortex driven by subcortical noise*, Phys. Rev. E (3), 76 (2007), 011916.
[78] A. Tonnelier, *The McKean's caricature of the FitzHugh–Nagumo model* I. *The space-clamped system*, SIAM J. Appl. Math., 63 (2002), pp. 459–484.
[79] A. Tonnelier, *McKean model*, Scholarpedia, 2 (2007), p. 12071.
[80] A. Tonnelier and W. Gerstner, *Piecewise linear differential equations and integrate-and-fire neurons: Insights from two-dimensional membrane models*, Phys. Rev. E (3), 67 (2003), 021908.
[81] C. van Vreeswijk, *Analysis of the asynchronous state in networks of strongly coupled oscillators*, Phys. Rev. Lett., 84 (2000), pp. 5110–5113.
[82] J. L. P. Velazquez and P. L. Carlen, *Gap junctions, synchrony and seizures*, Trends in Neurosciences, 23 (2000), pp. 68–74.
[83] W. P. Wang, *Multiple impulse solutions to McKean's caricature of the nerve equation.* I. *Existence*, Comm. Pure Appl. Math., 41 (1988), pp. 71–103.
[84] W. P. Wang, *Multiple impulse solutions to McKean's caricature of the nerve equation.* II. *Stability*, Comm. Pure Appl. Math., 41 (1988), pp. 997–1025.
[85] H. R. Wilson, *Simplified dynamics of human and mammalian neocortical neurons*, J. Theoret. Biol., 200 (1999), pp. 375–388.
[86] S. Yi, P. W. Nelson, and A. G. Ulsoy, *Delay differential equations via the Lambert W function and bifurcation analysis: Application to machine tool chatter*, Math. Biosci. Engrg., 4 (2007), pp. 355–368.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1131–1162
© 2008 Society for Industrial and Applied Mathematics

The Geometry of Slow Manifolds near a Folded Node*

M. Desroches†, B. Krauskopf†, and H. M. Osinga†

Abstract. This paper is concerned with the geometry of slow manifolds of a dynamical system with one fast and two slow variables. Specifically, we study the dynamics near a folded-node singularity, which is known to give rise to so-called canard solutions. Geometrically, canards are intersection curves of two-dimensional attracting and repelling slow manifolds, and they are a key element of slow-fast dynamics. For example, canard solutions are associated with mixed-mode oscillations, where they organize regions with different numbers of small oscillations. We perform a numerical study of the geometry of two-dimensional slow manifolds in the normal form of a folded node in $\mathbb{R}^3$. Namely, we view the part of a slow manifold that is of interest as a one-parameter family of orbit segments up to a suitable cross-section. Hence, it is the solution of a two-point boundary value problem, which we solve by numerical continuation with the package AUTO. The computed family of orbit segments is used to obtain a mesh representation of the manifold as a surface. With this approach we show how the attracting and repelling slow manifolds change in dependence on the eigenvalue ratio μ associated with the folded-node singularity. At $\mu = 1$ two primary canards bifurcate and secondary canards are created at odd integer values of μ. We compute 24 secondary canards to investigate how they spiral more and more around one of the primary canards. The first sixteen secondary canards are continued in μ to obtain a numerical bifurcation diagram.

Key words. slow-fast systems, singular perturbation, canard solution, boundary value problem, invariant manifolds

AMS subject classifications. 34E15, 34C30, 37C10, 65L10

DOI. 10.1137/070708810

1. Introduction.

Multiple time scale systems are characterized by the property that certain variables evolve on vastly different time scales, which means that the systems may display dynamics that is composed of slow and fast elements. The occurrence of different time scales is quite natural in many applications, including chemical reaction dynamics [10, 33, 35, 37, 39], cell modeling [15, 43, 44], electronic circuits [13, 48, 49], and laser dynamics [16, 21]. The first example of slow-fast dynamics was discovered by Van der Pol [48, 49] in the 1920s. He considered an electrical circuit with a triode valve where the current is a cubic function of the voltage. The mathematical model of this circuit is known today as the Van der Pol equations. It shows sustained, very nonharmonic oscillations when the nonlinear damping is large. In this case the periodic solution is composed of slow motion that closely follows attracting segments of the underlying cubic curve (which forms one of the nullclines), followed by fast jumps as the trajectory reaches either of the two folds of this curve. At the jumps one of

*Received by the editors November 20, 2007; accepted for publication (in revised form) by T. Kaper April 25, 2008; published electronically October 13, 2008.
http://www.siam.org/journals/siads/7-4/70881.html
†Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, United Kingdom (M.Desroches@bristol.ac.uk, B.Krauskopf@bristol.ac.uk, H.M.Osinga@bristol.ac.uk). The first author was supported by grant EP/C54403X/1 from the Engineering and Physical Sciences Research Council (EPSRC) and the third author by an EPSRC Advanced Research Fellowship grant.

the variables barely changes until the trajectory reaches another attracting segment of cubic curve. Van der Pol called these periodic solutions *relaxation oscillations*. Models showing relaxation oscillations quite similar to that of the Van der Pol equations have been found in other application areas. A well-known example is the FitzHugh–Nagumo system, which also has two time scales and a cubic nonlinearity. It was derived independently by FitzHugh [24] and Nagumo [38] as a simplified planar version of the famous Hodgkin–Huxley equations for the action potential of the giant axon of a squid in terms of transmembrane currents [31].

At the end of the 1970s, Benoît et al. [5] found and analyzed even more unusual periodic solutions in the Van der Pol equations with an added external constant forcing term; they called these periodic solutions *canards*. A canard orbit has the special property that it follows at least one unstable segment of the underlying cubic nullcline. In other words, the trajectory does not jump at the respective fold. Canard orbits account for a sudden increase in the amplitude of the attracting periodic orbit in the transition from harmonic oscillations to relaxation oscillations. This change is known as a *canard explosion*, a term that was first introduced by Brøns and Bar-Eli [7]. A canard explosion and the associated canards are extremely difficult to observe in two-dimensional slow-fast systems, because the transition happens in an exponentially small parameter interval.

In three-dimensional slow-fast systems canards may exist in much larger regions of the system parameters. They have been found in numerous slow-fast systems. For example, canards explain sudden changes in amplitude and period of oscillatory behavior in chemical reactions [10, 33, 35, 37, 39]; they organize the dynamics in models of coupled neurons and also play an important role in intracellular activities [15, 43, 44]; canards have been studied for their role in diffusion-induced instabilities [10, 41]; and Guckenheimer et al. [6, 29] recently performed an extensive study of a reduced hybrid model of the (periodically) forced Van der Pol equations which revealed relaxation oscillations and canard orbits of different types. A related phenomenon in slow-fast systems are *mixed-mode oscillations*, which consist of large-amplitude excursions followed by small-amplitude motions that are typically of (relatively) high frequency. This type of oscillation has been found in chemical and biological systems, and the connection between mixed-mode oscillations and canards has been clarified recently [9, 35, 44, 50]; see also the special issue [8].

For the theoretical study of canards in three-dimensional phase space, one considers a dynamical system with two slow and one fast variable of the form

$$\left\{\begin{array}{rcl} \dot{u}_1 &=& g_1(u_1,u_2,v,\varepsilon),\\ \dot{u}_2 &=& g_2(u_1,u_2,v,\varepsilon),\\ \varepsilon\dot{v} &=& f(u_1,u_2,v,\varepsilon). \end{array}\right. \tag{1.1}$$

Here g_1, g_2, and f are sufficiently smooth functions, and $\varepsilon > 0$ is a small parameter that separates the different time scales. Since ε is small, the variables u_1 and u_2 move on a slower time scale than the fast variable v. The equivalent of the cubic nullcline of the Van der Pol equations is now the surface—called the *critical manifold*—that is given as the v-nullcline of (1.1). The critical manifold has repelling parts and attracting parts, which meet along one-dimensional fold curves. As for the Van der Pol equations, the slow dynamics takes place close to the critical manifold. When a fold is reached two things can happen. The trajectory may jump at the fold curve toward another attracting sheet of the critical manifold. If a global

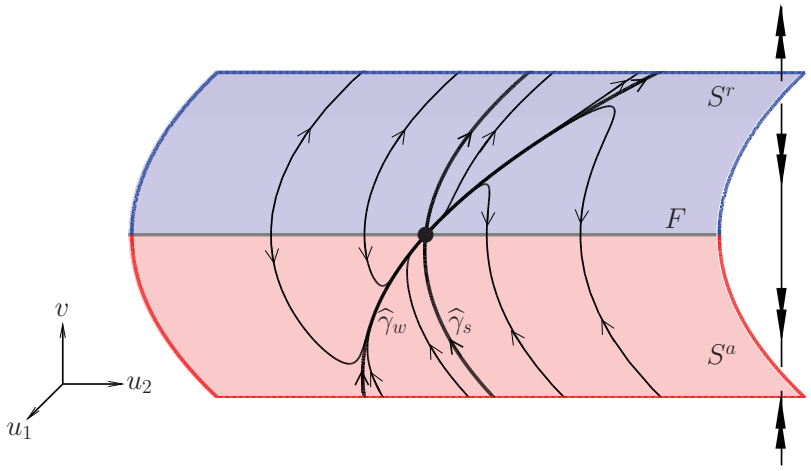


Figure 1. *Sketch of a folded critical manifold S consisting of an attracting sheet S^a (red) and a repelling sheet S^r (blue) that meet at a fold curve F. The sketched flow on S is the generic slow flow near a folded node (black dot); shown are two singular canards $\widehat{\gamma}_s$ and $\widehat{\gamma}_w$ and some other trajectories of the slow flow. The action of the fast flow is shown along one fast fiber.*

return mechanism is present, e.g., the critical manifold is S-shaped with two fold curves, such jumps give rise to classical relaxation oscillations [36, 46]. The other possibility is that the trajectory is a canard solution that does not jump at the fold curve but instead stays near the repelling part of the critical manifold for a certain amount of time.

Since the beginning of the 1980s different analytical techniques have been applied to the study of canard solutions. Initially, nonstandard analysis [3, 4, 5] and matched asymptotic expansions [19, 36] were used. Dumortier and Roussarie showed in their seminal work [17, 18, 42] that (standard) nonlinear analysis can be applied to the study of canards. Underlying this approach is the realization that canards can be understood via the dynamics of the system near folds of the critical manifold. Of specific interest here are (isolated) points on fold curves where the direction of flow changes from pointing toward the fold to pointing away from the fold. These points, called *folded singularities*, are key to the understanding of canard solutions.

We are concerned here with the case of a *folded node*, which is a type of folded singularity that has been identified as an organizing center for the creation of canard solutions. Figure 1 shows the dynamics for $\varepsilon = 0$ near a folded node on a regular fold curve F along which the attracting sheet S^a and the repelling sheet S^r of the critical manifold join. At the folded node (black dot), the direction of the flow on the critical manifold changes. This allows for the existence of canard solutions on the critical manifold that cross F at the folded node with nonzero speed and then follow the repelling sheet S^r. These canards for $\varepsilon = 0$ are referred to as *singular canards*, and they occur in an entire region that is bounded by the repelling part of the fold curve F (to the left of the folded node) and the special canard solution $\widehat{\gamma}_s$ in Figure 1. The main question now is: What can be said about the dynamics near a folded node when $\varepsilon > 0$? According to a well-known result by Fenichel [22, 23], away from the fold curve F the attracting and repelling sheets give rise to an *attracting slow manifold* S^a_ε and a *repelling slow manifold* S^r_ε for $\varepsilon > 0$, respectively. Importantly, for $\varepsilon > 0$ the two surfaces S^a_ε and S^r_ε do not

connect along the fold curve, but rather intersect, generically transversely, in one-dimensional solution curves. These curves are referred to as *maximal canards*, as they stay close to the repelling part of the critical manifold for a certain amount of time. In other words, finding the structure of canard solutions near the folded node is equivalent to understanding the geometry of the two-dimensional surfaces S^a_ε and S^r_ε.

Canards near a folded node are best analyzed in a normal form setting. Normal forms for all types of folded singularities, including the folded node, were derived by Il'yashenko; see Arnol′d et al. [2, part I, chapter 4] and also the topological classification of Takens [47]. Benoît [4] completely analyzed the case of a folded saddle and also considered the case of a folded node. He proved the existence of two maximal canards under a nonresonance condition with tools of nonstandard analysis and found secondary canards with spiraling behavior by numerical integration. Szmolyan and Wechselberger [45] considered all cases of folded singularities. For the folded node they proved the existence of primary canards by using geometric singular perturbation theory in combination with blow-up transformations. The existence of secondary canards depends on the eigenvalue ratio of the Jacobian at the folded node, which we denote μ with $1 \leq \mu < \infty$; precise details are given in section 2. Guckenheimer and Haiduc [28] proved that for any fixed μ there is a finite number of secondary canards near a folded node, and this number goes to infinity as both $\mu \to \infty$ and $\varepsilon \to 0$. Wechselberger [50] studied how secondary canards bifurcate from the weak primary canard, and he sketched the underlying bifurcation structure. Both Guckenheimer and Haiduc [28] and Wechselberger [50] find canards numerically by computing the one-dimensional intersection curves of the attracting and repelling slow manifolds in a two-dimensional cross-section with a shooting approach (integration from a line of initial conditions far from the fold curve); the canards are then identified as crossings of the two curves.

Our goal is to compute and visualize the slow manifolds as surfaces and the canards as trajectories in $\mathbb{R}^3$ to obtain additional geometric insight into the dynamics near a folded node. To this end, we consider the normal form as used in [50], which is given as the three-dimensional vector field

$$\left\{ \begin{array}{rcl} \dot{x} &=& \frac{1}{2}\mu y - (\mu+1)z, \\ \dot{y} &=& 1, \\ \dot{z} &=& x + z^2. \end{array} \right. \tag{1.2}$$

In (1.2) the folded node is at the origin and the parameter μ is the ratio of the eigenvalues of the Jacobian matrix of the so-called desingularized reduced flow evaluated at the folded node. Note that, as a result of a parameter dependent blow-up procedure, (1.2) does not depend on ε. Nevertheless, this normal form describes how the attracting and repelling slow manifolds intersect near a folded node. Specifically, for any μ system (1.2) has an attracting slow manifold C^- and a repelling slow manifold C^+ that intersect in the maximal canards one wants to study; see section 2 for more information on the derivation and meaning of the normal form.

Specifically, we compute in this paper the slow manifolds C^- and C^+ of (1.2) as surfaces in $\mathbb{R}^3$. This is achieved by representing a piece of interest of a slow manifold as a one-parameter family of orbit segments that satisfy suitably chosen boundary conditions. The resulting boundary value problem is solved with the continuation and collocation routines of

the package AUTO [14]. This setup is very flexible and accurate, because it is based on the continuation of two-point boundary value problems [34]. It allows us to adjust the boundary conditions to emphasize certain local features of the underlying dynamics in the vicinity of the folded node. With our method we obtain a very precise visualization of the behavior of the slow manifolds. In particular, we are able to capture in detail the complexity of their intersections, that is, the canards. For increasing μ, the two surfaces rotate more and more around the weak primary canard (the perturbation of $\widehat{\gamma}_w$ in Figure 1), which leads to the creation of secondary canards at odd integer values of μ. Our method allows us to detect canards and to continue them as solutions of a boundary value problem in the parameter μ. We compute up to 24 secondary canards and show how they spiral around the weak primary canard. Furthermore, we continue sixteen secondary canards in μ to obtain a numerical bifurcation diagram.

Note that very few attempts have been made so far to produce accurate computations of slow manifolds as surfaces. Milik et al. [35] visualize slow manifolds in the normal form of a folded saddle-node; in this special case (of codimension one) the system decouples into a one-parameter family of two-dimensional systems, so that the slow manifolds can be built up from individual one-dimensional stable and unstable manifolds (which can be found by integration). Ginoux and Rossetto derive an implicit equation that provides local approximations of slow manifolds as the locus where the torsion vanishes [25, 26]. Both methods are quite different from our approach. Our boundary value problem formulation allows one not only to compute slow manifolds as global objects in a specified region of interest but also to detect canards and follow them in system parameters. More generally, computing invariant manifolds by continuation and collocation as solution families subject to suitable boundary conditions is a very flexible method [34] which is particularly well suited for slow-fast systems [20]. Our method is not restricted to the normal form setting but can be used more widely, for example, for the study of canards in the self-coupled FitzHugh–Nagumo system [11].

The outline of this paper is as follows. In the next section we present some technical background material. We briefly review slow-fast systems in section 2.1, explain the blow-up method in section 2.2, and then discuss some known properties of the normal form (1.2) in section 2.3. Section 3 explains in detail how to compute slow manifolds as collections of orbit segments. In section 4 we show how the slow manifolds change with μ, and section 5 is devoted to the detection and continuation of the secondary canards. In section 6 we show that the geometry of the slow manifolds does not change topologically when the normal form is perturbed. We conclude with a summary and outlook in section 7.

2. Background on the folded node.

In this section we recall some basic facts about singularly perturbed dynamical systems, folded singularities, and the blow-up method to analyze them. We consider the three-dimensional normal form (1.2), as studied, for example, by Benoît [3, 4], Guckenheimer and Haiduc [28], and Wechselberger [50]. Following [50], we show how the normal form is derived from a generic three-dimensional system with a folded node at the origin and present some useful properties of (1.2).

2.1. Slow-fast dynamical systems.

The *slow-time system* (1.1) defines a vector field using the slow time t. An alternative way of writing (1.1) is to introduce the fast time $\tau = \frac{t}{\varepsilon}$, which

gives the *fast-time system*

$$\left\{\begin{array}{rcl} u_1' &=& \varepsilon g_1(u_1,u_2,v,\varepsilon),\\ u_2' &=& \varepsilon g_2(u_1,u_2,v,\varepsilon),\\ v' &=& f(u_1,u_2,v,\varepsilon), \end{array}\right. \tag{2.1}$$

where the prime indicates the derivative with respect to the fast time τ. This rescaling of time is valid only for $\varepsilon > 0$ and does not modify the geometry of the trajectories. The main question is whether it is possible to understand the dynamics for small $\varepsilon > 0$ by considering the two limits of (1.1) and (2.1) given by $\varepsilon = 0$.

The limit of the slow-time system (1.1) for $\varepsilon = 0$ is known as the *reduced system* or *slow subsystem*

$$\left\{\begin{array}{rcl} \dot{u}_1 &=& g_1(u_1,u_2,v,0),\\ \dot{u}_2 &=& g_2(u_1,u_2,v,0),\\ 0 &=& f(u_1,u_2,v,0). \end{array}\right. \tag{2.2}$$

System (2.2) is a set of differential algebraic equations on the slow time scale, namely, two differential equations constrained by the algebraic equation $f = 0$. This condition defines the *critical manifold*

$$S := \left\{(u_1,u_2,v) \in \mathbb{R}^3 \mid f(u_1,u_2,v,0) = 0\right\}, \tag{2.3}$$

on which the dynamics of the reduced system (2.2) takes place. Typically, it is not possible to solve for S as a function of u_1 and u_2, but we may assume that S is locally a graph over, say, u_2 and v. Then we can describe the flow of the reduced system (2.2) on S as a projection onto the (u_2, v)-plane. Namely, by differentiating the algebraic equation $f = 0$ with respect to time, we obtain

$$\left\{\begin{array}{rcl} \dot{u}_2 &=& g_2(u_1,u_2,v,0),\\ -f_v\dot{v} &=& f_{u_1}g_1 + f_{u_2}g_2, \end{array}\right. \tag{2.4}$$

where $u_1 = s(u_2, v)$ is uniquely defined by the condition $(u_1, u_2, v) \in S$. We then rescale (2.4) by $-f_v$ to obtain the vector field

$$\left\{\begin{array}{rcl} \dot{u}_2 &=& -f_v g_2(u_1,u_2,v,0),\\ \dot{v} &=& f_{u_1}g_1 + f_{u_2}g_2, \end{array}\right. \tag{2.5}$$

with $u_1 = s(u_2, v)$, which generates a *slow flow* on S.

The limit of the fast-time system (2.1) for $\varepsilon = 0$ is known as the *layer system* or *fast subsystem*

$$\left\{\begin{array}{rcl} u_1' &=& 0,\\ u_2' &=& 0,\\ v' &=& f(u_1,u_2,v,0). \end{array}\right. \tag{2.6}$$

The variables u_1 and u_2 are constants in system (2.6) and enter the equation for v as parameters. Hence, the layer system is a two-parameter family of differential equations on the fast

time scale. The critical manifold S also plays a key role in the layer system, namely, as a manifold of equilibria.

The idea of geometric singular perturbation theory [23, 32] is to understand the dynamics of system (1.1) with $\varepsilon > 0$ sufficiently small by splitting the motion into its fast and slow components. The fast dynamics of (1.1) is described by the layer system (2.6), meaning that trajectories behave like solutions of (2.6) until they get close to S. On the slow time scale, solutions are well approximated by the reduced system (2.2); in particular, trajectories remain confined to an ε-neighborhood of S. The overall dynamics can indeed be understood in this way if the critical manifold S is *normally hyperbolic*. This means that the dynamics in the direction normal to the manifold dominates the dynamics in the tangent direction [30]. Due to results by Fenichel [22, 23], a normally hyperbolic critical manifold S persists under small perturbations as a nearby normally hyperbolic invariant manifold S_ε for the singularly perturbed system (1.1). Since S is a manifold of equilibria for (2.6), the normally hyperbolic points on S are those points for which all eigenvalues associated with the directions normal to S do not lie on the imaginary axis.

In the vicinity of points where normal hyperbolicity fails, the singularly perturbed problem can give rise to very complex dynamics. In order to study what happens when S is not normally hyperbolic, we consider the projection of S onto the (u_1, u_2)-plane of slow variables. The critical manifold S consists of regular points where $f_v \neq 0$ and critical points of the projection where $f_v = 0$. According to singularity theory [1], regular points are generic and they correspond to points where S is normally hyperbolic. Moreover, a generic critical point is a fold point, and together the fold points form a codimension-one submanifold F of S. Along F two sheets of S meet. In $\mathbb{R}^3$ there may be cusp points (degenerate folds), but they are generically isolated. We focus here on fold points and their influence on the dynamics of system (1.1).

Specifically, we consider a critical manifold S with (locally) a nonempty fold curve that does not contain cusp points. Therefore, S can be written as $S = S^a \cup F \cup S^r$, where S^a and S^r refer to the attracting and repelling sheets of S, respectively, that meet at F; formally

$$\begin{aligned} S^a &= \{(u_1, u_2, v) \in S \mid f_v(u_1, u_2, v) < 0\}, \\ F &= \{(u_1, u_2, v) \in S \mid f_v(u_1, u_2, v) = 0\}, \\ S^r &= \{(u_1, u_2, v) \in S \mid f_v(u_1, u_2, v) > 0\}. \end{aligned} \tag{2.7}$$

System (2.4) is singular along F, while the desingularized system (2.5) governs the dynamics in the vicinity of the critical manifold S. Note that the rescaling by $-f_v$ that achieves this desingularization changes the direction of time where $f_v > 0$, that is, on the repelling sheet S^r of the critical manifold S.

Roughly speaking, the original system (1.1) is governed by system (2.5) when it evolves almost on the attracting sheet S^a. The situation changes when a trajectory reaches the fold curve F, that is, when f_v becomes zero. In the generic situation, that is, if $\dot{v} \neq 0$, the prominent dynamics switches to the fast dynamics (2.6) and the trajectory escapes from S along a fast fiber parallel to the v-axis. The condition $\dot{v} = f_{u_1} g_1 + f_{u_2} v_2 \neq 0$ is called the *normal switching condition* [36] and means geometrically that the reduced flow projected onto the (u_1, u_2)-plane is not tangent to the fold curve F. The point on the fold curve F where the change of dynamics occurs is called a *jump point*. If S is S-shaped, that is, there are

two separate attracting sheets and two fold curves connected to a repelling sheet, then the presence of jump points typically leads to the existence of relaxation oscillations [36, 46]; a detailed discussion of this phenomenon is given in [29] for the forced Van der Pol system.

A point on F with $f_{u_1} g_1 + f_{u_2} v_2 = 0$ is called a *folded singularity*. At such a point it is possible for a trajectory to pass through F and not escape along a fast fiber. According to the topological type of the singularity as an equilibrium of system (2.6), one has generically *folded nodes*, *folded saddles*, and *folded foci*. Locally, the dynamics near a folded singularity can be described by normal forms. The case of a folded-node singularity is sketched in Figure 1, and several trajectories of the slow flow associated with the normal form (1.2) are shown. Notice the change of direction of the slow flow across the fold curve F. As a result, some initial conditions on S^a are attracted to regular points on F, which leads to a jump. However, an entire wedge exists, bounded by the singular canard $\widehat{\gamma}_s$ and a half-line on F that ends at the folded node, where trajectories converge to the folded node and pass through to follow the repelling sheet S^r. This wedge is called the *funnel region* [47], and it is responsible for the generic existence of canard solutions in systems with $\varepsilon > 0$.

2.2. Blow-up of the folded node. One can apply the method of blow-up in the setting of geometric singular perturbation theory [17, 18, 42]. The general idea is to rescale the variables of the original problem together with the singular parameter ε. In this way, one can transform a singularly perturbed system into a regularly perturbed system that is defined on a higher-dimensional phase space. In what follows we assume that the fold curve F is the u_2-axis, which can be achieved by a (nonlinear) coordinate transformation.

The slow manifolds S^a_ε and S^r_ε correspond to ε-leaves of three-dimensional attracting and repelling center manifolds M_a and M_r of the extended system

$$\left\{ \begin{array}{rcl} u_1' &=& \varepsilon g_1(u_1, u_2, v, \varepsilon), \\ u_2' &=& \varepsilon g_2(u_1, u_2, v, \varepsilon), \\ v' &=& f(u_1, u_2, v, \varepsilon), \\ \varepsilon' &=& 0. \end{array} \right. \tag{2.8}$$

The linearization of the extended system (2.8) has all eigenvalues equal to zero at the folded node. Hence, one cannot apply center manifold theory at points on F and describe the behavior of the slow manifolds in a neighborhood of F. A good way of overcoming this difficulty is to apply a blow-up transformation at the folded node, which is a degenerate singularity of system (2.8). Roughly speaking, the blow-up method is a well-chosen coordinate transformation that desingularizes such a degenerate singularity. It was originally developed for planar vector fields [18] but has been adapted to the case of three-dimensional singularly perturbed systems [45, 50]. The change of coordinates transforms the degenerate singularity at the origin into a sphere $\mathbb{S}^3$ that contains points with (at least) one nonzero eigenvalue. Then the general methods of dynamical systems are applicable—in particular, center manifold theory; see [45] for a detailed exposition of the blow-up in this specific context.

In the case of a three-dimensional singularly perturbed system with a folded node at the origin, the desingularizing transformation is defined by

$$u_1 = \rho^2 x, \quad u_2 = \rho y, \quad v = \rho z, \quad \varepsilon = \rho^2 \bar{\varepsilon},$$

where $(x, y, z, \bar{\varepsilon}) \in \mathbb{S}^3$ and $\rho \in [0, \rho_0]$ is a new radial parameter. The analysis on the sphere is now done using charts that yield so-called *directional rescalings* obtained by setting one coordinate in $\mathbb{S}^3$ equal to ± 1 and desingularizing the vector field in these charts. As is explained in [50], the possible intersections between the slow manifolds, that is, the existence of maximal canards, are best studied using the chart $\bar{\varepsilon} = 1$, denoted κ_2, which describes the situation on the blown-up locus. As a main result, the blow-up extends the normal hyperbolicity of the slow manifolds S^a_ε and S^r_ε to the blown-up sphere. After desingularization, the system in κ_2 is given by

$$\left\{\begin{array}{rcl} \dot{x} &=& \frac{1}{2}\mu y - (\mu+1)z + O(\rho), \\ \dot{y} &=& 1, \\ \dot{z} &=& x + z^2 + O(\rho). \end{array}\right. \tag{2.9}$$

By definition of the directional rescaling in chart κ_2 we have $\rho = \sqrt{\varepsilon}$.

The role of (2.9) is as follows. By Fenichel's theorems, for any fixed $\varepsilon > 0$ sufficiently small, the slow manifolds S^a_ε and S^r_ε exist outside a neighborhood of the fold curve F. These manifolds have extensions into the neighborhood of F, that is, into the blown-up locus; see [9, 45, 50] for the technical details. The great benefit of the blow-up transformation in chart κ_2 is that for $\varepsilon \to 0$ the extensions on the blown-up locus tend to invariant manifolds $C^\pm$ of (2.9) for $\rho = 0$, which is the normal form (1.2). Therefore, the normal form acts as a "germ" in the sense that intersections of C^+ and C^- are "seeds" that give rise to actual canards of (1.1) for $\varepsilon > 0$. Consequently, the geometry of the manifolds $C^\pm$ contains all the relevant information regarding the corresponding "true" slow manifolds S^r_ε and S^a_ε of system (1.1), respectively. Indeed, the study of the geometry of $C^\pm$ is the main topic of this paper. Note that from now on we refer to $C^\pm$ simply as repelling and attracting slow manifolds and to their intersections as canards (rather than their seeds).

2.3. Properties of the normal form. An important property of the normal form (1.2) is its invariance under the time-reversing symmetry $(x, y, z, t) \mapsto (x, -y, -z, -t)$. Therefore, it suffices to concentrate on the attracting slow manifold C^-; the repelling slow manifold C^+ is given by the symmetry. The manifolds $C^\pm$ connect at infinity (the boundary of the blown-up locus) with the respective sheets of the critical manifold of the underlying slow-fast system (1.1). Therefore, in the limit $x \to -\infty$, the manifold C^- converges to the upper sheet of the parabolic cylinder

$$S := \{(x, y, z) \in \mathbb{R}^3 \mid x + z^2 = 0\}. \tag{2.10}$$

Furthermore, the interactions of C^+ with C^- take place in the vicinity of the origin, that is, near the fold curve $F := \{(0, y, 0) \in \mathbb{R}^3\}$ of S.

A main advantage of (1.2) is that it possesses two explicit canard solutions of algebraic growth, γ_s and γ_w, given as

$$\begin{array}{rcl} \gamma_s(t) &=& \left(-\frac{\mu^2}{4}t^2 + \frac{\mu}{2},\ t,\ \frac{\mu}{2}t\right), \\ \gamma_w(t) &=& \left(-\frac{1}{4}t^2 + \frac{1}{2},\ t,\ \frac{1}{2}t\right). \end{array} \tag{2.11}$$

We refer to γ_s as the *strong canard* and to γ_w as the *weak canard*, because they correspond to the strong and the weak eigendirections of the linearization of system (1.2) at the folded node, respectively. The maximal canards γ_s and γ_w are extensions on the blown-up locus of the singular canards $\widehat{\gamma}_s$ and $\widehat{\gamma}_w$ shown in Figure 1, respectively. Note that the geometry of $C^{\pm}$ for $\mu > 1$ and for $1/\mu \in [0,1]$ is topologically the same (where the roles of γ_s and γ_w are interchanged); recall that μ denotes the ratio of the eigenvalues of system (2.5) projected onto the (y,z)-plane and linearized at the origin. Therefore, we consider here the changes of the slow manifolds $C^{\pm}$ as a function of μ only for $\mu \geq 1$.

It has been proved that $C^{\pm}$ intersect transversely along γ_s and γ_w when μ is not an integer [4, 45]. For integer values of μ the manifolds C^+ and C^- intersect transversely along γ_s and tangentially along γ_w. A new canard is created from the weak canard γ_w at every odd integer value of the parameter μ (for $\mu \geq 3$) [50]. The bifurcating canards η_i are called *secondary canards*. It was analytically proved in [45] that the slow manifolds C^- and C^+ spiral $\lfloor \mu \rfloor$ times around the weak canard γ_w (here $\lfloor q \rfloor$ denotes the integer part of the real number q). Due to the time-reversing symmetry of the normal form, this implies that C^- and C^+ make $\lfloor \frac{\mu-1}{2} \rfloor$ full rotations around each other. Each one of these full rotations ends in a transverse intersection along a secondary canard. Hence, away from the resonances $\mu \in \mathbb{N}$, there are $\lfloor \frac{\mu-1}{2} \rfloor$ secondary canards that successively make one additional complete revolution around γ_w. More precisely, η_i makes $i + \frac{1}{2}$ rotations around γ_w; see section 5. Geometrically, we can think of the strong canard γ_s as the secondary canard η_0, because γ_s makes a half-rotation around γ_w. The order is such that γ_w is always located between $\eta_{\lfloor \frac{\mu-1}{2} \rfloor}$ and $\eta_{\lfloor \frac{\mu-1}{2} \rfloor - 1}$.

3. Computing slow manifolds.

The main underlying idea of our approach is that one can compute (a finite part of) a two-dimensional invariant manifold of a system of ordinary differential equations as a collection of orbit segments by numerical continuation of a one-parameter family of two-point boundary value problems. This approach can be applied in a wide variety of contexts [34]. Below we explain how we use it to compute $C^{\pm}$ for the normal form; see [11] for details on how to compute $C^{\pm}$ for systems not in the normal form.

3.1. Slow manifolds as collections of orbit segments.

As is common in numerical continuation, we consider a vector field of the form

$$\dot{\mathbf{u}} = T\mathbf{g}(\mathbf{u}, \lambda), \tag{3.1}$$

where $\mathbf{g} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ is sufficiently smooth and $T \in \mathbb{R}$ and $\lambda \in \mathbb{R}^p$ are parameters. The parameter T is the total integration time. It appears explicitly as a free parameter on the right-hand side of (3.1), so that an orbit segment $\mathbf{u}(t)$ is always represented over the interval $[0,1]$. By imposing suitable boundary conditions on solutions of (3.1) we can characterize any k-dimensional invariant submanifold in $\mathbb{R}^n \times \mathbb{R}^p$. To be more precise, we consider the boundary conditions

$$\begin{cases} \mathbf{u}(0) \in L, \\ \mathbf{u}(1) \in \Sigma, \end{cases} \tag{3.2}$$

where L is a one-dimensional submanifold and Σ is a codimension-one submanifold of $\mathbb{R}^n$. One needs $(n-1)$ boundary conditions to restrict $\mathbf{u}(0)$ to the curve L and one boundary

condition to restrict $\mathbf{u}(1)$ to the $(n-1)$-dimensional manifold Σ. Hence, the total number of boundary conditions in (3.2) is n. That is, with T as a free parameter, (3.1)–(3.2) define a one-parameter family of well-posed two-point boundary value problems that represent orbit segments starting at L and ending in the section Σ; see, for example, [12]. The family is parametrized by the position of $\mathbf{u}(0)$ on L, and T is the integration time to reach Σ from L. Depending on the choice of L and Σ, this general setup can be used to compute different types of dynamical objects, including two-dimensional invariant manifolds [34] and their one-dimensional intersection curves with the section Σ [20].

In the present setting the phase space is three-dimensional, and our goal is to find appropriate definitions for L and Σ so that (3.1)–(3.2) define $C^{\pm}$ as surfaces in a region of interest. Since C^+ can be found from C^- by symmetry in (1.2), we explain here only the computation of C^-. The family of orbit segments obtained by continuation of (3.1)–(3.2) defines (part of) the two-dimensional manifold C^-, provided the one-dimensional submanifold L satisfies $L \subset C^-$ [23, 50]. We do not know C^-, but we do know that in the half-space $\{x \leq -\xi\}$ with $\xi > 0$ sufficiently large, C^- is well approximated by

$$S^a = \{(x, y, z) \in \mathbb{R}^3 \mid x + z^2 = 0,\ z < 0\},$$

that is, by the lower sheet of the parabolic cylinder S given in (2.10). Therefore, we define

$$L = L^-_\xi := \{(-\xi, s, -\sqrt{\xi}) \mid s \in \mathbb{R}\}, \tag{3.3}$$

which is the line on S^a with $x = -\xi$. The interesting dynamics takes place near the origin on the fold curve F of S, so a suitable choice for Σ is a plane transverse to F. We define

$$\Sigma = \Sigma_\alpha := \{y = \alpha\}, \tag{3.4}$$

where $\alpha \geq 0$. The two-point boundary value problem (3.1)–(3.2) for the choices (3.3) and (3.4) defines a one-parameter family of orbit segments that lie on C^- in good approximation, provided ξ is large enough.

3.2. Finding a first orbit segment on the slow manifold. To start the continuation we must provide a first orbit segment that solves (3.1) subject to the boundary conditions (3.2). For the normal form (1.2) two explicit canard solutions are known, which we can use to start a computation. Note that neither one of the two explicit solutions has a point in common with the lower sheet S^a of the parabolic cylinder S, which means that it is not possible to choose ξ such that the two explicit canard solutions contain segments that start on L^-_ξ and solve the boundary value problem (3.1)–(3.2). However, we use the explicit solutions only to start the Newton iteration; that is, we select a suitable explicit solution segment such that Newton's method converges to a solution of (3.1)–(3.2).

To be concrete, we start from the strong canard γ_s given in (2.11) and consider the initial orbit segment

$$\mathbf{u}(t) = \gamma_s(tT + t_0), \qquad 0 \leq t \leq 1, \tag{3.5}$$

for some start time t_0 and total integration time T. We choose $t_0 < 0$ such that the x-coordinate of $\gamma_s(t_0)$ is equal to $-\xi$, that is,

$$-\frac{\mu^2}{4}t_0^2 + \frac{\mu}{2} = -\xi.$$

In order to satisfy the second boundary condition $\mathbf{u}(1) \in \Sigma_\alpha$, we need

$$\gamma_s(T + t_0) \in \Sigma_\alpha \Leftrightarrow T + t_0 = \alpha.$$

Note that the start time t_0 must be negative, because the y-coordinate acts as time in the normal form (1.2), and we wish to preserve the direction of time. Therefore, we have

$$t_0 = -\sqrt{\frac{2\mu + 4\xi}{\mu^2}} \quad \text{and} \quad T = \alpha + \sqrt{\frac{2\mu + 4\xi}{\mu^2}}. \tag{3.6}$$

The solution segment (3.5) only approximately satisfies the boundary condition $\mathbf{u}(0) \in L_\xi^-$; namely, the difference between the z-coordinates of $\mathbf{u}(0) = \gamma_s(t_0)$ and the point on L_ξ^- at $s = t_0$ is

$$\frac{\mu}{2}\sqrt{\frac{2\mu + 4\xi}{\mu^2}} - \sqrt{\xi} = \frac{\mu}{2\left(\sqrt{\frac{\mu}{2} + \xi} + \sqrt{\xi}\right)}. \tag{3.7}$$

This difference is small, provided $\xi \gg \mu$, and decreases as $\xi \to \infty$. Hence, if ξ is large enough, we expect that Newton's method converges, and the first correction step of the continuation leads to a solution of (3.1)–(3.2). We remark that for a slow-fast system that is not in normal form an explicit solution is generally not known. This difficulty can be overcome with a homotopy approach, as is demonstrated in [11].

3.3. Computation of C^- from L_ξ^- to Σ_α. The computed part of C^- depends on the two user-specified parameters ξ and α that define L_ξ^- and Σ_α, respectively. The parameter ξ controls the accuracy of the computation in that it determines the initial distance between C^- and S^a. By construction, an orbit segment satisfying (3.1)–(3.2) converges to an actual orbit segment on C^- in the limit $\xi \to \infty$. It is a very difficult task beyond the scope of this paper to find an explicit ξ-dependent error bound for the approximation of C^- and how it depends on μ. To derive a practical measure for the accuracy of the computations presented here, we make use of the fact that the strong canard γ_s is given as an explicit solution (2.11). As mentioned in section 3.2, the orbit segment (3.5) of γ_s is the start solution for Newton's method, and we use the difference between $\gamma_s(tT + t_0)$ and the approximate solution $\mathbf{u}^*(t)$ as an indication of the overall approximation error. Namely, we consider the pointwise difference between $\gamma_s(tT + t_0)$ and $\mathbf{u}^*(t)$ with $0 \le t \le 1$ and ensure that it is sufficiently small. At $t = 0$ this difference is given by (3.7), and it decreases exponentially for $t > 0$; this decrease is particularly fast due to the difference in time scales. Specifically, we consider only the difference in the z-coordinate and require that

$$|\, \gamma_s(tT + t_0) - \mathbf{u}^*(t) \,|_z < 10^{-5}$$

for all t with $\mathbf{u}_x^*(t) > -\frac{1}{2}\xi$. In other words, this condition ensures the accuracy of the second and relevant part of the orbit segment, and we found that it is satisfied with $\xi = 100$ for all $\mu \le 8.5$ that we consider in this paper. For larger μ, also ξ needs to be increased; namely, we use $\xi = 200$ for $\mu = 14.5$, $\xi = 400$ for $\mu = 25.5$, and $\xi = 1000$ for $\mu = 49.5$.

The parameter α determines the location of the section Σ_α. Its choice depends on which aspect of $C^\pm$ is of interest. A natural choice is $\alpha = 0$, such that the folded node (the origin) is contained in Σ_α. This means that C^- and C^+ are computed both up to Σ_0, which emphasizes their intersection curves $C_0^\pm = C^\pm \cap \Sigma_0$. Note that these intersection curves have been computed before by shooting methods (numerical integration of initial values of S^a); see, for example, [28, 27, 50]. By contrast, we compute the curves $C_0^\pm$ as well as the surfaces $C^\pm$ themselves with the collocation and continuation routines of the package AUTO [14]. The main advantage of using collocation, as opposed to a shooting method, is that the size of the continuation step is determined by taking the entire orbit segment into account instead of the initial condition alone. This feature is particularly useful for slow-fast systems, which are extremely sensitive to changes of the initial condition [20]. Specifically, we compute the one-dimensional curves $C_0^\pm$ with an adaptation of the software MANBVP [20], where orbit segments are generated according to the local curvature of C_0^-. The two-dimensional surfaces $C^\pm$ are computed with an AUTO run with a fixed continuation step size, which ensures a uniform distribution of mesh points on the surface.

In order to investigate how C^- and C^+ intersect near the folded node at the origin, we consider orbit segments computed up to Σ_α with $\alpha > 0$. By symmetry, C^+ ends in $\Sigma_{-\alpha}$, so that the two slow manifolds are seen to interact in the region $-\alpha \le y \le \alpha$. To visualize the geometry of this interaction it is convenient to show only the "ribbons" of C^- and C^+ in between the planes $\Sigma_{-\alpha}$ and Σ_α. To this end, we clip each orbit segment of the computed manifold C^- where it intersects $\Sigma_{-\alpha}$. We then determine a mesh with a fixed number of mesh points that are uniformly distributed according to arclength along clipped orbit segments. The resulting ribbons of $C^\pm$ can readily be visualized; see section 4.

3.4. Illustration of the method. Figure 2 illustrates our method for the normal form (1.2) with $\mu = 1.2$. Figure 2(a) shows an approximation of the attracting slow manifold C^- for $\alpha = 0$; that is, we computed a collection of orbit segments that start on L_{100}^-, the straight red line in Figure 2(a), and end in Σ_0, the (x, z)-plane shown in green. The bold red curve in Σ_0 is the intersection C_0^- of C^- with Σ_0, and a small segment of its symmetrical image C_0^+ (blue curve, not labeled) is also shown. We started the continuation from the solution segment (bold black curve) $\gamma_s(tT + t_0)$, $t \in [0, 1]$, of the explicitly known strong canard γ_s with t_0 and T as defined in (3.6). The continuation is done in two directions parametrized by the y-coordinate along L_{100}^-, where we start at $y = t_0 < 0$.

Let us first focus on the continuation run where y increases, because this part generates most of the intersection curve C_0^-. Note that it is natural to stop the continuation when $y = 0$ is reached. (This is detected by a user-defined function in AUTO; the solution family exists for $y > 0$, but then T becomes negative.) Notice from Figure 2(a) that the curve C_0^- is very close to the lower branch of the parabola $S \cap \Sigma_0 = \{x + z^2 = 0\}$. At the scale of Figure 2(a) it is difficult to see what happens near the folded node; an enlarged view of $C_0^\pm$ is presented in the next section.

During the second continuation run, when y is decreasing, we encounter the orbit segment between L_{100}^- and Σ_0 that corresponds to the weak canard γ_w, for which we have an explicit expression given in (2.11). Note that, due to the symmetry of the normal form, any intersection of the curve C_0^- with the line $z = 0$ in the section Σ_0 corresponds to a canard. Hence, by imposing the user-defined function $\mathbf{u}_z(1) = 0$ as part of the continuation of the one-parameter

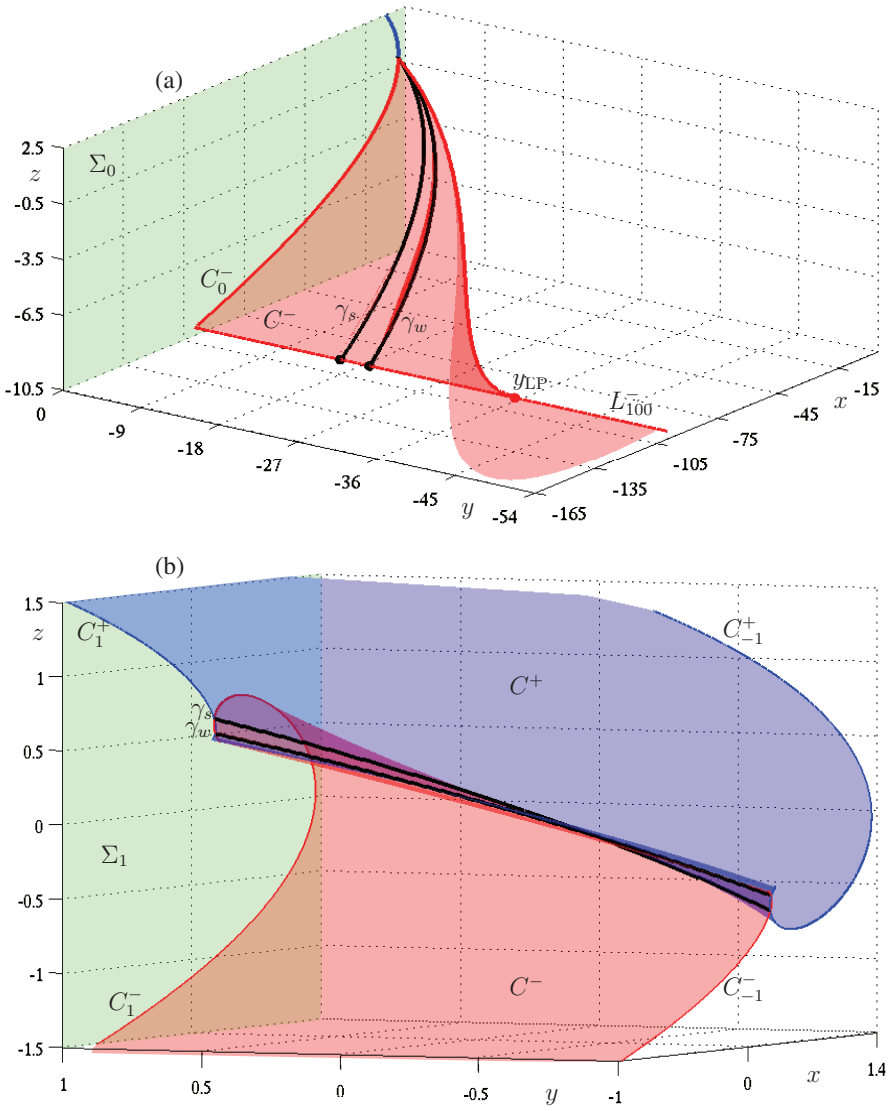


Figure 2. *Global overview of the slow manifolds for $\mu = 1.2$. Panel* (a) *shows the attracting slow manifold C^- (red surface) computed from the line of initial conditions L_{100}^- (red line) up to section Σ_0; panel* (b) *shows the parts of C^- and C^+ (red and blue surfaces, respectively) in between the sections Σ_{-1} and Σ_1. The two primary canards γ_s and γ_w have been highlighted as bold black curves; the red bold curve (panel* (a) *only) is the orbit starting at the intersection point $y_{\rm LP}$ between L_{100}^- and the x-nullcline. We also show the intersection curves $C_\alpha^\pm$ of the slow manifolds $C^\pm$ with the sections Σ_0 and $\Sigma_{\pm 1}$. See also the accompanying animation* (70881_01.gif [5.9MB]) *of the computation of C^- for $\mu = 8.5$.*

family that solves (3.1)–(3.2), AUTO [14] automatically detects these canard solutions of the normal form (1.2); a more detailed discussion is provided in section 5. The actual canard solutions are obtained by concatenating the detected orbit segment on C^-, which ends in Σ_0 on the line $z = 0$, with its symmetrical copy on C^+ on the other side of Σ_0.

As y decreases further, we encounter another special solution during the continuation, which is shown as the bold red curve in Figure 2(a) that starts at the point labeled y_{LP} on L_{100}^{-}. The point

$$y_{\mathrm{LP}} = -\frac{2(\mu+1)}{\mu}\sqrt{\xi}$$

is the unique point where the x-direction of the vector field (1.2) vanishes on L_{100}^{-}; that is, $y_{\mathrm{LP}} \approx 36.777$ is the intersection of L_{100}^{-} with the x-nullcline for $\mu = 1.2$. If L_{100}^{-} were exactly on C^{-}, then all initial conditions on L_{100}^{-} beyond y_{LP}, that is, with y-coordinates less than y_{LP}, would lie on (backward-extended) orbit segments that intersect L_{100}^{-} at y-coordinates larger than y_{LP}. This behavior corresponds exactly to the case that an (un)stable manifold in a Poincaré section crosses the locus where the flow is tangent to the section; see [20] for more details.

In practice L_{100}^{-} lies only approximately on C^{-} and solutions beyond y_{LP} do not lie exactly on the computed approximation of C^{-} but still very close to it. Hence, it appears as though a new part of C^{-} is obtained, which manifests itself as a very sharp fold, or "crease," on the approximation of C^{-}. From a computational point of view, the continuation makes sense only for $\mathbf{u}(0) \in [0, y_{\mathrm{LP}}]$, because continuation beyond y_{LP} produces a second approximation of the same part of C^{-}. Figure 2(a) does show a computation of C^{-} past y_{LP} to illustrate what happens. Note that $y_{\mathrm{LP}} \to -\infty$ as $\xi \to \infty$, that is, in the limit where L_{ξ}^{-} converges to a line on C^{-}, the point y_{LP} no longer exists.

Figure 2(b) demonstrates how the ribbons of $C^{\pm}$ in between Σ_1 and Σ_{-1} can be used as a means of visualizing the interaction of the two manifolds. For clarity, the intersection curves $C_1^{\pm}$ and $C_{-1}^{\pm}$ are shown as well. The geometry of $C^{\pm}$ is further enhanced by including the strong and weak canards γ_s and γ_w, respectively.

4. Geometry of the slow manifolds. We now study the slow manifolds $C^{\pm}$ for different values of the parameter μ. We use both $\alpha = 0$ and $\alpha > 0$ in the method from section 3 to illustrate not only the intersection curves $C_0^{\pm}$ of $C^{\pm}$ with Σ_0, as was done in [50], but also the geometry of the two-dimensional slow manifolds $C^{\pm}$ themselves. A main goal is to see how maximal canards arise as new intersection curves between C^{-} and C^{+}. In all figures the attracting slow manifold C^{-} is colored red, the repelling slow manifold C^{+} is blue, the section Σ_α is green, and the strong canard γ_s and the weak canard γ_w are black. As μ is increased, secondary canards appear, which we label successively as η_i. We adopted a particular color coding for these secondary canards: η_1 is orange, η_2 is magenta, η_3 is cyan, and we repeat these successive colors for each group of three consecutive secondary canards after η_3.

4.1. Geometry of $C^{\pm}$ up to Σ_0. We begin with a series of images for $\mu = 1$, $\mu = 1.2$, $\mu = 2.5$, $\mu = 3.5$, and $\mu = 8.5$ that illustrate the behavior of $C^{\pm}$ up to the section Σ_0; see Figures 3(a),(b), 4(a),(b), and 5(a), respectively. Each figure shows a three-dimensional view in a neighborhood of the folded node of C^{+} and C^{-}, computed up to Σ_0, together with the corresponding intersections in the plane Σ_0. To facilitate comparison and analysis, the viewpoint and aspect ratio are identical for all three-dimensional pictures, although the ranges along the axes vary.

Figure 3(a) shows the case $\mu = 1$, which acts as the starting point where $C^{\pm}$ have a nontransverse, tangent intersection along a single orbit. The first two (maximal) canards γ_s

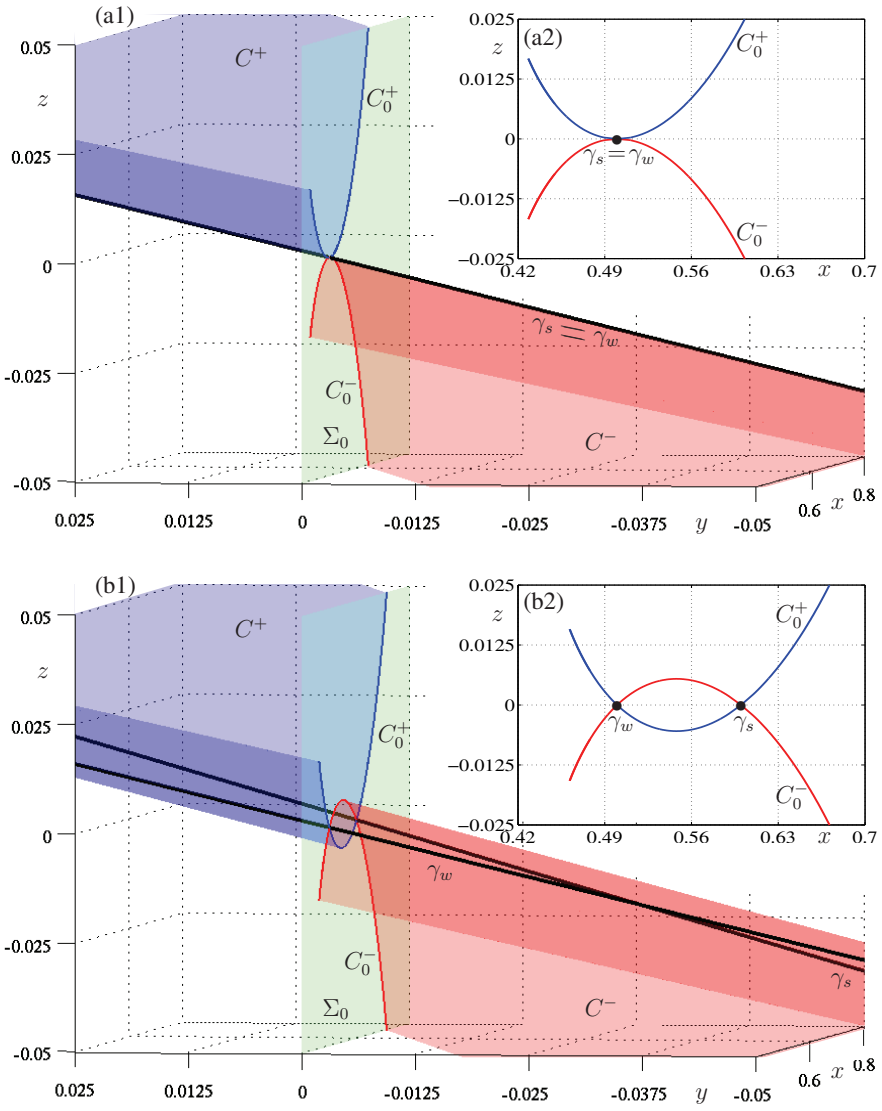


Figure 3. *Local three-dimensional views of the slow manifolds $C^{\pm}$ computed up to section Σ_0; the inset panels show their intersections $C_0^{\pm}$ with Σ_0. Panels* (a) *are for $\mu = 1$ and panels* (b) *are for $\mu = 1.2$. The weak canard γ_w and the strong canard γ_s are shown as black curves and their intersections with Σ_0 are denoted by black dots. See also Figures 4 and 5.*

and γ_w are created as μ is increased. They are shown in Figure 3(b) for $\mu = 1.2$; see also Figure 2. Note that γ_s and γ_w are now two distinct orbits of the normal form (1.2) in which the slow manifolds $C^{\pm}$ intersect transversely. The case $\mu = 1.2$ is representative of all values $1 < \mu < 2$. Notice in Figure 3 the parts of $C_0^{\pm}$ to the left of γ_w, which correspond to the part of $C^{\pm}$ between γ_w and the orbit of the point y_{LP}. We remark that these parts, which we call the tips of $C_0^{\pm}$, exist for all values of μ, but they rapidly become so small that they are invisible in Figures 4 and 5(a).

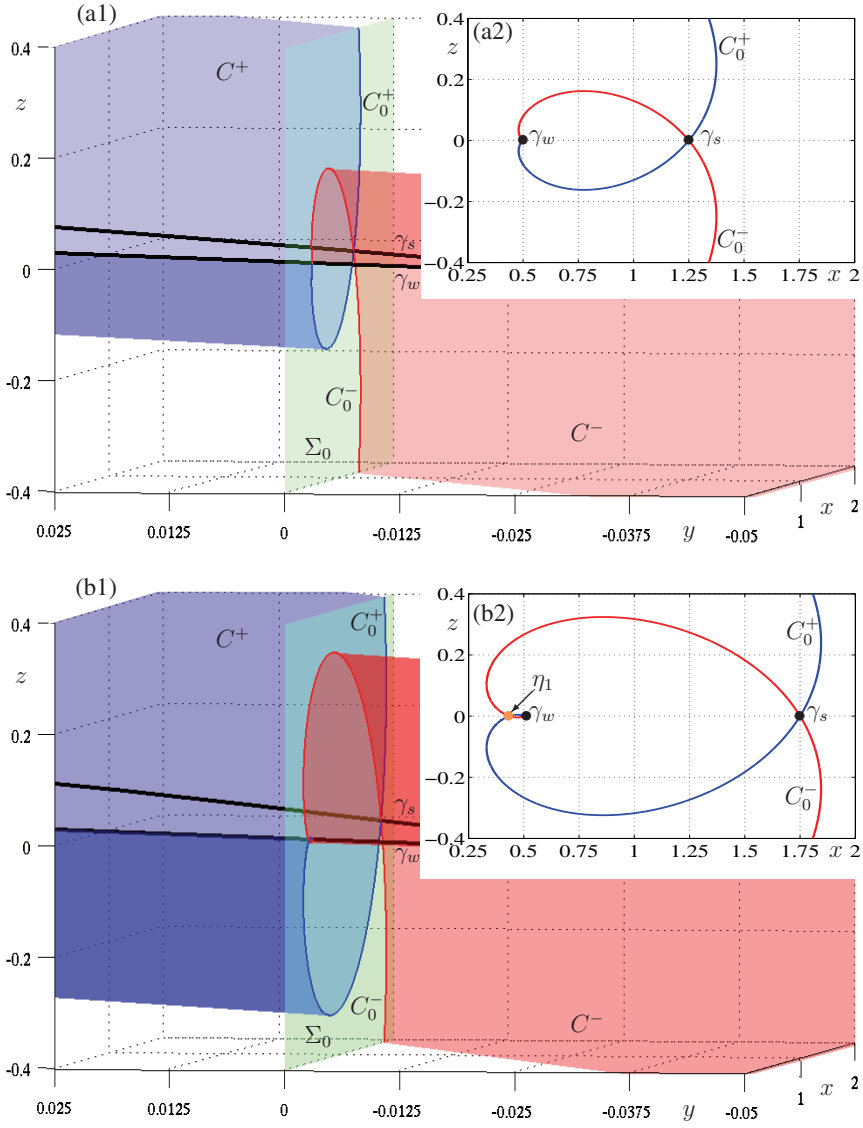


Figure 4. *Local three-dimensional views of the slow manifolds $C^{\pm}$ computed up to section Σ_0; the inset panels show their intersections $C_0^{\pm}$ with Σ_0. Panels* (a) *are for $\mu = 2.5$ and panels* (b) *are for $\mu = 3.5$. Note the existence of the first secondary canard η_1 (orange dot) in panel* (b2)*, which appears in a transcritical bifurcation at $\mu = 3$. See also Figures 3 and 5.*

A qualitative change occurs at $\mu = 2$. Figure 4(a) shows the situation for $\mu = 2.5$. As can be observed particularly in panel (a2), the tips of $C_0^{\pm}$ have rotated around so that they now point inside the region delimited by γ_s and γ_w. Indeed, these tips rotate continuously with μ. At $\mu = 2$ the tangent bundles $T_{\gamma_w} C^{\pm}$ coincide and the directions in Σ are parallel to the z-axis; that is, $C_0^{\pm}$ both have a vertical tangency vector at $\gamma_w(0)$. Numerical evidence in [50] suggests that near $\mu = 2$, and indeed near all even μ, there are (very) short branches of canards that

correspond to intersection points of the tip of C_0^- with C_0^+ and vice versa. For $\mu = 3$ the tips have rotated so that the tangent bundles $T_{\gamma_w}C^\pm$ coincide again, but $C_0^\pm$ now both have a tangency vector parallel to the x-axis, and a transcritical bifurcation occurs that results in a secondary canard (again, there are very short branches associated with intersections of the tips of $C_0^\pm$). As shown in [50], transcritical bifurcations occur for all odd integer values of $\mu \geq 3$. Figure 4(b) shows a phase portrait for $\mu = 3.5$ with the first secondary canard η_1 (orange curve). As can be seen clearly in Figure 4(b2), there are now three intersection points of $C_0^\pm$ in Σ_0. Note that γ_w is located between η_1 and γ_s on the x-axis.

Figure 5(a) shows the situation for $\mu = 8.5$, where we have three secondary canards, denoted η_1 (orange), η_2 (magenta), and η_3 (cyan). It gives an idea of how the secondary canards appear as a result of the spiraling motion of $C^\pm$ around the weak canard γ_w; this is particularly visible for $C_0^\pm$ in Σ_0, shown in Figure 5(a2) and in the accompanying animation (70881_02.gif [3.7MB]). The figure also illustrates the fact that $C^\pm$ make $\lfloor \frac{\mu-1}{2} \rfloor$ full rotations around γ_w.

4.2. Ribbons of $C^\pm$ near the folded singularity. Figures 3, 4, and 5(a) give a good insight into the topological changes of the geometry of $C^\pm$. After the bifurcation of maximal canards at $\mu = 1$, all secondary maximal canards bifurcate from γ_w in transcritical bifurcations at odd integer values of μ. To bring out this behavior more clearly, we also compute ribbons of $C^\pm$ in between $\Sigma_{-\alpha}$ and Σ_α for suitable $\alpha > 0$. Figure 5(b) shows $C^\pm$ in between Σ_{-1} and Σ_1, along with the intersection curves γ_s, γ_w (black curves), η_1 (orange), and η_2 (magenta); note that η_3 is not shown in this picture, because it cannot be distinguished from γ_w at this scale. The intersection curves $C_{-1}^\pm$ and $C_1^\pm$ in the bounding sections Σ_{-1} and Σ_1 give an idea of how C^- and C^+ spiral out past the origin in forward and backward time, respectively.

Figure 6 shows $C^\pm$ for much larger values of μ, namely, for $\mu = 25.5$ and $\mu = 49.5$ in panels (a) and (b), respectively. The slow manifolds spiral out faster as μ increases; hence, we now show ribbons of $C^\pm$ only in between $\Sigma_{-0.5}$ and $\Sigma_{0.5}$. The figure shows the increased complexity of $C^\pm$ with many more intersection curves that form additional secondary canards. For $\mu = 25.5$ there are twelve secondary canards that wind around γ_w; only the first four are labeled in Figure 6(a). For $\mu = 49.5$ there are 24 secondary canards, but again only η_1, η_2, η_3, and η_4 are labeled in Figure 6(b). Due to the increased spiraling amplitude of the intersection curves $C_{-0.5}^\pm$ and $C_{0.5}^\pm$, the strong canard and the secondary canards now lie further away from the weak canard than for $\mu = 25.5$ (or any $\mu < 49.5$); see also the accompanying animation (70881_03.gif [1.3MB]).

Figure 7 illustrates the increasing complexity with μ by showing $C^\pm$ for $\mu = 8.5$, $\mu = 14.5$, $\mu = 25.5$, and $\mu = 49.5$. Here we rotated the slow manifolds $C^\pm$ about the z-axis with the visualization package Geomview [40] to generate common enlarged views centered around the weak canard γ_w. In this way, one obtains a good impression of the spiraling behavior of the secondary canards around γ_w and how their positions and distances to γ_w change with μ. Figure 7(a) shows that for $\mu = 8.5$ the manifolds C^- and C^+ intersect in the two secondary canards η_1 and η_2. For $\mu = 14.5$ there are a total of six secondary canards, four of which, η_1 to η_4, are shown in Figure 7(b). Note how the distance of η_1 and η_2 from the central weak canard γ_w (black curve) is now much larger for $\mu = 14.5$; in a way, this creates space for η_3 and η_4 to spiral around γ_w as well. For $\mu = 25.5$ there are twelve secondary canards in total, but only η_1 to η_4 are labeled in Figure 7(c). Similarly, for $\mu = 49.5$ in Figure 7(d) there

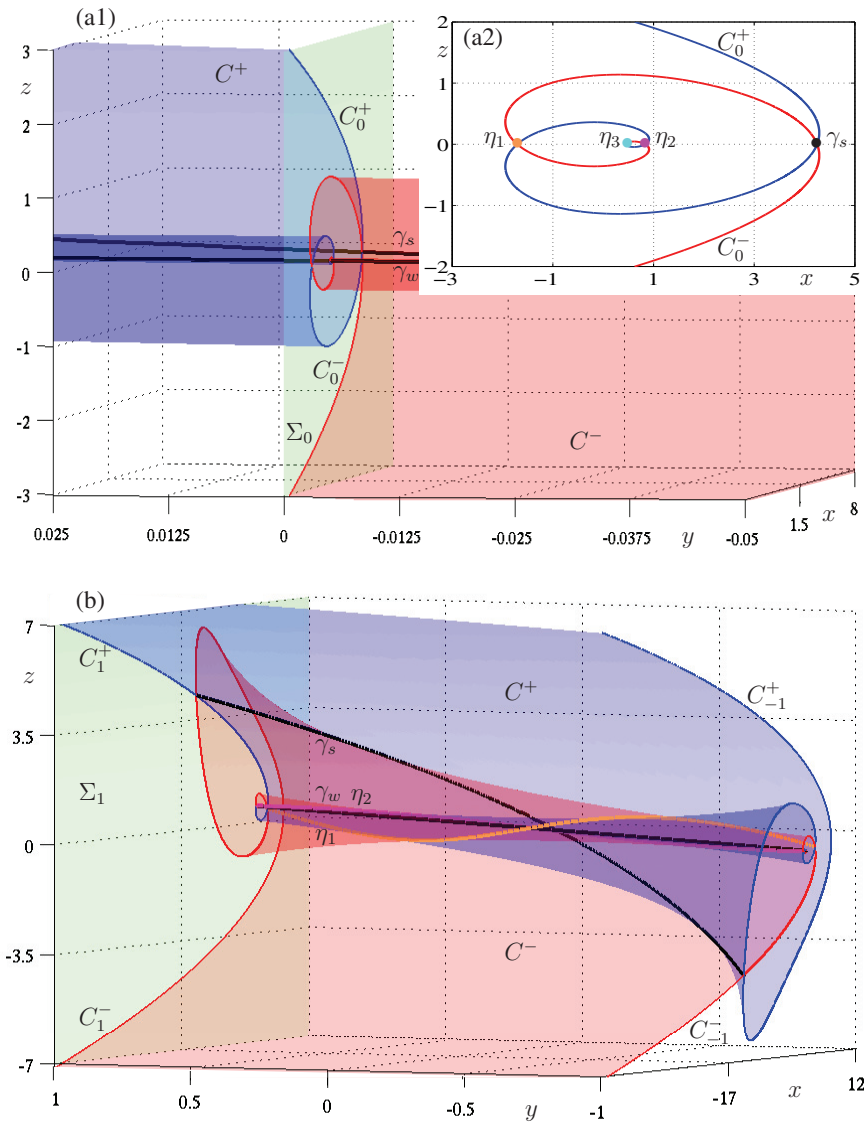


Figure 5. *Local three-dimensional views of the slow manifolds $C^\pm$ for $\mu = 8.5$ computed up to section Σ_0 in panels* (a)*; the inset panel* (a2) *shows their intersections $C_0^\pm$ with Σ_0. See also Figures 3 and 4 and the accompanying animation* (70881_02.gif [3.7MB]). *Panel* (b) *shows the ribbons of $C^\pm$ in between Σ_{-1} and Σ_1 along with the corresponding intersection curves $C_{-1}^\pm$ and $C_1^\pm$. There are three secondary canards η_1, η_2, and η_3, indicated by colored dots in Σ_0 in panel* (a2).

are 24 secondary canards, of which the first nine are clearly visible, while only η_1 to η_4 are labeled.

5. Geometric study of the secondary canards. As we have seen, the secondary canards arise as intersections of the slow manifolds C^- and C^+. We now find them directly as special orbits within the boundary value problem setup in section 3. This allows us to visualize and

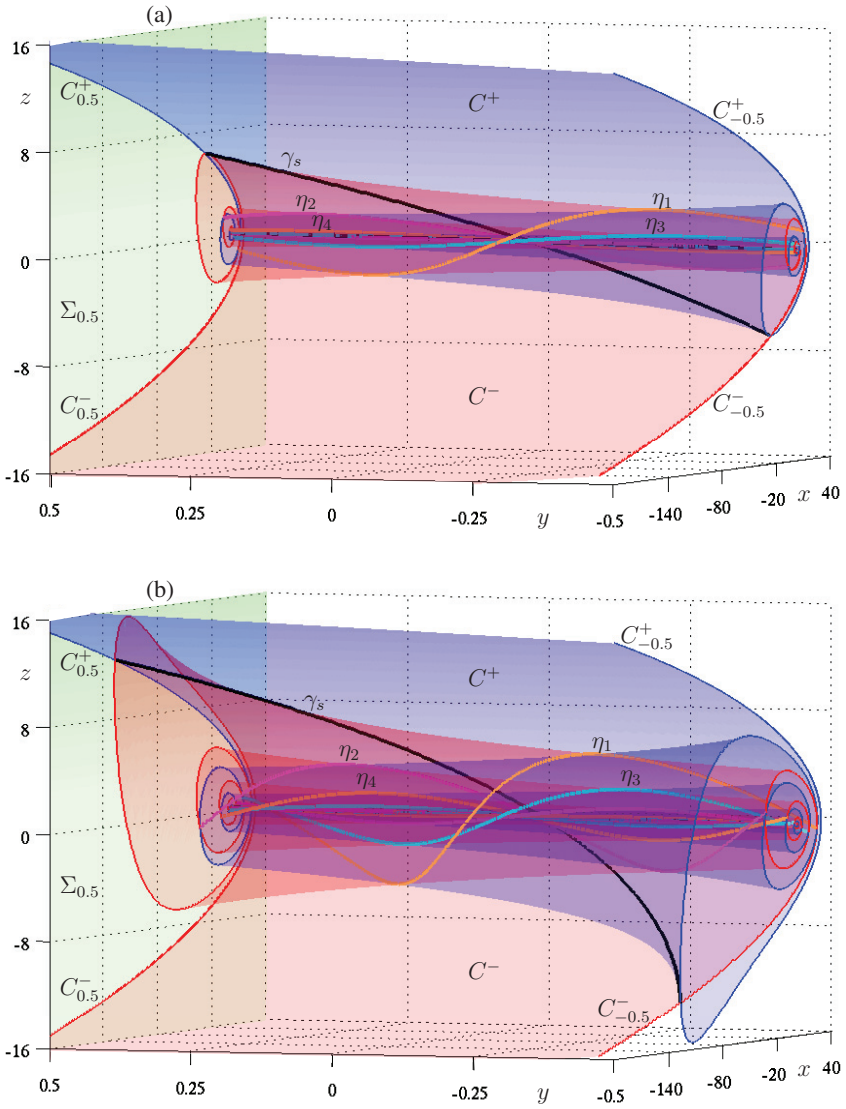


Figure 6. *Three-dimensional views of the ribbons of $C^{\pm}$ in between $\Sigma_{-0.5}$ and $\Sigma_{0.5}$, together with all maximal canards and the intersection curves $C^{\pm}_{\pm 0.5}$. Panel* (a) *is for $\mu = 25.5$, with twelve secondary canards, and panel* (b) *is for $\mu = 49.5$, with 24 secondary canards. See also Figures* 2(a) *and* 5(b) *and the accompanying animation* (70881_03.gif [1.3MB]).

discuss their spiraling behavior with respect to the weak canard γ_w. Furthermore, we continue the secondary canards in the parameter μ to reveal an overall bifurcation diagram.

5.1. Detection of secondary canards. During the continuation of (3.1)–(3.2) for fixed μ, the end points $\mathbf{u}(1) \in \Sigma_0$ of the computed orbit segments oscillate about the z-axis; see, for example, Figure 5(a2). Secondary canards are detected by the condition that the z-coordinate $\mathbf{u}_z(1)$ satisfies $\mathbf{u}_z(1) = 0$, which is done during the continuation by monitoring a user-defined function in AUTO. Recall that the point $\mathbf{u}_z(1)$ is a function of the y-coordinate $\mathbf{u}_y(0)$ of the

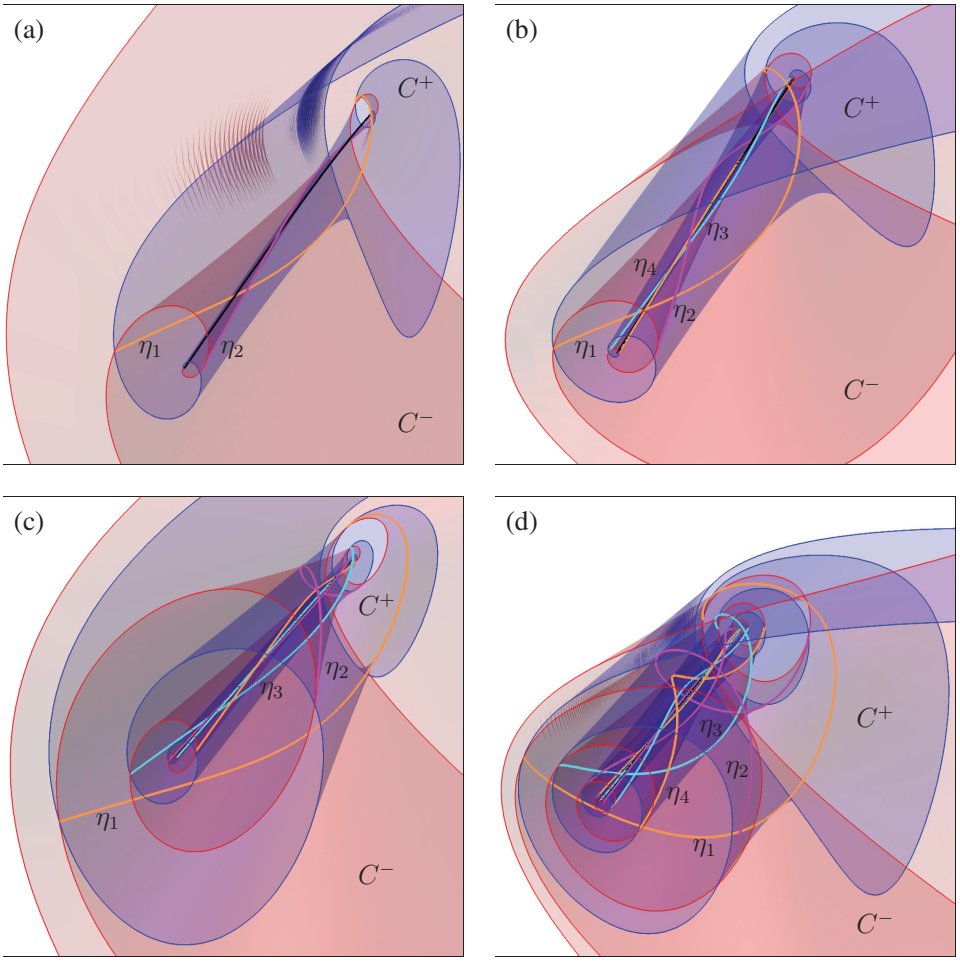


Figure 7. *Different views of the attracting and repelling slow manifolds C^- and C^+, respectively, to illustrate the spiraling dynamics of the secondary canards around the weak canard γ_w (black curve). The parameter μ is 8.5 in panel* (a)*, $\mu = 14.5$ in panel* (b)*, $\mu = 25.5$ in panel* (c)*, and $\mu = 49.5$ in panel* (d)*.*

point $\mathbf{u}(0)$ that varies along L_ξ^-.

Figure 8 shows the graph of $\mathbf{u}_z(1)$ as a function of $\mathbf{u}_y(0)$ for $\mu = 49.5$, where we show data for the run that starts from the strong canard γ_s for which we have $\mathbf{u}_y(0) = t_0 \approx -1.29$. Due to the spiraling nature of C_0^-, the graph oscillates with a rapidly decreasing amplitude; note that the continuation is in the direction of negative $\mathbf{u}_y(0)$. The enlargement in the inset of Figure 8 shows the oscillation of $\mathbf{u}_z(1)$ in the region where η_{12} to η_{15} are detected and the oscillation amplitude has decreased to values of order 10^{-4}. Numerically it becomes increasingly difficult to detect where $\mathbf{u}_z(1)$ changes sign when the oscillation amplitude becomes very small. In other words, for large μ, as in Figure 8, for $\mu = 49.5$, it is quite a challenge to detect the secondary canards that lie very close to γ_w.

For all values of μ in this paper we start the continuation with the AUTO accuracy settings

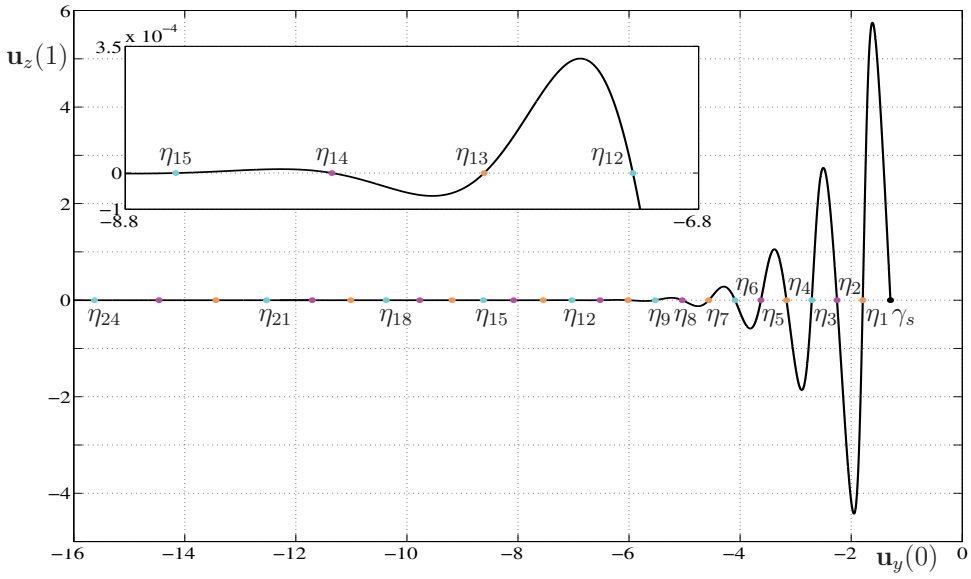


Figure 8. *Graph of the z-coordinate* $\mathbf{u}_z(1)$ *of the end point* $\mathbf{u}(1) \in \Sigma_0$ *of the computed orbit segments on* C^- *for* $\mu = 49.5$ *as a function of the y-coordinate* $\mathbf{u}_y(0)$ *of the point* $\mathbf{u}(0) \in L^-_\xi$*, where we used* $\xi = 1000$*. The continuation starts from the strong canard* γ_s*, and secondary canards* η_i *are detected when* $\mathbf{u}_z(1) = 0$*.*

Table 1

AUTO accuracy parameters as used during the detecting the secondary canards of (1.2)*. Row* (a) *is our regular accuracy setting, and row* (b) *is the increased accuracy as used for* $\mu = 49.5$*. Here* NTST *is the number of mesh points,* NCOL *the number of collocation points,* DSMIN *and* DSMAX *are the minimal and maximal stepsizes for the continuation, and* EPSS *is the relative arclength convergence criterion for the detection of special solutions; all other AUTO accuracy parameters are set to their default values.*

	NTST	NCOL	DS	DSMIN	DSMAX	EPSS
(a)	200	4	0.001	5×10^{-4}	0.01	10^{-4}
(b)	400	6	0.001	10^{-7}	0.01	10^{-7}

as shown in row (a) of Table 1. This is sufficient for the reliable detection of the η_i even for $\mu = 25.5$, but for the case $\mu = 49.5$ shown in Figure 8 the detection stops when extrema of $\mathbf{u}_z(1)$ are less than 10^{-9} in modulus. At this stage the secondary canards η_1 to η_{17} have been detected reliably. The next four secondary canards η_{18} to η_{21} are found in a second run, where we increase the accuracy parameters to the settings given in row (b) of Table 1. Nevertheless, the detection of η_{22} to η_{24} is very difficult even with the increased accuracy settings, because $\mathbf{u}_z(1)$ is now consistently below 10^{-15} in modulus. Since we are reaching the limit of machine precision, spurious roots of $\mathbf{u}_z(1)$ are reported, out of which we need to select η_{22} to η_{24}. This can be done by taking into consideration the distance between roots in $\mathbf{u}_y(0)$, which leads to a selection that is consistent with the detected secondary canards η_1 to η_{21}; see Figure 11(d).

Once a secondary canard has been detected for a fixed value of μ as a zero of $\mathbf{u}_z(1)$, it can be continued in the parameter μ by imposing $\mathbf{u}_z(1) = 0$ as an additional boundary condition. In this way, we can compute μ-dependent families of detected secondary canards.

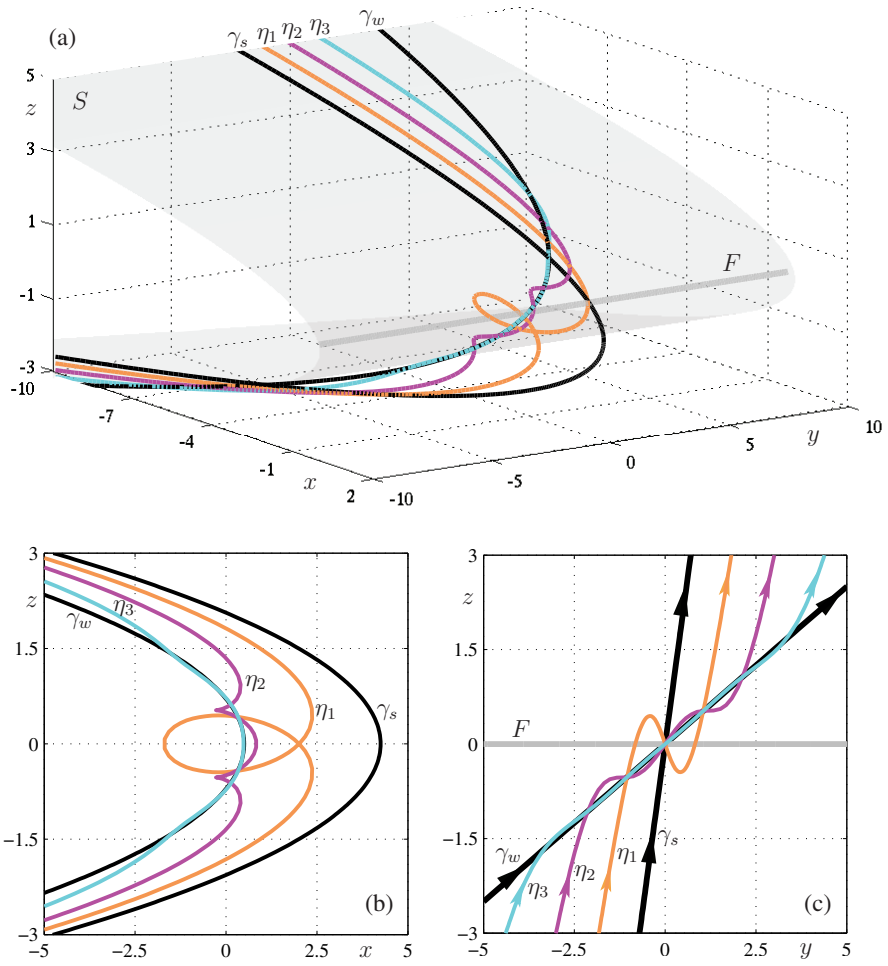


Figure 9. *Spiraling behavior of the secondary canards of* (1.2) *with* $\mu = 8.5$. *Panel* (a) *shows the primary canards* γ_s *and* γ_w *(black) and the three secondary canards* η_1, η_2, *and* η_3; *also shown for orientation is the parabolic cylinder* S *with its fold curve* F. *Panels* (b) *and* (c) *show the projections onto the* (x, z)- *and* (y, z)-*planes, respectively.*

We remark here that the oscillations of $\mathbf{u}_z(1)$ near a fixed canard increase as μ is increased. Hence, canards η_i for large i can be detected reliably for larger μ and then continued back into the range of lower values of μ.

5.2. Spiraling behavior of the secondary canards. To explain the spiraling of the secondary canards around the weak canard, we concentrate on the case $\mu = 8.5$, for which there are three secondary canards, η_1 to η_3. They are shown in Figure 9 together with the primary canards γ_s and γ_w (black curves). Figure 9(a) is a three-dimensional view of the canards, where we also show for orientation the parabolic cylinder S (grey) with its fold curve F (thick grey line). The secondary canards η_i lie seemingly parallel to γ_s for $|x|$ large but follow γ_w near F. With increasing i the η_i lie closer to γ_w as they spiral increasingly around it. Figures

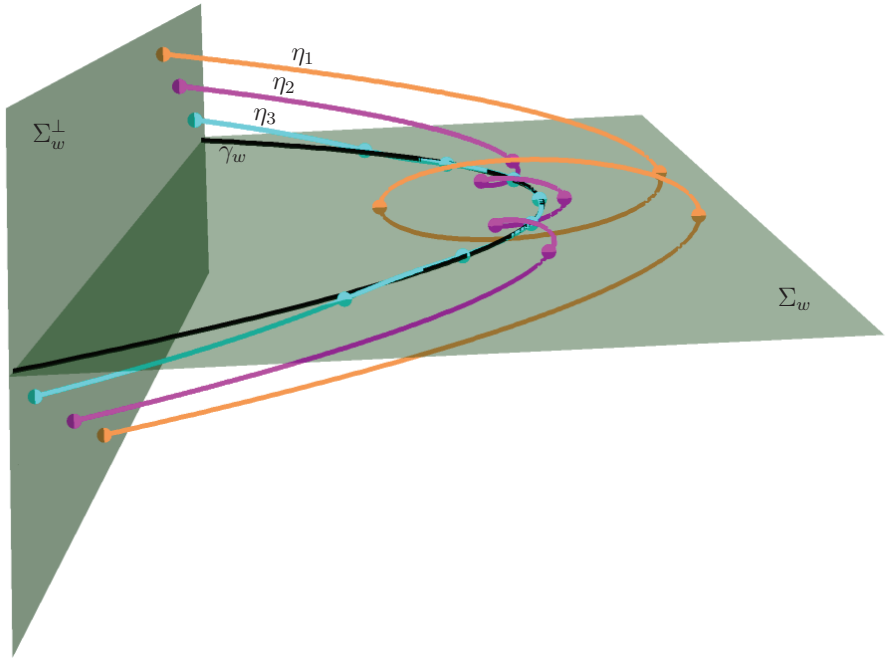


Figure 10. *Detailed view of the three secondary canards and γ_w for $\mu = 8.5$. The weak canard γ_w lies in the plane Σ_w and the three secondary canards start and end in a suitable plane $\Sigma_w^\perp$ perpendicular to Σ_w. The intersections of the secondary canards with Σ_w are marked by dots to emphasize the rotations around γ_w.*

9(b) and (c) are projections of the primary and secondary canards onto the (x, z)- and (y, z)-planes, respectively. Figure 9(b) illustrates how each new secondary canard makes one more full rotation around the weak canard γ_w. Figure 9(c) focuses on the (slow) dynamics of the secondary canards in the neighborhood of the fold.

Figure 10 is a visualization with the package Geomview [40] in the spirit of a wire and cardboard model to bring out the spiraling of the secondary canards η_1 to η_3 around γ_w. Namely, shown are the plane $\Sigma_w = \{y = 2z\}$ that contains γ_w and the plane $\Sigma_w^\perp = \{x = -10\}$ that is perpendicular to Σ_w chosen so that all spiraling behavior is captured. The secondary canards η_1 to η_3 start on $\Sigma_w^\perp$ below Σ_w and return to $\Sigma_w^\perp$ above Σ_w. Notice that η_1 has three intersection points (yellow dots), η_2 has five intersection points (magenta dots), and η_3 has seven intersection points (blue dots) with Σ_w. This illustrates the theoretical results that η_i makes $\lfloor \frac{\mu-1}{2} \rfloor$ rotations around γ_w. Figure 10 also illustrates that the secondary canards η_i lie successively closer to γ_w in the region of the fold.

The spiraling character of a secondary canard η_i does not depend on the value of μ. When a secondary canard is created in a transcritical bifurcation closest to γ_w at an odd integer value of μ, its rotating property is fixed. This is illustrated in Figure 11 with projections onto the (y, x)-plane of all canards for $\mu = 8.5$, $\mu = 14.5$, $\mu = 25.5$, and $\mu = 49.5$, respectively. For each case we chose a region of the (y, x)-plane that allows for a comparison between the panels; specifically, the x-maximum of γ_s is fixed, and the y-range is adjusted so that the last secondary canard is seen to "leave" γ_w. Figure 11(a) for $\mu = 8.5$ should be compared directly

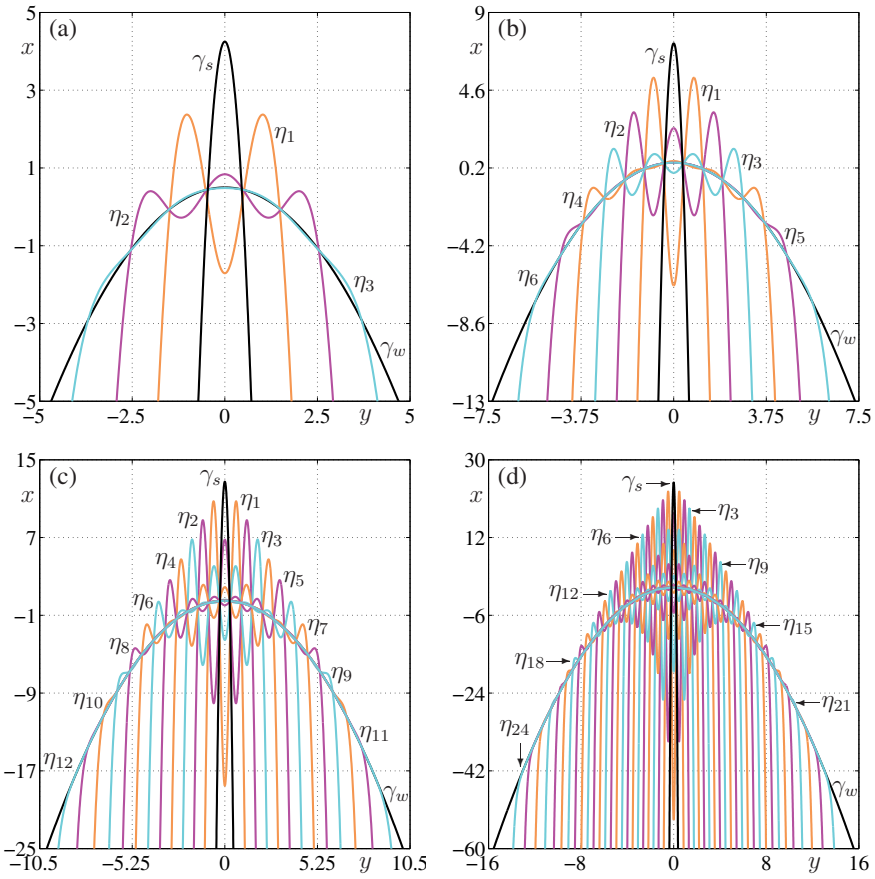


Figure 11. *Representation in the (y, x)-plane of the primary and secondary canards for four different values of μ. From panel* (a) *to panel* (d) *$\mu = 8.5$, $\mu = 14.5$, $\mu = 25.5$, and $\mu = 49.5$, respectively; the shown parts of the (y, x)-plane are chosen so that the maximum of γ_s appears in the same location and all the computed canards are covered.*

with Figure 9. There are six secondary canards in Figure 11(b), twelve in Figure 11(c), and 24 in Figure 11(d). Overall, the secondary canards run parallel to (have the same slope as) the strong canard γ_s for large $|x|$ and then spiral around γ_w. Observe that the maxima of the η_i appear to line up along a curve that connects γ_w to the maximum of γ_s; an initial investigation showed that this curve is not a straight line. It would be an interesting challenge to study limiting features of the canards in a suitably rescaled (y, x)-plane for μ tending to ∞.

5.3. Continuation of the secondary canards in μ. It is a particular advantage of our boundary value problem setup that secondary canards can be continued in the parameter μ. Figure 12 shows the result of the continuation of the secondary canards η_1 to η_{16}, where we plot the x-coordinate $\mathbf{u}_x(1)$ of the end point in Σ_0. Also shown in this bifurcation diagram are the primary canards γ_w and γ_s. They are determined from (2.11) as the straight lines $\mathbf{u}_x(1) = \frac{1}{2}$ and $\mathbf{u}_x(1) = \frac{\mu}{2}$, respectively, which intersect transversely at $\mu = 1$. The continuation was

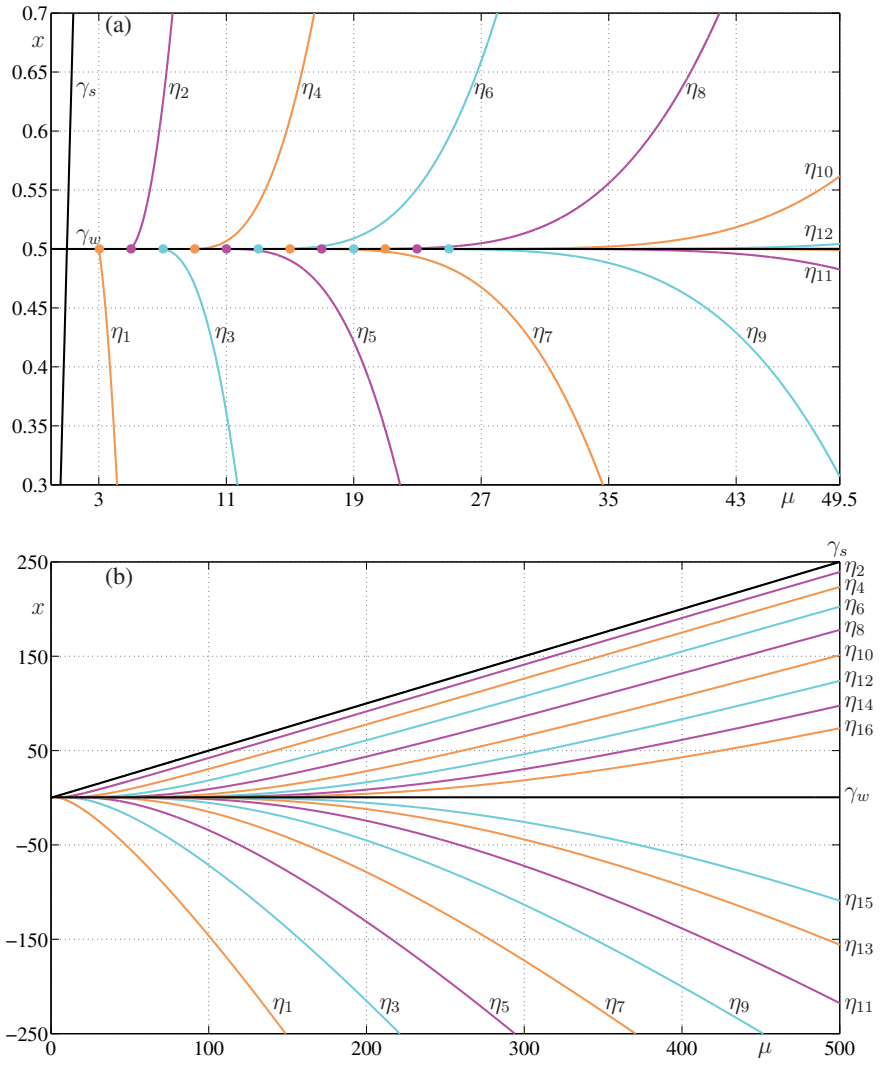


Figure 12. *Continuation in μ of η_i for $1 \le i \le 16$. Shown are the projections of η_i onto the (μ, x)-plane with γ_s and γ_w included for reference. Panel* (a) *shows how the first twelve secondary canards bifurcate from γ_w at odd integer values of μ; the bifurcation points for η_1 to η_{12} are indicated by thick colored dots. Panel* (b) *shows a larger view of the continuation up to $\mu = 500$.*

started by detecting all twelve secondary canards for $\mu = 25.5$, where we used $\xi = 1000$ to ensure sufficient accuracy for their continuation for $\mu > 25.5$. Figure 12(a) shows how η_1 to η_{12} bifurcate from γ_w at odd integer values. Figure 12(b) shows all sixteen branches η_i for the much larger μ-range up to $\mu = 500$.

The computed Figure 12 should be compared with the sketch provided in [50, Fig. 17]. Wechselberger showed in [50] that the bifurcation diagram is organized by transcritical bifurcations at odd μ-values. Notice that, on the scale of Figure 12(a), the branches η_i for $i > 3$

appear to be tangent to the branch γ_w, rather than making an angle with it as one would expect in a transcritical bifurcation. This illustrates the fact that the angle of the branch η_i with the branch γ_w decreases extremely rapidly with i; see the proofs of [50, Propositions 3.2 and 3.3]. It may be of interest to note that on the scale of Figure 12(b) the branches η_i for even i appear to have the same slope for large μ as the branch γ_s. By contrast, on this scale one cannot conjecture any convergence of the slopes of the branches η_i for odd i.

According to [50], there are two quite subtle features of the bifurcation diagram whose computation is beyond the scope of this paper. First, there are very short branches η_i that exist for μ just before the transcritical bifurcations. We found that the numerical calculation becomes so sensitive that it already stops before the associated transcritical bifurcation point is reached. Second, numerical evidence in [50] suggests that there are also very small branches of secondary canards associated with pitchfork bifurcations at even μ-values. In contrast to all other canards, these secondary canards occur in symmetric pairs and do not intersect the x-axis in Σ_0. Hence, they would need to be detected directly as an intersection of C_0^- and C_0^+.

6. Beyond the normal form. It is in the nature of a normal form that (1.2) has special properties. Specifically C^- and C^+ are each other's images under a symmetry operation. From a computational point of view, this means that only C^- needs to be computed. Furthermore, secondary canards can be detected and continued by considering the condition $\mathbf{u}_z(1) = 0$. However, for a system that is not in normal form the symmetry of the normal form is typically lost. Hence, in general the attracting and repelling slow manifolds must be computed separately as the solution families of two different two-point boundary value problems. As a consequence, the primary and the secondary canards must be detected as intersection points of the curves C_0^- and C_0^+.

As an example, we show here what the slow manifolds look like in the perturbation of the normal form (1.2) that is given by (2.9) for small nonzero ρ, that is,

$$\left\{\begin{array}{rcl} \dot{x} &=& \frac{1}{2}\mu y - (\mu+1)z + \rho, \\ \dot{y} &=& 1, \\ \dot{z} &=& x + z^2 + \rho. \end{array}\right. \tag{6.1}$$

Note that the slow manifolds of (6.1) can be computed with the same boundary conditions as those of (1.2). As before, we require that the orbit segments start on the line L_ξ^- defined by (3.3). For the computation of the repelling slow manifold, we reverse time in (6.1) and consider

$$L_\xi^+ := \{(-\xi, s, \sqrt{\xi}) \mid s \in \mathbb{R}\},$$

which lies on the upper sheet of the parabolic cylinder S. Furthermore, we require the orbit segments to end in a transverse section Σ_α given by (3.4) with $\alpha = 0$ or $\alpha = \pm 1$.

We consider here the case $\mu = 8.5$ and $\rho = 0.316$ and fix $\xi = 100$ as before. For this value of ρ, the explicitly known solution γ_s of the unperturbed system (1.2) is not a suitable start solution for Newton's method. Therefore, starting from γ_s, we first compute the ρ-dependent family of orbit segments that solve (6.1) subject to the boundary conditions (3.2). Here we fix the parameter s that defines the position on L_ξ^- to the value (given by γ_s) of

$$s = t_0 = -\sqrt{\frac{2\mu + 4\xi}{\mu^2}}$$

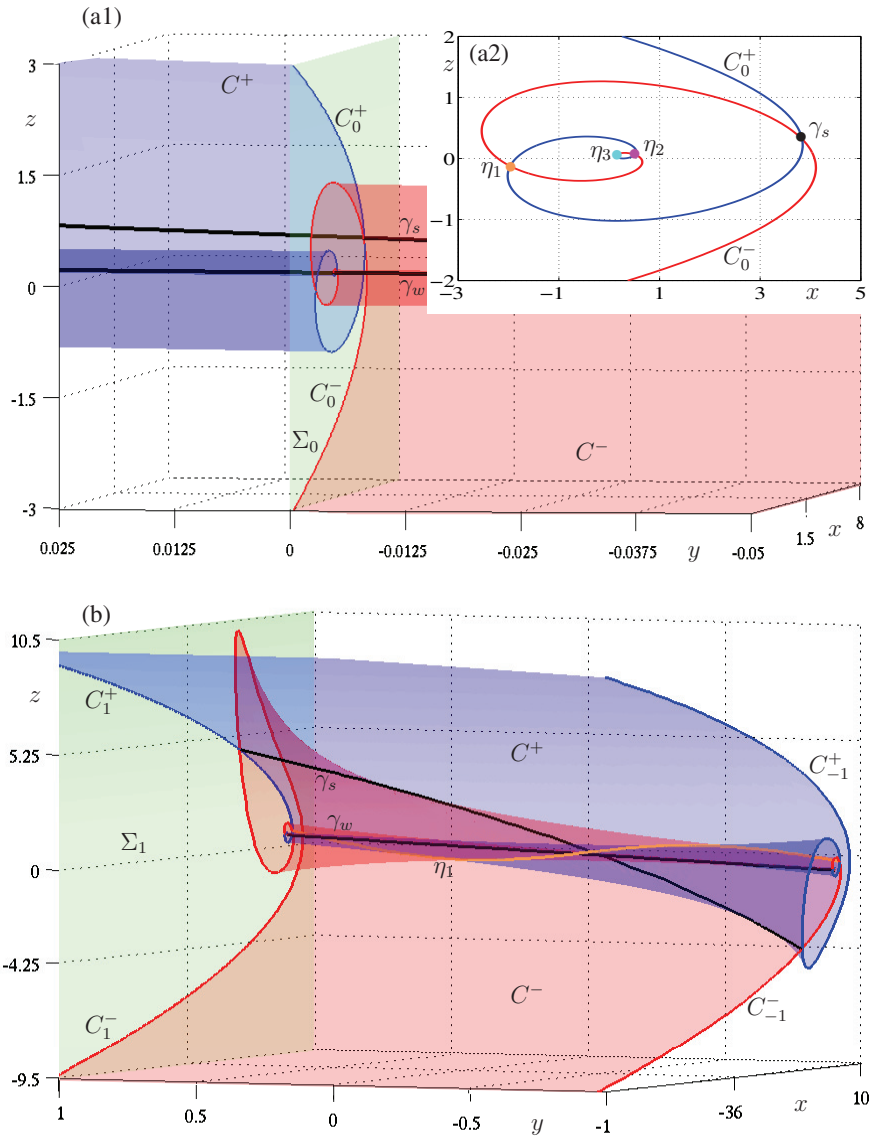


Figure 13. *Slow manifolds and canard solutions of (2.9) with $\mu = 8.5$ and $\rho = 0.316$. The manifolds are smooth deformations of the equivalent manifolds for $\rho = 0$, but the attracting and repelling slow manifolds are no longer related by symmetry; compare with Figure 5.*

as defined in (3.6). When $\rho = 0.316$ is reached, a first orbit segment on C^- has been found; a first orbit segment on C^+ is found similarly by starting a continuation in ρ from the part of γ_s that connects L_ξ^+ to Σ_α. We now fix ρ and continue in s to sweep out C^- and C^+, respectively.

Figure 13 shows the slow manifolds of (6.1) for $\mu = 8.5$ and $\rho = 0.316$. This figure should be compared with Figure 5 for $\rho = 0$; for ease of comparison we use the same viewpoints in both figures. In Figure 13 the slow manifolds C^- and C^+ have deformed and are no longer

each other's images under a symmetry operation. Nevertheless, the situation is topologically the same as that for $\rho = 0$ in Figure 5. Namely, C^- and C^+ intersect in the same way in the primary canards γ_s and γ_w and the secondary canards η_1 to η_3; see Figure 13(a2) and (b). In particular, the rotating behavior of η_1 to η_3 around γ_w is preserved. The canards are found by detecting orbit segments on C^- and C^+ that end at the same point in Σ_0 (within the accuracy of the computation). Concatenation of the two respective orbit segments results in the representation of the secondary canard as a solution that starts on L_ξ^- and ends at L_ξ^+. After applying a Newton step to get an exact solution to this boundary value problem, detected canards can be continued in a system parameter.

7. Conclusion. We performed a study of slow manifolds and associated canard solutions in a three-dimensional normal form of a slow-fast system with a folded node. Specifically, we computed the two-dimensional attracting and repelling slow manifolds as one-parameter families of orbit segments that satisfy appropriate boundary conditions. This approach also allows us to detect and continue the canard solutions themselves. The visualization of these geometric objects for different values of the normal form parameter μ (the ratio of eigenvalues at the folded node) provided unprecedented insight into the geometry of the dynamics near a folded node. We discussed in detail how the secondary canards spiral around the weak primary canard and presented the first computed bifurcation diagram showing branches of the secondary canards as a function of μ.

The numerical continuation of solution families of a well-posed boundary value problem can be performed very accurately. In our computations we use the continuation and boundary value solver routines of AUTO, which uses pseudoarclength continuation and collocation with piecewise-polynomial approximations. Hence, the boundary value problems we define are solved subject to established error bounds. Therefore, the accuracy of our calculations of slow manifolds and canard solutions comes down to determining how the choice of boundary condition influences the distance of the approximation from the real object. In our setup we define approximating orbit segments by requiring that they start on a suitable line (sufficiently far away from the origin) on a parabolic cylinder. Numerical checks ensured that the pointwise distance to the true slow manifolds along selected orbit segments is sufficiently small. The detection and continuation of canard solutions generally works very well with our method, but there remain numerical difficulties near bifurcation points. A more detailed error analysis, which would be very useful in this context, remains a challenging subject for further investigation.

While this paper concentrates on the normal form of a folded node, our boundary value problem approach to computing slow manifolds and canard solutions can be applied more widely. This was demonstrated with the example of a perturbation of the normal form that breaks the underlying symmetry. In [11] we computed slow manifolds and canard solutions in the self-coupled FitzHugh–Nagumo model, which required the implementation of a homotopy approach to generating initial approximate orbits on the attracting and repelling slow manifolds. In this way, we were able to identify sectors between different secondary canards that correspond to mixed-mode oscillations with different numbers of small oscillations.

In the near future we plan to use our computational approach to investigate other slow-fast systems arising in applications, especially those showing mixed-mode oscillations. This

is relatively straightforward for the case of three-dimensional vector field models with a clear splitting of the phase space into slow and fast variables, such as the self-coupled FitzHugh–Nagumo model [11] or the forced Van der Pol system [6, 29]. However, we believe that the computation of invariant manifolds would also be a very helpful tool in situations where there is no obvious split of the system into slow and fast variables. The goal here would be to identify slow and fast components of the dynamics numerically and to use this knowledge to unravel the geometry of slow manifolds and associated canard solutions.

Acknowledgments. The authors thank John Guckenheimer and Martin Wechselberger for helpful discussions.

REFERENCES

[1] V. I. Arnol′d, *Singularity Theory*, London Math. Soc. Lecture Note Ser. 53, Cambridge University Press, Cambridge, UK, 1981.

[2] V. I. Arnol′d, V. S. Afrajmovich, Yu. S. Il′yashenko, and P. L. Shil′nikov, *Dynamical Systems V: Bifurcation Theory and Catastrophe Theory*, Encyclopaedia Math. Sci., Springer-Verlag, Berlin, 1994.

[3] E. Benoît, *Systèmes lents-rapides dans* $\mathbb{R}^3$ *et leurs canards*, in Troisième rencontre du Schnepfenried, Astérisque 109–110, Soc. Math. France, Paris, 1983, pp. 159–191.

[4] E. Benoît, *Canards et enlacements*, Inst. Hautes Études Sci. Publ. Math., 72 (1990), pp. 63–91.

[5] E. Benoît, J.-L. Callot, F. Diener, and M. Diener, *Chasse au canard*, Collect. Math., 31–32 (1981), pp. 37–119.

[6] K. Bold, C. Edwards, J. Guckenheimer, S. Guharay, K. Hoffman, J. Hubbard, R. Oliva, and W. Weckesser, *The forced Van der Pol equation* II: *Canards in the reduced system*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 570–608.

[7] M. Brøns and K. Bar-Eli, *Canard explosion and excitation in a model of the Belousov-Zhabotinskii reaction*, J. Phys. Chem., 95 (1991), pp. 8706–8713.

[8] M. Brøns, T. J. Kaper, and H. G. Rotstein, *Special issue on mixed-mode oscillations*, Chaos, 18 (1) (2008).

[9] M. Brøns, M. Krupa, and M. Wechselberger, *Mixed mode oscillations due to the generalized canard phenomenon*, in Bifurcation Theory and Spatio-Temporal Pattern Formation, Fields Inst. Commun. 49, AMS, Providence, RI, 2006, pp. 39–63.

[10] F. Buchholtz, M. Dolnik, and I. R. Epstein, *Diffusion-induced instabilities near a canard*, J. Phys. Chem., 99 (1995), pp. 15093–15101.

[11] M. Desroches, B. Krauskopf, and H. M. Osinga, *Mixed-mode oscillations and slow manifolds in the self-coupled FitzHugh–Nagumo system*, Chaos, 18 (2008), 015107.

[12] E. J. Doedel, *Lecture notes on numerical analysis of nonlinear equations*, in Numerical Continuation Methods for Dynamical Systems: Path Following and Boundary Value Problems, B. Krauskopf, H. M. Osinga, and J. Galán-Vioque, eds., Springer-Verlag, New York, 2007, pp. 1–50.

[13] E. J. Doedel, E. Freire, E. Gamerio, and A. J. Rodríguez-Luis, *An analytical and numerical study of a modified Van der Pol oscillator*, J. Sound Vibration, 256 (2002), pp. 755–771.

[14] E. J. Doedel, R. C. Paffenroth, A. R. Champneys, T. F. Fairgrieve, Yu. A. Kuznetsov, B. E. Oldeman, B. Sandstede, and X. J. Wang, *AUTO2000: Continuation and Bifurcation Software for Ordinary Differential Equations*, available via http://cmvl.cs.concordia.ca/.

[15] M. Domijan, R. Murray, and J. Sneyd, *Dynamical probing of the mechanisms underlying calcium oscillations*, J. Nonlinear Sci., 16 (2006), pp. 483–506.

[16] J. L. A. Dubbeldam and B. Krauskopf, *Self-pulsations in lasers with saturable absorber: Dynamics and bifurcations*, Opt. Commun., 159 (1999), pp. 325–338.

[17] F. Dumortier, *Techniques in the theory of local bifurcations: Blow-up, normal forms, nilpotent bifurcations, singular perturbations*, in Bifurcations and Periodic Orbits of Vector Fields, D. Szlomiuk, ed., Kluwer Academic, Dordrecht, The Netherlands, 1993, pp. 19–73.

[18] F. Dumortier and R. Roussarie, *Canard cycles and center manifolds*, Mem. Amer. Math. Soc., 121 (1996).
[19] W. Eckhaus, *Relaxation oscillations including a standard chase on French ducks*, in Asymptotic Analysis II, Lecture Notes in Math. 958, Springer-Verlag, New York, 1983, pp. 449–494.
[20] J. P. England, B. Krauskopf, and H. M. Osinga, *Computing one-dimensional global manifolds of Poincaré maps by continuation*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 1008–1041.
[21] T. Erneux, *Q-switching bifurcation in a laser with saturable absorber*, J. Opt. Soc. Amer. B Opt. Phys., 5 (1988), pp. 1063–1069.
[22] N. Fenichel, *Persistence and smoothness of invariant manifolds*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.
[23] N. Fenichel, *Geometric singular perturbation theory*, J. Differential Equations, 31 (1979), pp. 53–98.
[24] R. FitzHugh, *Thresholds and plateaus in the Hodgkin–Huxley nerve equations*, J. Gen. Physiol., 43 (1960), pp. 867–896.
[25] J.-M. Ginoux and B. Rossetto, *Differential geometry and mechanics: Applications to chaotic dynamical systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 16 (2006), pp. 887–910.
[26] J.-M. Ginoux and B. Rossetto, *Slow manifold of a neuronal bursting model*, in Emergent Properties in Natural and Artificial Dynamical Systems, M. A. Aziz-Alaoui and C. Bertelle, eds., Springer-Verlag, New York, 2006, pp. 119–128.
[27] J. Guckenheimer, *Return maps of folded nodes and folded saddle-nodes*, Chaos, 18 (2008), 015108.
[28] J. Guckenheimer and R. Haiduc, *Canards at folded node*, Mosc. Math. J., 5 (2005), pp. 91–103.
[29] J. Guckenheimer, K. Hoffman, and W. Weckesser, *The forced Van der Pol equation* I: *The slow flow and its bifurcations*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 1–35.
[30] M. W. Hirsch, C. C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.
[31] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Physiology, 117 (1952), pp. 500–544.
[32] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme, 1994), Lecture Notes in Math. 1609, Springer-Verlag, New York, 1995, pp. 44–118.
[33] M. T. M. Koper, *Bifurcations of mixed-mode oscillations in a three-variable autonomous Van der Pol–Duffing model with a cross-shaped phase diagram*, Phys. D, 80 (1995), pp. 72–94.
[34] B. Krauskopf and H. M. Osinga, *Computing invariant manifolds via the continuation of orbit segments*, in Numerical Continuation Methods for Dynamical Systems: Path Following and Boundary Value Problems, B. Krauskopf, H. M. Osinga, and J. Galán-Vioque, eds., Springer-Verlag, New York, 2007, pp. 117–154.
[35] A. Milik, P. Szmolyan, H. Löffelmann, and E. Gröller, *Geometry of mixed-mode oscillations in the* 3*-D autocatalator*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 8 (1998), pp. 505–519.
[36] E. F. Mishchenko, Yu. S. Kolesov, A. Yu. Kolesov, and N. Kh. Rhozov, *Asymptotic Methods in Singularly Perturbed Systems*, Monogr. Contemp. Math., Consultants Bureau, New York, 1994.
[37] J. Moehlis, *Canards in a surface oxidation reaction*, J. Nonlinear Sci., 12 (2002), pp. 319–345.
[38] J. S. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, Proc. IRE, 50 (1962), pp. 2061–2070.
[39] V. Petrov, S. K. Scott, and K. Showalter, *Mixed-mode oscillations in chemical systems*, J. Chem. Phys., 97 (1992), pp. 6191–6198.
[40] M. Phillips, S. Levy, and T. Munzner, *Geomview: An interactive geometry viewer*, Notices Amer. Math. Soc., 40 (1993), pp. 985–988; also available online from http://www.geomview.org/.
[41] H. G. Rotstein and R. Kuske, *Localized and asynchronous patterns via canards in coupled calcium oscillators*, Phys. D, 215 (2006), pp. 46–61.
[42] R. Roussarie, *Techniques in the theory of local bifurcations: Cyclicity and desingularisation*, in Bifurcations and Periodic Orbits of Vector Fields, D. Szlomiuk, ed., Kluwer Academic, Dordrecht, The Netherlands, 1993, pp. 347–382.
[43] J. Rubin and D. Terman, *Geometric singular perturbation analysis for neuronal dynamics*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 93–146.
[44] J. Rubin and M. Wechselberger, *Giant squid—hidden canard: The* 3*D geometry of the Hodgkin–Huxley model*, Biol. Cybernet., 97 (2007), pp. 5–32.

[45] P. Szmolyan and M. Wechselberger, *Canards in* $\mathbb{R}^3$, J. Differential Equations, 177 (2001), pp. 419–453.
[46] P. Szmolyan and M. Wechselberger, *Relaxation oscillations in* $\mathbb{R}^3$, J. Differential Equations, 200 (2004), pp. 69–104.
[47] F. Takens, *Constrained equations: A study of implicit differential equations and their discontinuous solutions*, in Structural Stability: The Theory of Catastrophes and Applications in the Sciences, Lecture Notes in Math. 525, Springer-Verlag, New York, 1976, pp. 143–234.
[48] B. van der Pol, *A theory of the amplitude of free and forced triode vibrations*, Radio Review, 1 (1920), pp. 701–710, 754–762.
[49] B. van der Pol, *On "relaxation oscillations"* I, Phil. Mag., 2 (1926), pp. 978–992.
[50] M. Wechselberger, *Existence and bifurcation of canards in* $\mathbb{R}^3$ *in the case of a folded node*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 101–139.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1163–1190
© 2008 Society for Industrial and Applied Mathematics

Asymptotics of a Slow Manifold*

J. Vanneste†

Abstract. Approximately invariant elliptic slow manifolds are constructed for the Lorenz–Krishnamurthy model of fast-slow interactions in the atmosphere. As is the case for many other two-time-scale systems, the various asymptotic procedures that may be used for this construction diverge, and there are no exactly invariant slow manifolds. Valuable information can however be gained by capturing the details of the divergence: this makes it possible to define exponentially accurate slow manifolds, identify one of these as optimal, and predict the amplitude and phase of the fast oscillations that appear for trajectories started on it. We demonstrate this for the Lorenz–Krishnamurthy model by studying the slow manifolds obtained using a power-series expansion procedure. We develop two distinct methods to derive the leading-order asymptotics of the late coefficients in this expansion. Borel summation is then used to define a unique slow manifold, regarded as optimal, which is piecewise analytic in the slow variables. This slow manifold is not analytic on a Stokes surface: when slow solutions cross this surface, they switch on exponentially small fast oscillations through a Stokes phenomenon. We show that the form of these oscillations can be recovered from the Borel summation. The approach that we develop for the Lorenz–Krishnamurthy model has a general applicability; we sketch how it generalizes to a broad class of two-time-scale systems.

Key words. slow manifold, exponential asymptotics, atmospheric waves

AMS subject classifications. 37N10, 76B15, 76U05, 37K05

DOI. 10.1137/070710081

1. Introduction.

Dynamical systems with two time scales appear in a wide variety of applications, particularly in physics and chemistry. A central concept in their analysis is that of a slow manifold [28, 18, 19]. Slow manifolds are nearly invariant submanifolds of the state space of these systems near which the dynamics is slow; their dimension is the number of slow variables, and they are defined by constraints slaving fast variables to slow ones (see, e.g., [25]). The advantages of identifying slow manifolds in two-time-scale systems are obvious: projecting the dynamics onto a slow manifold leads to a dynamical system of reduced dimensionality; this system approximates the full dynamics while filtering out the fast behavior, and it can therefore be integrated efficiently.

The fast behavior we are concerned with in this paper consists of rapid undamped oscillations. In this case, the slow manifolds are elliptic and hence fragile. Specifically, if $\epsilon \ll 1$ is the small parameter characterizing the separation between fast and slow time scales, the slow manifold that exists for $\epsilon = 0$ cannot be expected to persist as an invariant object when $\epsilon \neq 0$. This is of course in contrast to the normally hyperbolic case, for which persistence can be established [13]. Even though elliptic slow manifolds are generally not exactly invariant, they

*Received by the editors December 4, 2007; accepted for publication (in revised form) by T. Kaper June 3, 2008; published electronically October 13, 2008.
http://www.siam.org/journals/siads/7-4/71008.html
†School of Mathematics and Maxwell Institute, University of Edinburgh, King's Buildings, Edinburgh EH9 3JZ, UK (J.Vanneste@ed.ac.uk).

can be approximately so to a very high degree of accuracy. Indeed, systematic asymptotic procedures make it possible to improve this accuracy, estimated by the angle between the vector field and the manifold, systematically order-by-order in ϵ. For analytic vector fields, it is possible to construct slow manifolds with $O(\epsilon^n)$ accuracy for arbitrary $n \in \mathbb{N}$ and even, by optimal truncation, to achieve exponential accuracy [15, 11, 35, 25, 12, 27].

The nonexistence of exactly invariant slow manifolds reflects the fact that fast activity cannot be completely filtered out by a suitable projection of the initial conditions. In other words, in the elliptic situation, fast oscillations are typically generated by the slow dynamics, however well the initial data are prepared. The noninvariance is also manifested by the divergence of the asymptotic procedures used in the construction of slow manifolds. The two aspects are related: the nature of the divergence—the manner in which the coefficients of ϵ^n in the power-series expansions defining slow manifolds grow with n—encodes the generation of (exponentially small) oscillations. Thus, capturing the details of the divergence provides a means of describing these oscillations. It also gives a way of analyzing the differences between the various slow manifolds that are obtained near optimal truncation. One of the motivations here is to distinguish, among these slow manifolds differing by exponentially small terms, a unique one, enjoying special properties.

The exponential accuracy of elliptic slow manifolds, the divergence of the asymptotic procedures used in their construction, and the connection between this divergence and the generation of fast oscillations are the themes of this paper. Although these have a general appeal for a broad class of two-time-scale systems, we mainly explore them in a specific context, and for a specific model. The context is geophysical fluid dynamics. Because of the fast rotation of the earth, the midlatitude atmosphere and oceans are typical two-time-scale systems, with the corresponding small parameter—the Rossby number—taking values of the order of 0.1 and 0.01 in the atmosphere and the oceans, respectively. Furthermore, the nature of the forcing is such that the fast degrees of freedom, consisting of inertia-gravity waves, are often only weakly excited. As a result, the notion of a slow manifold is eminently relevant. (See, e.g., [34, 31] and references therein for more background.)

The specific model that we analyze is the Lorenz five-component model [21], often referred to as the Lorenz–Krishnamurthy (LK) model [23]. This model, governed by five ordinary differential equations, was devised by Lorenz in order to explore the concept of slow manifolds and study their invariance. Since it was proposed, it has become one of the main testbeds for the study of slow manifolds, reduced models (termed "balanced models" in this context), spontaneous wave generation, etc., in geophysical fluid dynamics [21, 22, 23, 9, 10, 14, 32, 6, 7, 8, 17, 29, 30].

Several asymptotic procedures have been proposed for the derivation of slow manifolds in the LK model (e.g., [20, 32]). Since their divergence properties are identical, we concentrate here on a particularly simple one (see [34] for a detailed discussion). Specifically, a slaving relation, which defines the slow manifold by relating the fast variables to the slow ones, is postulated and introduced into the dynamical equations. An approximate solution of the resulting partial differential equation is then sought as a series expansion in powers of the small parameter ϵ. The coefficients in this series, which we term "slaving coefficients," are functions of the slow variables. Our main aim is to capture their late form, that is, to obtain the asymptotics of the coefficient of ϵ^n as $n \to \infty$. Two alternative approaches are

discussed. These are interchangeable in the case of the LK model, but one or the other may be preferable for more complicated models. Remarkably, the leading-order asymptotics of the slaving coefficients can be determined in closed form, up to a single constant which is readily estimated by solving a recurrence relation numerically. The accuracy of the asymptotic result is established by a comparison with the slaving coefficients computed numerically for a range of values of the slow variables.

We emphasize that our asymptotic results give a precise description of the manner in which the power-series expansion defining the slow manifolds diverges as $n \to \infty$. This makes it possible to go beyond the standard optimal truncation arguments (e.g., [11, 35]), which only provide bounds on the accuracy of the slow manifold, and delve into the dynamics of the exponentially small terms. Specifically, we use the Borel summation of the divergent power series [3, 2] to define a unique manifold, which we term the "optimal slow manifold." This is defined in a piecewise manner, with discontinuities across codimension-one surfaces. Trajectories started on the optimal slow manifold move away from it by an exponentially small distance when they cross these surfaces, and fast oscillations develop. The amplitude and phase of these oscillations can be determined from the late behavior of the slaving coefficients. In previous work [29], we derived this amplitude and phase by considering the dynamics along specific trajectories. The present approach recovers these results by taking a more geometric perspective, which views the slow manifold as a single object rather that a collection of slow trajectories.

The analysis we carry out for the LK model is representative of a more general treatment applicable to more complicated two-time-scale systems. We make this plain by also considering a broad class of such systems and sketching how the theory developed for the LK model generalizes to this class. The results presented are largely formal, and they make a number of simplifying assumptions, in particular about the nature of the singularities of slow trajectories in the complex time plane. Nevertheless, they provide a first glimpse into the relationship between these singularities, the divergence of the asymptotic procedures used for constructing slow manifolds, and the generation of fast oscillations.

This paper is organized as follows. The LK model is introduced in section 2. There we discuss a systematic approach for the construction of slow manifolds of increasing accuracy. As mentioned, this approach relies on expanding in power-series of ϵ the relations which define the slow manifolds by slaving fast variables to slow variables. The coefficients of ϵ^n in this expansion—the slaving coefficients—satisfy recurrence relations involving partial derivatives with respect to the slow variables. The nonlinearity of the LK model, involving only quadratic terms, is simple enough that the slaving coefficients are homogeneous polynomials in the slow variables. It is therefore easy to derive them by solving simple algebraic recurrences for the coefficients of these polynomials. Section 3 focuses on these coefficients, which we refer to as polynomial coefficients to distinguish them from the slaving coefficients. Specifically, we examine the form of the polynomial coefficients for large n. Two asymptotic results are obtained. The first result gives a Gaussian approximation to the slaving coefficients. The second, more general, result improves on this approximation in the same manner as the large-deviation theory, for the probability density of sums of random numbers improves on the central-limit theorem. Section 4 then considers the large-n asymptotics of the slaving coefficients themselves. The asymptotic behavior is obtained using two different approaches, one based on

the polynomial coefficients and the other directly considering the partial differential equation satisfied by the slaving coefficients. The results are exploited in section 5, where we discuss the resummation of the divergent series defining the slow manifolds. There, we use Borel summation (e.g., [3, 2]) to define a unique slow manifold which we regard as optimal. The slaving relation for this slow manifold is given as an integral which is clearly discontinuous across certain surfaces in the slow space. Examining the dynamics across these surfaces, we demonstrate that it is characterized by the generation of exponentially small fast oscillations whose form is encoded in the Borel sum. General two-time-scale systems with elliptic slow manifolds are considered in section 6. The paper concludes with a discussion in section 7.

2. Formulation.

2.1. Model. We consider the model devised by Lorenz [21] and variously referred to as the Lorenz five-component model or as the LK model [23]. In its conservative form, on which we will focus, it can be written as the set of five ordinary differential equations

$$\dot{u} = -vw + \epsilon bvy, \tag{2.1}$$

$$\dot{v} = uw - \epsilon buy, \tag{2.2}$$

$$\dot{w} = -uv, \tag{2.3}$$

$$\epsilon\dot{x} = -y, \tag{2.4}$$

$$\epsilon\dot{y} = x + buv \tag{2.5}$$

for the five dependent variables (u, v, w, x, y). This model, obtained by truncation of the rotating shallow-water equations, governs the dynamics of a triad of vortical modes, with amplitudes (u, v, w), coupled to a gravity mode described by (x, y). The two parameters b and ϵ of the model control the strength of the coupling and the gravity-wave frequency, respectively.

Following Camassa [9] and Bokhove and Shepherd [6], we note that the constancy of the $u^2 + v^2$, obvious from (2.1)–(2.2), can be used to reduce the dimension of the LK model from 5 to 4. Specifically, letting

$$u = u_0 \cos\phi \quad \text{and} \quad v = u_0 \sin\phi \tag{2.6}$$

reduces (2.1)–(2.5) to the two degree-of-freedom Hamiltonian system

$$\dot{\phi} = w - \epsilon by, \tag{2.7}$$

$$\dot{w} = -u_0^2 \sin(2\phi)/2, \tag{2.8}$$

$$\epsilon\dot{x} = -y, \tag{2.9}$$

$$\epsilon\dot{y} = x + bu_0^2 \sin(2\phi)/2. \tag{2.10}$$

Here, $u_0^2 = u^2 + v^2$ is a constant which could be set to 1 by scaling.

In the form (2.7)–(2.10), the LK model can be recognized as describing the dynamics of a pendulum (making an angle 2ϕ with the vertical), coupled in some way to a spring of extension x. This interpretation is useful in developing some intuition about the dynamics of the model; it also makes transparent the relationship between the LK model and mechanical

models such as the swinging spring (or elastic pendulum; see, e.g., [24]). In what follows, we mostly use the original formulation (2.1)–(2.5), which gives a more compact form to various mathematical expressions; however, we often use the variable ϕ in place of (u, v) to display functions of (u, v, w) in the reduced, two-dimensional space (ϕ, w).

We are interested in the dynamics of the LK model when $\epsilon \ll 1$. In this regime, there is a large separation between the $O(1)$ time scale of evolution of the slow variables (u, v, w) and the $O(\epsilon)$ time scale of the fast variables (x, y). We also assume that $b = O(1)$. In the geophysical context, these assumptions correspond to the quasi-geostrophic regime, in which fast gravity waves interact only weakly with the much slower vortical motion, termed "balanced motion." In the mechanical interpretation of the LK model, $\epsilon \ll 1$ indicates that the spring is stiff, so that its frequency ϵ^{-1} far exceeds that of the pendulum.

The large time-scale separation implies the existence of slow manifolds. For the LK model, these are three-dimensional submanifolds of the state space, parameterized by (u, v, w), which are nearly invariant and near which the motion is slow. The dynamics in the neighborhood of such slow manifolds is approximately devoid of the fast oscillations which characterize the dynamics elsewhere in the state space. The slow manifolds are elliptic, since linearizing the fast dynamics gives the purely imaginary eigenvalues $\pm i\epsilon^{-1}$. Therefore, they cannot be expected to be invariant when $\epsilon \neq 0$. Nevertheless, their accuracy, measured by the difference between the angle made by the vector field $(\dot{u}, \dot{v}, \dot{w}, \dot{x}, \dot{y})$ and the slow manifold, can be very high indeed: systematic improvement procedures make it possible to define slow manifolds with exponentially small errors.

The main interest of slow manifolds is that they allow a simplified description of the dynamics. Projecting the vector field onto a slow manifold leads to a reduced system of slow equations for (u, v, w) which approximates well the full dynamics for initial conditions near the slow manifold. Reduced models obtained in this manner are termed "balanced models" in the geophysical context, where they have proved highly successful.

It is clear from (2.4)–(2.5) that a slow manifold for the LK model can be defined as the graph

$$x = -buv \quad \text{and} \quad y = 0. \tag{2.11}$$

The corresponding balanced model is then given by (2.1)–(2.3) with $y = 0$. The slow manifold (2.11) is only a leading-order approximation; starting with Lorenz [21], many authors have considered how this can be improved. In the next sections, we examine in detail a simple asymptotic procedure of the type described by Warn et al. [34] which leads to an arbitrary $O(\epsilon^n)$ accuracy. Our aim is to capture the manner in which this procedure diverges so as to define a slow manifold with a better-than-exponential accuracy.

2.2. Slow manifolds. Slow manifolds can be sought by introducing the so-called slaving relations

$$x = X(u, v, w; \epsilon) \quad \text{and} \quad y = Y(u, v, w; \epsilon) \tag{2.12}$$

for unknown functions X and Y into (2.4)–(2.5). Eliminating the time derivatives by means of (2.1)–(2.3) gives what Lorenz [20] termed the "superbalance equation," namely,

$$\epsilon\left[\frac{\partial X}{\partial u}(-vw+\epsilon bvy)+\frac{\partial X}{\partial v}(uw-\epsilon buy)-\frac{\partial X}{\partial w}uv\right]=-Y, \tag{2.13}$$

$$\epsilon\left[\frac{\partial Y}{\partial u}(-vw+\epsilon bvy)+\frac{\partial Y}{\partial v}(uw-\epsilon buy)-\frac{\partial Y}{\partial w}uv\right]=X+buv. \tag{2.14}$$

These are two coupled partial differential equations for X and Y for which approximate solutions can be found using iteration or expansion in powers of ϵ. Here we employ the latter method which is more suited to deriving explicit results. To some extent the method used is irrelevant, since the slow manifolds obtained by different means coincide up to terms smaller than the accuracy of the methods. Nevertheless, specific methods may have some advantage: for instance, the iterative procedure proposed in [25] guarantees that all equilibria of the system near the slow manifold lie exactly on it. The expansion used here does not have this property.

Inspection of (2.13)–(2.14) indicates that power-series expansions of the slaving relations (2.12) take the form

$$x=\sum_{n=0}^{N}\epsilon^{2n}X_n(u,v,w)\quad\text{and}\quad y=\sum_{n=0}^{N}\epsilon^{2n+1}Y_n(u,v,w), \tag{2.15}$$

where the functions of the slow variables X_n and Y_n are termed slaving coefficients. These are homogenous polynomials in u, v, and w of degree $2n+2$ and $2n+3$, respectively. We make their specific form explicit by writing

$$X_n(u,v,w)=(2n)!\sum_{i,j=0}C_{ij}^n\,u^{2i+1}v^{2j+1}w^{2k}, \tag{2.16}$$

with $k=n-i-j\geq 0$, and

$$Y_n(u,v,w)=(2n+1)!\sum_{i,j=0}D_{ij}^n\,u^{2i}v^{2j}w^{2k+1}, \tag{2.17}$$

with $k=n+1-i-j\geq 0$. In defining the coefficients C_{ij}^n and D_{ij}^n, we have introduced the normalization factors $(2n)!$ and $(2n+1)!$ which roughly capture the dominant growth of X_n and Y_n with n. We refer to C_{ij}^n and D_{ij}^n as polynomial coefficients and emphasize that, unlike the slaving coefficients X_n and Y_n which they generate, they are simply numbers (for fixed b).

Substituting (2.15)–(2.17) into (2.13)–(2.14) leads to the following recurrence relations for C_{ij}^n and D_{ij}^n:

$$\begin{aligned}(2n+1)D_{ij}^n&=(2i+1)C_{i(j-1)}^n-(2j+1)C_{(i-1)j}^n+(2k+2)C_{(i-1)(j-1)}^n\\&\quad-b\sum_{m=0}^{n-1}\sum_{p,q=0}^{m}\frac{(2m)!(2n-2m-1)!}{(2n)!}C_{pq}^m\\&\qquad\times\left[(2p+1)D_{(i-p)(j-q-1)}^{n-m-1}-(2q+1)D_{(i-p-1)(j-q)}^{n-m-1}\right],\end{aligned} \tag{2.18}$$

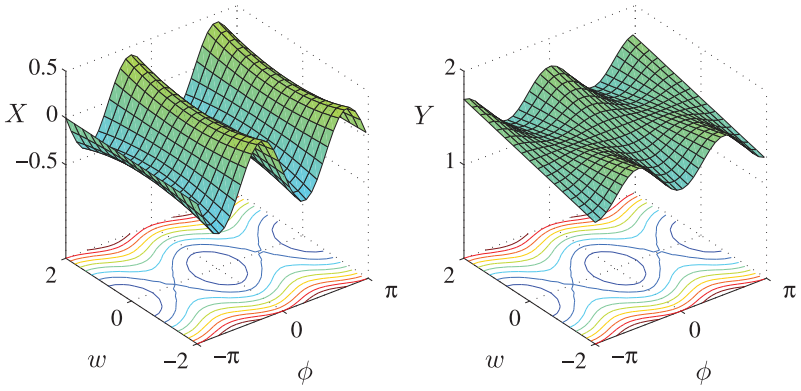


Figure 1. *Approximate slow manifold with $O(\epsilon^3)$ accuracy: The slaving functions $X(u,v,w)$ and $Y(u,v,w)$ are plotted as functions of ϕ (with $u=\cos\phi$ and $v=\sin\phi$) and w for $\epsilon=0.2$ and $b=0.5$. Approximate slow trajectories are plotted in the (ϕ,w)-plane.*

where $k=n+1-i-j$, and

$$
\begin{aligned}
2nC_{ij}^n &= -2(i+1)D_{(i+1)j}^{n-1} + 2(j+1)D_{i(j+1)}^{n-1} - (2k+1)D_{ij}^{n-1} \\
&\quad + b\sum_{m=0}^{n-2}\sum_{p,q=0}^{m+1} \frac{(2m+1)!(2n-2m-3)!}{(2n-1)!} D_{pq}^m \\
&\qquad \times \left[2pD_{(i-p+1)(j-q)}^{n-m-2} - 2qD_{(i-p)(j-q+1)}^{n-m-2}\right],
\end{aligned}
\tag{2.19}
$$

where $k=n-i-j$. The initial condition for this iteration is provided by the leading-order slow manifold (2.11) which gives

$$
C_{00}^0 = -b. \tag{2.20}
$$

The successive D_{ij}^n and C_{ij}^n are then calculated from (2.18)–(2.19), with the convention that $D_{ij}^n=0$ for $i<0$, $j<0$, or $i+j>n+1$, and $C_{ij}^n=0$ for $i<0$, $j<0$, or $i+j>n$. The first few coefficients are

$$
D_{00}^0 = 0, \quad D_{10}^0 = b, \quad D_{01}^0 = -b \tag{2.21}
$$

and

$$
C_{00}^1 = -2b, \quad C_{10}^1 = -b/2, \quad C_{01}^1 = b/2. \tag{2.22}
$$

For larger n, the coefficients are easily computed numerically for fixed b. The numerical results presented in this paper rely on such computations carried out for n up to 100.

The slow manifold corresponding to (2.20)–(2.22), for which the superbalance equation is approximated within an $O(\epsilon^3)$ error, is shown in Figure 1. The approximations to X and Y are shown as a function of ϕ and w, with $u_0=1$ in (2.6), $\epsilon=0.2$, and $b=0.5$. The figure also shows approximate trajectories in the plane of the slow variables (ϕ,w); lifting them to the slow manifold gives an approximation to full trajectories.

The series (2.15) diverge as $N \to \infty$. In this paper we examine more precisely the nature of this divergence by considering the late behavior of the slaving coefficients X_n and Y_n as $n \to \infty$. A possible approach, attempted by Warn [33] for a simplifed model, consists in deriving approximations for the polynomial coefficients C_{ij}^n and D_{ij}^n as $n \to \infty$ from the recurrence relations (2.18)–(2.19). This is carried out in the next section.

3. Late behavior of C_{ij}^n and D_{ij}^n. We consider the behavior of C_{ij}^n and D_{ij}^n for large n. Numerical computations of these coefficients suggest the asymptotic forms

$$C_{ij}^n \sim (-1)^{j+1} f(\xi,\eta) \quad \text{and} \quad D_{ij}^n \sim (-1)^j g(\xi,\eta), \tag{3.1}$$

where

$$\xi = n^{-1/2}(i - n/3) \quad \text{and} \quad \eta = n^{-1/2}(j - n/3). \tag{3.2}$$

The two functions $f(\xi,\eta)$ and $g(\xi,\eta)$ introduced in (3.1) are smooth and localized, peaking at $(\xi,\eta) = (0,0)$ and decreasing rapidly for $|\xi| \to \infty$ and $|\eta| \to \infty$. Thus, the coefficients C_{ij}^n and D_{ij}^n are maximum for $i \approx n/3$ and $j \approx n/3$, and $O(1)$ only in a "core" region where $\xi, \eta = O(1)$. As we now show, it is not difficult to derive explicit expressions for $f(\xi,\eta)$ and $g(\xi,\eta)$ in this core region.

3.1. Core: $\xi, \eta = O(1)$. We first note that the nonlinear terms in the recurrence relations (2.18)–(2.19) (the last two lines in each of these equations) can be neglected in the limit $n \to \infty$; provided that C_{ij}^n and D_{ij}^n remain $O(1)$ as $n \to \infty$, this is a valid approximation because of the rapid decrease of the ratios of factorials. Neglecting the nonlinear terms, we obtain two sets of first-order linear recurrence relations and, by elimination of D_{ij}^n, a single set of second-order recurrence relations for C_{ij}^n. Substituting the form (3.1) and using Taylor expansions to write, for instance,

$$\begin{aligned} C_{(i+1)j}^n &\sim (-1)^{j+1} f(\xi + n^{-1/2}, \eta) \\ &\sim (-1)^{j+1} \left[f(\xi,\eta) + n^{-1/2} \frac{\partial f}{\partial \xi}(\xi,\eta) + \frac{1}{2n} \frac{\partial^2 f}{\partial \xi^2}(\xi,\eta) + \cdots \right] \end{aligned}$$

leads to a partial differential equation for $f(\xi,\eta)$. The first nontrivial term appears at order $O(n^{-1})$ and is given by

$$4\left(\frac{\partial^2 f}{\partial \xi^2} + \frac{\partial^2 f}{\partial \eta^2} - \frac{\partial^2 f}{\partial \xi \partial \eta} \right) + 45\left(\xi \frac{\partial f}{\partial \xi} + \eta \frac{\partial f}{\partial \eta} \right) + 90 f = 0.$$

Separating variables, it is easily verified that the only solution decreasing to 0 for large $|\xi|$ and $|\eta|$ is the Gaussian

$$f(\xi,\eta) = \Lambda \mathrm{e}^{-15(\xi^2 + \xi\eta + \eta^2)/2}, \tag{3.3}$$

where the constant Λ remains to be determined. From the linearization of (2.18) and from (3.1), we also deduce that

$$g(\xi,\eta) = f(\xi,\eta). \tag{3.4}$$

With these results, the form of C_{ij}^n and D_{ij}^n for large n is known up to the single number Λ which depends solely on b and needs to be determined numerically. This is conveniently done by considering the behavior of the solutions $(u(t), v(t), w(t), x(t), y(t))$ of the LK model near their poles in the complex t-plane. This approach makes contact with the exponential-asymptotics treatment of solutions of the LK models in [29].

Let $t_* \in \mathbb{C}$ be one of the poles of the solutions (as discussed below, these are poles of Jacobi elliptic functions, but their location is unimportant at this point). At a distance from such a pole, the dependent variables can be expanded in inverse powers of $t - t_*$ as

$$u = \sum_{n=0}^{\infty} \frac{\epsilon^{2n} \hat{U}_n}{(t-t_*)^{2n+1}}, \quad v = \sum_{n=0}^{\infty} \frac{\epsilon^{2n} \hat{V}_n}{(t-t_*)^{2n+1}}, \quad w = \sum_{n=0}^{\infty} \frac{\epsilon^{2n} \hat{W}_n}{(t-t_*)^{2n+1}}, \tag{3.5}$$

$$x = \sum_{n=0}^{\infty} \frac{\epsilon^{2n} \hat{X}_n}{(t-t_*)^{2n+2}}, \quad \text{and} \quad y = \sum_{n=0}^{\infty} \frac{\epsilon^{2n+1} \hat{Y}_n}{(t-t_*)^{2n+3}}. \tag{3.6}$$

Note that the coefficients $\hat{X}_n$ and $\hat{Y}_n$ are just (complex) numbers, unlike X_n and Y_n, which are functions of (u, v, w). Substituting (3.5)–(3.6) into (2.1)–(2.5) gives a set of five first-order recurrence relations for the coefficients $(\hat{U}_n, \hat{V}_n, \hat{W}_n, \hat{X}_n, \hat{Y}_n)$. Observing that

$$u \sim -\mathrm{i}/(t-t_*), \quad v \sim 1/(t-t_*), \quad w \sim -\mathrm{i}/(t-t_*) \tag{3.7}$$

is a possible leading-order behavior near t_*, we find the initial conditions

$$\hat{U}_0 = -\mathrm{i}, \quad \hat{V}_0 = 1, \quad \hat{W}_0 = -\mathrm{i}, \quad \hat{X}_0 = \mathrm{i}b, \quad \text{and} \quad \hat{Y}_0 = 2\mathrm{i}b \tag{3.8}$$

for these recurrence relations. There are other possible behaviors near the poles that are alternatives to (3.7). These are obtained by changing the signs of a pair of (u, v, w) and hence of $(\hat{U}_0, \hat{V}_0, \hat{W}_0)$ and correcting the signs of $(\hat{X}_0, \hat{Y}_0)$ accordingly. (Such an alternative choice is made in [29].)

With the initial conditions (3.8) and for fixed b, it is straightforward to compute $(\hat{U}_n, \hat{V}_n, \hat{W}_n, \hat{X}_n, \hat{Y}_n)$ numerically. The value of Λ can then be inferred from their behavior for $n \gg 1$. Specifically, the late form of $\hat{X}_n$ can be verified to be

$$\hat{X}_n \sim \mathrm{i}(-1)^n (2n+1)! \kappa \tag{3.9}$$

for some constant κ. This constant is easily estimated from the $\hat{X}_n$ obtained numerically by approximating the relation

$$\kappa = \lim_{n\to\infty} \frac{\mathrm{i}(-1)^{n+1} \hat{X}_n}{(2n+1)!}$$

for a large but finite n (cf. [29]).

Now, the asymptotic form (3.1)–(3.3) of C_{ij}^n provides an alternative expression for the right-hand side of (3.9). To obtain it, we substitute the leading-order behavior of u, v, and w near t_* given in (3.7) into the expansion (2.15)–(2.16) of the slaving relation $x = X(u, v, w; \epsilon)$. This reduces to

$$x \sim \mathrm{i} \sum_{n=0}^{\infty} \frac{(-1)^{n+1} (2n)! \epsilon^{2n}}{(t-t_*)^{2n+2}} \sum_{i,j=0}^{n} (-1)^j C_{ij}^n.$$

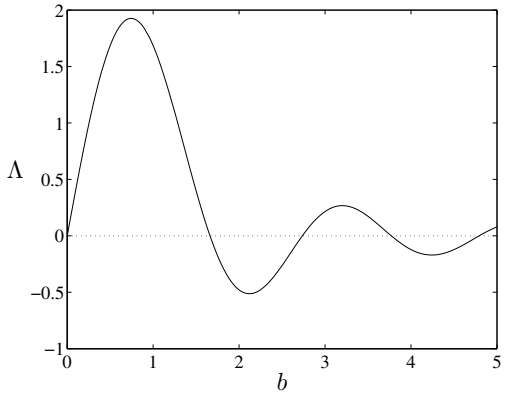


Figure 2. *Prefactor* Λ *in the asymptotics* (3.1)–(3.3) *of* C_{ij}^n *as a function of* b.

Comparing with (3.6) then gives

$$\hat{X}_n \sim \mathrm{i}(-1)^{n+1}(2n)! \sum_{i,j=0}^{n} (-1)^j C_{ij}^n.$$

The right-hand side can now be evaluated using the form (3.1)–(3.3) of C_{ij}^n. Approximating the sums by integrals, we obtain

$$\begin{aligned}\hat{X}_n &\sim \mathrm{i}(-1)^n(2n)!n\Lambda \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-15(\xi^2+\eta^2+\xi\eta)/2}\,\mathrm{d}\xi\,\mathrm{d}\eta \\ &\sim \mathrm{i}(-1)^n(2n+1)!\frac{2\pi\Lambda}{15\sqrt{3}}.\end{aligned}$$

This is a second expression for the late behavior of $\hat{X}_n$. Identifying with (3.9) leads to the relation

$$\Lambda = \frac{15\sqrt{3}}{2\pi}\kappa. \tag{3.10}$$

Thus, like κ, Λ can be obtained numerically by computing the coefficients $\hat{X}_n$ for $n \gg 1$ from the five recurrence relations for $(\hat{U}_n, \hat{V}_n, \hat{W}_n, \hat{X}_n, \hat{Y}_n)$. The results of this computation carried out for values of b in the range $(0, 5)$ are shown in Figure 2. For the value $b = 0.5$ which we use often in what follows, we find that $\Lambda = 1.6858\cdots$. Note that Λ vanishes for certain values of b; for these, the growth of the functions X_n and Y_n is slower than in the generic case $\Lambda \neq 0$, and it can be captured only by continuing the expansion beyond the leading-order term considered here.

With our estimate for the prefactor Λ, we now have the complete form of the leading-order asymptotics of C_{ij}^n and D_{ij}^n for $n \gg 1$ in the core region $\xi, \eta = O(1)$. This is compared in Figure 3 with the values of $(-1)^{j+1}C_{ij}^n$ computed numerically from the recurrence relations (2.16)–(2.19) for $n = 40$, 70, and 100. The figure confirms the asymptotic results and illustrates how the discrete dependence of C_{ij}^n on i and j asymptotes to the continuous dependence

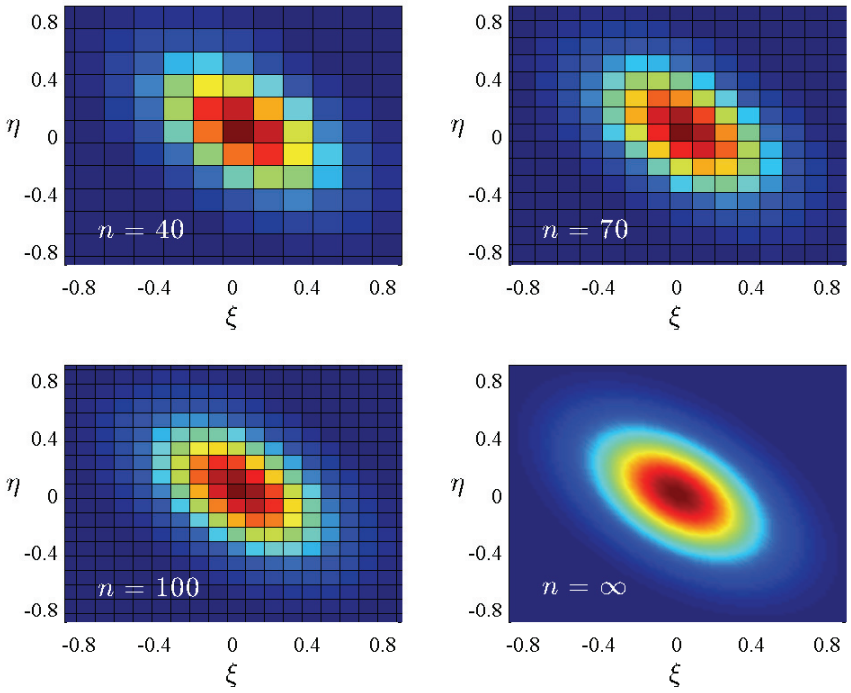


Figure 3. *Coefficients* $(-1)^{j+1}C_{ij}^n$ *as functions of* $\xi = n^{-1/2}(i - n/3)$ *and* $\eta = n^{-1/2}(j - n/3)$ *for* $n = 40$, 70, *and* 100, *and for* $b = 0.5$. *The last panel shows the asymptotic form for* $n \to \infty$.

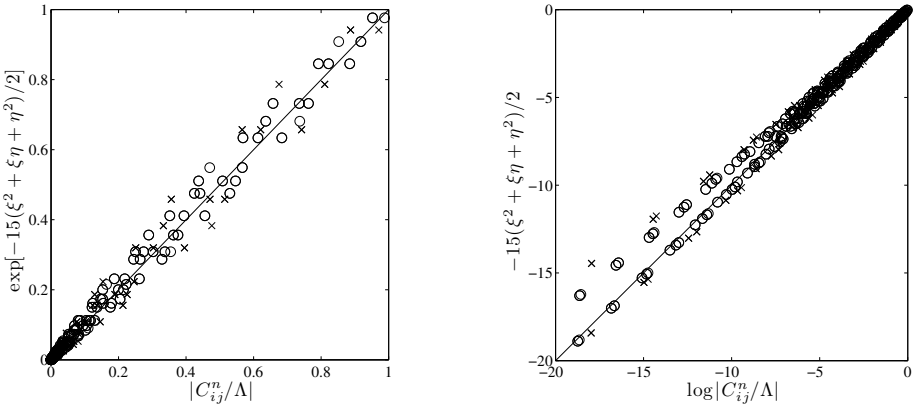


Figure 4. *Scatter plot of* $(-1)^{j+1}C_{ij}^n/\Lambda$, *with* Λ *estimated numerically, against* $\exp[-15(\xi^2 + \xi\eta + \eta^2)/2]$ *for* $b = 0.5$, *and* $n = 40$ $(\times)$ *and* $n = 100$ $(\circ)$. *The same data are plotted in linear coordinates (left panel) and in logarithmic coordinates (right panel).*

on ξ and η as $n \to \infty$. To give a more precise comparison between numerical and asymptotic results than afforded by the color-scale Figure 3, we show in Figure 4 scatter plots of $(-1)^{j+1}C_{ij}^n$, normalized by Λ, against its asymptotic limit $\exp[-15(\xi^2+\xi\eta+\eta^2)/2]$. This con-

firms the match between asymptotic and numerical results. It also shows that the convergence toward the asymptotic behavior is rather slow.

A noticeable feature of Figure 4 is the cloud of points for small values of C_{ij}^n. These correspond to indices i and j far from the core values $i, j \approx n/3$ or, in other words, to $|\xi|, |\eta| \gg 1$. In this tail region, we cannot expect the asymptotics (3.1)–(3.3) to be valid. Even though C_{ij}^n and D_{ij}^n are exponentially small there, the tail region is important for the evaluation of the coefficients X_n and Y_n from the sums (2.16)–(2.17). Indeed, for $u, v, w \in \mathbb{R}$, the largest terms in these sums are not those for which $\xi, \eta = O(1)$ but rather those for which $\xi, \eta = O(n^{1/2})$. This is because the factors $u^{2i+1}v^{2j+1)}w^{2k}$ in the sum (2.16), for instance, depend exponentially on n in the core region, since $i, j, k = O(n)$ there. Thus the asymptotic results derived so far, although providing valid estimates for the coefficients C_{ij}^n and D_{ij}^n where they are $O(1)$, are not sufficiently accurate to estimate X_n and Y_n from (2.16)–(2.17).

The situation is analogous to that encountered in probability theory when studying the distribution of the sum of random variables. The central-limit theorem provides a Gaussian approximation for the core of the distribution, but this approximation fails in the tails. These can be essential, however, for instance, if the expectation of the exponential of the sum is to be estimated. It is therefore necessary to go beyond the central-limit theorem and use the theory of large deviations, which gives an estimate of the distribution valid in the tails. Here, similarly, it is necessary to derive an approximation for C_{ij}^n and D_{ij}^n for $n \gg 1$ when $\xi, \eta = O(n^{1/2})$. This is done next.

3.2. Tail: $\xi, \eta = O(n^{1/2})$.

We start with the "large-deviation" ansatz

$$C_{ij}^n = (-1)^{j+1} A(a, b, n) \mathrm{e}^{-nG(a,b)}, \tag{3.11}$$

where

$$a = \frac{i}{n} - \frac{1}{3} \quad \text{and} \quad b = \frac{j}{n} - \frac{1}{3}.$$

Here the functions A and G need to be determined to satisfy the recurrence relations (2.18)–(2.19). The dependence of A on n is assumed to be such that its partial derivatives (denoted by subscripts) satisfy $A_a, A_b = O(1)$ as $n \to \infty$. We will be concerned only with determining the function G which governs the dominant, or controlling, behavior of C_{ij}^n. This function satisfies

$$G(0,0) = G_a(0,0) = G_b(0,0) = 0.$$

This is necessary to recover the Gaussian form given in (3.1) and (3.3)–(3.4) when $a = n^{-1/2}\xi = O(n^{-1/2})$ and $b = n^{-1/2}\eta = O(n^{-1/2})$. More specifically,

$$G(a, b) \sim \frac{15}{2}\left(a^2 + ab + b^2\right) \quad \text{as} \quad a, b \to 0. \tag{3.12}$$

Introducing (3.11) into (2.18)–(2.19) and retaining only the leading-order term yields a nonlinear differential equation for G which is too lengthy to reproduce here. It can, however, be much simplified by introducing the Legendre transform

$$S(p, q) = \sup_{a,b} \left(ap + bq - G(a, b)\right), \tag{3.13}$$

with $p = G_a(a,b)$ and $q = G_b(a,b)$. In terms of S, with p and q as independent variables, the equation satisfied by G takes the form

$$\left[(1-\mathrm{e}^{-p})S_p + (1-\mathrm{e}^{-q})S_q - \frac{1}{3}(1+\mathrm{e}^{-p}+\mathrm{e}^{-q})\right]^2 = \mathrm{e}^{S-2p/3-2q/3}.$$

Taking the square root, we obtain

$$(1-\mathrm{e}^{-p})S_p + (1-\mathrm{e}^{-q})S_q = \frac{1}{3}(1+\mathrm{e}^{-p}+\mathrm{e}^{-q}) - \mathrm{e}^{S/2-p/3-q/3}. \tag{3.14}$$

The sign choice is justified by considering this equation for small p and q. Assuming that S is quadratic, (3.14) reduces to

$$pS_p + qS_q + \frac{S}{2} = \frac{1}{9}\left(p^2 - pq + q^2\right) + \cdots,$$

where $\cdots$ denotes cubic- and higher-order terms. Solving gives

$$S(p,q) \sim \frac{2}{45}\left(p^2 - pq + q^2\right) \quad \text{as} \quad p,q \to 0, \tag{3.15}$$

which is the Legendre transform of (3.12), as expected.

The nonlinear equation (3.14) can be solved explicitly. Let

$$P = \frac{1}{2}\log\left[(\mathrm{e}^p - 1)(\mathrm{e}^q - 1)\right], \quad Q = \frac{1}{2}\log\left(\frac{\mathrm{e}^p - 1}{\mathrm{e}^q - 1}\right), \tag{3.16}$$

and

$$S(p,q) = \hat{S}(P,Q) + P - \frac{1}{3}(p+q). \tag{3.17}$$

We note that the branches of the logarithms in (3.16) need to be specified: a suitable choice takes $-\pi/2 < \arg(\mathrm{e}^p - 1) \le 3\pi/2$ and $-3\pi/2 < \arg(\mathrm{e}^q - 1) \le \pi/2$ so that $P(-p,-q) = P(p,q)$ and $Q(-p,-q) = Q(p,q) + \mathrm{i}\pi$ for $p,q > 0$. Introducing the variable transformation (3.16)–(3.17) into (3.14) leads to the simpler equation

$$\hat{S}_P = -\frac{\mathrm{e}^{(\hat{S}+P)/2}}{\left[\left(1+\mathrm{e}^{P+Q}\right)\left(1+\mathrm{e}^{P-Q}\right)\right]^{1/2}}$$

involving a P-derivative only. Integrating gives the solution

$$\hat{S}(P,Q) = -2\log\left(\frac{1}{2}\int^P \frac{\mathrm{e}^{P'/2}}{\left[\left(1+\mathrm{e}^{P'+Q}\right)\left(1+\mathrm{e}^{P'-Q}\right)\right]^{1/2}}\,\mathrm{d}P' + C(Q)\right), \tag{3.18}$$

where the function $C(Q)$ remains to be determined. It can be shown that $C(Q) = 0$ if the lower limit of integration in (3.18) is taken as $-\infty$ so that

$$\hat{S}(P,Q) = -2\log\left(\frac{1}{2}\int_{-\infty}^P \frac{\mathrm{e}^{P'/2}}{\left[\left(1+\mathrm{e}^{P'+Q}\right)\left(1+\mathrm{e}^{P'-Q}\right)\right]^{1/2}}\,\mathrm{d}P'\right). \tag{3.19}$$

Indeed, this choice ensures that the limiting behavior (3.15) is recovered for $p, q \to 0$. To verify this, note that $P \to -\infty$ and hence $\exp(P) \to 0$ as $p, q \to 0$. The denominator of the integrand in (3.19) can then be expanded, leading to

$$\begin{aligned}\hat{S}(P,Q) &= -2\log\left(\frac{1}{2}\int_{-\infty}^{P} \mathrm{e}^{P'/2}\left[1-\mathrm{e}^{P'}\cosh Q+\mathrm{e}^{2P'}\left(\frac{1}{4}+\frac{3}{4}\cosh(2Q)\right)+\cdots\right]\mathrm{d}P'\right)\\ &= -P+\frac{2}{3}\mathrm{e}^{P}\cosh Q-\frac{2}{5}\mathrm{e}^{2P}\left(\frac{1}{4}+\frac{3}{4}\cosh(2Q)\right)+\frac{1}{9}\mathrm{e}^{2P}\cosh^2 Q+\cdots.\end{aligned}$$

On using the approximations $\exp(P+Q) = p + p^2/2 + \cdots$, $\exp(P-Q) = q + q^2/2 + \cdots$, $\exp(2P) = pq + \cdots$, and $\exp(2Q) = p/q + \cdots$, this further simplifies to

$$\hat{S}(P,Q) = -P + \frac{1}{3}(p+q) + \frac{2}{45}\left(p^2 - pq + q^2\right) + \cdots.$$

Introducing this result into (3.17) reduces $S(p,q)$ to the form (3.15), as required.

With (3.19) established, the derivation of the large-n behavior of C_{ij}^n for $\xi, \eta = O(n^{-1/2})$ is complete: $S(p,q)$ can be calculated from (3.17), and the function $G(a,b)$ follows by inverting the Legendre transform (3.13). If the asymptotic form of C_{ij}^n is used only to approximate the coefficients $X_n(u,v,w)$, as is done in the next section, the inversion step is in fact not necessary since the X_n can be expressed directly in terms of $S(p,q)$.

4. Late behavior of X_n and Y_n. In this section, we present two approaches for the derivation of the asymptotic form of the slaving coefficients X_n and Y_n for $n \gg 1$. One approach relies on our approximation (3.11) for C_{ij}^n; the other considers the superbalance equation (2.13)–(2.14) directly. We start with the latter approach, which turns out to be somewhat simpler.

4.1. From the superbalance equation. From (2.13)–(2.14), and assuming that the linear terms dominate for large n, we have that

$$X_{n+1} \sim -\left(vw\frac{\partial}{\partial u} + uw\frac{\partial}{\partial v} - uv\frac{\partial}{\partial w}\right)^2 X_n \tag{4.1}$$

as $n \to \infty$. This recurrence relation can be solved using characteristics: let $(U,V,W)(t)$ be the solutions of the leading-order slow equations, namely,

$$\dot{U} = -VW, \tag{4.2}$$

$$\dot{V} = UW, \tag{4.3}$$

$$\dot{W} = -UV, \tag{4.4}$$

with initial conditions $(U,V,W)(0) = (u,v,w)$. Then the solution of (4.1) can be written as

$$X_n(u,v,w) \sim (-1)^n \left.\frac{\mathrm{d}^{2n}}{\mathrm{d}t^{2n}}\right|_{t=0} \tilde{X}_0(U(t),V(t),W(t)). \tag{4.5}$$

Here $\tilde{X}_0$ is an unknown polynomial, determined by the early iterations when the nonlinear terms neglected in (4.1) are significant. The details of $\tilde{X}_0$ do not matter: what controls the

right-hand side of (4.5) for large n are the singularities of $(U,V,W)(t)$ nearest the origin of the complex t-plane (e.g., [5]). Let t_* and $\bar{t}_*$, where the overbar denotes complex conjugation, be these singularities, and assume they are poles of order r. These poles should be thought of as functions of the slow variables: $t_* = t_*(u,v,w)$. Then, as $t \to t_*$,

$$\tilde{X}_0 \sim \frac{C}{(t-t_*)^r}, \tag{4.6}$$

where C is a constant which may depend on t_*. The complex conjugate behavior holds as $t \to \bar{t}_*$. These behaviors control the late asymptotics of X_n. From (4.5)–(4.6), we obtain the asymptotics in the explicit form

$$X_n \sim \frac{(-1)^n(2n+r-1)!C}{(r-1)!(-t_*)^{2n+r}} + \text{c.c.}$$

Comparison with (3.6) and (3.9) then shows that

$$r = 2 \quad \text{and} \quad C = \mathrm{i}\kappa$$

for trajectories $(U(t),V(t),W(t))$ consistent with (3.7). It follows that

$$X_n \sim \frac{(-1)^n(2n+1)!\mathrm{i}\kappa}{t_*^{2n+2}} + \text{c.c.} \tag{4.7}$$

The relationship $t_* = t_*(u,v,w)$ can be made completely explicit using the solution of (4.2)–(4.4). In what follows, we assume that $|w| \geq |u|$. This means that we consider the open trajectories in the (ϕ,w)-plane represented in Figure 1. (The case $|w| < |u|$, corresponding to closed trajectories, is treated similarly, by swapping the roles of u and w.) Defining

$$u_0 = \pm\sqrt{u^2+v^2} \quad \text{and} \quad w_0 = \pm\sqrt{v^2+w^2},$$

with the signs those of u and w, respectively, the solution of (4.2)–(4.4) can be written in terms of Jacobi elliptic functions as

$$U(t) = u_0\,\mathrm{cn}(w_0(t-t_0);k), \tag{4.8}$$

$$V(t) = u_0\,\mathrm{sn}(w_0(t-t_0);k), \tag{4.9}$$

$$W(t) = w_0\,\mathrm{dn}(w_0(t-t_0);k), \tag{4.10}$$

where the modulus $k = u_0/w_0 \leq 1$ (e.g., [1, Ch. 16]). The constant t_0 is determined by the initial conditions $(U,V,W)(0) = (u,v,w)$. With $\phi(u,v)$ defined as in (2.6) and taken in $(-\pi,\pi)$, we find that

$$t_0 = -\frac{1}{w_0}F(\phi(u,v);k), \tag{4.11}$$

where F denotes the elliptic integral of the first kind, defined as

$$F(\phi;k) = \int_0^\phi (1-k^2\sin^2\sigma)^{-1/2}\,\mathrm{d}\sigma$$

(e.g., [1, Ch. 17]).

Now, the poles of the elliptic functions in (4.8)–(4.10) are located on the lattice

$$t_* = t_0 + 2r\frac{K(k)}{w_0} + \mathrm{i}(2s+1)\frac{K(k')}{w_0}, \quad r, s \in \mathbb{Z}, \tag{4.12}$$

where $k'^2 = 1 - k^2$ and $K(k) = F(\pi/2; k)$ denotes the complete elliptic integral of the first kind ([1, Ch. 17]). The poles nearest the origin clearly have $s = 0$ or $s = -1$. We choose to denote by t_* the pole corresponding to $s = 0$; the pole corresponding to $s = -1$ is its complex conjugate $\bar{t}_*$. With this convention, the poles nearest the origin are given by t_* and $\bar{t}_*$, with

$$t_* = t_0 + 2r\frac{K(k)}{w_0} + \mathrm{i}\frac{K(k')}{w_0}. \tag{4.13}$$

Here,

$$r = \begin{cases} -1 & \text{for} \quad \phi \in (-\pi, -\pi/2), \\ 0 & \text{for} \quad \phi \in (-\pi/2, \pi/2), \\ 1 & \text{for} \quad \phi \in (\pi/2, \pi), \end{cases}$$

since, according to (4.11), t_0 is a monotonic function of ϕ in $(-\pi, -\pi/2)$, $(-\pi/2, \pi/2)$, and $(\pi/2, \pi)$, which satisfies $t_0 = \mp K(k)/w_0$ for $\phi = \pm\pi/2$ and $t_0 = \mp 2K(k)/w_0$ for $\phi = \pm\pi$. Substituting (4.13) into (4.7) gives the completely explicit form

$$X_n \sim \frac{(-1)^n(2n+1)!\mathrm{i}\kappa w_0^{2n}}{\left(F(\phi(u,v);k) - 2rK(k) - \mathrm{i}K(k')\right)^{2n+2}} + \text{c.c.} \tag{4.14}$$

for the late asymptotics of X_n. An analogous expression can be derived for Y_n.

Three remarks are in order. The first concerns the sign of the right-hand side of (4.14). Near the pole with $r = 0$, the behavior of the solution (4.8)–(4.10) is consistent with (3.7), as assumed in the derivation. Near the poles with $r = \pm 1$, the signs of $U(t)$ and $V(t)$ are opposite those in (3.7), but the sign of X_n remains unchanged because the transformation $(u, v, w, x, y) \mapsto (-u, -v, w, x, y)$ leaves (2.1)–(2.5) invariant. The second remark concerns the discontinuous behavior of X_n at $\phi(u, v) = \pm\pi/2$, that is, for $u = 0$. This is immediately remedied by noting that the two pairs of complex-conjugate poles with $r = 0$ and $r = \mp 1$ both contribute to X_n at the same order in a neighborhood of size $O(n^{-1})$ of $\phi = \pm\pi/2$. Adding the two contributions then leads to an approximation for X_n that is continuous at $\phi = \pm\pi/2$. The third remark is that the factorial growth of X_n described by (4.14) means that the asymptotic series (2.15) defining the slow manifold is of Gevrey type of order 1; the divergence of this type of series and their resummability is well understood (e.g., [2]).

The asymptotic result (4.14) is illustrated by Figure 5, which compares the asymptotic and numerical estimates of $X_n/(2n)!$ as a function of ϕ for fixed $u_0 = 1$, $w = 2$, and $b = 0.5$. Since X_n is a π-periodic function of ϕ, it is plotted only for $\phi \in [0, \pi)$. The upper panel of the figure corresponds to $n = 5$ and the lower panel to $n = 20$. For $n = 5$, we show two asymptotic estimates: the first takes into account only the pair of complex-conjugate poles nearest the origin; the second takes into account the two nearest pairs. As remarked above, the latter approximation eliminates the discontinuity at $\phi = \pi/2$; it also matches the

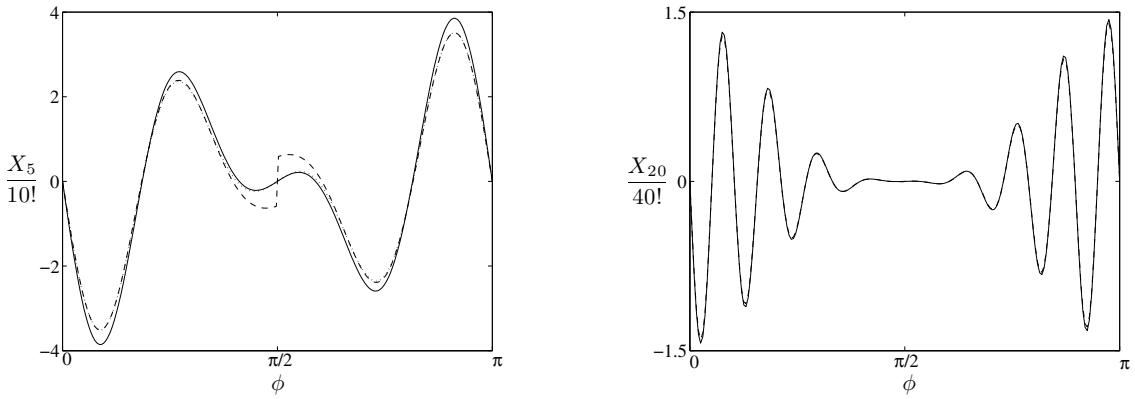


Figure 5. *Estimates of $X_n(u, v, w)$ as a function of ϕ for $w = 2$ and $b = 0.5$: Numerical results (solid curves) are compared with asymptotic results (dashed curve) for $n = 5$ and $n = 20$. For $n = 5$ the asymptotic result obtained by taking into account two poles is also shown (dotted curve).*

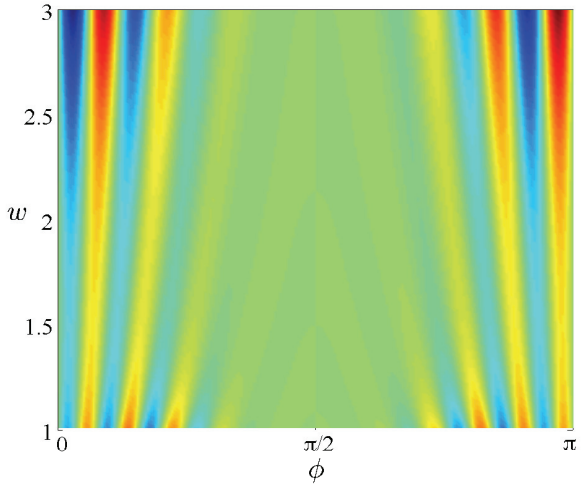


Figure 6. *X_{20} as a function of ϕ and w for $u_0 = 1$ and $b = 0.5$. (X_{20} has been normalized by w^{23} for the clarity of the picture.)*

numerical results remarkably well. For $n = 20$, the match is already excellent with a single pair of complex-conjugate poles, and the curves are indistinguishable.

To illustrate further the manner in which the coefficients X_n depend on ϕ and w, we show in Figure 6 results of the numerical computation of X_{20} for $\phi \in [0, \pi)$ and $w \in [1, 3)$. The values of X_{20} increase rapidly with w; in order to make the dependence on ϕ visible in the color scale for the smaller values of w, we have plotted X_n/w^α rather than X_n, with the parameter α chosen as $\alpha = 23$ to minimize the variations in color in the w direction. A completely indistinguishable picture would have been produced had we used the asymptotic estimate for X_{20} in place of the numerical results.

4.2. From the coefficients C_{ij}^n. We now present a derivation of the asymptotics (4.14) alternative to that of the previous section. This relies on the results of section 3 providing the asymptotics of the polynomial coefficients C_{ij}^n with sufficient accuracy that the sums (2.16)–(2.17) can be estimated. A possible advantage of this approach is that it is less dependent on the exact solution of the leading-order balanced model (4.2)–(4.4) and hence on the integrability of that model.

To obtain the asymptotics of X_n from that of C_{ij}^n, we introduce (3.11) into (2.16) to find that

$$X_n(u,v,w) \asymp (2n)! \sum_{i,j=0} A(a,b)(-1)^{j+1} \mathrm{e}^{-nG(a,b)} u^{2i+1} v^{2j+1} w^{2k}, \tag{4.15}$$

with $k = n - i - j$. In this section, we concentrate on the controlling behavior of X_n for $n \gg 1$ and ignore order-one prefactors. This is indicated by the symbol $\asymp$. The oscillations introduced by the factor $(-1)^{j+1}$ in the sum make the validity of the expression questionable for $v \in \mathbb{R}$. However, one can use it safely, for instance, if $\nu = \mathrm{i}v \in \mathbb{R}$, and then use an analytic continuation argument for $v \in \mathbb{R}$. We proceed in this formal manner. Approximating the sums in (4.15) by integrals over a and b gives

$$X_n \asymp (2n)! n^2 (u\nu w)^{2n/3} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(a,b) \mathrm{e}^{n[2a\log(u/w) + 2b\log(\nu/w) - G(a,b)]} \,\mathrm{d}a\,\mathrm{d}b. \tag{4.16}$$

The integrals can be approximated by Laplace's method to obtain

$$X_n \asymp (2n)! n (u\nu w)^{2n/3} \mathrm{e}^{S(p,q)}, \tag{4.17}$$

where S is the Legendre transform of G defined in (3.13),

$$p = \log\left(\frac{u}{w}\right)^2 \quad \text{and} \quad q = \log\left(\frac{\nu}{w}\right)^2.$$

Several simplifications occur upon using the variable transformation (3.16)–(3.17): this reduces (4.17) to the form

$$X_n \asymp (2n+1)! \left[(w^2 - u^2)(w^2 - \nu^2)\right]^{n/2} \mathrm{e}^{n\hat{S}(P,Q)}, \tag{4.18}$$

where

$$P = \frac{1}{2} \log \frac{(w^2 - u^2)(w^2 - \nu^2)}{w^4} \quad \text{and} \quad Q = \frac{1}{2} \log \frac{w^2 - u^2}{w^2 - \nu^2}. \tag{4.19}$$

At this point, we can reintroduce $\mathrm{i}v$ in place of ν and take $v \in \mathbb{R}$. Doing so, we analytically continue the function given by the integral in (4.16) for $v \in \mathrm{i}\mathbb{R}$ to $v \in \mathbb{R}$. This provides an approximation to at least one branch of (4.15) thought of as an analytic function of v in the complex plane minus possible branch cuts. Note that the arguments of the logarithm in P and Q are both positive if we assume, as in section 4.1, that $|w| \geq |u|$. However, for our choice of branch for the definition of P and Q, $\arg(\mathrm{e}^p - 1) = \arg(u^2/w^2 - 1) = \pi$, $\arg(\mathrm{e}^q - 1) = \arg(-v^2/w^2 - 1) = -\pi$, and hence

$$P = \frac{1}{2} \log \frac{(w^2 - u^2)(w^2 + v^2)}{w^4} \quad \text{and} \quad Q = \frac{1}{2} \log \frac{w^2 - u^2}{w^2 + v^2} + \mathrm{i}\pi. \tag{4.20}$$

We now consider the integral appearing in the expression (3.18) for $\hat{S}$, namely,

$$I = \int_{-\infty}^{P} \frac{\mathrm{e}^{P'/2}}{\left[\left(1+\mathrm{e}^{P'+Q}\right)\left(1+\mathrm{e}^{P'-Q}\right)\right]^{1/2}}\,\mathrm{d}P', \tag{4.21}$$

and note that the factor $1+\mathrm{e}^{P'-Q}$ in the denominator changes sign in the integration range for $P' = Q - \mathrm{i}\pi < P$. We introduce the change integration variable from P' to z', with

$$z'^2 = \frac{1+\mathrm{e}^{Q-P'}}{1-\mathrm{e}^{2Q}},$$

which maps $P' = -\infty$ to $z' = \infty$, $P' = Q - \mathrm{i}\pi$ to $z' = 0$, and $P' = P$ to $z' = z$, where

$$z^2 = \frac{1+\mathrm{e}^{Q-P}}{1-\mathrm{e}^{2Q}} = (v/u_0)^2, \tag{4.22}$$

with the last equality following from (4.20). The change of variables makes it possible to express I in terms of elliptic integrals as

$$\begin{aligned} I &= \left(\int_{Q-\mathrm{i}\pi}^{P} + \int_{-\infty}^{Q-\mathrm{i}\pi}\right) \frac{\mathrm{e}^{P'/2}}{\left[\left(1+\mathrm{e}^{P'+Q}\right)\left(1+\mathrm{e}^{P'-Q}\right)\right]^{1/2}}\,\mathrm{d}P' \\ &= -\mathrm{e}^{Q/2}\left[F(\phi(u,v);k) \pm \mathrm{i}K'(k)\right], \end{aligned}$$

where the $\pm$ sign depends on a branch choice and $\phi(u,v) = \sin^{-1}(v/u_0)$ is assumed to be in $(-\pi/2, \pi/2)$. In writing this expression we recover the parameter k appearing in section 4.1 from the computation

$$1 - \mathrm{e}^{2Q} = \frac{u^2+v^2}{v^2+w^2} = \frac{u_0^2}{w_0^2} = k^2 \tag{4.23}$$

using (4.20). Similarly, we compute $\mathrm{e}^{Q/2} = \mathrm{i}(w^2-u^2)^{1/4}(w^2+v^2)^{-1/4}$ and finally reduce (4.18) to

$$X_n \asymp \frac{(2n+1)!(-1)^n w_0^{2n}}{\left[F(\phi(u,v);k) \pm \mathrm{i}K'(k)\right]^{2n}}. \tag{4.24}$$

Once the two complex-conjugate contributions are taken into account, this is consistent with (4.14) when $-\pi/2 < \phi < \pi/2$. For $-\pi < \phi < -\pi/2$ and $\pi/2 < \phi < \pi$, other branch choices must be made in (4.21) to recover (4.14).

5. Resummation. Our main aim for examining the late asymptotics of the coefficients X_n and Y_n is to control the divergence of the power-series expansion of the slaving relation (2.12) defining the slow manifold. This makes it possible to ascertain how a unique slow manifold can be defined, which, although not invariant, is optimal in a certain sense. A natural way of achieving this is by using Borel resummation. As we now show, the Borel summation [2] of the divergent series (2.15) provides a natural definition of a unique, piecewise-continuous slow manifold, with discontinuities across what might be termed Stokes surface.

We define the Borel transform of $X(u,v,w;\epsilon)$ by the series

$$B_X(u,v,w;\xi) = \sum_{n=0}^{\infty} \frac{X_n(u,v,w)}{(2n+1)!}\xi^{2n+1}. \tag{5.1}$$

The asymptotics (4.7) of X_n ensures that this series converges for $|\xi| < |t_*|$. Analytic continuation can then be used to define B_X for larger $|\xi|$. Formally, X can be recovered from its Borel transform by Laplace transform, according to

$$X(u,v,w;\epsilon) = \epsilon^{-2}\int_0^{\infty} \mathrm{e}^{-\xi/\epsilon} B_X(u,v,w;\xi)\,\mathrm{d}\xi, \tag{5.2}$$

as a term-by-term integration indicates. We now propose to define the optimal slow manifold for the LK model by this relation and its counterpart for $Y(u,v,w;\epsilon)$. A crucial point is that we choose the integration contour in (5.2) to be the positive real line for all values of (u,v,w). A consequence is that the optimal slow manifold defined in this manner is not analytic and indeed is not even continuous in (u,v,w). This is unavoidable since the analytic continuation of (5.2) that may be obtained by suitably deforming the contour of integration in the complex plane picks up fast oscillations across certain surfaces in the (u,v,w)-space. We discuss this next.

The loss of analyticity in X arises when singularities of B_X in the ξ-plane cross the positive real axis. The singularities of B_X are, in turn, controlled by the asymptotics of X_n for $n \gg 1$. Taking (4.7) into account, we observe that B_X has poles for $\xi = \pm \mathrm{i}t_*$, with the behavior

$$B_X(u,v,w;\xi) \sim \pm\sum_{n=0}^{\infty} (-1)^n \mathrm{i}\kappa \frac{\xi^{2n+1}}{t_*^{2n+2}} = \pm\frac{\mathrm{i}\kappa\xi}{\xi^2+t_*^2} \tag{5.3}$$

near these. Here, as in section 4, t_* is a function of the slow variables: $t_* = t_*(u,v,w)$. The sign in (5.3) should be taken as $+$ if the behavior of x near the pole is in agreement with (3.7) and as $-$ if the sign of x is opposite. (Note that all poles t_* of $(u(t),v(t),w(t),x(t),y(t))$, not only those nearest the origin, lead to contributions of this form, although the latter have a dominant role.) Thus, we conclude from (5.3) that the optimal slow manifold is discontinous for values of (u,v,w) such that there are poles t_* with $\operatorname{Re} t_* = 0$. Taking the location (4.12) of the poles t_* into account, this is seen to occur for $\phi = -\pi, 0, \pi$, that is, when $v = 0$. A simple picture therefore emerges of an optimal slow manifold analytic everywhere in (u,v,w) except on the surface $v = 0$, which can be termed Stokes surface. Across this surface, a Stokes phenomenon occurs, and $X(u,v,w;\epsilon)$ and $Y(u,v,w;\epsilon)$ jump. Not surprisingly, the jumps are associated with the generation of fast oscillations.

Let us examine this more closely by considering a trajectory of the slow system crossing the Stokes surface $v = 0$. For definiteness, we consider the crossing corresponding to $\dot v > 0$, i.e., $\phi = 0$ (the crossing with $\dot v < 0$, i.e., $\phi = \pm\pi$, is identical modulo a few sign changes). For $v < 0$, the relevant poles of the slow solution (4.8)–(4.10) have $\operatorname{Re} t_* > 0$. Considering only the poles closest to the real axis, and taking $\operatorname{Im} t_* > 0$ by convention, the location of the poles of the function B_X in the ξ-plane (the Borel plane) is as represented on the left panel of Figure 7, with four poles at $\pm\mathrm{i}t_*$ and $\pm\mathrm{i}\bar t_*$. As v increases toward 0, $\operatorname{Re} t_*$ decreases, and the

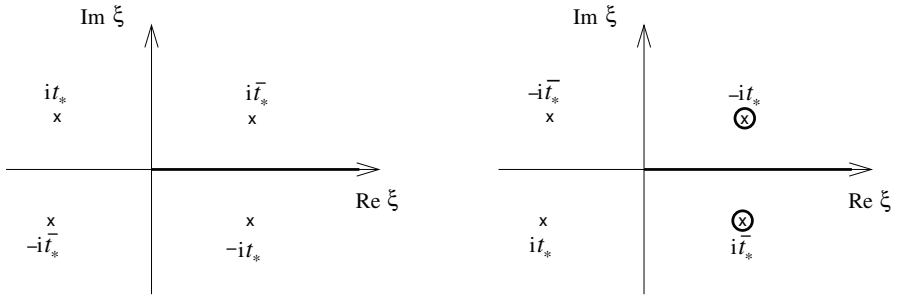


Figure 7. *Location of the poles of B_X in the complex ξ-plane for $v < 0$ (left panel), and for $v > 0$ (right panel). The optimal slow manifold, defined by as the integral of $\exp(-\xi/\epsilon)B_X$ along the positive real axis of ξ, is not analytic for $v = 0$: The optimal slow manifold for $v > 0$ differs from the analytic continuation of the manifold defined for $v < 0$ by the contributions of the two poles encircled in the right panel.*

poles move toward the real line, which they cross when $v = 0$. Thereafter, there is a difference between the function X defined by (5.2) for $v > 0$ and the function obtained by analytically continuing X from $v < 0$. The difference is the contribution of the two poles $-\mathrm{i}t_*$ and $\mathrm{i}\bar{t}_*$ that have crossed the integration contour. This contribution, computed from (5.2)–(5.3) (with the + sign) as the residue

$$X_{\mathrm{pole}}(u, v, w; \epsilon) = \frac{\pi\kappa}{\epsilon^2}\mathrm{e}^{\mathrm{i}t_*/\epsilon} + \text{c.c.} = \frac{2\pi\kappa}{\epsilon^2}\mathrm{e}^{-\operatorname{Im} t_*/\epsilon}\cos(\operatorname{Re} t_*/\epsilon), \tag{5.4}$$

corresponds to fast gravity oscillations. This is made obvious by evaluating (5.4) along the slow trajectory. To leading order in ϵ, the slow trajectory is given by

$$u(t) \sim u_0\,\mathrm{cn}(w_0 t; k), \quad v(t) \sim u_0\,\mathrm{sn}(w_0 t; k), \quad w(t) \sim w_0\,\mathrm{dn}(w_0 t; k). \tag{5.5}$$

Introducing this into (5.4) leads to

$$x_{\mathrm{pole}}(t) = X_{\mathrm{pole}}(u(t), v(t), w(t); \epsilon) = \frac{2\pi\kappa}{\epsilon^2}\mathrm{e}^{-K(k')/(\epsilon w_0)}\cos(t/\epsilon), \tag{5.6}$$

since $-\pi/2 < \phi < \pi/2$ and $t_* = -t + \mathrm{i}K(k')/w_0$, in the simple case considered with $\dot{v} > 0$. This pole contribution clearly corresponds to fast oscillations that appear when v goes through 0 and have exponentially small amplitudes, proportional to $\exp(-K(k')/(\epsilon w_0))$. The expression (5.6) coincides with that obtained in [29] using a different approach (up to a sign change arising from a different sign combination in (5.5)). The computations carried out in that paper, comparing (5.6) with results of the numerical integration of the LK system, confirms the validity of this expression.

Physically, the pole contribution represents gravity waves that are generated spontaneously by the slow balanced motion and cause the exact trajectories to move away from the optimal slow manifold by an exponentially small amount. Note that we have considered only the leading-order contribution associated with the pole t_*. In the full problem, there are not only corrections to the amplitude and phase in (5.6), but also terms with higher frequencies n/ϵ, $n > 1$, which appear as a result of the nonlinearities of the LK model.

6. General slow manifold. In this section, we briefly discuss how some of the results obtained above for the LK model generalize to a broad class of two-time-scale systems. The systems that we consider can be written in the form

$$\frac{\partial \mathsf{s}}{\partial t} = \mathsf{N}_\mathsf{s}(\mathsf{s},\mathsf{f}), \tag{6.1}$$

$$\frac{\partial \mathsf{f}}{\partial t} + \frac{1}{\epsilon}\mathcal{L}(\mathsf{s})\mathsf{f} = \mathsf{N}_\mathsf{f}(\mathsf{s},\mathsf{f}), \tag{6.2}$$

where s denotes the vector of slow variables and f the vector of fast variables. Here $\mathsf{N}_\mathsf{s}(\mathsf{s},\mathsf{f})$ and $\mathsf{N}_\mathsf{f}(\mathsf{s},\mathsf{f})$ are vector-valued functions of s and f, analytic in finite regions around $\operatorname{Im}\mathsf{s} = 0$ and $\operatorname{Im}\mathsf{f} = 0$. They are assumed to be of order one, and could depend on ϵ, but we have ignored this dependence. The matrix $\mathcal{L}(\mathsf{s})$ governs the linear dynamics of the fast variables; it is assumed to be analytic in s and skew symmetric. The eigenvalues $\pm\mathrm{i}\omega_k(\mathsf{s})$, $k = 1, 2, \ldots$, of $\mathcal{L}(\mathsf{s})$ are assumed to satisfy

$$1 < \omega_1(\mathsf{s}) < \omega_2(\mathsf{s}) < \cdots < \omega_n(\mathsf{s}).$$

The boundedness from below by a constant, which can be set to 1 by suitably defining ϵ, is crucial to ensure the time-scale separation between the variables s and f. Note that the fact that $\omega_k \neq 0$ implies that the dimension of f is even. The LK model is of the form (6.1)–(6.2), with $\mathsf{s} = (u, v, w)$, $\mathsf{f} = \epsilon(x, y)$, and $\mathcal{L}(\mathsf{s})$ given by the 2×2 canonical symplectic matrix.

The system (6.1)–(6.2) clearly has an elliptic slow manifold which, to leading order, is simply given by $\mathsf{f} = 0$. More accurate slow manifolds can be obtained by seeking a relationship

$$\mathsf{f} = \mathsf{F}(\mathsf{s};\epsilon) \tag{6.3}$$

slaving the fast variables to the slow ones. Introducing (6.3) into (6.2) and using (6.1) to eliminate the time derivative leads to the superbalance equation

$$\epsilon\mathsf{N}_\mathsf{s}(\mathsf{s},\mathsf{F}(\mathsf{s}))\cdot\partial_\mathsf{s}\mathsf{F}(\mathsf{s}) + \mathcal{L}(\mathsf{s})\mathsf{F}(\mathsf{s}) = \epsilon\mathsf{N}_\mathsf{f}(\mathsf{s},\mathsf{F}(\mathsf{s})), \tag{6.4}$$

where $\cdot$ denotes summation over the components of s. An approximation solution F can be derived by iteration or expansion in powers of ϵ. Here we use the latter procedure and write F as the formal series

$$\mathsf{F}(\mathsf{s};\epsilon) = \sum_{n=0}^{\infty} \epsilon^{n+1}\mathsf{F}^{(n)}(\mathsf{s}). \tag{6.5}$$

The successive $\mathsf{F}^{(n)}$ are then determined from a recurrence relation, starting with $\mathsf{F}^{(0)}(\mathsf{s}) = \mathcal{L}(\mathsf{s})^{-1}\mathsf{N}_\mathsf{f}(\mathsf{s},0)$.

6.1. Late behavior of $\mathsf{F}^{(n)}$. We now consider the asymptotics of $\mathsf{F}^{(n)}$ for $n \gg 1$. In the absence of detailed information on the nature of the terms N_s and N_f in (6.1)–(6.2), we cannot write $\mathsf{F}^{(n)}$ as polynomials in s, as is the case for the LK model (and, more generally, for any model where N_s and N_f are polynomials in s and f). However, it remains possible to infer the late behavior of $\mathsf{F}^{(n)}$ directly from the superbalance equation (6.4) following the approach taken for the LK model in section 4.1.

Introducing the expansion (6.5) into (6.4) and considering the coefficient of ϵ^{n+2}, say, leads to a recurrence relation for $\mathsf{F}^{(n)}$. It would be very tedious to write down this recurrence explicitly; however, our interest is in the behavior of the solution $\mathsf{F}^{(n)}$ for $n \gg 1$ only. It is therefore sufficient to consider the dominant terms in the recurrence relation; these correspond to the balance

$$\mathcal{L}(\mathsf{s})\mathsf{F}^{(n+1)}(\mathsf{s}) \sim \mathsf{N}_{\mathsf{s}}(\mathsf{s},0)\cdot\partial_{\mathsf{s}}\mathsf{F}^{(n)}(\mathsf{s}). \tag{6.6}$$

To see this, assume that $\mathsf{F}^{(n)}$ depends on n like $(n+r-1)!/a^n$ for some n-independent parameters $r \geq 0$ and $a(\mathsf{s})$, as is confirmed below. The controlling behavior of the terms retained in (6.6) (i.e., the fastest dependence on n) is then proportional to $(n-r)!$. One of the terms neglected in (6.6) is $\mathsf{F}^{(n)}(\mathsf{s})\cdot\partial_{\mathsf{f}}\mathsf{N}_{\mathsf{f}}(\mathsf{s},0)$, with controlling behavior $(n+r-1)!$, and hence smaller by a factor $1/n$ than the terms retained. All the other terms are nonlinear in $\mathsf{F}^{(n)}$ and give contributions also behaving like $(n+r-1)!$ or smaller. We demonstrate this for the quadratic terms that arise in the expansion of the first term in (6.4). Ignoring irrelevant constants, these give a contribution at $O(\epsilon^{n+2})$ of the form

$$\sum_{k=0}^{n-1}\mathsf{F}^{(n-1-j)}(\mathsf{s})\partial_{\mathsf{s}}\mathsf{F}^{(j)}(\mathsf{s}) = \mathsf{F}^{(0)}(\mathsf{s})\partial_{\mathsf{s}}\mathsf{F}^{(n-1)}(\mathsf{s}) + \sum_{k=0}^{n-2}\mathsf{F}^{(n-1-j)}(\mathsf{s})\partial_{\mathsf{s}}\mathsf{F}^{(j)}(\mathsf{s}).$$

The first term on the right-hand side behaves like $(n+r-1)!$. The controlling behavior of each of the terms in the remaining sum can be bounded by $(n+r-2)!$, so that, together, they also yield a contribution bounded by (a multiple of) $(n+r-1)!$. All the nonlinear terms neglected in (6.6) can be treated using a similar argument relying on the fact that multiple sums of powers of $\mathsf{F}^{(j)}$ are dominated by the terms involving the coefficients $\mathsf{F}^{(j)}$ with the largest possible indices j.

Now, the late behavior of $\mathsf{F}^{(n)}$ can be captured by solving the approximate recurrence relation (6.6). To do this, we define

$$\mathbf{v} = \mathsf{N}_{\mathsf{s}}(\mathsf{s},0)\cdot\partial_{\mathsf{s}}, \tag{6.7}$$

which we will think of either as a differential operator or as a vector field in the space of the slow variables s. The dynamics associated with this vector field is that of the simplest balanced model, obtained by substituting the lowest-order slaving relation $\mathsf{f}=0$ into (6.1). The approximate recurrence relation (6.6) can be rewritten in terms of $\mathbf{v}$ as

$$\mathcal{L}(\mathsf{s})\mathsf{F}^{(n+1)}(\mathsf{s}) \sim -\mathbf{v}\mathsf{F}^{(n)}(\mathsf{s}). \tag{6.8}$$

This can be solved using the method of characteristics. We denote by

$$\mathsf{S}(t) = \exp(t\mathbf{v})\mathsf{s}$$

the solution of

$$\dot{\mathsf{S}} = \mathbf{v}(\mathsf{S}) \quad \text{with} \quad \mathsf{S}(0) = \mathsf{s}.$$

Thus $\exp(t\mathbf{v})$ gives the approximate slow trajectory obtained on the leading-order slow manifold $\mathsf{f} = 0$. In terms of this trajectory, we integrate (6.8) as

$$\mathsf{F}^{(n+1)}(\mathsf{s}) \sim -\mathcal{L}(\mathsf{s})^{-1} \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathsf{F}^{(n)}(\mathrm{e}^{t\mathbf{v}}\mathsf{s}).$$

This gives the general solution of (6.6) and hence the leading-order form of the late coefficients as

$$\mathsf{F}^{(n)}(\mathsf{s}) \sim (-1)^n \mathcal{L}(\mathsf{s})^{-n} \left.\frac{\mathrm{d}^n}{\mathrm{d}t^n}\right|_{t=0} \tilde{\mathsf{F}}^{(0)}(\mathrm{e}^{t\mathbf{v}}\mathsf{s};\mathsf{s}), \tag{6.9}$$

where $\tilde{\mathsf{F}}_0$ is an unknown (vector) function, determined by the early behavior of the recurrence, when the approximation (6.6) does not hold.

As in the case of the LK model, the large-n behavior of $\mathsf{F}^{(n)}$ is controlled by the singularities of the function $\psi(t;\mathsf{s}) = \tilde{\mathsf{F}}^{(0)}(\mathrm{e}^{t\mathbf{v}}\mathsf{s};\mathsf{s})$ nearest the origin of the complex t-plane. It is difficult to make general statements about the nature of these singularities, since $\mathrm{e}^{t\mathbf{v}}\mathsf{s}$ is the solution of a nonlinear, typically nonintegrable system of ordinary differential equations. Poles, branch points, essential singularities, but also more complicated behavior such as natural boundaries are all possible. Here, we restrict our attention to the simplest situation, where the singularities nearest the real t-axis are a pair of complex-conjugate poles t_* and $\bar{t}_*$. It should be emphasized that these poles depend on s, though we do not make this explicit. Near t_*, ψ takes the form

$$\psi(t;\mathsf{s}) = \tilde{\mathsf{F}}^{(0)}(\mathrm{e}^{t\mathbf{v}}\mathsf{s};\mathsf{s}) \sim \frac{\mathsf{g}}{(t-t_*)^r}, \tag{6.10}$$

where g is a time-independent vector, depending on s only through t_*. In a manner similar to that used to determine κ (or Λ) for the LK model, it should be relatively easy to determine g by considering solutions of (6.1)–(6.2) in the limit $t \to t_*$ as expansions in powers of $(t-t_*)^{-1}$.

Introducing (6.10) into (6.9) and taking the complex-conjugate pole into account give

$$\mathsf{F}^{(n)}(\mathsf{s}) \sim \frac{(n+r-1)!}{(r-1)!(-t_*)^{n+r}} \mathcal{L}(\mathsf{s})^{-n}\mathsf{g} + \text{c.c.} \tag{6.11}$$

Now, for generic g,

$$\mathcal{L}(\mathsf{s})^{-n}\mathsf{g} \sim \frac{\alpha \mathsf{e}_1}{(\mathrm{i}\omega_1)^n} + \frac{\beta \bar{\mathsf{e}}_1}{(-\mathrm{i}\omega_1)^n}. \tag{6.12}$$

Here $\pm\mathrm{i}\omega_1$ are the lowest eigenvalues of $\mathcal{L}(\mathsf{s})$, and e_1 and its complex conjugate $\bar{\mathsf{e}}_1$ are the associated eigenvectors, normalized so that $\bar{\mathsf{e}}_1 \cdot \mathsf{e}_1 = 1$, where $\cdot$ denotes the (non-Hermitian) scalar product. The constants α and β are given by $\alpha = \bar{\mathsf{e}}_1 \cdot \mathsf{g}$ and $\beta = \mathsf{e}_1 \cdot \mathsf{g}$. Taking (6.12) into account reduces (6.11) to

$$\mathsf{F}^{(n)}(\mathsf{s}) \sim \frac{(\mathrm{i})^n(-1)^r(n+r-1)!}{(r-1)!\omega_1^n t_*^{n+r}} \left(\alpha \mathsf{e}_1 + (-1)^n \beta \bar{\mathsf{e}}_1\right) + \text{c.c.} \tag{6.13}$$

In this expression, the dependence on s of the right-hand side is through that of t_*, α, β, ω_1, and e_1. If $\mathcal{L}$ is independent of s, however, ω_1 and e_1 are constant, and α and β depend on s

through t_* only. This is the situation of the LK model. Note that (6.13) indicates that the slow manifold is again determined by an asymptotic series of Gevrey type of order 1. Note also that (6.13) has the form $(n+r-1)!/a^n$ assumed to obtain the approximate recurrence relation (6.6).

6.2. Resummation. Once the asymptotic behavior (6.13) is determined, it is possible to use the Borel summation of the divergent series (6.5) to define a unique optimal slow manifold piecewise. Specifically, we define

$$\mathsf{B}_\mathsf{F}(\mathsf{s};\xi) = \sum_{n=0}^{\infty} \frac{\mathsf{F}^{(n)}(\mathsf{s})}{(n+r-1)!}\xi^{n+r-1}, \tag{6.14}$$

which, according to (6.13), converges for $|\xi| < |\omega_1 t_*|$. The formal inversion is given by

$$\mathsf{F}(\mathsf{s};\epsilon) = \frac{1}{\epsilon^r}\int_0^\infty \mathrm{e}^{-\xi/\epsilon}\mathsf{B}_\mathsf{F}(\mathsf{s};\xi)\,\mathrm{d}\xi, \tag{6.15}$$

as is readily verified. Like for the LK model, we can choose to define an optimal slow manifold by this expression, insisting that the contour of integration be the positive real line. This slow manifold is discontinuous for the values of s such that the poles of $\mathsf{B}_\mathsf{F}(\mathsf{s};\xi)$ in ξ lie on the positive real line. The poles of $\mathsf{B}_\mathsf{F}(\mathsf{s};\xi)$ are found from (6.13) to be located at $\pm\mathrm{i}t_*$ and $\pm\mathrm{i}\bar{t}_*$. Thus the Stokes surfaces, across which the optimal slow manifold is discontinuous, are simply defined by the condition $\operatorname{Re} t_* = 0$.

The analytic continuation of (6.15) across the Stokes surface includes fast oscillations, as we now demonstrate. From (6.13), we obtain that the behavior of $\mathsf{B}_\mathsf{F}(\mathsf{s};\xi)$ near the poles $\xi = \pm\mathrm{i}t_*$ is of the form

$$\mathsf{B}_\mathsf{F}(\mathsf{s},\xi) \sim \frac{(-1)^r\mathrm{i}\omega_1}{(r-1)!}\left(\frac{\xi}{t_*}\right)^{r-1}\left(\frac{\alpha}{\xi+\mathrm{i}\omega_1 t_*}\mathsf{e}_1 - \frac{\beta}{\xi-\mathrm{i}\omega_1 t_*}\bar{\mathsf{e}}_1\right). \tag{6.16}$$

A similar expression gives the behavior near the complex-conjugate poles $\xi = \pm\mathrm{i}\bar{t}_*$. When a Stokes surface is crossed, the difference between the value of F on the optimal slow manifold and that on the full trajectory is given by the contribution of the poles which cross the positive real axis when $\operatorname{Re} t_* = 0$. Taking $\omega_1 \operatorname{Im} t_* > 0$ for definiteness, these poles are $-\mathrm{i}\omega_1 t_*$ and $\mathrm{i}\omega_1\bar{t}_*$. Computing their contribution using the residue theorem gives

$$\mathsf{F}_{\mathrm{pole}}(\mathsf{s};\epsilon) = \pm\frac{2\pi\mathrm{i}(\mathrm{i}\omega_1)^r}{\epsilon^r(r-1)!}\mathrm{e}^{\mathrm{i}\omega_1 t_*/\epsilon}\alpha\mathsf{e}_1 + \mathrm{c.c.}, \tag{6.17}$$

where the sign depends on the direction in which $-\mathrm{i}t_*$ crosses the positive real axis when $\operatorname{Re} t_* = 0$. Evaluating this expression along the approximate slow solutions $\mathsf{s}(t) = \mathrm{e}^{\mathbf{v}t}\mathsf{s}(0)$ confirms that the pole contribution corresponds to fast oscillations. Their amplitude is the exponentially small $\epsilon^{-r}\exp(-\omega_1 \operatorname{Im} t_*)$, their frequency is the lowest frequency ω_1/ϵ, and their polarization (relative size of the various components of f) is fixed by the eigenvector of $\mathcal{L}(s)$ associated with the eigenvalue $\mathrm{i}\omega_1$. The other frequencies $\omega_k > 0$, $k \neq 1$, give contributions that are smaller than (6.17) by the exponentially small factors $\exp[(\omega_1 - \omega_k)\operatorname{Im} t_*/\epsilon]$.

7. Discussion. In this paper, we have examined in detail the divergence of the asymptotic procedures leading to approximately invariant slow manifolds for the Lorenz–Krishnamurthy (LK) model. This divergence was initially observed by Lorenz [20] and was considered in some of the subsequent literature [32, 33]. Here we derive an explicit expression for the leading-order behavior of the slaving coefficients X_n as $n \to \infty$. This makes it possible to employ Borel summation to define a unique slow manifold. In this manner, we resolve the ambiguity that exists for finite-accuracy slow manifolds which are not uniquely defined even for fixed accuracy ϵ^n. The Borel summation requires a choice of integration contour in the Laplace integral that defines it. Our choice is the obvious one which minimizes the oscillations that appear in the slaved fast variables along slow trajectories. The oscillations are not completely eliminated, however: as the trajectories approach the Stokes surface, oscillations appear with the characteristic error-function switching on, which characterizes the Stokes phenomenon [4]. The definition of the slow manifold that we propose ensures that the oscillations are reduced to subexponential levels away from the Stokes surfaces.

We emphasize that our optimal slow manifold differs from Lorenz's slowest invariant manifold [20]. The latter is a truly invariant manifold consisting of periodic orbits. These periodic orbits exist because the slow system with $\epsilon = 0$ is integrable: its phase space is foliated by periodic orbits, most of which persist when $\epsilon \neq 0$ (see [6, 25, 16]). The periodic orbits with $\epsilon \neq 0$ contain exponentially small fast oscillations, with the same amplitude as the oscillations switched on by the Stokes phenomenon that we consider here. In fact, it is easy to obtain an approximation for these orbits from our results. This is achieved by adding oscillations to a solution which starts on the optimal slow manifold. The amplitude and phase of these oscillations are chosen such that after a period of the slow solution, when extra oscillations have been switched on by (two) Stokes phenomena, the complete solution returns to its initial value. This is a consistent approximation because, to leading order, the added oscillations are an approximate solution of the LK equations, and because their (exponentially) small amplitudes make their superposition possible.

We have limited our computations to the leading asymptotics of the slaving coefficients X_n and Y_n for $n \gg 1$. As a result, our estimate for the pole contributions associated with the spontaneous generation of oscillations approximates only the leading-order part of these oscillations. In other words, our computations are carried out to an exponential accuracy that is sufficient to capture the dominant part of the oscillations only. Higher accuracy would require obtaining several terms in the large-n expansion of X_n. A complete expansion for X_n includes terms of different origins. In particular, it includes contributions from all the poles t_* associated with the slow dynamics rather than from those nearest the real axis only. Furthermore, because of the nonlinearity of the recurrence relations for X_n, contributions mixing the different poles arise.

We present two approaches for the determination of the asymptotics of the slaving coefficients. The first, which applies the method of characteristics to the superbalance equation, is readily generalized formally to a large class of two-time-scale systems. This approach makes plain the connection between the exponential asymptotics carried out in [29] for solutions of the LK model and that carried out here for the slow manifold as a whole. In essence, it treats slow manifolds as unions of slow trajectories; in doing so, it turns the problem of exponential asymptotics for the partial differential equation that is the superbalance equation into a

problem of exponential asymptotics for ordinary differential equations. Practical use of this approach requires computing the location of the poles of the leading-order slow trajectories in the complex time plane. For integrable systems such as the LK model, this is possible very explicitly; however, this can be much more problematic for more complex, nonintegrable models. Our second approach relies on the observation that the slaving coefficients X_n and Y_n are polynomials in the slow variables; it concentrates then on first obtaining the asymptotics for the corresponding polynomial coefficients and then summing these. It is clear that the integrability of the LK model underlies the fact that the polynomial coefficients can be obtained in quite an explicit form. It is however possible that this type of approach remains useful in nonintegrable problems, provided that the slaving coefficients continue to take polynomial forms. Of course, some numerical work may be required—for instance, evaluating the function G (or its Legendre transform) governing the asymptotics.

We conclude by noting that the control of late coefficients together with Borel summation have been used for two-time-scale systems in the context of averaging [26]. Slow manifolds are of course closely related to averaging, and averaging order-by-order provides a means of constructing slow manifolds, at least for single-frequency systems, when the difficulties associated with resonances do not arise [15]. In this context, the control of the divergence of asymptotics series can be used as an alternative to the more standard iterative approach, with an incomplete Laplace transform in the Borel summation used in place of the optimal truncation argument to bound error terms by exponentially small quantities. In this paper, we use Borel summation as a practical tool in situations simple enough that the late coefficients in the asymptotic expansions not only can be bounded but also can be approximated accurately.

Acknowledgments. The author thanks J. G. Byatt-Smith, T. N. Bailey, A. M. Davie, and A. B. Olde Daalhuis for contributions to this work.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.
[2] W. Balser, *Formal Power Series and Linear Systems of Meromorphic Ordinary Differential Equations*, Universitext, Springer-Verlag, New York, 2000.
[3] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Springer-Verlag, New York, 1999.
[4] M. V. Berry, *Uniform asymptotic smoothing of Stokes's discontinuities*, Proc. Roy. Soc. London Ser. A, 422 (1989), pp. 7–21.
[5] M. V. Berry, *Universal oscillations of high derivatives*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 1735–1751.
[6] O. Bokhove and T. G. Shepherd, *On Hamiltonian balanced dynamics and the slowest invariant manifold*, J. Atmospheric Sci., 53 (1996), pp. 276–297.
[7] J. P. Boyd, *The slow manifold of a five-mode model*, J. Atmospheric Sci., 51 (1994), pp. 1057–1064.
[8] J. P. Boyd, *Eight definitions of the slow manifold: Seiches, pseudoseiches, and exponential smallness*, Dyn. Atmos. Oceans, 22 (1995), pp. 49–75.
[9] R. Camassa, *On the geometry of an atmospheric slow manifold*, Phys. D, 84 (1995), pp. 357–397.
[10] R. Camassa and S.-K. Tin, *The global geometry of the slow manifold in the Lorenz–Krishnamurthy model*, J. Atmospheric Sci., 53 (1996), pp. 3251–3264.
[11] C. J. Cotter, *Model Reduction for Shallow Water Dynamics: Balance, Adiabatic Invariance and Subgrid Modelling*, Ph.D. thesis, Imperial College London, London, 2004.
[12] C. J. Cotter and S. Reich, *Semigeostrophic particle motion and exponentially accurate normal forms*, Multiscale Model. Simul., 5 (2006), pp. 476–496.

[13] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[14] A. C. Fowler and G. Kember, *The Lorenz–Krishnamurthy slow manifold*, J. Atmospheric Sci., 53 (1996), pp. 1433–1437.
[15] V. Gelfreich and L. Lerman, *Almost invariant elliptic manifold in a singularly perturbed Hamiltonian system*, Nonlinearity, 15 (2002), pp. 447–457.
[16] V. Gelfreich and L. Lerman, *Long periodic orbits and invariant tori in a singularly perturbed Hamiltonian system*, Phys. D, 176 (2003), pp. 125–146.
[17] S. J. Jacobs, *Existence of a slow manifold in a model system of equations*, J. Atmospheric Sci., 48 (1991), pp. 893–901.
[18] H.-O. Kreiss, *Problems with different time scales for ordinary differential equations*, SIAM J. Numer. Anal., 16 (1979), pp. 980–998.
[19] H.-O. Kreiss and J. Lorenz, *On the existence of slow manifolds for problems with different time scales*, Philos. Trans. Roy. Soc. London Ser. A, 346 (1994), pp. 159–171.
[20] E. N. Lorenz, *Attractor sets and quasi-geostrophic equilibrium*, J. Atmospheric Sci., 37 (1980), pp. 1685–1699.
[21] E. N. Lorenz, *On the existence of a slow manifold*, J. Atmospheric Sci., 43 (1986), pp. 1547–1557.
[22] E. N. Lorenz, *The slow manifold—what is it?*, J. Atmospheric Sci., 49 (1992), pp. 2449–2451.
[23] E. N. Lorenz and V. Krishnamurthy, *On the nonexistence of a slow manifold*, J. Atmospheric Sci., 44 (1987), pp. 2940–2950.
[24] P. Lynch, *The swinging spring: A simple model for atmospheric balance*, in Large-Scale Atmosphere-Ocean Dynamics, Vol. II: Geometric Methods and Models, I. Roulstone and J. Norbury, eds., Cambridge University Press, Cambridge, UK, 2002, pp. 64–108.
[25] R. S. MacKay, *Slow manifolds*, in Energy Localisation and Transfer, T. Dauxois, A. Litvak-Hinenzon, R. S. MacKay, and A. Spanoudaki, eds., World Scientific, New York, 2004, pp. 149–192.
[26] J. Ramis and R. Schäfke, *Gevrey separation of fast and slow variables*, Nonlinearity, 9 (1996), pp. 353–384.
[27] R. Temam and D. Wirosoetisno, *Exponential approximations for the primitive equations of the ocean*, Discrete Contin. Dyn. Syst. Ser. B, 7 (2007), pp. 425–440.
[28] N. G. van Kampen, *Elimination of fast variables*, Phys. Rep., 124 (1985), pp. 69–160.
[29] J. Vanneste, *Inertia-gravity-wave generation by balanced motion: Revisiting the Lorenz-Krishnamurthy model*, J. Atmospheric Sci., 61 (2004), pp. 224–234.
[30] J. Vanneste, *Wave radiation by balanced motion in a simple model*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 783–807.
[31] J. Vanneste, *Exponential smallness of inertia-gravity-wave generation at small Rossby number*, J. Atmospheric Sci., 65 (2008), pp. 1622–1637.
[32] R. Vautard and B. Legras, *Invariant manifolds, quasi-geostrophy and initialization*, J. Atmospheric Sci., 43 (1986), pp. 565–584.
[33] T. Warn, *Nonlinear balance and quasi-geostrophic sets*, Atmos.-Ocean, 35 (1997), pp. 135–145.
[34] T. Warn, O. Bokhove, T. G. Shepherd, and G. K. Vallis, *Rossby number expansions, slaving principles, and balance dynamics*, Quart. J. R. Met. Soc., 121 (1995), pp. 723–739.
[35] D. Wirosoetisno, *Exponentially accurate balance dynamics*, Adv. Differential Equations, 9 (2004), pp. 177–196.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1191–1212
© 2008 Society for Industrial and Applied Mathematics

Novel Vehicular Trajectories for Collective Motion from Coupled Oscillator Steering Control*

Margot Kimura† and Jeff Moehlis†

Abstract. We consider a model for vehicle motion coordination for three vehicles that uses coupled oscillator steering control. Prior work on such models has focused primarily on sinusoidal coupling functions, which typically give behavior in which individual vehicles move either in straight lines or in circles. We show that other, more exotic trajectories are possible when more general coupling functions are considered. Such trajectories are associated with periodic orbits in the steering control subsystem. The proximity of these periodic orbits to heteroclinic bifurcations allows for a detailed characterization of the properties of the vehicular trajectories.

Key words. collective motion, coupled oscillators

AMS subject classifications. 34C15, 93C15, 37C27, 37C29

DOI. 10.1137/070704496

1. Introduction.

Many organisms display ordered collective motion [7], such as geese flying in a Chevron-shaped formation [22], wildebeests herding on the Serengeti plains of Africa [32], locusts swarming in sub-Saharan Africa [34], and fish schooling [28]. Collective motion is also of great interest and importance for engineering applications such as formation control of unmanned vehicles and spacecraft [18, 29, 31], cooperative robotics [8], and sensor networks [9]. Much recent work in the engineering community involves formulating and studying interaction rules which allow a population to operate in a particular collective motion state; e.g., [10, 14, 17, 18, 21, 24, 31].

In the present paper, we consider the "LPS model" for vehicle motion coordination developed by Leonard, Paley, and Sepulchre [25, 26, 27, 29, 30, 31]; cf. [19]. This considers N Dubins-type vehicles [11] which are identical, move with constant unit speed, and are globally (all-to-all) coupled:

$$\begin{aligned} \dot{r}_n &= e^{i\theta_n}, \\ \dot{\theta}_n &= u_n(r,\theta), \qquad n = 1,\dots,N. \end{aligned} \tag{1.1}$$

Here the complex vector r_n denotes the position of vehicle n with respect to the origin, while the angle θ_n denotes the orientation of its (unit) velocity vector with respect to the positive real axis. Since $r_n = x_n + iy_n$, with $(x_n, y_n) \in \mathbb{R}^2$, we will hereafter use the following equivalent

*Received by the editors October 4, 2007; accepted for publication (in revised form) by R. Murray July 9, 2008; published electronically October 24, 2008. This work was supported by National Science Foundation grants NSF-0547606 and NSF-0434328 and an Alfred P. Sloan Research Fellowship in Mathematics.
http://www.siam.org/journals/siads/7-4/70449.html
†Department of Mechanical Engineering, University of California, Santa Barbara, CA 93106 (kimura@engineering.ucsb.edu, moehlis@engineering.ucsb.edu).

equations for the velocity of each vehicle:

$$\begin{aligned} \dot{x}_n &= \cos(\theta_n), \\ \dot{y}_n &= \sin(\theta_n). \end{aligned} \tag{1.2}$$

It can be shown that the system in (1.1) is invariant to rigid group rotation and translation for controllers $u_n(r,\theta)$ that are functions of only the relative positions and headings of the vehicles, defined as $r_m - r_n$ and $\theta_m - \theta_n$, respectively [25, 26, 27, 29, 30, 31]; cf. [19].

The steering control $u_n(r,\theta)$ of the vehicles can be decomposed as

$$u_n(r,\theta) = \underbrace{\omega_0 + u_n^{head}(\theta)}_{u_n^{phase}(\theta)} + u_n^{spac}(r,\theta), \qquad n = 1,\dots,N, \tag{1.3}$$

where $\omega_0 \in \mathbb{R}$ is a constant, the heading controller u_n^{head} depends only on the relative orientation of the vehicles and governs the relative directions, and the spacing controller u_n^{spac} is used to attract the vehicles to a given spatial formation. Following [27, 29, 31], we call $u_n^{phase}(\theta)$ the *phase controller*. When the phase controller depends only on the differences $\theta_m - \theta_n$, a useful connection with the coupled oscillator literature (e.g., [5, 6, 33]) is possible.

Constructing the spacing controller is more challenging in general, since it must be designed to stabilize a specific formation. In [26, 27, 29, 30, 31], a controller that stabilizes a circular formation and a proof of stabilization are given. The basic idea is to design a potential function which is minimized when the vehicles are in the desired configuration. Then, for $u_n^{head} = 0$, it is possible to construct a Lyapunov function to demonstrate that the desired formation is asymptotically stable. For the overall system, one can use a composite Lyapunov function, made up of a linear combination of the Lyapunov functions used for the spacing and heading controls, to prove the stability of the overall desired configuration [25, 26, 27, 29, 30, 31].

The benefits of this type of model for controlling the motion of a group of vehicles are clear: the model takes advantage of results from research on coupled oscillators and translates them into a simple but robust law governing individual vehicle motion that produces the desired overall group motion.

Most of the previous work on the LPS model has assumed a sinusoidal coupling function for the phase controller:

$$u_n^{phase} = \omega_0 + \frac{k}{N}\sum_{j=1}^{N} \sin(\theta_m - \theta_n). \tag{1.4}$$

With this phase controller alone (i.e., $u_n^{spac} = 0$), the system converges asymptotically to a synchronized phase arrangement for $k > 0$, and a phase-balanced solution for $k < 0$ [23, 25, 27, 29]. Both of these phase-space solutions lead to vehicular trajectories that are either straight lines or circles, depending on the value of ω_0: for $\omega_0 \neq 0$, the trajectories converge to circles, and for $\omega_0 = 0$, the trajectories converge to straight lines.

In this paper, we explore the effects of more general coupling functions to see what other types of coordinated motion are possible for this model using the phase controller alone. We

will show that one can get trajectories that are much more exotic than straight lines or circles, and which may be advantageous in situations where one wants a relatively complicated trajectory that is a natural result of the phase controller, rather than having to piece a similar trajectory together with existing methods. The trajectories from our phase controller are characterized by almost regular, Spirograph-like shapes, where the vehicles spend some time circling one section of space before moving on to another area, eventually tracing out an annulus, which may be useful in applications where one wants a robot to patrol an appropriately shaped space while periodically doing a more careful search of a subsection of that space. These trajectories are related to heteroclinic cycles for the coupled oscillator system; see [3, 16] for related heteroclinic orbits in systems of N coupled identical oscillators; cf. [4, 5]. To simplify our analysis, we will restrict the system to three vehicles. (A discussion of more general coupling functions for two vehicles is given in [20].)

We begin with an analysis of the general phase control and then present a detailed analysis of the resulting trajectories for a specific coupling function. Sections 2 and 3 consider the case of all-to-all coupling, while section 4 considers a different coupling topology. Our conclusion is given in section 5.

2. Identical all-to-all coupling: Phase dynamics.

2.1. Equations and symmetry. A system of three identical oscillators with all-to-all identical phase-difference coupling is given by

$$\dot{\theta}_n = \omega_0 + k \sum_{m \neq n} f(\theta_m - \theta_n), \qquad n = 1, 2, 3, \tag{2.1}$$

where $\theta_n \in [0, 2\pi)$ and the coupling function f is 2π-periodic. This system of equations is equivariant with respect to the group $S_3 \times T^1$, where S_3 is the six-element permutation group generated by

$$\begin{aligned} \sigma_1 &: (\theta_1, \theta_2, \theta_3) \to (\theta_2, \theta_1, \theta_3), \\ \sigma_2 &: (\theta_1, \theta_2, \theta_3) \to (\theta_2, \theta_3, \theta_1), \end{aligned} \tag{2.2}$$

and T^1 is the circle group with action

$$\tau_\phi : (\theta_1, \theta_2, \theta_3) \to (\theta_1 + \phi,\, \theta_2 + \phi,\, \theta_3 + \phi) \tag{2.3}$$

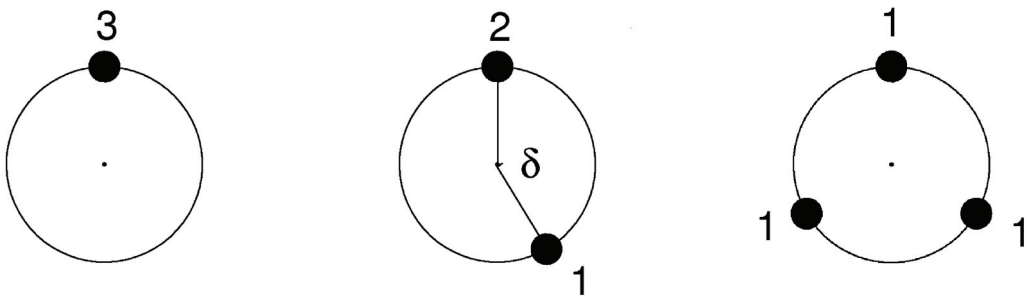
for all $\phi \in [0, 2\pi)$. This means that if $(\theta_1(t), \theta_2(t), \theta_3(t))$ is a solution to (2.1), then, for any $\gamma \in S_3 \times T^1$, so is $\gamma \cdot (\theta_1(t), \theta_2(t), \theta_3(t))$.

Equation (2.1) can be reduced to a two-dimensional system by introducing the 2π-periodic variables $\psi_1 = \theta_1 - \theta_2$ and $\psi_2 = \theta_1 - \theta_3$:

$$\begin{aligned} \dot{\psi}_1 &= \dot{\theta}_1 - \dot{\theta}_2 = k[f(-\psi_1) + f(-\psi_2) - f(\psi_1) - f(\psi_1 - \psi_2)], \\ \dot{\psi}_2 &= \dot{\theta}_1 - \dot{\theta}_3 = k[f(-\psi_1) + f(-\psi_2) - f(\psi_2) - f(\psi_2 - \psi_1)]. \end{aligned} \tag{2.4}$$

Equation (2.4) inherits equivariance with respect to the actions obtained from (2.2) and (2.3) on the ψ variables:

$$\begin{aligned} \hat{\sigma}_1 &: (\psi_1, \psi_2) \to (-\psi_1,\, \psi_2 - \psi_1), \\ \hat{\sigma}_2 &: (\psi_1, \psi_2) \to (\psi_2 - \psi_1,\, -\psi_1). \end{aligned} \tag{2.5}$$



(a) S_3 Solution (b) $S_2 \times S_1$ Solution (c) Z_3 Solution

Figure 1. *Phase-locked solutions guaranteed to exist for any coupling function f. The locations of the dots on the phase circle are determined by the values of θ for the oscillators, with the number indicating how many oscillators share the same phase. These solutions are labeled according to their isotropy subgroup, as described in the text.*

Note that $\hat{\tau}_\phi : (\psi_1, \psi_2) \to (\psi_1, \psi_2)$ acts as the identity for all ϕ. The actions $\hat{\sigma}_1$ and $\hat{\sigma}_2$ generate the permutation group S_3. We will sometimes find it convenient to think of ψ_1 and ψ_2 as being restricted to $[0, 2\pi)$, and other times it will be useful to allow them to take any real value.

2.2. Solutions and bifurcations. Phase-locked solutions are characterized by each pair of θ variables always differing by a fixed value. Thus in the ψ variables, phase-locked solutions correspond to fixed points. The symmetry and stability properties of phase-locked solutions are discussed below. As convenient, we will discuss these solutions in either the θ or the ψ variables. The three types of phase-locked solutions shown in Figure 1 are guaranteed to exist for any coupling function f of the form of (2.1), given a simple nondegeneracy condition [4, 5, 6]. These are labeled according to their isotropy subgroup, which is the set of elements of $S_3 \times T^1$ that leave the solution unchanged [15]. We note that the existence of a fixed point at (ψ_1^*, ψ_2^*) implies the existence of fixed points at $(\psi_1^* + 2\pi j,\ \psi_2^* + 2\pi m)$ for all $j \in \mathbb{Z}$ and $m \in \mathbb{Z}$.

The S_3 solutions: Fixed points at $(\psi_1^*, \psi_2^*) = (0, 0)$.

Symmetry. This phase-locked solution is invariant under the symmetry $S_3 = \langle \sigma_1, \sigma_2 \rangle$ in the θ variables, and $S_3 = \langle \hat{\sigma}_1, \hat{\sigma}_2 \rangle$ in the ψ variables; hence it has the name "S_3 solution." Since it corresponds to $\theta_1 = \theta_2 = \theta_3$, it is also sometimes referred to as the "in-phase" or "synchronous" solution.

Stability analysis and bifurcations. The Jacobian for (2.4) at the fixed point $(\psi_1^*, \psi_2^*) = (0, 0)$ has a double eigenvalue $\lambda_{1,2} = -3kf'(0)$. Thus, the stability of the fixed point depends solely on the sign of the real part of $kf'(0)$: if $kf'(0)$ is positive (resp., negative), then the S_3 solution is stable (resp., unstable).

Suppose that there is a bifurcation parameter which causes the shape of the coupling function f to change. It is immediately evident that the stability of the S_3 fixed point changes if the value $kf'(0)$ passes through zero as this parameter is varied. Because the fixed point at $(\psi_1^*, \psi_2^*) = (0, 0)$ will persist for all f, this corresponds to an S_3-symmetric transcritical

bifurcation. Assuming that there are no fixed points on the invariant lines $\psi_1 = 0$, $\psi_2 = 0$, or $\psi_1 = \psi_2$, for $(\psi_1, \psi_2) \in [0, 2\pi)$, at this bifurcation, a triangular heteroclinic connection appears between the fixed points at $(\psi_1^*, \psi_2^*) = (0, 0)$, $(2\pi, 0)$, and $(0, 2\pi)$. Since these points are identified by the 2π-periodicity of ψ_1 and ψ_2, this can also be referred to as a homoclinic connection. Thus, the authors of [4] call this an $\mathbb{S}_3$ transcritical/homoclinic bifurcation, or $\mathbb{S}_{3\text{THB}}$. If the heteroclinic loop is attracting at the bifurcation, the system will have a stable limit cycle very close to the triangle on the side of the bifurcation where the S_3 solution is unstable. Such a bifurcation will occur in the example below.

The $S_2 \times S_1$ solutions: Fixed points at $(\psi_1^*, \psi_2^*) = (0,\ 2\pi - \delta)$, $(2\pi - \delta,\ 0)$, and (δ, δ) for $\delta \in (0, 2\pi)$.

Symmetry. Arguments in [5, 6] imply that, provided $f'(0) \neq 0$, there must exist a $\delta \in (0, 2\pi)$ such that there is a phase-locked solution with two oscillators in phase and one oscillator shifted by the phase δ. The phase-locked solution corresponding to $(\psi_1^*, \psi_2^*) = (0,\ 2\pi - \delta)$ is invariant under the group $S_2 = \langle \sigma_1 \rangle$ in the θ variables, and $S_2 = \langle \hat{\sigma}_1 \rangle$ in the ψ variables. Following [5], this is referred to as an $S_2 \times S_1$ solution: the S_2 corresponds to the permutation just mentioned, and the S_1 refers to the identity permutation acting on the other oscillator. The other phase-locked solutions are related to this one by symmetry and are invariant under conjugate subgroups.

Stability analysis and bifurcations. The Jacobian at the fixed point $(2\pi - \delta,\ 0)$ has eigenvalues $\lambda_1 = k[-f'(\delta) - 2f'(-\delta)]$ and $\lambda_2 = k[-2f'(0) - f'(\delta)]$. Note that the symmetry-related fixed points at $(2\pi - \delta,\ 0)$ and (δ, δ) have the same stability. These points can be sinks, sources, or saddles.

Bifurcations occur when either $f'(\delta) + 2f'(-\delta) = 0$ or $f'(\delta) + 2f'(0) = 0$. Depending on the relative values of $f'(\delta)$, $f'(-\delta)$, and $f'(0)$ for different parameters of f, the fixed points' stability can change to or from a sink, source, or saddle in a pitchfork or saddle-node bifurcation; cf. [4]. Such solutions are involved in the $\mathbb{S}_{3\text{THB}}$ bifurcation described above, and can also be involved in the related global saddle-node heteroclinic bifurcation identified in [2].

The Z_3 solutions: Fixed points at $(\psi_1^*, \psi_2^*) = (\frac{2\pi}{3}, \frac{4\pi}{3})$ and $(\frac{4\pi}{3}, \frac{2\pi}{3})$.

Symmetry. The fixed point $(\psi_1^*, \psi_2^*) = (\frac{2\pi}{3}, \frac{4\pi}{3})$ corresponds to a solution for which $\theta_1 = \theta_2 + \frac{2\pi}{3}$ and $\theta_2 = \theta_3 + \frac{2\pi}{3}$. This is typically called the "splay state" because θ_1, θ_2, and θ_3 are equally spaced around the unit circle. This solution is invariant under the three-element cyclic group Z_3 generated by

$$(\theta_1, \theta_2, \theta_3) \to \left(\theta_2 + \frac{2\pi}{3},\ \theta_3 + \frac{2\pi}{3},\ \theta_1 + \frac{2\pi}{3}\right) \tag{2.6}$$

and hence is called the "Z_3 solution." In terms of the ψ variables, this solution is invariant under $\langle \hat{\sigma}_2 \rangle$, which is isomorphic to the group Z_3. The fixed point $(\psi_1^*, \psi_2^*) = (\frac{4\pi}{3}, \frac{2\pi}{3})$ is invariant under the group Z_3 generated by

$$(\theta_1, \theta_2, \theta_3) \to \left(\theta_3 + \frac{2\pi}{3},\ \theta_1 + \frac{2\pi}{3},\ \theta_2 + \frac{2\pi}{3}\right) \tag{2.7}$$

in the θ variables and $\langle \hat{\sigma}_2 \hat{\sigma}_1 \rangle$ in the ψ variables.

Stability analysis and bifurcations. The Jacobian at this fixed point $(\frac{2\pi}{3}, \frac{4\pi}{3})$ has eigenvalues $\lambda_{1,2} = k[-\frac{3}{2}(f'(\frac{2\pi}{3}) + f'(\frac{4\pi}{3})) \pm \frac{3i}{2}|f'(\frac{4\pi}{3}) - f'(\frac{2\pi}{3})|]$. Thus, unless $f'(\frac{4\pi}{3}) = f'(\frac{2\pi}{3})$,

this fixed point will be either a spiral sink or a spiral source. At $f'(\frac{2\pi}{3})+f'(\frac{4\pi}{3})=0$, the fixed point switches between a spiral sink and a spiral source, which is an indication of a Hopf bifurcation, as found in [4].

2.3. An example. As an example, we now consider the coupling function

$$f(\varphi)=\mu_1\sin(\varphi)+\mu_2\cos(\varphi)+\mu_3\sin(2\varphi), \tag{2.8}$$

which will provide a spectrum of novel trajectories when applied to vehicle motion coordination using the LPS model. While the coupling function given by (2.8) provides a nice example for our analysis of these interesting trajectories, the phenomena that produce the trajectories we consider are fairly generic, and so we expect to see similar bifurcations in the phase space and trajectories for the vehicles for other appropriate coupling functions [3, 16].

The above analysis predicts that both a $\mathbb{S}_{\text{3THB}}$ bifurcation involving the S_3 and $S_2\times S_1$ solutions and, independently, a Hopf bifurcation involving the Z_3 solutions will occur at $\mu_1+2\mu_3=0$ for the system (1.1) with coupling function (2.8). Numerical bifurcation analysis using XPPAUT [12] shows that for $\mu_2=1$, $\mu_3=-0.06$, and $k=1$ and when treating μ_1 as the bifurcation parameter, the Hopf bifurcation is subcritical, and that the branch of unstable periodic orbits turns around in a saddle-node bifurcation of periodic orbits to give stable periodic orbits; see Figure 2. This figure also illustrates that the phase space for the system can be divided into two triangles bounded by the invariant lines $\psi_1=0$, $\psi_1=2\pi$, $\psi_2=0$, $\psi_2=2\pi$, and $\psi_1=\psi_2$. Trajectories in these triangles are related by symmetry, and the resulting vehicular trajectories are identical. Thus, without loss of generality, we will assume that all initial conditions are chosen such that the system moves in the lower right triangle.

3. Identical all-to-all coupling: Vehicular trajectories. We now illustrate the richness of possible vehicular trajectories for (1.1) with identical all-to-all phase-difference steering control by considering the coupling function given in (2.8) with parameters $\mu_1=0.1$, $\mu_2=1$, and $\mu_3=-0.06$; see Figure 2(c) for the corresponding reduced phase-space system. If the system converges to the stable Z_3 solution, then the vehicles will move either in circles or in straight lines, depending on the value of ω_0, with each instantaneously moving in a direction at an angle of $\pm\frac{2\pi}{3}$ with respect to the others. Such motion has been found for the LPS model with the coupling function $f(\theta)=\sin(\theta)$ [25, 26, 27, 29, 30, 31]. However, if the system converges to the stable limit cycle, then the vehicles can display more exotic trajectories, such as the trajectory shown in Figure 3. Thus, we will focus our analysis on the solutions that converge to the stable limit cycle. As we will demonstrate later, these exotic trajectories are products of a stable limit cycle in the reduced phase system, so one can expect to see qualitatively similar trajectories for other appropriate coupling functions and coupling topologies.

Motion along the limit cycle is not uniform: the system slows near each of the fixed points and moves quickly in regions away from a fixed point. As will be explained in the following, it is from this nonuniform motion that the trajectories get their peculiar shapes. We first present an explanation of the vehicular motion in an intuitive way, and then validate the intuition with results from numerical simulations, which were done using a fourth-order variable-timestep Runge–Kutta algorithm. Without loss of generality, we will restrict discussion to the motion of vehicle 1 (denoted v_1) only. The motion of vehicle 2 (v_2) and vehicle 3 (v_3) is identical to but out of phase with the motion of v_1; this is summarized in Table 1.

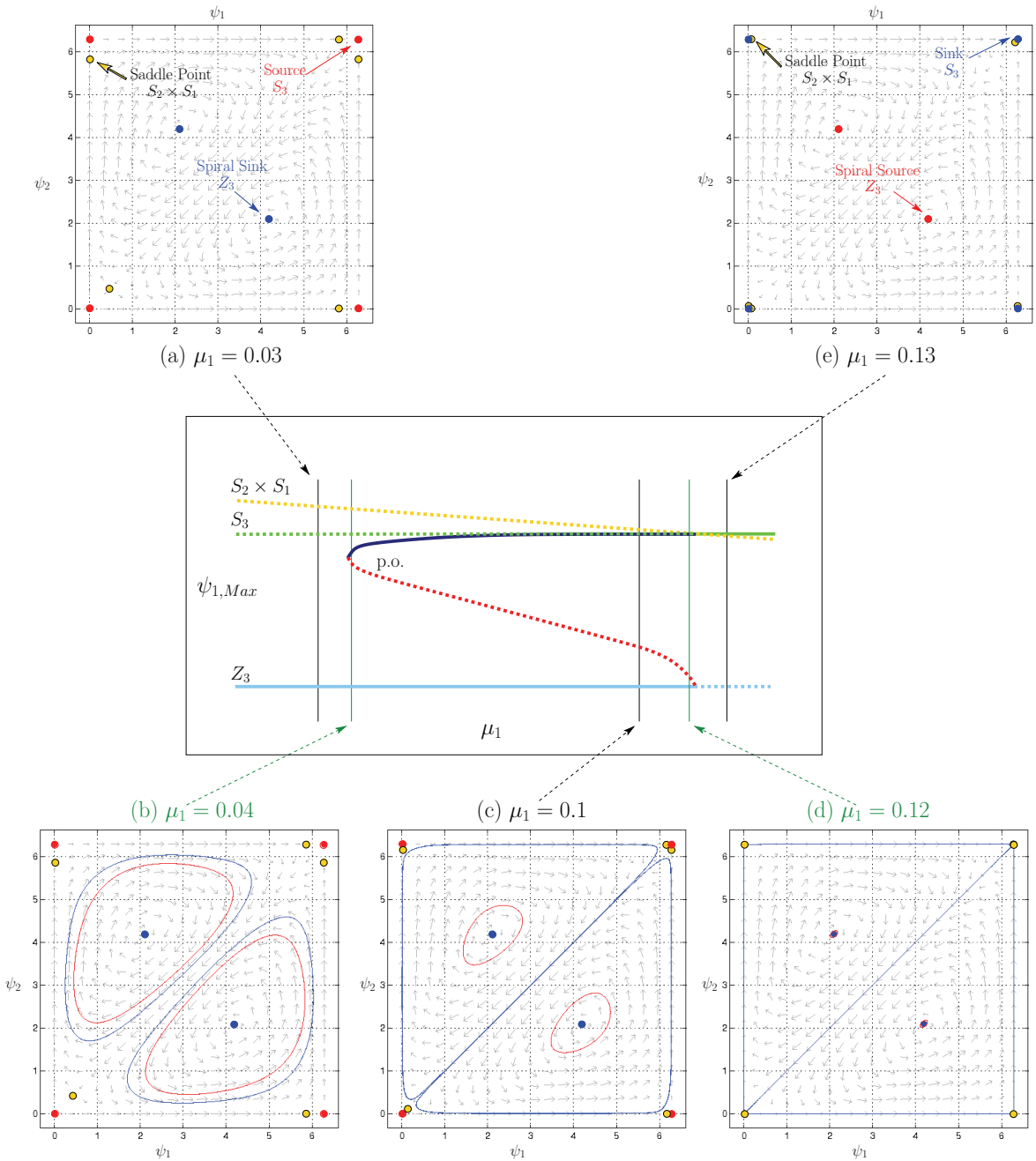


Figure 2. *The bifurcation diagram in terms of μ_1, showing the phase portraits at several values of μ_1 of interest for $\mu_2 = 1$ and $\mu_3 = -0.06$. In the (ψ_1, ψ_2) plane, yellow dots represent saddle points, red shows sources or unstable periodic orbits, and blue represents sinks or stable periodic orbits. Solid (resp., dashed) lines in the bifurcation diagram indicate stable (resp., unstable) solutions.*

3.1. The intuitive description. The overall vehicle motion in Figure 3 can be decomposed into identical units, each of which contains a cluster and a tail. We will name the tail

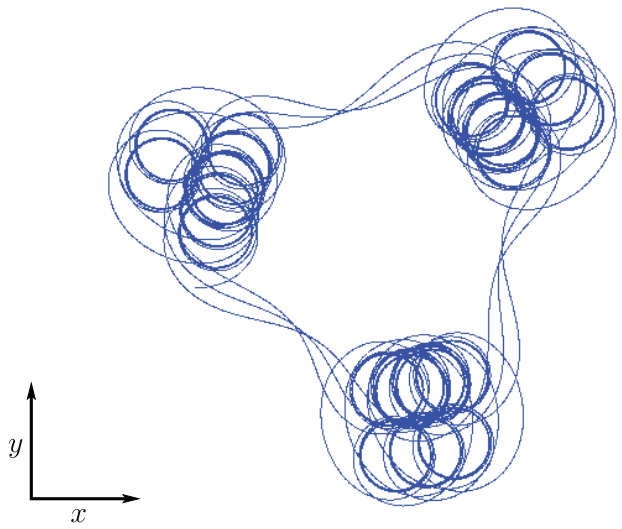


Figure 3. *An example trajectory for vehicle 1 (v_1) with parameters $\mu_1 = 0.1$, $\mu_2 = 1$, $\mu_3 = -0.06$, $\omega_0 = k = 1$. This trajectory is taken over many cycles of the periodic orbit in the (ψ_1, ψ_2) plane.*

Table 1

Relative phase and resulting behavior of all three vehicles in terms of position in the (ψ_1, ψ_2) plane. Here, $\uparrow$ means "increase(s)," and $\downarrow$ means "decrease(s)." The definition of "excursion" is given in the text.

Box	ψ behavior	θ behavior	Vehicle motion
1	$\psi_1 \uparrow$ to $\approx 2\pi$ $\psi_2 \approx 0$	θ_1 & $\theta_3 \uparrow$ at the same rate θ_2 temporarily $\downarrow$	v_1 & v_3: short excursion v_2: long excursion
2	$\psi_2 \uparrow$ to $\approx 2\pi$ $\psi_1 \approx 2\pi$	θ_1 & $\theta_2 \uparrow$ at the same rate θ_3 temporarily $\downarrow$	v_1 & v_2: short excursion v_3: long excursion
3	$\psi_1 \approx \psi_2 \downarrow$ together to ≈ 0	θ_2 & $\theta_3 \uparrow$ at the same rate θ_1 temporarily $\downarrow$	v_2 & v_3: short excursion v_1: long excursion

connecting the units a *long excursion*. Each cluster can be further broken down to show two general types of behavior: small approximately circular orbits, which we will call *small orbits*, and the roughly semicircular excursions that connect the small orbits, which we will refer to as *short excursions*. The vehicle path in a single unit can be described as a cycle through a small orbit followed by a short excursion to another small orbit, followed by a second short excursion to a third small orbit, followed by a long excursion to the next cluster. This is illustrated in Figure 4.

We can understand this behavior by dividing the periodic orbit into six boxes, as labeled in Figure 4. Simulations show that when the system in the (ψ_1, ψ_2) plane is in a lettered box (i.e., near a fixed point), the vehicles move in a small orbit, and when the system is in a numbered box, the vehicles undergo an excursion. This is expected, since the vehicles would move in a circle if the system were actually at the fixed point (i.e., generically, at a fixed point, $\dot{\theta}_j = \text{constant} \neq 0$). Therefore, one can intuitively expect the vehicles to show switching behavior between small orbits and excursions as the system moves in the (ψ_1, ψ_2) plane.

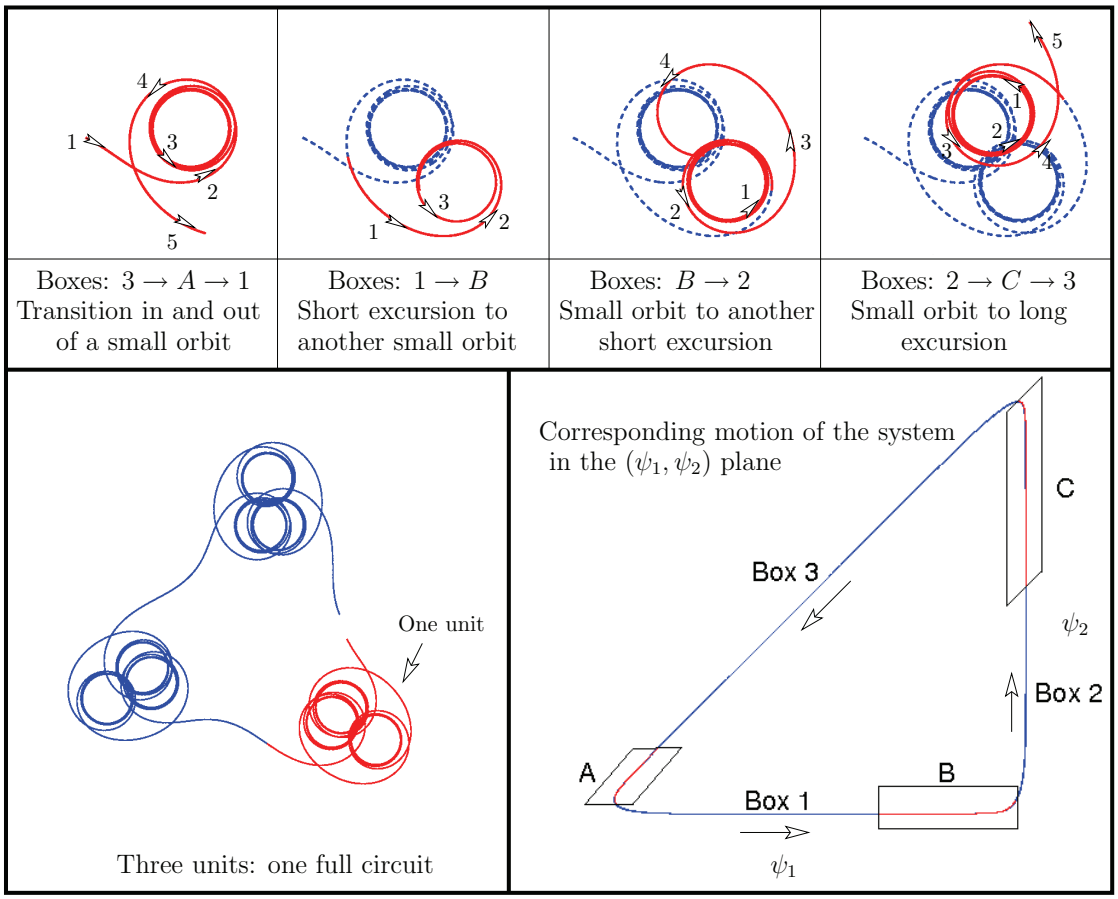


Figure 4. *Behavior of v_1 in the (x, y) plane with corresponding position of the system in the (ψ_1, ψ_2) plane. The top explains the motion of v_1 within one unit: Follow the ordered arrows in the time-series of pictures. The bottom-left panel shows one full circuit of vehicle motion, and the bottom-right panel shows the various boxes in the (ψ_1, ψ_2) plane.*

3.2. Numerical analysis and validation.

3.2.1. Box definition. To validate the above intuition, we need to be more precise about the boundaries of the boxes. Since the vehicles are always moving in a smooth and roughly circular trajectory, it is natural to define the boxes in terms of the instantaneous radius of curvature of the vehicles' trajectories. This was calculated from simulation data for each point by finding the radius of the circle defined by that point and its two neighboring points; see Figure 5.

The lettered boxes were chosen by calculating where the radius of curvature for v_1 was within 0.01 of the minima of each trough, as seen in Figure 6. Boxes 1, 2, and 3 are then defined as the intervening lengths of the periodic orbit in the (ψ_1, ψ_2) plane.

3.2.2. Approximate solutions. Within each box, we present an approximate solution with a few simplifying assumptions.

Near a fixed point (i.e., in a lettered box), the behavior of the system is approximately

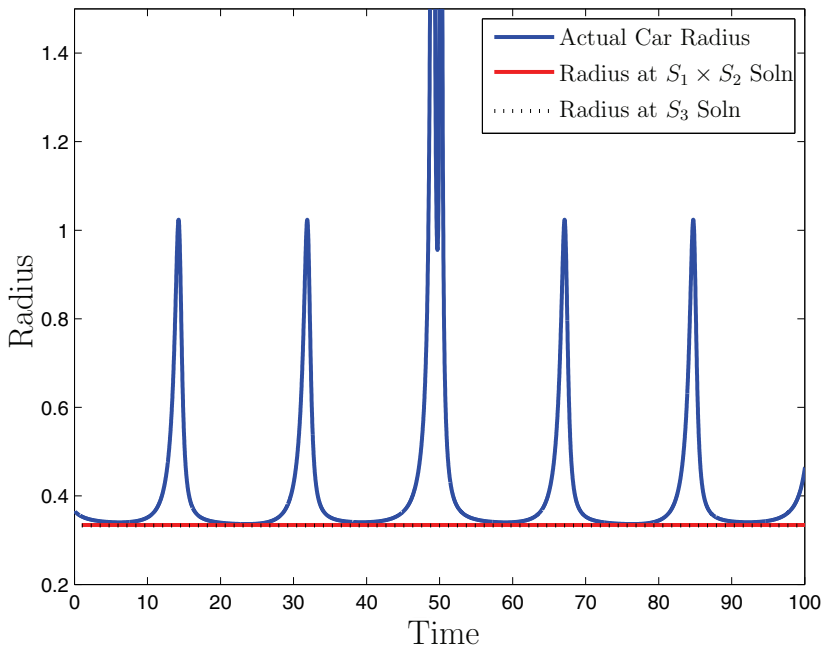


Figure 5. *Measurements of the radius of curvature for v_1 moving in the trajectory shown in Figure 3 with the approximations at each nearby fixed point. It is evident from the periodic flat troughs that the radius of curvature of the vehicles' motion spends a significant amount of time at an approximately constant value. Moreover, the value of that constant value is very close to the radius of curvature the vehicles' motion would have if the system were at the $S_2 \times S_1$ solution. See Figure 6 for an enlargement.*

the same as if the system were actually at the fixed point. At a fixed point, we have $\dot{\theta}_1 = \dot{\theta}_2 = \dot{\theta}_3 \equiv \varpi$, where ϖ is a constant. This is easily integrated, giving

$$\theta_i(t) = \varpi t + \theta_{0i}.$$

This corresponds to the following equations in the (x, y) plane:

$$\dot{x}_i = \cos(\varpi t + \theta_{0i}),$$
$$\dot{y}_i = \sin(\varpi t + \theta_{0i}).$$

These equations can also be integrated, yielding

$$x_i = \tfrac{1}{\varpi} \sin(\varpi t + \theta_{0i}),$$
$$y_i = -\tfrac{1}{\varpi} \cos(\varpi t + \theta_{0i}),$$

corresponding to motion in a circle of radius $\frac{1}{\varpi}$.

For the particular coupling function discussed in the example above, $\varpi = \omega_0 + 2k\mu_2$. Plugging in $\omega_0 = k = \mu_2 = 1$, we find that the vehicles move in circles with radius $\frac{1}{3}$ if the system is at an S_3 solution. When the system is at one of the $S_2 \times S_1$ solutions, found for

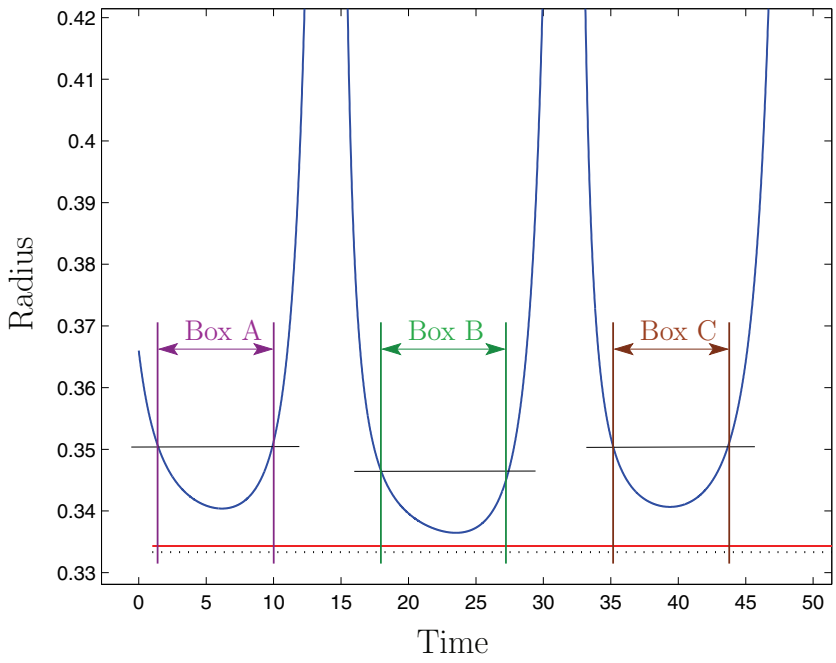


Figure 6. *An enlargement of Figure 5, showing how close the actual instantaneous radius of curvature of v_1 comes to the approximated values, and how the radius of curvature defines the location of the lettered boxes. The dotted line represents what the radius of curvature would be at the S_3 solution, and the red solid line represents the radius at the $S_2 \times S_1$ solution. The line segments show where the radius of curvature of v_1 is within 0.01 of its minimum for each box. The edges of the boxes correspond to the intersections of these line segments with the radius of curvature of v_1. The numbered boxes are then assigned as the intervening spaces between lettered boxes.*

these parameters to be at $(0.11511, 0.11511)$, $(0, 2\pi - 0.11511)$, or $(2\pi - 0.11511, 0)$, the radius of the motion of v_1 is approximately 0.334317. As one can see in Figure 6, the approximation that the system is at an $S_2 \times S_1$ fixed point is very close to the results obtained from the actual simulation.

In the numbered boxes, we can approximate the behavior of the system by noting that in Box 1, $\psi_2 \approx 0$, in Box 2, $\psi_1 \approx 0$, and in Box 3, $\psi_1 \approx \psi_2$ and both decrease from a value close to 2π to a value close to 0 at about the same rate.

Taking $\psi_2 = 0$ (which is approximately true in Box 1) in (2.4), we obtain $\dot{\psi}_2 = 0$ and

$$\dot{\psi}_1 = \dot{\theta}_1 - \dot{\theta}_2 = k[f(-\psi_1) + f(0) - 2f(\psi_1)], \tag{3.1}$$

a one-dimensional differential equation. Similarly, taking $\psi_1 = 2\pi = 0$ (which is approximately true in Box 2) in (2.4) gives the same formula as (3.1) but with $\psi_1 \to \psi_2$. Finally, taking $\psi_1 = \psi_2 \equiv \psi$ (which is approximately true in Box 3), we obtain

$$\dot{\psi}_1 = \dot{\psi}_2 = \dot{\psi} = k[2f(-\psi) - f(0) - f(\psi)], \tag{3.2}$$

which is related to (3.1) through $\psi_1 \to -\psi$.

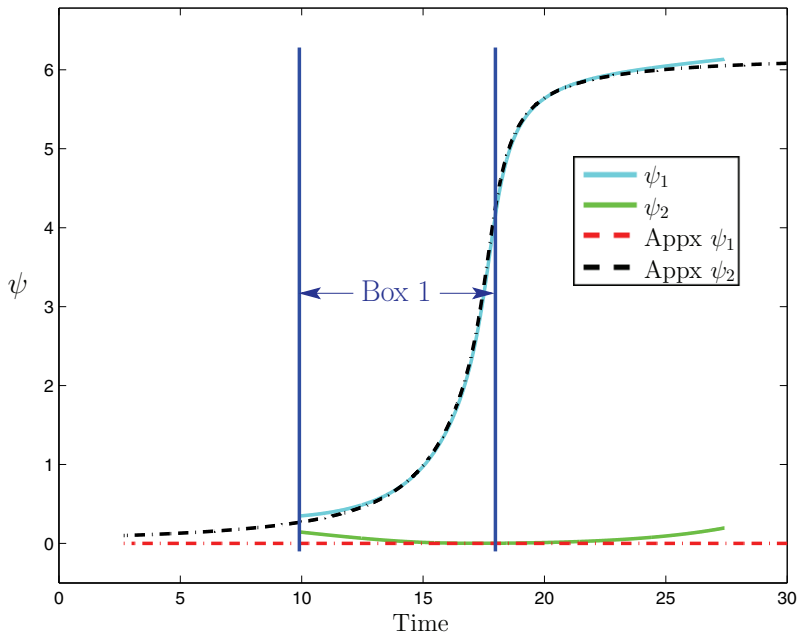


Figure 7. *Demonstration of the validity of the approximation leading to (3.1): The graphs of the approximate solutions in Box 1 and actual simulation data show that the assumptions made are reasonable.*

Numerical integration of the approximate equations very closely matches the data from simulation in all three boxes; see Figures 7 and 8 for Boxes 1 and 3, respectively. (The approximate solutions are nearly identical in Boxes 1 and 2, so only the simulation for Box 1 is shown.)

3.2.3. The Spirograph kaleidoscope. The ω_0 and k terms effectively control the curvature of the individual trajectories and the speed at which the system moves through the (ψ_1, ψ_2) plane, respectively. The shape of the vehicular trajectories, even in transients, depends only on the ratio $\frac{\omega_0}{k}$, as can be seen most easily in an equivalent form of (2.1):

$$\dot{\theta}_n = k\left(\frac{\omega_0}{k} + \sum_{m \neq n} f(\theta_m - \theta_n)\right), \qquad n = 1, 2, 3. \tag{3.3}$$

In this form, it is clear that the variable k simply scales time, while the actual dynamics depend only on the constant $\frac{\omega_0}{k}$, which can be thought of as the effective natural frequency. Since we have constrained our vehicles to have constant unit velocity, the only way that the vehicles can compensate for a larger (resp., smaller) k (with appropriately scaled ω_0), which would make the vehicles move more quickly (resp., slowly), is to produce a smaller (resp., larger), scaled, version of the exact same pattern, even in transients. This effect is demonstrated in Figure 9.

There are many possible trajectories found by varying the $\frac{\omega_0}{k}$ ratio, which have a base shape

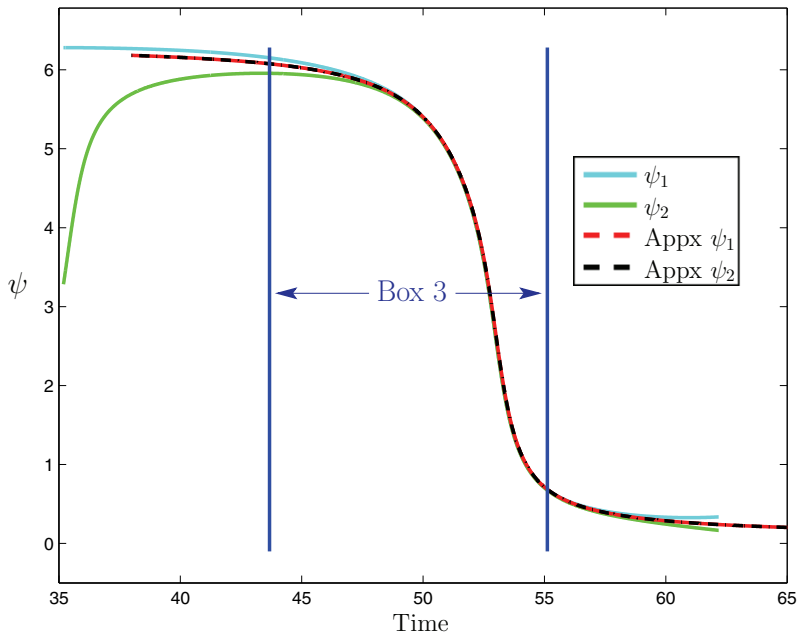


Figure 8. *Demonstration of the validity of the approximation leading to (3.2): The graphs of the approximate solutions in Box 3 and actual simulation data show that the assumptions made are reasonable.*

resembling a pattern from a Spirograph.[1] It is possible to obtain a regular overall trajectory (global) shape with any number of sides that either passes through the approximate center of the polygon, or travels exclusively along the edges. In other words, the radius of the global shape can be made to be anywhere between zero and infinity. Moreover, as one steps through the possible values of $\frac{\omega_0}{k}$, the radius runs continuously from zero through infinity and back to zero again, providing a kaleidoscope-like effect. Recognition of this trend allows one to look at a trajectory for a given set of parameters, and to be able to expect roughly what the trajectories will look like for neighboring values of $\frac{\omega_0}{k}$.

To sample over the different types of trajectories possible for $\omega_0 > 0$ and $k > 0$, we first held $\omega_0 = 1$ and varied k from 0 to 1, and then held $k = 1$ and varied ω_0 from 0 to 1. Some example trajectories are shown in Figures 10 and 11. From simulations, we have found that the global radius goes to infinity when $\frac{\omega_0}{k} \approx 0.1292 + 0.1189n$, where n is an integer.

4. The Arbiter configuration. We have also found interesting phase dynamics and vehicular trajectories for coupling topologies other than all-to-all. Here we focus on the coupling topology shown in Figure 12, which we have nicknamed the "Arbiter" configuration.

[1]A "Spirograph" is a toy invented by Denys Fisher and was first introduced to the United States in 1966 by Kenner, Inc. The name "Spirograph" is a trademark of Hasbro, Inc. The toy allows the user to create intricate designs: The user puts a pen on a point within a circle, which rotates around the inside or outside of another shape, typically also a circle. The geometric curves produced by a Spirograph are mathematically known as hypotrochoids and epitrochoids [1]. An interactive applet demonstrating what patterns are possible with a Spirograph can be found at [13].

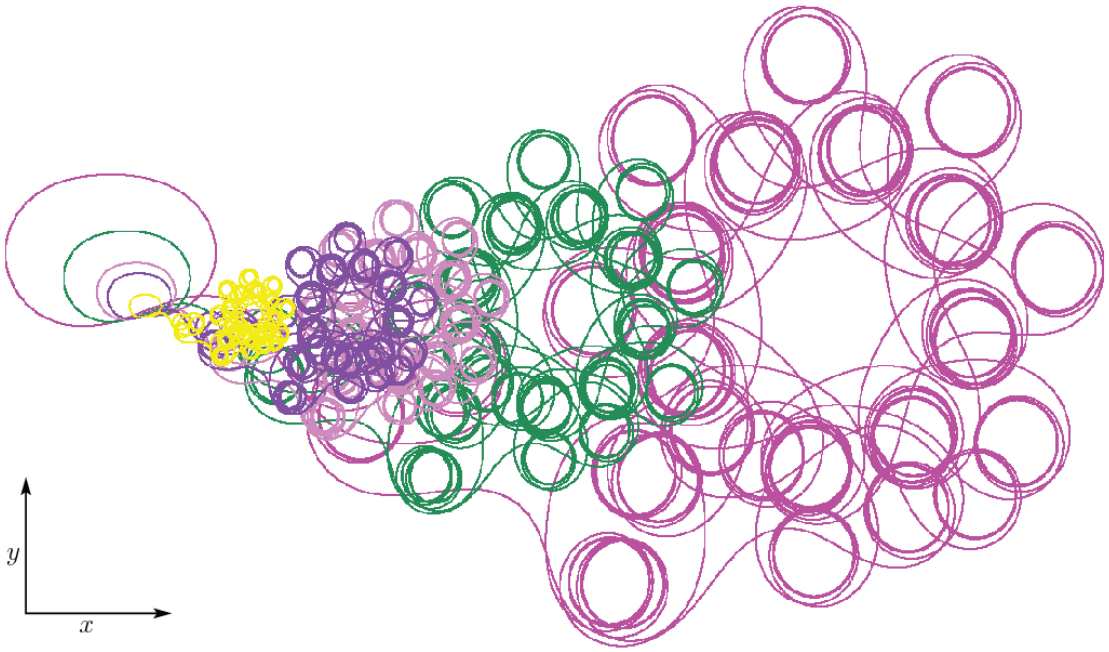


Figure 9. *Five trajectories with the same initial conditions in (x, y) and (ψ_1, ψ_2), and with the same ratio $\frac{\omega_0}{k}$, but with different values of k (and appropriately scaled ω_0).*

The equations for the Arbiter configuration for $N = 3$ are

$$\begin{aligned} \dot{\theta}_1 &= \omega_0 + k[f(\theta_2 - \theta_1) + f(\theta_3 - \theta_1)], \\ \dot{\theta}_2 &= \omega_0 + kf(\theta_1 - \theta_2), \\ \dot{\theta}_3 &= \omega_0 + kf(\theta_1 - \theta_3). \end{aligned} \tag{4.1}$$

Transforming (4.1) into the ψ coordinates as in section 2.3 gives

$$\begin{aligned} \dot{\psi}_1 &= k[f(-\psi_1) + f(-\psi_2) - f(\psi_1)], \\ \dot{\psi}_2 &= k[f(-\psi_1) + f(-\psi_2) - f(\psi_2)]. \end{aligned} \tag{4.2}$$

It is evident that the (ψ_1, ψ_2) equations are equivariant under permutation of ψ_1 and ψ_2, and that the lines $\psi_1 = 2\pi n$ and $\psi_2 = 2\pi n$, where n is an integer, are no longer invariant. The system does have an invariant line at $\psi_1 = \psi_2$. Along this line, $\psi_1 = \psi_2 \equiv \psi$, and we see that if there exists a δ^* such that $2f(-\delta^*) - f(\delta^*) = 0$, then there will be at least one fixed point on the invariant line at $(\psi_1, \psi_2) = (\delta^*, \delta^*)$. An argument for the existence of such a δ^* under quite general conditions follows.

4.1. Existence of $S_2 \times S_1$ solutions with $\psi_1 = \psi_2$. Letting

$$c_1(\delta) = 2f(-\delta), \qquad c_2(\delta) = f(\delta), \tag{4.3}$$

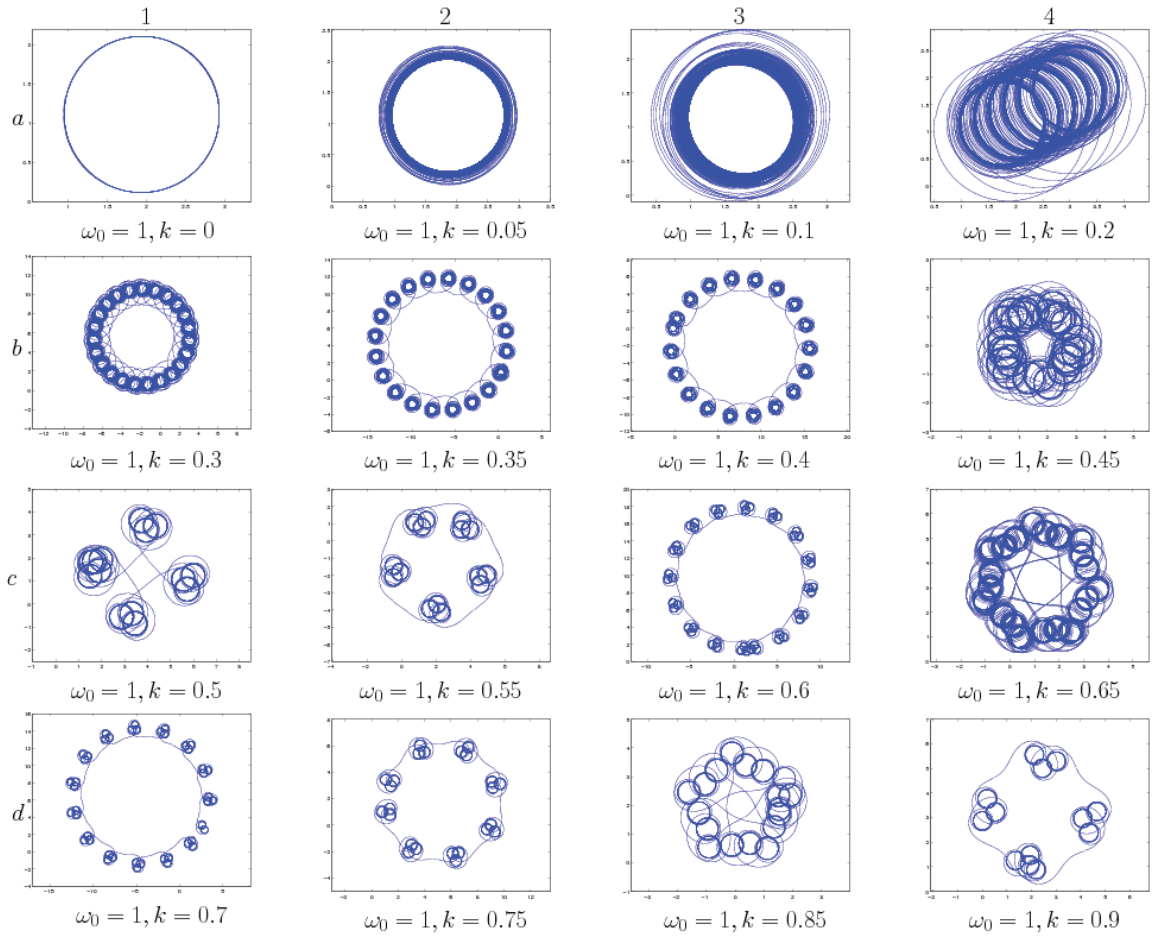


Figure 10. *A few examples of vehicular trajectories for v_1 from coupling function (2.8) with $\mu_1 = 0.1$, $\mu_2 = 1$, and $\mu_3 = -0.06$, while holding $\omega_0 = 1$ and varying k from 0 to a value close to 1.*

a valid δ^* will satisfy

$$c_1(\delta^*) = c_2(\delta^*). \tag{4.4}$$

If $|f(\delta)| \geq 0$ for all δ, it is possible that no such δ exists: for example, take $f(\delta) = 1$. Therefore, we assume that there exists a $\phi_1 \neq 0$ such that $f(\phi_1) = 0$, but $f'(\phi_1) \neq 0$. Then, by periodicity of f, there must be a $\phi_2 \neq 0$ such that $f(\phi_2) = 0$ but $f'(\phi_2) \neq 0$.

If $f(0) \neq 0$, without loss of generality, we can assume that $c_1(0) > c_2(0) > 0$. This implies that $c_1(2\pi) > c_2(2\pi) > 0$. Now,

$$\begin{aligned}
&\min[c_2(\delta)] = \min[f(\delta)] \equiv \beta,\\
&\min[c_1(\delta)] = \min[2f(-\delta)] = \min[2f(\delta)] = 2[\min f(\delta)] = 2\beta,
\end{aligned}$$

where $\beta < 0$, as shown in Figure 13. This implies that there exists a δ^{**} such that $c_1(\delta^{**}) < c_2(\delta^{**})$. Therefore, by the intermediate value theorem, there must be at least two valid values

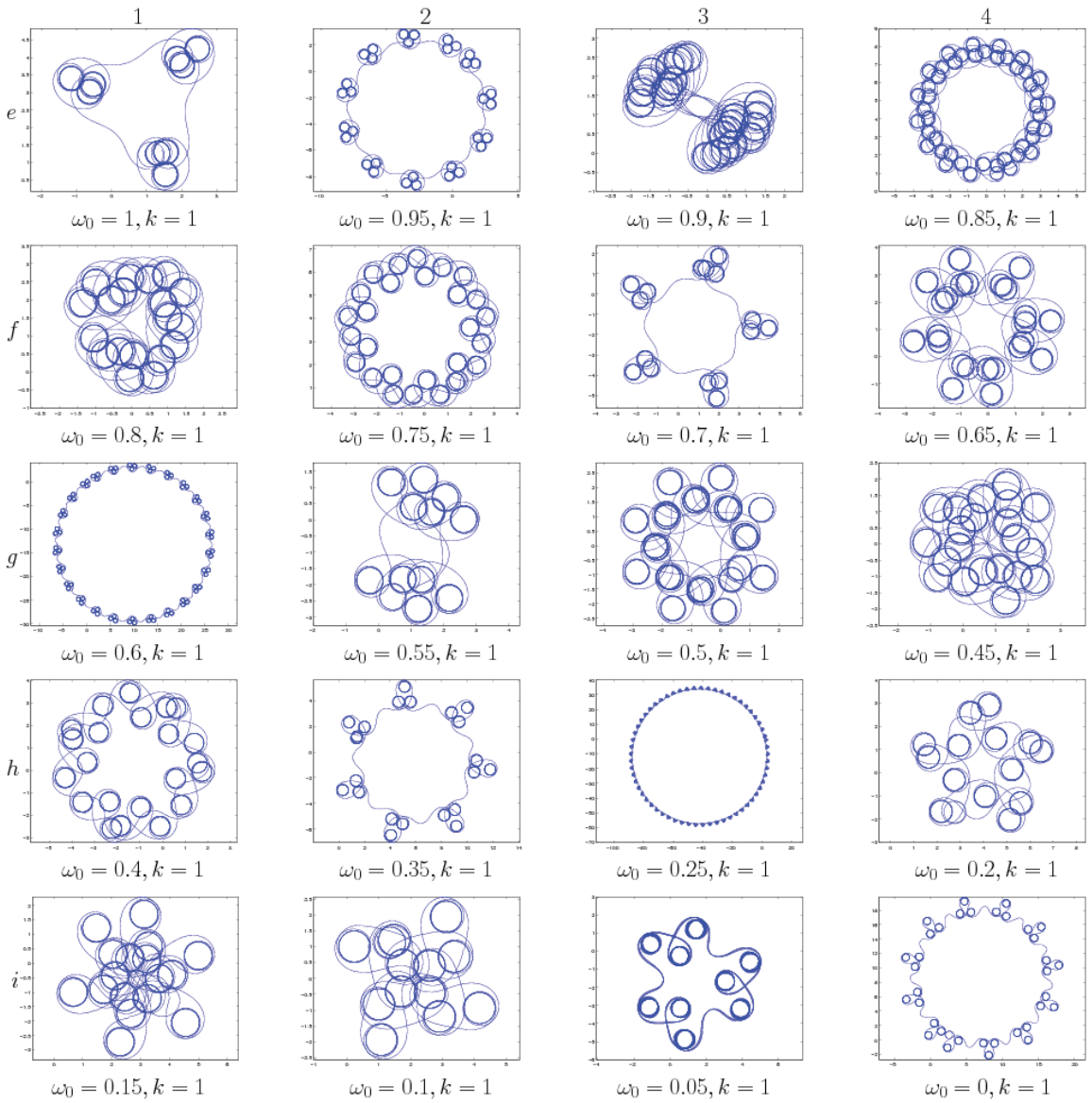


Figure 11. *A continuation of Figure 10: A few example vehicular trajectories for v_1 holding $k = 1$ and varying ω_0 from 1 to 0.*

$\delta^*_{1,2}$ such that $c_1(\delta^*_1) = c_2(\delta^*_1)$ and $c_1(\delta^*_2) = c_2(\delta^*_2)$. Furthermore, all further viable values for δ^* will occur in pairs.

A similar argument can be made to prove the existence of a $\delta^* \in (0, 2\pi)$ if $f(0) = 0$ (corresponding to the existence of an S_3-symmetric fixed point) provided that there exists a $\phi_1 \neq 0$ such that $f(\phi_1) = 0$ and $f'(\phi_1) \neq 0$.

4.2. Example. Using the example coupling function (2.8) with $\mu_1 = 0.1$, $\mu_2 = 1$, $\mu_3 = -0.06$, and $k = 1$, we find that the system has saddle points at $(\psi^*_1, \psi^*_2) = (4.29213, 4.29213)$

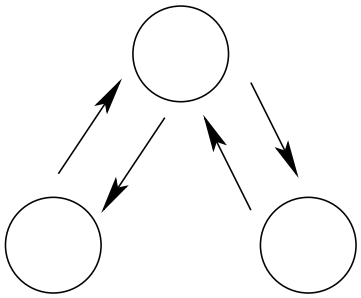


Figure 12. *The Arbiter configuration. Here the arrows indicate the coupling between agents.*

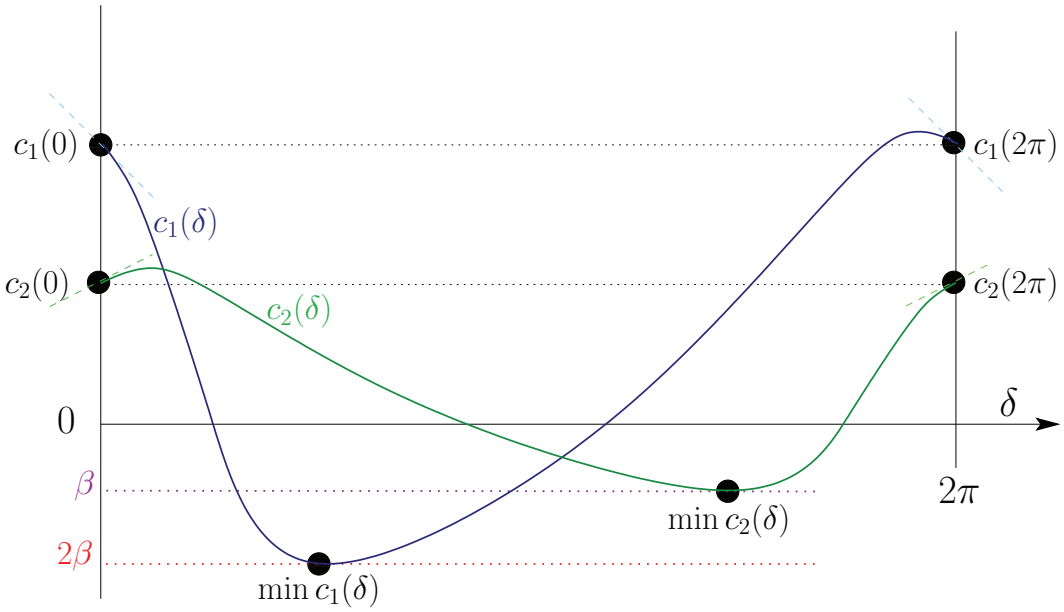


Figure 13. *Illustration of the argument that given the constraints mentioned in the text, there must be at least two possible values for δ^*. Without loss of generality, we can set $c_1(0) > c_2(0) > 0$, which gives $c_1(2\pi) > c_2(2\pi) > 0$ by periodicity. However, by noting that $\min[f(-\delta)] = \min[f(\delta)]$, it is clear that $\min[c_1(\delta)] = 2\min[c_2(\delta)]$. Therefore, $c_1(\delta)$ and $c_2(\delta)$ must cross in at least two points. The points where the two functions cross are viable values for δ^*, and this proves the existence of the $S_2 \times S_1$ solutions.*

and $(1.35235, 1.35235)$, which are guaranteed to exist from the above argument, and spiral sinks at $(\psi_1^*, \psi_2^*) = (4.8432, 1.63105)$ and $(1.63105, 4.8432)$. For these parameters, there are also two symmetry-related stable periodic orbits in the (ψ_1, ψ_2) coordinates; see Figure 14. The vehicular trajectories corresponding to motion along several cycles of one of the stable periodic orbits is shown in Figure 15, which is reminiscent of the trajectories found in section 2.3. For the same reasons as in section 3.2.3, one can also produce a variety of trajectories by varying the values of ω_0 and k, as shown in Figure 16.

We find that at least one stable periodic orbit exists in the (ψ_1, ψ_2) system between the saddle-node bifurcations of limit cycles at $\mu_1 = \pm 0.115681$. (For some parameters, there are two symmetry-related periodic orbits.) Within this range, there are several global bifurcations

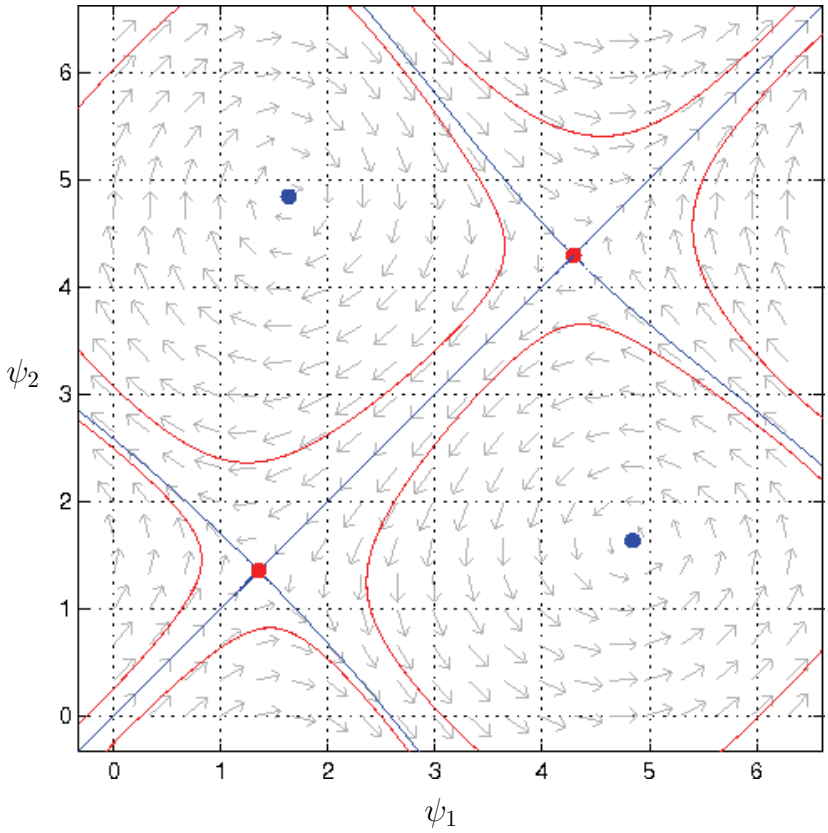


Figure 14. *The (ψ_1, ψ_2) plane for the Arbiter coupling topology with $N = 3$, and coupling function (2.8) with $\mu_1 = 0.1$, $\mu_2 = 1$, $\mu_3 = -0.06$, $\omega_0 = 1$, $k = 1$. The existence of stable periodic orbits suggests that this system may provide interesting patterns of vehicular motion.*

involving the $S_2 \times S_1$ fixed points on the line $\psi_1 = \psi_2$. The details of these bifurcations are outside of the scope of our present study, but we do note that it would be possible to interpret the vehicular motion in terms of visits near and between the fixed points, as was done in section 3.

5. Conclusion. In this paper, we considered a model for vehicle motion coordination developed by Leonard, Paley, and Sepulchre which uses coupled oscillator steering control. We showed that novel trajectories are possible using only the phase controller when coupling functions more general than sinusoidal are considered. Such trajectories are associated with periodic orbits in the steering control subsystem, and the proximity of these periodic orbits to heteroclinic bifurcations allowed a detailed characterization of the properties of the vehicular trajectories.

Similar trajectories are expected to be possible for such systems with $N > 3$ vehicles. An attempt to understand the details of such trajectories would likely benefit from previous studies of phase-locked solutions for coupled oscillator systems with phase-difference coupling [5, 6, 33] and heteroclinic orbits for such systems [3, 16].

The trajectories described in this paper may have applications in sensor area covering

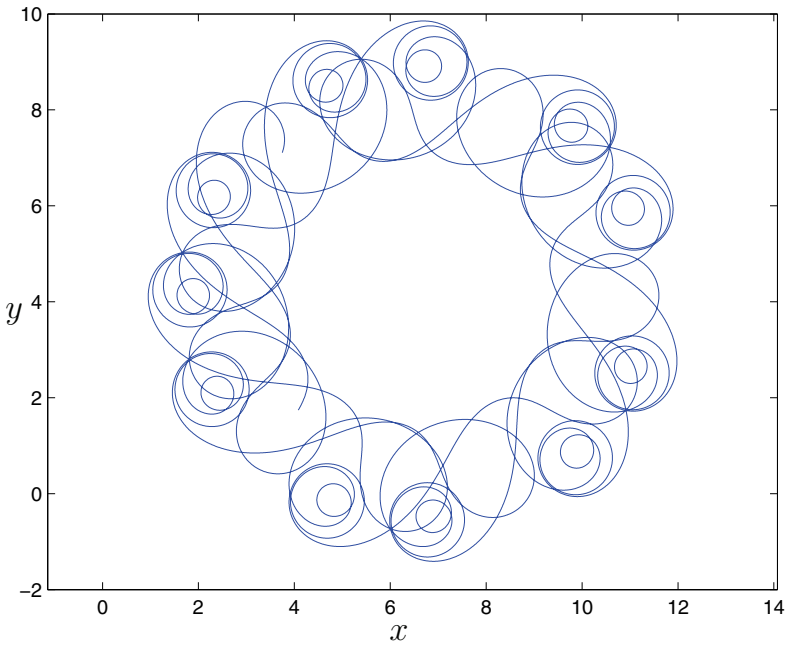


Figure 15. *Motion of v_1 using the Arbiter coupling topology with $N = 3$ corresponding to the motion of the system along a stable periodic orbit in Figure 14.*

problems in which one is particularly interested in certain regions of an annulus in the plane, with the option of either passing through the center or moving along the circumference of the area to be covered. For example, the trajectory shown in 1c of Figure 10 may be useful for the case where one wants agents to carefully patrol four evenly distributed areas as well as check the area in the center of those four areas periodically. If one desires to check sections of a circular area but is not interested in the area in the center of the sections, a trajectory such as 2e of Figure 11 may be appropriate. Should the areas inside the circular area be of higher interest than the perimeter, then a trajectory such as 2n of Figure 16 may be of interest. If one desires to patrol an annulus evenly in sections, a trajectory similar to 4e or 2f of Figure 11 may be useful. Most parameter values provide trajectories where an almost regular polygon-like global trajectory drifts around the center of pattern; thus, over time, the trajectories eventually cover an annulus. An example of this can be found in plot 3e of Figure 11—this is a "polygon" with slightly more than 2 sides, that is drifting around, and will eventually fill out an annulus. These patterns may be useful for applications where it is desirable for a robot patrolling an annulus-shaped space to not only periodically thoroughly investigate subregions of the space, but to also be relatively difficult to predict.

Despite the fact that the system is very stable in the reduced phase space, the trajectories described here are quite sensitive to variations in the parameters of the coupling equations. Should these trajectories prove to be potentially useful for a particular area coverage problem, it may be worthwhile to investigate the use of spacing control, and to make the global behavior robust to uncertainty and perturbations in the parameters.

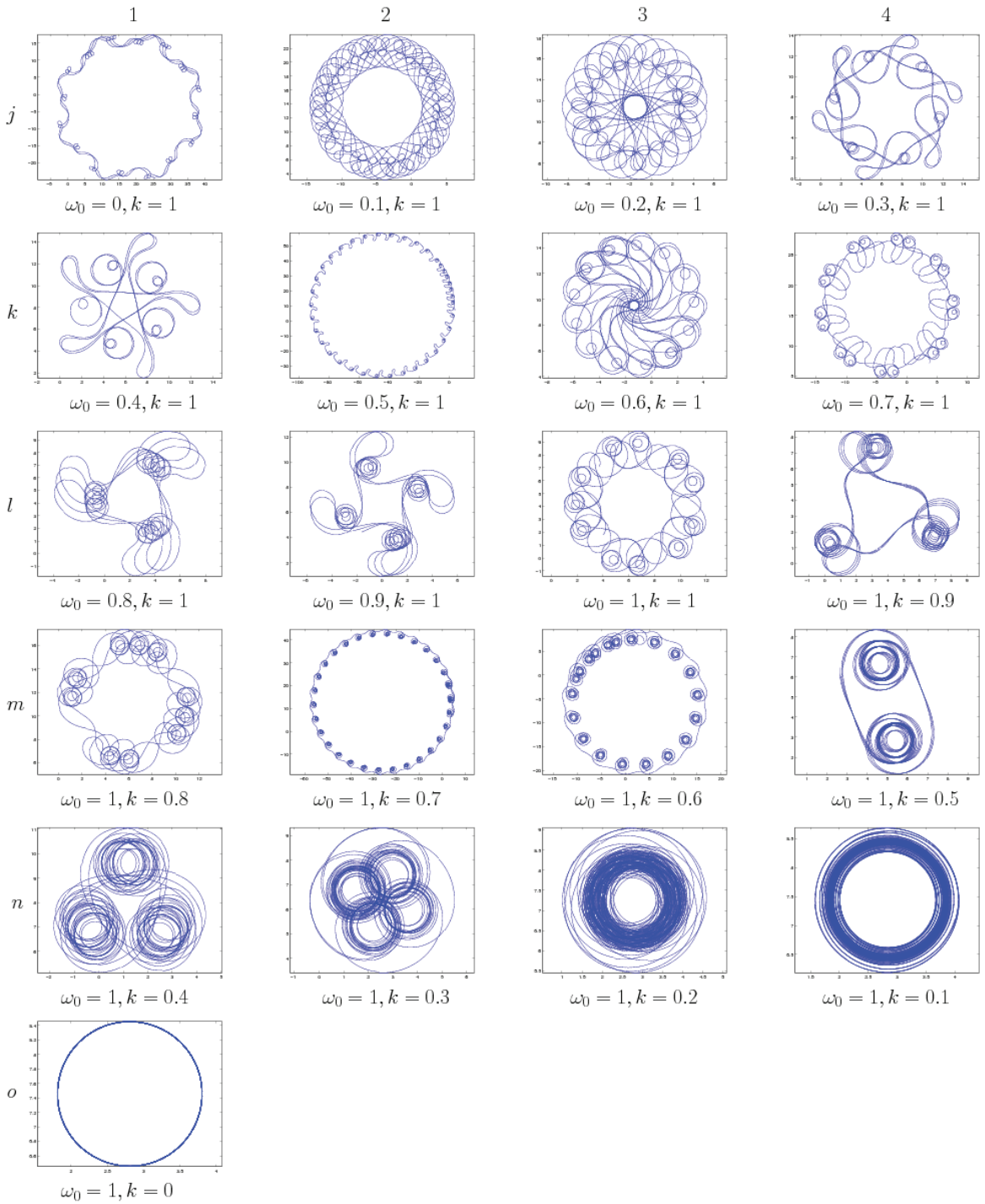


Figure 16. *Various vehicular trajectories generated using the Arbiter coupling topology and the example coupling function* (2.8) *with* $\mu_1 = 0.1$, $\mu_2 = 1$, $\mu_3 = -0.06$, *while varying the values of* ω_0 *and* k, *as was done in Figures* 10 *and* 11.

REFERENCES

[1] Wikipedia, *Spirograph*, http://en.wikipedia.org/wiki/Spirograph.

[2] P. Ashwin, O. Burlko, P. Maistrenko, and O. Popvych, *Extreme sensitivity to detuning for globally coupled phase oscillators*, Phys. Rev. Lett., 96 (2006), 054102.

[3] P. Ashwin, O. Burylko, and Y. Maistrenko, *Bifurcation to heteroclinic cycles and sensitivity in three and four coupled phase oscillators*, Phys. D, 237 (2007), pp. 454–466.

[4] P. Ashwin, G. P. King, and J. W. Swift, *Three identical oscillators with symmetric coupling*, Nonlinearity, 3 (1990), pp. 585–601.

[5] P. Ashwin and J. W. Swift, *The dynamics of n weakly coupled identical oscillators*, J. Nonlinear Sci., 2 (1992), pp. 69–108.

[6] E. Brown, P. Holmes, and J. Moehlis, *Globally coupled oscillator networks*, in Perspectives and Problems in Nonlinear Science: A Celebratory Volume in Honor of Larry Sirovich, E. Kaplan, J. Marsden, and K. R. Sreenivasan, eds., Springer-Verlag, New York, 2003, pp. 183–215.

[7] S. Camazine, J. L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau, *Self-Organization in Biological Systems*, Princeton University Press, Princeton, NJ, 2003.

[8] L. Chaimowicz, V. Kumar, and F. M. Campos, *A paradigm for dynamic coordination of multiple robots*, Autonomous Robots, 17 (2004), pp. 7–21.

[9] J. Cortès, S. Martìnes, T. Karatas, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Trans. Robot. Automat., 20 (2004), pp. 243–255.

[10] J. P. Desai, J. P. Ostrowski, and V. Kumar, *Modeling and control of formations of nonholonomic mobile robots*, IEEE Trans. Robot. Automat., 17 (2001), pp. 905–908.

[11] L. E. Dubins, *On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents*, Amer. J. Math., 79 (1957), pp. 497–516.

[12] G. B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.

[13] A. Garg, *Spirograph: Spirographs*, http://wordsmith.org/~anu/java/spirograph.html.

[14] V. Gazi and K. M. Passino, *Stability analysis of swarms*, IEEE Trans. Automat. Control, 48 (2002), pp. 692–696.

[15] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory, Vol.* 2, Springer-Verlag, New York, 1988.

[16] D. E. Hansel, G. Mato, and C. Meunier, *Clustering and slow switching in globally coupled phase oscillators*, Phys. Rev. E (3), 48 (1993), pp. 3470–3477.

[17] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 692–696.

[18] E. Justh and P. Krishnaprasad, *Equilibria and steering laws for planar formations*, Systems Control Lett., 52 (2004), pp. 25–38.

[19] E. W. Justh and P. S. Krishnaprasad, *A Simple Control Law for UAV Formation Flying*, Technical report, Institute for Systems Research, University of Maryland, College Park, MD, 2002; available online from http://www.lib.umd.edu/drum/handle/1903/6274?mode=simple.

[20] A. Kolpas and J. Moehlis, *Optimal switching between coexisting stable collective motion states*, Appl. Math. Lett., to appear.

[21] N. Leonard and E. Fiorelli, *Virtual leaders, artificial potentials and coordinated control of groups*, in Proceedings of the 40th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2001, pp. 2968–2973.

[22] P. B. S. Lissaman and C. A. Shollenberger, *Formation flight of birds*, Science, 168 (1970), pp. 1003–1005.

[23] B. Nabet, N. Leonard, I. Couzin, and S. Levin, *Leadership in animal group motion: A bifurcation analysis*, in Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems, Kyoto, Japan, 2006, pp. 1–14.

[24] R. Olfati-Saber, *Flocking for multi-agent dynamic systems: Algorithms and theory*, IEEE Trans. Automat. Control, 51 (2006), pp. 401–420.

[25] D. Paley, N. Leonard, and R. Sepulchre, *Collective motion: Bistability and trajectory tracking*, in Proceedings of the 43rd IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2004, pp. 1932–1937.

[26] D. Paley, N. Leonard, and R. Sepulchre, *Oscillator models and collective motion: Splay state stabilization of self-propelled particles*, in Proceedings of the 44th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 2005, pp. 3935–3940.

[27] D. A. Paley, N. E. Leonard, R. Sepulchre, D. Grunbaum, and J. Parrish, *Oscillator models and collective motion*, IEEE Control Systems Magazine, 27 (2007), pp. 89–105.

[28] B. L. Partridge, *The structure and function of fish schools*, Sci. Amer., 245 (1982), pp. 90–99.

[29] R. Sepulchre, D. Paley, and N. Leonard, *Collective motion and oscillator synchronization*, in Proceedings of the 2003 Block Island Workshop on Cooperative Control, V. Kumar, N. Leonard, and A. Morse, eds., Springer-Verlag, New York, 2003, pp. 1–17.

[30] R. Sepulchre, D. Paley, and N. Leonard, *Graph Laplacian and Lyapunov design of collective planar motions*, in Proceedings of the IEEE Conference on Nonlinear Theory and Its Applications, Bruges, Belgium, 2005, pp. 217–232.

[31] R. Sepulchre, D. Paley, and N. Leonard, *Stabilization of planar collective motion part* I: *All-to-all communication*, IEEE Trans. Automat. Control, 52 (2007), pp. 811–824.

[32] A. R. E. Sinclair and M. Norton-Griffiths, *Serengeti: Dynamics of an Ecosystem*, University of Chicago, Chicago, IL, 1979.

[33] S. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators*, Phys. D, 143 (2000), pp. 1–20.

[34] B. P. Uvarov, *Grasshoppers and Locusts*, Imperial Bureau of Entomology, London, 1928.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1213–1246
© 2008 Society for Industrial and Applied Mathematics

A Hamiltonian Analogue of the Meandering Transition*

Claudia Wulff†

Abstract. In this paper a Hamiltonian analogue of the well-known meandering transition from rotating waves to modulated rotating and modulated traveling waves in systems with the Euclidean symmetry of the plane is presented. In non-Hamiltonian systems, for example, in spiral wave dynamics, this transition is a Hopf bifurcation in a corotating frame, as external parameters are varied, and modulated traveling waves occur only at certain resonances. In Hamiltonian systems, for example, in systems of point vortices in the plane, the conserved quantities of the system, angular and linear momentum, are natural bifurcation parameters. Depending on the symmetry properties of the momentum map, either modulated traveling waves do not occur, or, in contrast to the dissipative case, modulated traveling waves are the typical scenario near rotating waves, as momentum is varied. Systems with the symmetry group of a sphere and with the Euclidean symmetry group of three-dimensional space are also treated.

Key words. meandering transition, resonance drift, symmetric Hamiltonian systems

AMS subject classifications. 37J15, 37J20, 53D20, 70H33

DOI. 10.1137/070704940

1. Introduction.

The meandering transition in spiral wave dynamics is a transition from rigidly rotating to meandering and drifting spiral waves. In symmetry terms, it is a bifurcation from rotating waves to modulated rotating and modulated traveling waves in systems with SE(2)-symmetry. Here SE(2) is the special Euclidean group of motions of the plane. Rotating waves are solutions which become stationary in a corotating frame and are examples of relative equilibria. Modulated rotating and modulated traveling waves are solutions which become periodic in a corotating/comoving frame and are examples of relative periodic orbits (RPOs). In non-Hamiltonian systems, the meandering bifurcation corresponds, in a rotating frame, to a Hopf bifurcation induced by changing an external parameter. Typically the bifurcating relative periodic orbits are modulated rotating waves, and modulated traveling waves occur only at certain resonances. See, for example, [3, 8, 10, 27, 30, 31] and the references therein.

In this paper the first ever analysis of the Hamiltonian analogue of this meandering transition is presented. Examples of Hamiltonian systems where such a transition occurs are rotating point vortices on the plane [1, 2, 21, 25, 29] or rotating rigid bodies in ideal fluids [15]. In a Hamiltonian system it is natural to study the persistence and bifurcation of the rotating wave to nearby momentum levels since the momentum map is a conserved quantity and hence an internal parameter of the system.

The differential equations near Hamiltonian relative equilibria in symmetry-adapted local

*Received by the editors October 9, 2007; accepted for publication (in revised form) by M. Golubitsky June 3, 2008; published electronically October 24, 2008. This work was partially supported by a grant from the Nuffield organization, by EPSRC grant EP/D063906/1, and by a Leverhulme Research Fellowship.
http://www.siam.org/journals/siads/7-4/70494.html
†Department of Mathematics, University of Surrey, Guildford GU2 7XH, UK (c.wulff@surrey.ac.uk).

coordinates from [26] are used to study the transition from rotating waves to modulated rotating and modulated traveling waves on nearby momentum levels in Hamiltonian systems with SE(2)-symmetry. Thereby a Hamiltonian analogue of the meandering transition of spiral waves is obtained.

It is shown that, depending on the symmetry properties of the momentum map, either modulated traveling waves are typical near rotating waves, as momentum is varied (cf. sections 4.2 and 4.3), or modulated traveling waves do not occur; see section 4.4 and in particular Proposition 4.10. As far as I am aware, for the first time, rotating waves and transitions to relative periodic orbits are continued in the cocycle parameter which determines the symmetry properties of the momentum map. These results hold under conditions which are generically satisfied.

The transition from rotating waves to modulated traveling waves occurring in the meandering transition is an example of resonance drift, as analyzed in [31]; see also [4] and [6]. Resonance drift occurs if there is a discontinuity of the average drift velocities of the bifurcating relative periodic orbits at the relative equilibrium. In the case of the meandering transition it is a discontinuous jump between a rotational and a translational velocity. This phenomenon is also discussed in systems with spherical symmetry SO(3) and in systems with the Euclidean symmetry SE(3) of motions in three-dimensional space; see sections 5.1 and 5.2.

The meandering transition is a transition from relative equilibria to relative periodic orbits. In non-Hamiltonian systems it is a Hopf bifurcation of the symmetry reduced dynamics. The Hamiltonian analogue of a Hopf bifurcation is a Lyapunov center bifurcation. In this paper Lyapunov center bifurcations for the reduced Hamiltonian system on the symplectic slice are proved to obtain families of RPOs nearby elliptic relative equilibria; see Proposition 4.6, Theorems 4.11 (a), 5.1 (b), and 5.2 (c), and Propositions 5.3 (b) and 5.6.

The technically most complicated parts of the paper are the results on bifurcation from relative equilibria to RPOs which lie outside the symplectic leaf of the original equilibrium of the reduced dynamics; see Theorems 4.3, 5.2 (b), and 5.5. Here Lyapunov center type theorems are proved for the symmetry reduced system which is a Poisson system and not a Hamiltonian system. It is shown that in this case resonance drift occurs.

Related results in the literature are the following: Persistence results for generic Hamiltonian relative equilibria and relative periodic orbits of noncompact group actions, extending earlier results for compact symmetry groups, can be found in [32, 33]. See also Ortega and Ratiu [23] and Montaldi and Tokieda [20] and references therein for results on bifurcations of Hamiltonian relative equilibria.

Relative Lyapunov center bifurcations from Hamiltonian relative equilibria with isotropy to RPOs, which lie on nearby energy level sets, have been obtained by Ginzburg and Lerman [9] (see also references therein). Ortega [22] studies persistence of the bifurcating RPOs to nearby energy level sets and to those nearby momentum values which correspond to the isotropy subgroup of the relative equilibrium. Instead, in this article, the group is assumed to act freely, and the main focus is the bifurcation of relative equilibria to RPOs on all nearby momentum level sets.

The paper is organized as follows: In section 2 the meandering transition for dissipative systems is reviewed. In section 3 symmetric Hamiltonian systems are introduced and the equations near relative equilibria from [26] are reviewed. In section 4 a Hamiltonian analogue

of the meandering transition is presented using the equations near Hamiltonian relative equilibria from section 3. First Euclidean symmetric Hamiltonian systems with an equivariant momentum map for the standard coadjoint action are studied. Then systems with Euclidean symmetry for which the momentum map has a cocycle are considered. Finally, in section 5, the Hamiltonian analogue of the meandering transition is discussed for systems with spherical symmetry and for systems with the Euclidean symmetry group of three-dimensional space.

2. Meandering transition for dissipative systems. In this section the notions of relative equilibria and relative periodic orbits of general symmetric differential equations are defined. Suitable symmetry-adapted coordinates near relative equilibria are introduced, and the differential equations are given in these coordinates. Then the results are applied to dissipative systems with the Euclidean symmetry of the plane, and the meandering transition for dissipative systems is reviewed. Note that in this paper the terms "dissipative systems," "non-Hamiltonian systems," and "general systems" are used interchangeably. Most of the material of this section is basically contained in [8, 10, 31]. Only Remark 2.2 (c) is a new result.

2.1. Relative equilibria and relative periodic orbits of general systems. Let us consider an ordinary differential equation on a finite-dimensional manifold $\mathcal{M}$

$$\dot{x}(t) = f(x(t)) \tag{2.1}$$

with flow $\Phi_t(x_0) = x(t; x_0)$, $x(0) = x_0$. Let a finite-dimensional Lie group Γ act properly and smoothly on $\mathcal{M}$. For simplicity it is assumed that the Γ-action is *free*, that is,

$$\Gamma_x = \{\gamma \in \Gamma,\ \ \gamma x = x\} = \{\mathrm{id}\}$$

for all $x \in \mathcal{M}$. The vectorfield (2.1) is taken to be Γ-*equivariant*, i.e.,

$$\gamma f(x) = f(\gamma x) \quad \text{for all } \gamma \in \Gamma.$$

A solution $x(t)$ with initial condition $x(0) = x_0$ lies on a *relative equilibrium* Γx_0 whenever the group orbit Γx_0 is invariant under the flow of (2.1), i.e., if $x(t; x_0) \in \Gamma x_0$ for all t. This means that

$$f(x_0) = \xi_0 x_0 := \left(\frac{\mathrm{d}}{\mathrm{d}s} \exp(s\xi_0).x_0\right)\Big|_{s=0}$$

for some $\xi_0 \in \mathbf{g}$. Here $\mathbf{g} = \mathcal{T}_{\mathrm{id}}\Gamma$ is the Lie algebra of Γ. The element ξ_0 is called the *drift velocity* of the relative equilibrium at x_0. Note that the trajectory through x_0 becomes an equilibium in a frame moving with velocity ξ_0. If ξ_0 is an infinitesimal rotation, then the relative equilibrium is called a *rotating wave* (RW). Note that at the point γx_0 of the relative equilibrium Γx_0 the drift velocity is determined by the equation

$$f(\gamma x_0) = \gamma f(x_0) = \gamma \xi_0 x_0 = (\mathrm{Ad}_\gamma \xi_0) x_0$$

and is therefore given by $\mathrm{Ad}_\gamma \xi_0$. Here $\mathrm{Ad}_\gamma : \mathbf{g} \to \mathbf{g}$ and

$$\mathrm{Ad}_\gamma \eta = \gamma \eta \gamma^{-1}, \quad \mathrm{ad}_\xi \eta = \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{Ad}_{\exp(t\xi)} \eta|_{t=0} = [\xi, \eta], \quad \gamma \in \Gamma,\ \eta, \xi \in \mathbf{g}, \tag{2.2}$$

are the *adjoint action* of Γ and $\mathbf{g}$ and the *infinitesimal adjoint action* of $\mathbf{g}$ on $\mathbf{g}$.

An example of such a finite-dimensional manifold $\mathcal{M}$ with SE(2)-equivariant vectorfield (2.1) on it is the center manifold near a rotating spiral $\mathrm{SE}(2)x_0$ in a reaction-diffusion system; see, e.g., [27]. Here $\mathrm{SE}(2) = \mathrm{SO}(2) \ltimes \mathbb{R}^2$ is the special Euclidean symmetry of rotations and translations in the plane with group multiplication defined in (2.5) below.

By the slice theorem of Palais [24] sufficiently small neighborhoods $\mathcal{U}$ of the group orbit Γx_0 have the bundle structure $\mathcal{U} = \Gamma \times \mathcal{N}$. Here $\mathcal{N} \subseteq \mathcal{T}_{x_0}\mathcal{M}$ is a local section, also called *slice*, transversal to Γx_0 at x_0; see Figure 1.

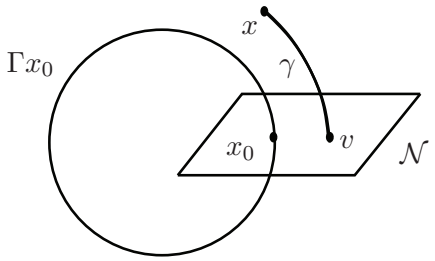


Figure 1. *Palais coordinates near* Γx_0.

To analyze the dynamics near, and bifurcations from, relative equilibria, it has proved very useful to model the flow in a Γ-invariant neighborhood $\mathcal{U}$ of the relative equilibrium by differential equations on the space $\Gamma \times \mathcal{N}$:

$$\dot{\gamma} = \gamma f_\Gamma(v), \quad \dot{v} = f_\mathcal{N}(v), \tag{2.3}$$

where $f_\Gamma : \mathcal{N} \to \mathbf{g}$ and $f_\mathcal{N} : \mathcal{N} \to \mathcal{N}$. Any $x \in \mathcal{U}$ takes the form $x \simeq (\gamma, v) \in \Gamma \times \mathcal{N}$, and the point x_0 corresponds to $x_0 \simeq (\mathrm{id}, 0)$. Then $f_\mathcal{N}(0) = 0$; i.e., the relative equilibrium Γx_0 of (2.1) becomes an equilibrium of the $\dot{v}$-equation. Moreover, $f_\Gamma(0) = \xi_0$. Note that the equations (2.3) have skew-product form: the $\dot{v}$-equation, which is called the *slice equation*, does not depend on the group variable γ. It describes the symmetry-reduced dynamics, whereas the $\dot{\gamma}$-equation describes the drift dynamics on the group Γ. These results are due to Krupa [14] for compact Lie groups and to Fiedler et al. [8] for noncompact Lie groups. For later use, the linearization $L_0 = \mathrm{D}f(x_0) - \xi_0$ of the relative equilibrium Γx_0 at x_0 in the frame moving with the velocity ξ_0 in symmetry-adapted coordinates is

$$L_0 = \begin{pmatrix} \mathrm{ad}_{\xi_0} & \mathrm{D}_v f_\Gamma(0) \\ 0 & \mathrm{D}_v f_\mathcal{N}(0) \end{pmatrix}. \tag{2.4}$$

The point $x_0 \in \mathcal{M}$ lies on a relative periodic orbit $\mathcal{P}$ of (2.1) if $x(t; x_0) = \Phi_t(x_0)$ is periodic in the space of group orbits $\mathcal{M}/\Gamma$. This means that there exist $T_0 > 0$ and $\gamma_0 \in \Gamma$ such that $\Phi_{T_0}(x_0) = \gamma_0 x_0$; see Figure 2. The infimum of the numbers T_0 with this property is the *relative period* of the RPO. The corresponding group element γ_0 is called the *drift symmetry* of the RPO with respect to x_0; cf. [33, 34]. The relative periodic orbit itself is defined to be the submanifold of $\mathcal{M}$ given by

$$\mathcal{P} \;=\; \{\gamma \Phi_t(x_0) \mid \gamma \in \Gamma,\, t \in \mathbb{R}\}.$$

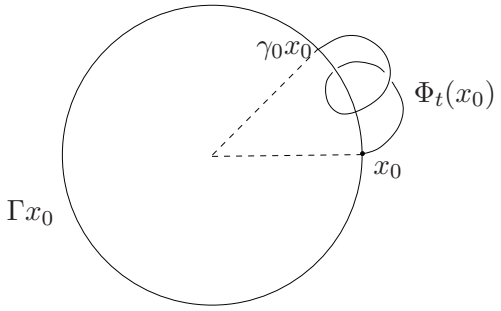


Figure 2. *A relative periodic orbit.*

If γ_0 is a translation, the RPO is called a *modulated traveling wave* (MTW); if γ_0 is a nonvanishing translation, it is a proper modulated traveling wave. If γ_0 is a (nonvanishing) rotation, the RPO is called a (proper) *modulated rotating wave* (MRW); see Figure 3. Any $\xi_0 \in \mathbf{g}$ such that $\gamma_0 = \exp(T_0\xi_0)$ is called an *average drift velocity* of the RPO at x_0. Note that the trajectory through x_0 becomes T_0-periodic in a frame moving with velocity ξ_0.

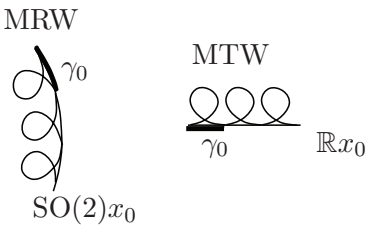


Figure 3. *Drift symmetries of modulated rotating and modulated traveling waves.*

2.2. The meandering transition for dissipative systems. Let Γ be the Euclidean symmetry of the plane consisting of rotations and translations

$$\Gamma = \mathrm{SE}(2) = \mathrm{SO}(2) \ltimes \mathbb{R}^2,$$

where the semidirect product is defined as

$$(\phi_1, a_1)(\phi_2, a_2) = (\phi_1 + \phi_2, a_1 + R_{\phi_1} a_2), \qquad \phi_i \in \mathrm{SO}(2),\ a_i \in \mathbb{R}^2,\ i = 1, 2. \tag{2.5}$$

Here R_ϕ is a rotation by ϕ in $\mathbb{R}^2$. Let us assume that the relative equilibrium $\mathrm{SE}(2)x_0$ is a rotating wave with rotation frequency ω_0^{rot}. Then the $\dot{\gamma}$-equation in (2.3), which models the drift dynamics near the rotating wave, takes the following form:

$$\dot{\phi} = f_\phi(v), \quad \dot{a} = R_\phi f_a(v). \tag{2.6}$$

Moreover, $f_\phi(0) = \omega_0^{\mathrm{rot}}$ is the rotation frequency of the rotating wave and $f_a(0) = 0$. As in the general case, the rotating wave $\mathrm{SE}(2)x_0$ becomes an equilibrium of the slice equation:

$f_{\mathcal{N}}(0) = 0$. These equations were first formulated by Barkley [3] and then derived by Fiedler et al. [8] and Golubitsky, LeBlanc, and Melbourne [10].

Let us now assume that both $f_{\mathcal{N}}(\cdot, \mu)$ and $f_\Gamma(\cdot, \mu) = (f_\phi(\cdot, \mu), f_a(\cdot, \mu))$ depend on an external parameter $\mu \in \mathbb{R}$. In a meandering transition the symmetry-reduced system undergoes a Hopf bifurcation. Suppose that this bifurcation occurs for $\mu = 0$; let $\pm i\omega_0^{\mathrm{Hopf}}$ be the Hopf eigenvalues of $\mathrm{D}_v f_{\mathcal{N}}(0, 0)$. Assume that $\pm i\omega_0^{\mathrm{Hopf}}$ are simple eigenvalues and that $\mathrm{D}f_{\mathcal{N}}(0, 0)$ has no other eigenvalues in $i\omega_0^{\mathrm{Hopf}}\mathbb{Z}$. Let $v_{\mathrm{RW}}(\mu) \approx 0$ be the equilibrium of $f_{\mathcal{N}}(\cdot, \mu)$, $\mu \approx 0$, such that $v_{\mathrm{RW}}(\mu)$ is smooth in μ and $v_{\mathrm{RW}}(0) = 0$. Then $x_{\mathrm{RW}}(\mu) \simeq (\mathrm{id}, v_{\mathrm{RW}}(\mu))$ lies on a rotating wave of (2.1). Let $\lambda(\mu)$ be the eigenvalue of $\mathrm{D}_v f_{\mathcal{N}}(v_{\mathrm{RW}}(\mu), \mu)$ such that $\lambda(\mu)$ is smooth in μ and $\lambda(0) = i\omega_0^{\mathrm{Hopf}}$. Assume that the usual transversality condition

$$\left(\mathrm{Re}\,\frac{\partial}{\partial \mu}\lambda(\mu)\right)\Big|_{\mu=0} \neq 0 \tag{2.7}$$

for a Hopf bifurcation is satisfied. Then there is a smooth path $v(s)$, $s \geq 0$, of points on periodic solutions of the $\dot{v}$-equation with period $T(s) \approx T_0^{\mathrm{Hopf}} = 2\pi/\omega_0^{\mathrm{Hopf}}$ and parameter $\mu(s)$ such that $v(0) = 0$, $T(0) = T_0^{\mathrm{Hopf}}$, $\mu(0) = 0$.

The periodic orbit through $v(s)$ of the slice equation corresponds to a relative periodic orbit $\mathcal{P}(s)$ through $x(s) \simeq (\mathrm{id}, v(s))$ of the original ODE (2.1) with drift symmetry $\gamma(s) = (\phi(s), a(s))$. Here $\phi(s)$ and $a(s)$ are obtained by integrating (2.6) from 0 to $T(s)$. There are two cases:

(a) If $\phi(s) \neq 0 \bmod 2\pi$, then $x(s)$ lies on a modulated rotating wave, and this is the typical case.

(b) If $\phi(s) = 0 \bmod 2\pi$, then $x(s)$ lies on a modulated traveling wave.

Note that

$$\phi(s) \approx \omega_0^{\mathrm{rot}} T_0^{\mathrm{Hopf}} = \frac{\omega_0^{\mathrm{rot}}}{\omega_0^{\mathrm{Hopf}}} 2\pi.$$

Hence case (b) occurs if $\omega_0^{\mathrm{rot}}/\omega_0^{\mathrm{Hopf}} \in \mathbb{Z}$, i.e., if there is a resonance between the rotation frequency ω_0^{rot} and the Hopf frequency ω_0^{Hopf} of the rotating wave Γx_0; see [3, 8, 10, 31]. In the case of two real parameters $\mu \in \mathbb{R}^2$ the following proposition holds true; see also Figure 4.

Proposition 2.1 (see [31, Example 3.6]). *Let $\mu \in \mathbb{R}^2$, and let $\mathrm{SE}(2)x_0$ be a rotating wave at $\mu = 0$ at which a resonant Hopf bifurcation occurs: $\omega_0^{\mathrm{rot}}/\omega_0^{\mathrm{Hopf}} \in \mathbb{Z}$. Then under some nondegeneracy conditions (detailed in the proof below) a path $\mathcal{P}_{\mathrm{MTW}}(s)$, $s \geq 0$, of modulated traveling waves at parameters $\mu_{\mathrm{MTW}}(s)$ bifurcates from the rotating wave $\mathrm{SE}(2)x_0$.*

Proof. Denote the rotation frequency of the rotating wave $\mathrm{SE}(2)x_{\mathrm{RW}}(\mu)$ at parameter μ by $\omega^{\mathrm{rot}}(\mu)$. The transversality condition (2.7) for Hopf bifurcation ensures that coordinates in parameter space $\mu \in \mathbb{R}^2$ can be chosen such that $\mu_1 = 0$ is the Hopf line near $\mu = 0$, i.e., such that $\lambda(0, \mu_2) = i\omega^{\mathrm{Hopf}}(\mu_2)$ for some smooth function $\omega^{\mathrm{Hopf}}(\mu_2)$ with $\omega^{\mathrm{Hopf}}(0) = \omega_0^{\mathrm{Hopf}}$. Periodic orbits bifurcating from this Hopf line are then parametrized by μ_2 and $s \geq 0$. Let $v(s, \mu_2)$ lie on a periodic orbit with parameters s, μ_2 such that $v(s, \mu_2)$ is smooth in its parameters and $v(0, \mu_2) = v_{\mathrm{RW}}(0, \mu_2)$. Let $T(s, \mu_2)$ be the period of the periodic orbit through $v(s, \mu_2)$. Modulated traveling waves satisfy $F(s, \mu_2) = \phi(T(s, \mu_2)) = 0$. This equation can be solved near 0 for $\mu_2(s)$ by the implicit function theorem if $\frac{\partial F}{\partial \mu_2}(0) \neq 0$. This condition holds

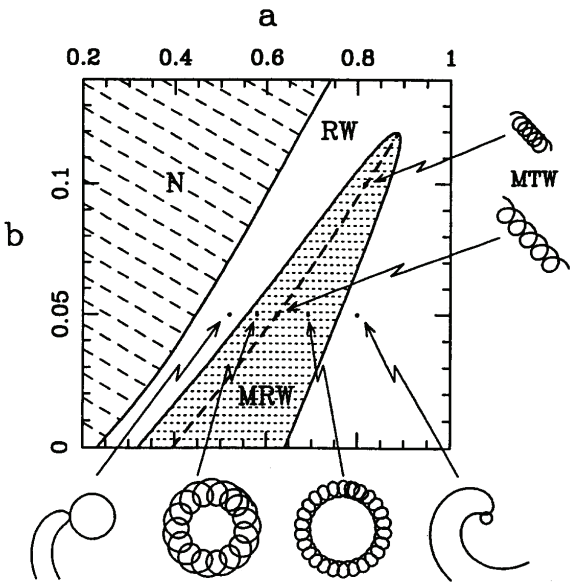


Figure 4. *Phase diagram for the spiral wave dynamics for a reaction-diffusion system depending on the parameters a, b. Reprinted Figure 1 from [3] with permission, copyright 1994 by the American Physical Society.*[1] *Shown are regions containing N: no spiral waves; RW: stable rigidly rotating waves; MRW: modulated rotating waves; MTW: modulated traveling waves (dashed curve). Spiral tip paths illustrate states at six points. Small portions of spiral waves are shown for the two rotating wave cases.*

true if the nondegeneracy condition

$$\frac{\partial}{\partial \mu_2}\left(\frac{\omega^{\mathrm{rot}}(\mu_2)}{\omega^{\mathrm{Hopf}}(\mu_2)}\right)\Big|_{\mu_2=0} \neq 0 \tag{2.8}$$

is satisfied. Then $v(s, \mu_2(s))$ lies on modulated traveling wave $\mathcal{P}_{\mathrm{MTW}}(s)$. ■

Remarks 2.2. (a) In [31] (see also [6] for compact groups) resonances of the form $\omega_\Gamma/\omega_\mathcal{N} = k \in \mathbb{Z}\backslash\{0\}$ between a nonvanishing imaginary eigenvalue $\pm i\omega_\Gamma$ of ad_{ξ_0} and an eigenvalue $\pm i\omega_\mathcal{N}$ of $\mathrm{D}_v f_\mathcal{N}(0)$ are shown to be necessary for *resonance drift* to occur. Resonance drift means that RPOs $\mathcal{P}(s)$ bifurcate with average drift velocities $\xi(s)$ at $x(s) \in \mathcal{P}(s)$ which cannot be chosen to converge to the drift velocity ξ_0 of the relative equilibrium, i.e., $x(s) \to x_0$ as $s \to 0$, but $\lim_{s\to 0} \xi(s) \neq \xi_0$. From the form of the linearization $L_0 = \mathrm{D}f(x_0) - \xi_0$ about a relative equilibrium Γx_0 in a corotating frame (see (2.4)), it follows that resonance drift is caused by resonances between drift dynamics and the symmetry-reduced dynamics. In Proposition 2.1 above, resonance drift occurs with $\omega_\Gamma = \omega_0^{\mathrm{rot}}$ and $\omega_\mathcal{N} = \omega_0^{\mathrm{Hopf}}$. Proposition 2.1 is a special case of [31, Proposition 3.4], which treats resonance drift for general Lie groups Γ.

(b) In the case of spherical symmetry $\Gamma = \mathrm{SO}(3)$ modeling, for example, rotating spiral waves of reaction-diffusion systems on the sphere, resonance drift caused by resonant Hopf

[1]Readers may view, browse, and/or download material for temporary copying purposes only, provided these uses are for noncommercial personal purposes. Except as provided by law, this material may not be further reproduced, distributed, transmitted, modified, adapted, performed, displayed, published, or sold in whole or part, without prior written permission from the American Physical Society.

bifurcation has been studied in [31, 7, 4]. In this case generically there is a path $(x(s), \mu(s))$ in two-dimensional parameter space $\mu(s) \in \mathbb{R}^2$ such that $\mathrm{SO}(3)x(s)$ is a modulated rotating wave at parameter $\mu(s)$ with an average drift velocity $\xi(s)$ at $x(s)$ which is orthogonal to the drift velocity $\xi_0 \in \mathrm{so}(3) \in \mathcal{T}_{\mathrm{id}}\mathrm{SO}(3)$ of the rotating wave $\mathrm{SO}(3)x_0$ at $x_0 = \lim_{s\to 0} x(s)$. The proof of this result is very similar to the proof of Proposition 2.1: For any $R \in \mathrm{SO}(3)$ write

$$R = \exp\left(\sum_{i=1}^{3} \phi_i \xi_i\right). \tag{2.9}$$

Here ξ_i, $i = 1, 2, 3$, are infinitesimal rotations such that $\exp(\phi_i \xi_i)$, $i = 1, 2, 3$, is a rotation by the angle ϕ_i around the e_i axis (often $\mathrm{so}(3)$ is identified with $\mathbb{R}^3$ and ξ_i with e_i, $i = 1, 2, 3$). Assume, as before, that the Hopf line is at $\mu_1 = 0$. Let $R(s, \mu_2)$ be the drift symmetry of the modulated rotating wave at $x(s, \mu_2)$. Assume, without loss of generality, that the rotating wave through $x_{\mathrm{RW}}(\mu)$ has a rotation velocity $\xi_{\mathrm{RW}}(\mu) || \xi_3$ so that $\xi_{\mathrm{RW}}(\mu) = \omega^{\mathrm{rot}}(\mu) e_3$. Then the modulated rotating waves to be found satisfy the equation $F(s, \mu_2) = \phi_3(s, \mu_2) = 0$. This equation can be solved if (2.8) holds. The bifurcating modulated rotating waves $\mathcal{P}(s)$ have average drift velocities in the (x_1, x_2)-plane. For a Hamiltonian analogue see section 5.1.

(c) Resonance drift also occurs for relative equilibria of systems with the Euclidean symmetry group $\Gamma = \mathrm{SE}(3) = \mathrm{SO}(3) \ltimes \mathbb{R}^3$ of rotations and translations in three-dimensional space; cf. [5]. An example would be a Hopf bifurcation from a rigidly rotating and translating scroll wave $\mathrm{SE}(3)x_0$ of a reaction-diffusion system on $\mathbb{R}^3$; see, e.g., [30]. The group multiplication on $\Gamma = \mathrm{SE}(3) = \mathrm{SO}(3) \ltimes \mathbb{R}^3$ is analogous to (2.5): For $(R_1, a_1), (R_2, a_2) \in \mathrm{SO}(3) \ltimes \mathbb{R}^3$ it is given by

$$(R_1, a_1)(R_2, a_2) = (R_1 R_2, a_1 + R_1 a_2), \qquad R_1, R_2 \in \mathrm{SO}(3),\ a_1, a_2 \in \mathbb{R}^3.$$

Note that (R, a) is a rotation around the rotation axis of R about the point

$$c = (\mathrm{id} - R)^+ a \in \mathbb{R}^3$$

combined with a translation along the axis of R. Here A^+ denotes the Moore–Penrose pseudoinverse of A; i.e., $x = A^+ b$ satisfies $\|Ax - b\|_2 = \min$, $A \in \mathrm{Mat}(n)$, $x, b \in \mathbb{R}^n$. Let $\xi_0 = (\xi_0^r, \xi_0^a)$ be the drift velocity of the relative equilibrium $\mathrm{SE}(3)x_0$ at x_0 and assume that $\xi_0^r \neq 0$. Without loss of generality, choose x_0 in its group orbit such that $\xi_0^r \in \mathrm{so}(3) \simeq \mathbb{R}^3$ (see part (b) for this identification) is parallel to ξ_0^a and to e_3, and write $\xi_0^r = \omega_0^{\mathrm{rot}} e_3$, where $\omega_0^{\mathrm{rot}} \neq 0$. Align the family of relative equilibria $\mathrm{SE}(3)x_{\mathrm{RE}}(\mu)$ with $x_{\mathrm{RE}}(0) = x_0$ such that their drift velocity $\xi_{\mathrm{RE}}(\mu)$ at $x_{\mathrm{RE}}(\mu)$ also satisfies $\xi_{\mathrm{RE}}^r(\mu) = \omega^{\mathrm{rot}}(\mu) e_3$. If the Hopf frequency ω_0^{Hopf} satisfies $\omega_0^{\mathrm{rot}}/\omega_0^{\mathrm{Hopf}} \in \mathbb{Z}$ and this resonance is passed transversely as in (2.8), then, as in part (b), there is a curve $\mathcal{P}(s)$ of relative periodic orbits through $x(s) \approx x_0$ with drift symmetry $\gamma(s) = (R(s), a(s))$ at $x(s)$ satisfying $\phi_3(s) = 0 \bmod 2\pi$. Here $R(s)$ is determined by $\phi_i(s)$, $i = 1, 2, 3$, as in (2.9). These RPOs rotate and translate along a vector in the (x_1, x_2)-plane. The point around which they rotate approaches infinity as $s \to 0$.

(d) Note that for the groups $\Gamma = \mathrm{SE}(2)$, $\Gamma = \mathrm{SO}(3)$, and $\Gamma = \mathrm{SE}(3)$ considered above, resonance drift can occur only near relative equilibria with nonvanishing rotational velocity. Otherwise, the linear map ad_{ξ_0} has no eigenvalues in $\mathrm{i}\mathbb{R}\backslash\{0\}$, but this is necessary for resonance drift; cf. part (a) and [31].

3. Dynamics near Hamiltonian relative equilibria. As in the dissipative case, the meandering transition in Hamiltonian systems is studied by analyzing the equations near relative equilibria (2.3). Therefore, in this section symmetric Hamiltonian systems and the structure of the equations (2.3) for Hamiltonian systems are reviewed. Then these results are applied to Hamiltonian systems with Euclidean symmetry for later use in the analysis of the Hamiltonian meandering transition. Most of the material of this section is taken from [12, 18, 26, 28].

3.1. Symmetric Hamiltonian systems. In this section a brief introduction to symmetric Hamiltonian differential equations is given (see, e.g., [12, 18] for more details). The starting point is a Hamiltonian ordinary differential equation on a smooth finite-dimensional symplectic manifold $\mathcal{M}$ with a symplectic form (i.e., a nondegenerate, closed 2-form) Ω_x, $x \in \mathcal{M}$. A Hamiltonian vector field

$$\dot{x} = f_H(x) \tag{3.1}$$

is generated by a smooth function (the *Hamiltonian*) $H : \mathcal{M} \mapsto \mathbb{R}$ via the relationship

$$\Omega_x(f_H(x), v) = \mathrm{D}H(x)v, \qquad x \in \mathcal{M},\ v \in \mathcal{T}_x\mathcal{M}. \tag{3.2}$$

Example 3.1. The simplest example is a Hamiltonian system

$$\dot{x} = \mathbb{J}\mathrm{D}_x H(x)$$

on $\mathcal{M} = \mathbb{R}^{2n}$, where

$$\mathbb{J} = \begin{pmatrix} 0 & \mathrm{id} \\ -\mathrm{id} & 0 \end{pmatrix}$$

and $H : \mathcal{M} \to \mathbb{R}$ is a smooth Hamiltonian. Then the symplectic form Ω is the standard symplectic form given by $\Omega(u, v) = \langle \mathbb{J}^{-1}u, v\rangle$, and $\mathbb{J}$ is called the symplectic structure matrix. By the Darboux theorem (see, e.g., [18]), locally every Hamiltonian system has this form in suitable coordinates.

Let us assume that a finite-dimensional Lie group Γ acts *symplectically* on $\mathcal{M}$, i.e., that

$$\Omega_{\gamma x}(\gamma u, \gamma v) = \Omega_x(u, v) \qquad \text{for all}\ \ x \in \mathcal{M},\ \ \gamma \in \Gamma,\ \ u, v \in \mathcal{T}_x\mathcal{M}.$$

If H is invariant under the action of Γ, then the vector field f_H is Γ-equivariant.

Let $\mathbf{g}^*$ denote the dual of the Lie algebra $\mathbf{g}$ of Γ. By Noether's theorem, for each continuous symmetry $\xi \in \mathbf{g}$ locally there is a conserved quantity $\mathbf{J}(\xi)(\cdot)$ of (3.1). The function $\mathbf{J}(\xi)$ is linear in ξ, so that $\mathbf{J}$ maps into $\mathbf{g}^*$ (see, e.g., [18]). It is assumed that $\mathbf{J}$ exists globally on $\mathcal{M}$.

Example 3.2. The dynamics of N point vortices $(z_1, \dots, z_N) \in \mathbb{R}^{2N}$, $z_j = (x_j, y_j)$, $j = 1, \dots, N$, on the plane is given by the following Hamiltonian system [1, 2, 21]:

$$k_i\dot{x}_i = \frac{\partial H}{\partial y_i}, \quad k_i\dot{y}_i = -\frac{\partial H}{\partial x_i}, \quad i = 1, \dots, N, \tag{3.3}$$

where $k_i \neq 0$, $k = 1, \dots, N$. The Hamiltonian H

$$H(z_1, \dots, z_N) = -\frac{1}{\pi}\sum_{\substack{i,j=0 \\ i<j}}^{N} k_i k_j \ln|z_i - z_j|$$

of (3.3) is invariant under the action of the special Euclidean group of the plane $\Gamma = \mathrm{SE}(2) = \mathrm{SO}(2) \ltimes \mathbb{R}^2$ on $\mathbb{R}^{2N}$, given by

$$(R_\varphi, a) \cdot (z_1, \dots, z_N) := (R_\varphi z_1 + a, \dots, R_\varphi z_N + a)$$

for $R_\varphi \in \mathrm{SO}(2)$ and $a \in \mathbb{R}^2$. The symplectic form

$$\Omega(z_1, \dots, z_N) = \sum_{i=1}^{N} k_i \mathrm{d}x_i \wedge \mathrm{d}y_i$$

is SE(2)-invariant. The Hamiltonian system (3.3) can be obtained from Euler's equations for ideal fluids by modeling the point vortices as δ-distributions; see, e.g., [2]. In this example the space of momenta is $\mathbf{g}^* = \mathrm{se}(2)^* = \mathrm{so}(2)^* \oplus (\mathbb{R}^2)^*$. By Noether's theorem, $\mathbf{J}(x) = (\mathbf{J}^\phi(x), \mathbf{J}^a(x))$ is conserved. Here the angular momentum $\mathbf{J}^\phi$ and linear momentum $\mathbf{J}^a = (\mathbf{J}^{a_1}, \mathbf{J}^{a_2})$ are given by

$$\mathbf{J}^\phi(x) = -\frac{1}{2}\sum_{i=1}^{N} k_i |z_i|^2, \quad \mathbf{J}^{a_1} = \sum_{i=1}^{N} k_i y_i, \quad \mathbf{J}^{a_2} = -\sum_{i=1}^{N} k_i x_i. \tag{3.4}$$

In the following transitions from relative equilibria to relative periodic orbits are studied when the conserved quantities angular and linear momentums are varied. In contrast to dissipative systems, external parameters are not needed for the study of bifurcations. These transitions are studied by analyzing the symmetry-reduced equations (2.3) for Hamiltonian systems. As in the general case (see section 2.1), the reduction by the symmetry group is achieved by transforming the dynamics into a comoving frame. As a consequence, in the symmetry-reduced system the momentum is moving with the velocity of the comoving frame and might not be conserved anymore. Therefore, to compute the reduced system in the Hamiltonian case, first the action of the symmetry group on the space of momenta is investigated.

3.2. Symmetries of momentum maps. Let us assume that $\mathbf{J}$ commutes with $\gamma \in \Gamma$,

$$\mathbf{J}(\gamma x) = \gamma \mathbf{J}(x), \quad \gamma \in \Gamma, \tag{3.5}$$

and, unless otherwise stated, that the action on momentum space $\mathbf{g}^*$ is the coadjoint action, so that the momentum map is Ad^*-equivariant:

$$\gamma \mathbf{J}(x) = (\mathrm{Ad}^*_\gamma)^{-1} \mathbf{J}(x), \quad \gamma \in \Gamma. \tag{3.6}$$

The *coadjoint action* of Γ on $\mathbf{g}^*$ is given by $\gamma\mu := (\mathrm{Ad}^*_\gamma)^{-1}\mu$, where $\mathrm{Ad}_\gamma : \mathbf{g} \to \mathbf{g}$ from (2.2) is the adjoint action. The *infinitesimal coadjoint action* of $\mathbf{g}$ on $\mathbf{g}^*$ is defined by

$$\xi\mu = -\mathrm{ad}^*_\xi \mu, \tag{3.7}$$

with ad_ξ as in (2.2). The *isotropy subgroup* of $\mu \in \mathbf{g}^*$ is denoted by

$$\Gamma_\mu = \{\gamma \in \Gamma,\ \mathrm{Ad}^*_\gamma \mu = \mu\}$$

and its Lie algebra by $\mathbf{g}_\mu$.

Example 3.3. As an example the adjoint and coadjoint action for the Euclidean group are computed. They are needed later for the computation of the drift dynamics near Hamiltonian rotating waves.

Let $\gamma = (\phi, a)$, $\hat{\gamma} = (\hat{\phi}, \hat{a})$. Then

$$\begin{aligned}\gamma\hat{\gamma}\gamma^{-1} &= (\phi, a)(\hat{\phi}, \hat{a})(\phi, a)^{-1} = (\phi + \hat{\phi}, R_\phi \hat{a} + a)(\phi, a)^{-1} \\ &= (\phi + \hat{\phi}, R_\phi \hat{a} + a)(-\phi, -R_{-\phi} a) = (\hat{\phi}, -R_{\hat{\phi}} a + R_\phi \hat{a} + a) \\ &= (\hat{\phi}, R_\phi \hat{a} + (\mathrm{id} - R_{\hat{\phi}}) a).\end{aligned}$$

Letting $\hat{\phi} = \xi^\phi \epsilon$ and $\hat{a} = \xi^a \epsilon$ and differentiating with respect to ϵ at $\epsilon = 0$, one gets, with $\xi = (\xi^\phi, \xi^a) = (\xi^\phi, \xi^a_1, \xi^a_2) \in \mathbb{R}^3$,

$$\mathrm{Ad}_\gamma \xi = \gamma \xi \gamma^{-1} = (\phi, a)(\xi^\phi, \xi^a)(\phi, a)^{-1} = (\xi^\phi, (R_\phi \xi^a)_1 + \xi^\phi a_2, (R_\phi \xi^a)_2 - \xi^\phi a_1).$$

Using $\mathrm{ad}_\xi = \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{Ad}_{\exp(t\xi)}|_{t=0}$, the adjoint actions of SE(2) on se(2) and the infinitesimal adjoint action of se(2) on se(2) are obtained:

$$\mathrm{Ad}_\gamma = \begin{pmatrix} 1 & 0 & 0 \\ a_2 & \cos\phi & -\sin\phi \\ -a_1 & \sin\phi & \cos\phi \end{pmatrix}, \quad \mathrm{ad}_\xi = \begin{pmatrix} 0 & 0 & 0 \\ \xi^a_2 & 0 & -\xi^\phi \\ -\xi^a_1 & \xi^\phi & 0 \end{pmatrix}. \tag{3.8}$$

The coadjoint action of SE(2) and se(2) and the infinitesimal coadjoint action of se(2) on se(2)* are obtained by transposing and inverting Ad_γ and by transposition and multiplication by -1 of ad_ξ:

$$(\mathrm{Ad}^*_\gamma)^{-1} = \begin{pmatrix} 1 & -(R_{-\phi} a)_2 & (R_{-\phi} a)_1 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix}, \quad -\mathrm{ad}^*_\xi = \begin{pmatrix} 0 & -\xi^a_2 & \xi^a_1 \\ 0 & 0 & -\xi^\phi \\ 0 & \xi^\phi & 0 \end{pmatrix}. \tag{3.9}$$

From these equations it can be seen that the isotropy subgroup Γ_μ of $\mu \in \mathrm{se}(2)^*$ is $\Gamma_\mu = \Gamma = \mathrm{SE}(2)$ if and only if $\mu^a = 0$ and that $\Gamma_\mu \simeq \mathbb{R}$ for $\mu^a \neq 0$.

Remark 3.4. In the case of zero total circulation $\mathcal{K} := \sum_{i=1}^N k_i = 0$ the momentum map $\mathbf{J}$ for the planar vortex dynamics from (3.4) is Ad*-equivariant (see [2]); but, if $\mathcal{K} \neq 0$, then, instead of (3.6), the equivariance condition (3.5) now holds for the action

$$\gamma \cdot_\kappa \mu := \mathrm{Ad}^*_{\gamma^{-1}} \mu + \kappa(\gamma) \tag{3.10}$$

of Γ on $\mathbf{g}^*$. Here

$$\kappa(\phi, a) = \mathcal{K}(-\tfrac{1}{2}|a|^2, a_2, -a_1) \in \mathrm{se}(2)^* \tag{3.11}$$

is called a *cocycle*; see [18]. In other words, (3.5) now becomes

$$\mathbf{J}(\gamma x) = \gamma \cdot_\kappa \mathbf{J}(x) \quad \text{for all} \quad \gamma \in \Gamma. \tag{3.12}$$

The *infinitesimal cocycle* $K : \mathbf{g} \times \mathbf{g} \to \mathbb{R}$ corresponding to the cocycle κ is defined as

$$K(\xi) = \frac{\mathrm{d}}{\mathrm{d}t}\kappa(e^{t\xi})|_{t=0} \in \mathbf{g}^*,$$

and in this case it is given by

$$K(\xi,\eta) = \mathcal{K}\left\langle \xi^a, \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \eta_a \right\rangle = \mathcal{K}(-\xi_1^a\eta_2^a + \xi_2^a\eta_1^a). \tag{3.13}$$

The infinitesimal action of the Lie algebra $\mathbf{g}$ of Γ on $\mathbf{g}^*$ is defined as

$$\xi \cdot_K \mu = \frac{\mathrm{d}}{\mathrm{d}t}\exp(t\xi) \cdot_\kappa \mu|_{t=0} = -\mathrm{ad}^*_\xi\mu + K(\xi). \tag{3.14}$$

The isotropy subgroup of $\mu \in \mathbf{g}^*$ with respect to the cocycle action (3.10) is denoted by Γ^κ_μ. Its Lie algebra is denoted by $\mathbf{g}^K_\mu = \mathcal{T}_{\mathrm{id}}\Gamma^\kappa_\mu$. For later use, note that $\gamma = (\phi, a) \in \Gamma^\kappa_\mu$ for $\mu \in \mathrm{se}(2)^*$ if and only if

$$(R_\phi - \mathrm{id})\mu^a = \mathcal{K}\begin{pmatrix} -a_2 \\ a_1 \end{pmatrix}. \tag{3.15}$$

Hence for a nonvanishing cocycle the isotropy subgroup of every $\mu \in \mathbf{g}^*$ is conjugate to $\mathrm{SO}(2)$.

Let x_0 lie on a relative equilibrium Γx_0 with drift velocity $\xi_0 \in \mathbf{g}$ at x_0, so that $\Phi_t(x_0) = \exp(t\xi_0)x_0$. Since momentum is conserved,

$$\mu_0 = \mathbf{J}(x_0) = \mathbf{J}(\Phi_t(x_0)) = \mathbf{J}(\exp(t\xi_0)x_0) = \exp(t\xi_0)\mu_0,$$

and therefore μ_0 is fixed by ξ_0:

$$\xi_0\mu_0 = 0. \tag{3.16}$$

Such pairs $(\xi, \mu) \in \mathbf{g} \oplus \mathbf{g}^*$ are called *velocity-momentum pairs*. Note that the action of ξ_0 on μ_0 in (3.16) is the infinitesimal coadjoint action (3.7), or the infinitesimal action with cocycle (3.14), depending on the symmetry property of the momentum map.

Similarly, if $x_0 = \gamma_0^{-1}\Phi_{T_0}(x_0)$ lies on a relative periodic orbit with drift symmetry γ_0 and momentum $\mu_0 = \mathbf{J}(x_0)$, then μ_0 is fixed by γ_0:

$$\gamma_0\mu_0 = \mu_0. \tag{3.17}$$

Such pairs $(\gamma, \mu) \in \Gamma \times \mathbf{g}^*$ are called *drift-momentum pairs*.

3.3. Dynamics near Hamiltonian relative equilibria. For a symplectic manifold $\mathcal{M}$ with Ad^*-equivariant momentum map the normal space $\mathcal{N}$ to the group orbit Γx_0 at $x_0 \in \mathcal{M}$ from section 2.1 can be decomposed as

$$\mathcal{M}/\Gamma \simeq \mathcal{N} \;=\; \mathcal{N}_0 \oplus \mathcal{N}_1 \;\cong\; \mathbf{g}^*_{\mu_0} \oplus \mathcal{N}_1.$$

Here

$$\mathbf{g}_{\mu_0} = \mathcal{T}_{\mathrm{id}}\Gamma_{\mu_0} = \{\xi \in \mathbf{g} : \mathrm{ad}^*_\xi\mu_0 = 0\}$$

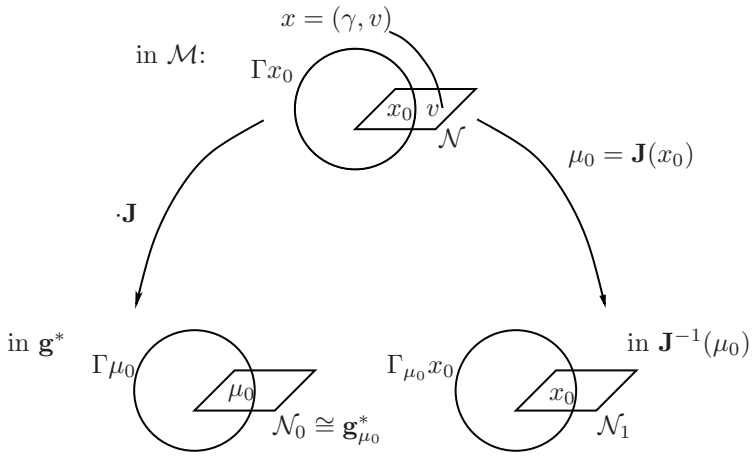


Figure 5. *Symplectic slice theorem.*

is the isotropy subalgebra of the momentum $\mu_0 = \mathbf{J}(x_0)$ of x_0. The space $\mathcal{N}_0$ is isomorphic to a section transverse to the momentum group orbit $\Gamma\mu_0$ at μ_0. To see that $\mathcal{N}_0 \simeq \mathbf{g}^*_{\mu_0}$, let $\mathbf{n}_{\mu_0}$ be a complement to $\mathbf{g}_{\mu_0}$ in $\mathbf{g}$ and let $\operatorname{ann}(\mathbf{n}_{\mu_0})$ denote the annihilator of $\mathbf{n}_{\mu_0}$ in $\mathbf{g}^*$. Then $\mathcal{T}_{\mu_0}\Gamma\mu_0 = \mathbf{g}\mu_0 = \operatorname{ann}(\mathbf{g}_{\mu_0})$, and so $\operatorname{ann}(\mathbf{n}_{\mu_0}) \cong \mathbf{g}^*_{\mu_0}$ is a section transverse to $\Gamma\mu_0$ at μ_0. The *symplectic normal space* or *symplectic slice* $\mathcal{N}_1$ at x_0 is a slice to the Γ_{μ_0} orbit of x_0 in the momentum level set $\mathbf{J}^{-1}(\mu_0)$; cf. Figure 5. Moreover, there is a choice of slice $\mathcal{N}$ such that the coordinate transformation $x \to (\gamma, \nu, w) \in \Gamma \times \mathcal{N}_0 \oplus \mathcal{N}_1$, where x lies in some Γ-invariant neighborhood $\mathcal{U}$ of Γx_0, is symplectic with symplectic form $\Omega_{\Gamma\times\mathcal{N}}$ on $\Gamma \times \mathcal{N}$ given by $\Omega_{\Gamma\times\mathcal{N}} = \Omega_{\Gamma\times\mathbf{g}^*_{\mu_0}} + \Omega_{\mathcal{N}_1}$. Here $\Omega_{\mathcal{N}_1}$ is the symplectic form on $\mathcal{N}_1$ and $\Omega_{\Gamma\times\mathbf{g}^*_{\mu_0}}$ the symplectic form on $\Gamma \times \mathbf{g}^*_{\mu_0}$, obtained by restriction of the symplectic form on $\mathcal{T}^*\Gamma \simeq \Gamma \times \mathbf{g}^*$. In these coordinates the momentum map becomes

$$\mathbf{J}(\gamma, \nu, w) = \gamma(\mu_0 + \nu); \tag{3.18}$$

see [11, 17] and also [26]. Let $\mathbb{J}_{\mathcal{N}_1}$ be the structure matrix of the symplectic form on $\mathcal{N}_1$.

One more technical assumption is needed: In this paper, unless otherwise stated, it is assumed that μ_0 is *split*; i.e., there is a $\Gamma^{\mathrm{id}}_{\mu_0}$-invariant complement to $\mathbf{g}_{\mu_0}$ in $\mathbf{g}$. Here $\Gamma^{\mathrm{id}}_{\mu_0}$ is the identity component of Γ_{μ_0}. This condition is always satisfied for compact groups and also for the special Euclidean group of the plane; see [26]. For the general case see [26].

Theorem 3.5 (see [26, Theorem 3.1]). *Let, as above, (γ, v), $v = (\nu, w) \in \mathcal{N}$, $\gamma \in \Gamma$, parametrize a Γ-invariant neighborhood $\mathcal{U}$ of the relative equilibrium Γx_0 with momentum $\mu_0 = \mathbf{J}(x_0)$. Let $h(\nu, w)$ be the restriction of the Hamiltonian H to the slice $\mathcal{N} = \mathbf{g}^*_{\mu_0} \oplus \mathcal{N}_1$, and let μ_0 be split. Assume that the momentum map is Ad^*-equivariant. Then $\gamma(t) \in \Gamma$, $\nu(t) \in \mathbf{g}^*_{\mu_0}$, $w(t) \in \mathcal{N}_1$, where $x(t) \simeq (\gamma(t), \nu(t), w(t)) \in \mathcal{U}$ solves (3.1), satisfy the differential equations*

$$\dot\gamma = \gamma\, \mathrm{D}_\nu h(\nu, w), \quad \dot\nu = \mathrm{ad}^*_{\mathrm{D}_\nu h(\nu,w)}\nu, \quad \dot w = \mathbb{J}_{\mathcal{N}_1}\mathrm{D}_w h(\nu, w). \tag{3.19}$$

As in the non-Hamiltonian case, the relative equilibrium Γx_0 corresponds to the equilibrium $v = (0, 0) \in \mathcal{N}$ of the slice equation on $\mathcal{N}$. The first equation describes the motion of

the body frame. Here $\mathrm{D}_\nu h$ is the velocity of the body frame, and $\mathrm{D}_\nu h(0,0) = \xi_0$ is the drift velocity of the relative equilibrium Γx_0 at x_0. The second equation describes the motion of the momenta in body coordinates, and the last equation models the shape dynamics.

From a comparison of (3.19) with the equations near relative equilibria in the general case (2.3), it becomes apparent that in the Hamiltonian case $v = (\nu, w)$ and

$$f_\Gamma(v) = \mathrm{D}_\nu h(\nu, w), \quad f_\mathcal{N}(v) = \begin{pmatrix} \mathrm{ad}^*_{\mathrm{D}_\nu h}(\nu) \\ \mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w h(\nu, w) \end{pmatrix}.$$

So the slice equation $\dot v = f_\mathcal{N}(v)$ now consists of the two differential equations for $\dot\nu$ and $\dot w$. Moreover,

$$\mathrm{D} f_\mathcal{N}(0) = \begin{pmatrix} \mathrm{ad}^*_{\xi_0}|_{\mathbf{g}^*_{\mu_0}} & 0 \\ \mathbb{J}_{\mathcal{N}_1} \mathrm{D}^2_{\nu w} h(0) & \mathbb{J}_{\mathcal{N}_1} \mathrm{D}^2_w h(0) \end{pmatrix}. \tag{3.20}$$

The energy $h(\nu, w)$ is a conserved quantity of the slice equation which is a *Poisson system*; see [18]. Any function $C(\nu, w)$ which is a conserved quantity of the slice equation for all choices of Hamiltonians $h(\nu, w)$ is called a *Casimir* of the slice equation. Note that the flow-invariant *symplectic leaves* of the slice equation are given by $\Gamma^{\mathrm{id}}_{\mu_0} \nu_0 \times \mathcal{N}_1$, where $\nu_0 \in \mathcal{N}_0$.

Note for later reference that (3.20) is true for nonsplit μ as well if the infinitesimal coadjoint action ad^*_ξ restricted to $\mathrm{ann}(\mathbf{n}_{\mu_0}) \simeq \mathbf{g}^*_{\mu_0}$ in $\mathrm{ad}^*_{\xi_0}|_{\mathbf{g}^*_{\mu_0}}$ is replaced by the corresponding action on $\mathbf{g}_{\mu_0}$. For split μ_0, both these actions coincide; see [26].

Remark 3.6. With the notation from (3.7), the $\dot\nu$-equation can be rewritten as

$$\dot\nu = -\mathrm{D}_\nu h(\nu, w)\nu.$$

As shown in [28], Theorem 3.5 remains true in the case of a momentum map which is symmetric with respect to a cocycle action if the infinitesimal coadjoint action in the $\dot\nu$-equation is replaced by the corresponding action (3.14) with a cocycle. Then the $\dot\nu$-equation becomes

$$\dot\nu = -\mathrm{D}_\nu h(\nu, w) \cdot_K \nu \tag{3.21}$$

provided that μ_0 is split for the action of Γ^κ on $\mathbf{g}$, i.e., if there is a $(\Gamma^\kappa_{\mu_0})^{\mathrm{id}}$-invariant complement $\mathbf{n}^K_{\mu_0}$ to $\mathbf{g}^K_{\mu_0}$ in $\mathbf{g}$, where $(\Gamma^\kappa_{\mu_0})^{\mathrm{id}}$ is the identity component of $\Gamma^\kappa_{\mu_0}$. Moreover, as before, $\mathcal{N}_0 \simeq \mathrm{ann}(\mathbf{n}^K_{\mu_0}) \simeq (\mathbf{g}^K_{\mu_0})^*$.

Remark 3.7. For later use, let us consider parameter dependent Hamiltonian systems

$$\dot x = f(x, \mathcal{K}), \tag{3.22}$$

where f is defined by

$$\Omega(x, \mathcal{K})(f(x, \mathcal{K}), v) = \mathrm{D}_x H(x, \mathcal{K})v, \qquad x \in \mathcal{M},\ v \in \mathcal{T}_x\mathcal{M}.$$

Assume that the symplectic form $\Omega(\mathcal{K})$, the Hamiltonian $H(\cdot, \mathcal{K})$, and the momentum map $\mathbf{J}(\cdot, \mathcal{K})$ depend smoothly on a parameter $\mathcal{K}$. Then Theorem 3.5 still applies, and the sections $\mathcal{N}_0(\mathcal{K})$, $\mathcal{N}_1(\mathcal{K})$, as well as the Hamiltonian $h(\nu, w, \mathcal{K})$, depend smoothly on $\mathcal{K}$, as long as the dimensions of $\mathcal{N}_0(\mathcal{K})$ and $\mathcal{N}_1(\mathcal{K})$ are constant. See the proofs in [11, 17] and [26].

Example 3.8. For later use, let us derive the differential equations (3.19) near rotating waves of Hamiltonian systems (3.1) with symmetry group $\Gamma = \mathrm{SE}(2)$, in the case of an Ad^*-equivariant momentum map. Let x_0 lie on a rotating wave so that $\Phi_t(x_0) = \exp(t\xi_0)x_0$, $\xi_0 \in \mathrm{so}(2)$, $\mu_0 = \mathbf{J}(x_0)$. From (3.16) it follows that $0 = -\mathrm{ad}^*_{\xi_0}\mu_0$, with $\mathrm{ad}^*_{\xi_0}$ from (3.9). Therefore, $\mu_0^a = 0$, so that the rotating wave through x_0 has vanishing linear momentum. Then $\Gamma_{\mu_0} = \mathrm{SE}(2)$; see Example 3.3. Hence $\mathcal{N}_0 \simeq \mathbf{g}^*_{\mu_0} \simeq \mathrm{se}(2)^*$, and so the equations (3.19) for SE(2)-equivariant Hamiltonian systems are

$$
\begin{array}{rcll}
\dot\phi &=& \mathrm{D}_{\nu^\phi}h, \quad \dot a = R_\phi \mathrm{D}_{\nu^a}h, & \Leftrightarrow \quad \dot\gamma = \gamma \mathrm{D}_\nu h(\nu, w), \\
\dot\nu^\phi &=& \nu_1^a \mathrm{D}_{\nu_2^a}h(\nu,w) - \nu_2^a \mathrm{D}_{\nu_1^a}h(\nu,w) & \\
\dot\nu_1^a &=& \nu_2^a \mathrm{D}_{\nu^\phi}h(\nu,w) & \Leftrightarrow \quad \dot\nu = \mathrm{ad}^*_{\mathrm{D}_\nu h}\nu, \\
\dot\nu_2^a &=& -\nu_1^a \mathrm{D}_{\nu^\phi}h(\nu,w) & \\
\dot w &=& \mathbb{J}_{\mathcal{N}_1}\mathrm{D}_w h(\nu,w). &
\end{array}
\tag{3.23}
$$

Example 3.9. Also, for later use, let us consider the equations (3.21) for Hamiltonian systems (3.22) which have a momentum map $\mathbf{J}(\cdot, \mathcal{K})$ with cocycle (3.11), $\mathcal{K} \neq 0$. From (3.10) and (3.11) it follows that the momentum μ_0 can always be translated such that $\mu_0^a = 0$. For $\mu_0^a = 0$ one has $\xi = (\xi^\phi, \xi^a) \in \mathbf{g}^K_{\mu_0}$ if and only if $\xi^a = 0$, and so

$$\Gamma^\kappa_{\mu_0} = \mathrm{SO}(2) \times \{0\} \subseteq \mathrm{SE}(2).$$

For $\xi^a = 0$ the infinitesimal cocycle (3.13) vanishes. Hence $\{(0, \xi^a) \in \mathrm{se}(2),\ \xi^a \in \mathbb{R}^2\}$ is a $\Gamma^\kappa_{\mu_0}$-invariant complement to $\mathbf{g}^K_{\mu_0} = \mathrm{so}(2)$ in se(2), and μ_0 is split. Consequently, $\mathcal{N}_0 \simeq (\mathbf{g}^K_{\mu_0})^* = \mathrm{so}(2)^*$, and $\dot\nu = 0$ in (3.21).

4. The meandering transition in Hamiltonian systems.

In this section a Hamiltonian analogue of the meandering transition for dissipative systems, which was described in section 2.2 above, is presented, using the equations near Hamiltonian relative equilibria (3.19) from section 3. First Euclidean symmetric Hamiltonian systems with an Ad^*-equivariant momentum map are studied (sections 4.1–4.3). Then systems with Euclidean symmetry for which the momentum map has a cocycle are considered (section 4.4).

4.1. Persistence of rotating waves.

In this section and in sections 4.2–4.3 it is assumed that (3.1) has the symmetry group $\Gamma = \mathrm{SE}(2)$ and an Ad^*-equivariant momentum map.

As a prerequisite for the analysis of the transition from rotating waves to modulated rotating waves and modulated traveling waves, the persistence of nondegenerate rotating waves to nearby momentum values is studied.

Definition 4.1. *A relative equilibrium* Γx_0 *of* (3.1) *is called* nondegenerate *if* $\mathrm{D}^2_w h(0)$ *is invertible. Here* $h(\nu, w)$ *is the Hamiltonian in the symmetry-adapted coordinates near* x_0 *from* (3.19).

Note that a relative equilibrium is typically nondegenerate. The next proposition shows that nondegenerate rotating waves persist to nearby angular momentum.

Proposition 4.2. *Let* $\mathrm{SE}(2)x_0$ *be a nondegenerate rotating wave of an* SE(2)*-equivariant Hamiltonian system* (3.1) *with* Ad^**-equivariant momentum map* $\mathbf{J}(\cdot)$*, and let* $\mu_0 = \mathbf{J}(x_0)$*. Then there is a one-parameter family* $\mathrm{SE}(2)x_{\mathrm{RW}}(\nu^\phi)$ *of rotating waves nearby parametrized*

by angular momentum $\mu^\phi = \mu_0^\phi + \nu^\phi$ *with vanishing linear momentum such that* $x_{\mathrm{RW}}(\nu^\phi) \simeq (\mathrm{id}, (\nu^\phi, 0, 0)^T, w_{\mathrm{RW}}(\nu^\phi))$ *is smooth in* ν^ϕ *and* $x(0) = x_0$, $w_{\mathrm{RW}}(0) = 0$.

Proof. Rotating waves are equilibria of the slice equation, i.e., of the (ν, w)-system of (3.23). Using the nondegeneracy assumption, the equation $0 = \dot{w} = \mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w h(\nu, w)$ can be solved by the implicit function theorem for $w_{\mathrm{RW}}(\nu)$ such that $w_{\mathrm{RW}}(0) = 0$. For rotating waves the linear momentum has to vanish; see Example 3.8. Hence $\nu^a = 0$ in any rotating wave. From the $\dot{\nu}$-equation of (3.23) then $\dot{\nu}^\phi = 0$, $\dot{\nu}^a = 0$ follows. Hence $x_{\mathrm{RW}}(\nu^\phi) \simeq (\mathrm{id}, (\nu^\phi, 0, 0)^T, w_{\mathrm{RW}}(\nu^\phi))$ lies on a rotating wave of (3.1) for all $\nu^\phi \approx 0$. By (3.18), $\mathbf{J}(x_{\mathrm{RW}}(\nu^\phi)) = \mu_0 + \nu = (\mu_0^\phi + \nu^\phi, 0)$. ∎

A persistence theory for generic nondegenerate relative equilibria of Hamiltonian systems with general noncompact symmetry group has been developed in [32]; see also [20] for an example of nonpersistence of rotating waves to nonvanishing linear momentum in point vortex dynamics. The above proposition could also be proved by applying the results of [32] to the example $\Gamma = \mathrm{SE}(2)$. However, the direct proof given above is more elementary.

4.2. Bifurcation of modulated traveling waves. The following theorem states that typically rotating waves of Euclidean equivariant Hamiltonian systems with Ad^*-equivariant momentum map persist to modulated traveling waves at nearby linear momenta $\mu^a \neq 0$. Consequently resonance drift occurs generically.

Theorem 4.3. *Let* $\mathrm{SE}(2)x_0$ *be a nondegenerate rotating wave of an* $\mathrm{SE}(2)$*-equivariant Hamiltonian system* (3.1) *with* Ad^**-equivariant momentum map. Denote its rotation frequency by* $\omega_0^{\mathrm{rot}} = \xi_0^\phi$ *and assume that* $\omega_0^{\mathrm{rot}} \neq 0$. *If all eigenvalues* $\mathrm{i}\omega_0$ *of* $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h(0)$ *satisfy*

$$\omega_0 / \omega_0^{\mathrm{rot}} \notin \mathbb{Z}, \tag{4.1}$$

then the rotating wave $\mathrm{SE}(2)x_0$ *persists as a modulated traveling wave* $\mathcal{P}_{\mathrm{MTW}}(\nu^\phi, r^a)$ *to all nearby momentum values* $\mu = (\mu_0^\phi + \nu^\phi, \nu^a)$, $r^a = \|\nu^a\|$. *Moreover, there is a smooth function* $x_{\mathrm{MTW}}(\nu^\phi, r^a) \in \mathcal{P}_{\mathrm{MTW}}(\nu^\phi, r^a)$ *such that* $x_{\mathrm{MTW}}(\nu^\phi, 0) = x_{\mathrm{RW}}(\nu^\phi)$. *Here* $\mathrm{SE}(2)x_{\mathrm{RW}}(\nu^\phi)$ *is the family of rotating waves from Proposition 4.2. The relative period* $T_{\mathrm{MTW}}(\nu^\phi, r^a)$ *of the modulated traveling wave* $\mathcal{P}_{\mathrm{MTW}}(\nu^\phi, r^a)$ *is close to* $T_{\mathrm{MTW}}(0) = \frac{2\pi}{|\omega_0^{\mathrm{rot}}|}$, *and the translation drift* $\gamma_{\mathrm{MTW}}(\nu^\phi, r^a) = (0, a_{\mathrm{MTW}}(\nu^\phi, r^a))$ *of the modulated traveling wave at* $x_{\mathrm{MTW}}(\nu^\phi, r^a)$ *satisfies* $a_{\mathrm{MTW}}(\nu^\phi, 0) = 0$.

Proof. The rotating wave $\mathrm{SE}(2)x_0$ is treated as a periodic orbit of period $T_0 = \frac{2\pi}{\omega_0^{\mathrm{rot}}}$. Introduce polar coordinates $\nu^a = (r^a \cos\phi^a, r^a \sin\phi^a)$. Then (3.23) implies that $r^a = \|\nu^a\|_2$ is a conserved quantity (Casimir) of the slice equation. Since the slice equation also conserves energy, the set

$$\mathcal{N}_{E,r^a} = \{(\nu, w) \in \mathcal{N},\ h(\nu, w) = E,\ \|\nu^a\|_2 = r^a\}$$

is flow-invariant. From $\dot{\phi}^a = \mathrm{D}_{\nu^\phi} h(\nu, w) \approx \omega_0^{\mathrm{rot}} \neq 0$ for $(\nu, w) \approx 0$ it can be deduced that for $E \approx E_0 = H(x_0)$, $r^a > 0$, $r^a \approx 0$, the section

$$\mathcal{S}_{E,r^a} = \{(\nu, w) \in \mathcal{N}_{E,r^a},\ \phi^a = 0,\ (\nu, w) \approx 0\}$$

is tranversal to the flow in $\mathcal{N}_{E,r^a}$. Let $\Pi(E, r^a, \cdot) : \mathcal{S}_{E,r^a} \to \mathcal{S}_{E,r^a}$ be the Poincaré map to the Poincaré section $\mathcal{S}_{E,r^a}$. Since $\mathrm{D}_{(\nu,w)} h(0) = (\xi_0, 0)$ with $\xi_0^\phi = \omega_0^{\mathrm{rot}} \neq 0$, the sections $\mathcal{S}_{E,r^a}$, $E \approx E_0$, $r^a \approx 0$, can be parametrized as

$$\mathcal{S}_{E,r^a} = \{(\nu^\phi, w),\ \nu^\phi = \nu^\phi(E, r^a, w),\ w \in \mathcal{N}_1\}.$$

Hence $\Pi(E, r^a, \cdot)$ can be considered as a map from $\mathcal{N}_1$ to itself. By assumption $ki\omega_0^{\mathrm{rot}}$, $k \in \mathbb{Z}$, is not an eigenvalue of $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h(0)$. Therefore,

$$\mathrm{D}_w \Pi(E_0, 0, 0) - \mathrm{id} = \exp\left(\frac{2\pi}{|\omega_0^{\mathrm{rot}}|} \mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h(0)\right) - \mathrm{id}$$

is invertible, and so there is a fixed point $w(E, r^a)$ of $\Pi(E.r^a, \cdot)$ for each $E \approx E_0$, $r^a \approx 0$. As $\mathrm{D}_{\nu^\phi} h(0) = \omega_0^{\mathrm{rot}} \neq 0$, this family of fixed points can be parametrized by ν^ϕ and r^a instead of E and r^a.

The periodic orbits of the slice equation through $v(\nu^\phi, r^a) = (\nu(\nu^\phi, r^a), w(\nu^\phi, r^a))$, where $\nu(\nu^\phi, r^a) = (\nu^\phi, r^a, 0)$, correspond to relative periodic orbits $\mathcal{P}_{\mathrm{MTW}}(\nu^\phi, r^a)$ of (3.1) through $x_{\mathrm{MTW}}(\nu^\phi, r^a) \simeq (\mathrm{id}, v(\nu^\phi, r^a))$ with momentum

$$\mathbf{J}(x_{\mathrm{MTW}}(\nu^\phi, r^a)) = \mu_0 + \nu(\nu^\phi, r^a) = (\mu_0^\phi + \nu^\phi, r^a, 0);$$

see (3.18). For vanishing linear momentum $r^a = 0$ they reduce to the rotating waves $\mathrm{SE}(2)x_{\mathrm{RW}}(\nu^\phi)$ from Proposition 4.2.

By (3.17), any RPO with drift $\gamma = (\phi, a)$ and momentum μ satisfies $(\mathrm{Ad}_\gamma^*)^{-1}\mu = \mu$, with $(\mathrm{Ad}_\gamma^*)^{-1}$ as in (3.9). Because of (3.9), the condition $\mu^a \neq 0$ implies $\phi = 0$ so that the RPOs $\mathcal{P}_{\mathrm{MTW}}(\nu^\phi, r^a)$ are modulated traveling waves for $r^a \neq 0$. ∎

4.3. Bifurcation of modulated rotating waves. In this section the existence of modulated rotating waves near elliptic rotating waves is proved by the Lyapunov center theorem.

Definition 4.4. *A relative equilibrium Γx_0, $x_0 \simeq (\mathrm{id}, (\nu, w) = (0, 0))$, of a Γ-equivariant Hamiltonian system (3.1) is called* elliptic *if all eigenvalues of the linearization $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h(0)$ of the $\dot{w}$-dynamics of (3.19) lie in $\mathrm{i}\mathbb{R} \setminus \{0\}$ and* nonresonant *if all its eigenvalues are simple and no eigenvalue $\mathrm{i}\omega_j$ is an integer multiple of another eigenvalue $\mathrm{i}\omega_k$ for $\omega_j \neq \omega_k$.*

Note that any stable relative equilibrium is elliptic and that relative equilibria are elliptic for an open range of parameters (until a Hamiltonian Hopf bifurcation of the $\dot{w}$-equation of (3.19) occurs).

Definition 4.5. *Let Γx_0, $x_0 \simeq (\mathrm{id}, (\nu, w) = (0, 0))$, be an elliptic relative equilibrium of (3.1), and denote the eigenvalues of $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h(0)$ by $\pm\mathrm{i}\omega_j$, $j = 1, \dots, d$, $d := \frac{1}{2} \dim \mathcal{N}_1$. The signs of the normal frequencies ω_j are chosen such that*

$$h(0, w) = \sum_{j=1}^{d} \frac{\omega_j}{2} \langle w_j, w_j \rangle + O(\|w\|^3). \tag{4.2}$$

Here $w = (w_1, \dots, w_d)$, $w_j \in \mathbb{R}^2$, are suitable coordinates on $\mathcal{N}_1$. The sign of ω_j is called the Krein signature *of ω_j. There is an $m : n$-*resonance *between the normal frequencies ω_j and ω_k if $m\omega_j = n\omega_k$, $m, n \in \mathbb{Z}$.*

Proposition 4.6. *Let $\mathrm{SE}(2)x_0$ be a nonresonant elliptic rotating wave of an $\mathrm{SE}(2)$-equivariant Hamiltonian system (3.1) with Ad^*-equivariant momentum map. Let ω_0^{rot} be its rotation frequency, let $H(x_0) = E_0$ be its energy, and let $\mu_0 = \mathbf{J}(x_0)$ be its momentum at x_0. Denote the eigenvalues of $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h(0)$ by $\pm\mathrm{i}\omega_j$, $j = 1, \dots, \frac{1}{2} \dim \mathcal{M} - 3$. Then there are $(\frac{1}{2} \dim \mathcal{M} - 3)$-many two-dimensional families $\mathcal{P}_j(\nu^\phi, s)$ of RPOs, $j = 1, \dots, \frac{1}{2} \dim \mathcal{M} - 3$, of (3.1), where*

$s \geq 0$, $\nu^\phi \approx 0$, with angular momentum $\mu_0^\phi + \nu^\phi$, with vanishing linear momentum, with energy $E = H(x_{\mathrm{RW}}(\nu^\phi)) \pm s^2$ (depending on the Krein signature of ω_j), and with relative periods $T_j(\nu^\phi, s)$ such that $T_j(0,0) = 2\pi/|\omega_j|$. Moreover, there are smooth functions $x_j(\nu^\phi, s)$ with $x_j(\nu^\phi, s) \in \mathcal{P}_j(\nu^\phi, s)$, $x_j(\nu^\phi, 0) = x_{\mathrm{RW}}(\nu^\phi)$. If $\omega_0^{\mathrm{rot}}/\omega_j \notin \mathbb{Z}$ for all normal frequencies ω_j, then all these RPOs are proper modulated rotating waves for $(\nu^\phi, s) \approx 0$. Proper modulated rotating waves do not persist to nonzero linear momenta.

Proof. First note that proper modulated rotating waves have a drift symmetry $\gamma = (\phi, a)$ with $\phi \neq 0 \bmod 2\pi$. This implies, because of (3.17) and (3.9), that the linear momentum ν^a of a proper modulated rotating wave vanishes. Equation (3.23) implies that $\dot\nu^\phi \equiv 0$, $\dot\nu^a \equiv 0$ at $\nu^a = 0$. So proper modulated rotating waves near x_0 correspond to nonlinear normal modes of the ν^ϕ-dependent $\dot w$-equation of (3.23) at $\nu^a = 0$. Note that $\mathcal{N}_1$ has dimension

$$\dim \mathcal{N}_1 = \dim \mathcal{M} - 2 \dim \mathrm{SE}(2) = \dim \mathcal{M} - 6.$$

By the Lyapunov center theorem (see, e.g., [19]) there are $d = \frac{\dim \mathcal{N}_1}{2}$ families of periodic orbits $w_j(\nu^\phi, s)$, $j = 1, \dots, d$, of the $\dot w$-equation of (3.23) such that $w_j(\nu^\phi, 0) = w_{\mathrm{RW}}(\nu^\phi)$, with $w_{\mathrm{RW}}(\nu^\phi)$ from Proposition 4.2. Let $\mathrm{i}\omega_j(\nu^\phi)$ be the eigenvalue of $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h((\nu^\phi, 0, 0), w_{\mathrm{RW}}(\nu^\phi))$ such that $\omega_j(0) = \omega_j$. Since $\partial_s w_j(\nu^\phi, 0)$ lies in the real eigenspace of $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h((\nu^\phi, 0, 0), w_{\mathrm{RW}}(\nu^\phi))$ to the eigenvalue $\mathrm{i}\omega_j(\nu^\phi)$ (see, e.g., [19]), and since $\mathrm{D}_w h((\nu^\phi, 0, 0), w_{\mathrm{RW}}(\nu^\phi)) = 0$, the energy of the periodic orbits is

$$h((\nu^\phi, 0, 0), w_j(\nu^\phi, s)) = H(x_{\mathrm{RW}}(\nu^\phi)) + \omega_j(\nu^\phi) s^2 + O(s^3).$$

Therefore, s can be rescaled to achieve that the periodic orbit $w_j(\nu^\phi, s)$ has energy $H(x_{\mathrm{RW}}(\nu^\phi))$ $\pm s^2$ depending on the sign of ω_j; see (4.2). Then $x_j(\nu^\phi, s) = (\mathrm{id}, (\nu^\phi, 0, 0), w_j(\nu^\phi, s))$ lies on an RPO $\mathcal{P}_j(\nu^\phi, s)$ of (3.1). Its momentum is

$$\mathbf{J}(x_j(\nu^\phi, s)) = (\mu_0^\phi + \nu^\phi, 0, 0)$$

by (3.18). The drift symmetry $\gamma_j(\nu^\phi, s) = (\phi_j(\nu^\phi, s), a_j(\nu^\phi, s))$ of the RPO at $x_j(\nu^\phi, s)$ satisfies $\phi_j(0,0) = 2\pi\omega_0^{\mathrm{rot}}/\omega_j$, and so $\phi_j(0,0) \neq 0 \bmod 2\pi$ if $\omega_0^{\mathrm{rot}}/\omega_j \notin \mathbb{Z}$. In this case $\phi_j(\nu^\phi, s) \neq 0 \bmod 2\pi$ for $(\nu^\phi, s) \approx 0$, and the RPOs $\mathcal{P}_j(\nu^\phi, s)$ are indeed proper modulated rotating waves for $(\nu^\phi, s) \approx 0$. ∎

Example 4.7. Let us now study a Hamiltonian analogue of the meandering transition for point vortices with vanishing total circulation $\mathcal{K} = 0$. In this case the momentum map of the point vortex system (3.3) is Ad^*-equivariant; cf. Remark 3.4. Let us start with a configuration of rigidly rotating point vortices. Such a configuration is a rotating wave of (3.3). Synge [29] and later Aref [1] showed the existence of rotating waves of 3 vortices with vanishing total circulation. Patrick [25] constructed rotating wave solutions with vanishing total circulation for $N \geq 3$ vortices. Let us assume that the rotating wave is nondegenerate and that the nonresonance condition (4.1) is satisfied (this assumption is trivially satisfied for 3 vortices, since then $\mathcal{N}_1 = \{0\}$). Then it persists as a translating and precessing configuration, i.e., as a modulated traveling wave, to nonzero linear momentum. Moreover if the rotating wave is elliptic and nonresonant, then there are $(\frac{1}{2} \dim \mathcal{M} - 3) = (N - 3)$ different 2-parameter families of rotating and precessing configurations of vortices nearby, which are modulated

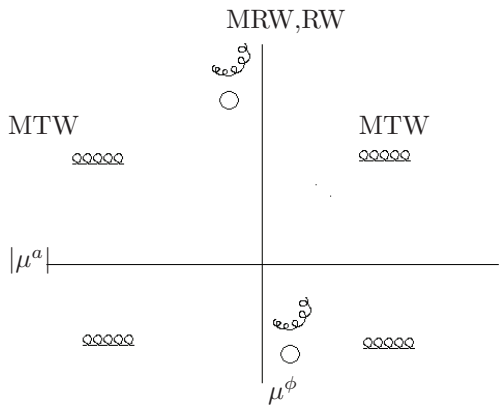


Figure 6. *Bifurcation diagram for the Hamiltonian meandering transition of point vortex dynamics in the case of vanishing circulation. RW: Rotating waves. MRW: Modulated rotating waves. MTW: Modulated traveling waves.*

rotating waves (Patrick [25] shows stability of the rotating waves of 4 vortices with vanishing total circulation, which he constructed. This implies that these rotating waves are elliptic, and, since $\mathcal{N}_1$ is 2-dimensional, they are also nonresonant).

Figure 6 shows the bifurcation diagram of the Hamiltonian meandering transition for point vortex dynamics in the case of an Ad^*-equivariant momentum map (for momentum maps with cocycle see section 4.4). This diagram should be compared with the corresponding bifurcation diagram of the dissipative meandering transition, Figure 4.

Note that, in contrast to the meandering/drifting transition in dissipative systems, here modulated traveling waves are the typical scenario as momentum is varied. Modulated rotating waves occur only for zero-linear momentum and so are a codimension two phenomenon in the three parameters' angular and linear momentums.

Example 4.8. Another example where a Hamiltonian meandering transition occurs is the Kirchhoff model of an underwater vehicle; see [13, 15]. In this case the configuration space is the Euclidean group $\mathrm{SE}(3) = \mathrm{SO}(3) \ltimes \mathbb{R}^3$ of three-dimensional space modeling the angle and position of the underwater vehicle, and the phase space $\mathcal{M} = \mathcal{T}^*\mathrm{SE}(3)$ is 12-dimensional. In the case of noncoincident centers of gravity and buoyancy the symmetry group is

$$\Gamma = \mathrm{SO}_3(2) \ltimes \mathbb{R}^3 = \mathrm{SE}(2) \times \mathbb{R}_3.$$

Here $\mathrm{SO}_3(2)$ denotes the group of rotations around the axis of gravity, which is chosen as the third coordinate axis (i.e., as the e_3-axis), and $\mathbb{R}_3$ is the group of translations along the e_3-axis. Near a vertically falling and spinning relative equilibrium the dynamics is given by the slice equations near a rotating wave of a Euclidean equivariant system (3.23), but now there is an additional equation

$$\dot{\nu}_3^a \equiv 0$$

in the slice and a corresponding equation

$$\dot{a}_3 = \mathrm{D}_{\nu_3^a} h(\nu, w)$$

for the group dynamics, where $\nu = (\nu^\phi, \nu^a)$, $\nu^a = (\nu^a_1, \nu^a_2, \nu^a_3)$. Assume that this relative equilibrium is nondegenerate and satisfies the nonresonance condition (4.1). Then it persists as translating RPO with nonvertical linear drift to horizontal linear momentum. Let the relative equilibrium be elliptic and nonresonant (from [15, section 4.4.2] it follows that these conditions are satisfied at least for an open range of parameters). Then there are two 3-parameter families of RPOs, parametrized by (ν^ϕ, ν^a_3, s), which fall, rotate, and precess.

4.4. Hamiltonian meandering transition for momentum maps with cocycle. In this section the analogue of the meandering transition of dissipative systems is considered for Hamiltonian systems where the momentum map has a nonvanishing cocycle of the form that occurs in point vortex dynamics. The limiting behavior for a vanishing cocycle is studied, and in this way the meandering transition for momentum maps with cocycle is related to the results of sections 4.1–4.3 on the meandering transition for Hamiltonian systems with Ad^*-equivariant momentum map.

Let us consider a parameter dependent SE(2)-symmetric Hamiltonian system (3.22). Assume that the symplectic form $\Omega(\mathcal{K})$, the Hamiltonian $H(\cdot, \mathcal{K})$, and the momentum map $\mathbf{J}(\cdot, \mathcal{K}) : \mathcal{M} \to \mathrm{se}(2)^*$ depend smoothly on a parameter $\mathcal{K}$. Moreover, assume that the momentum map is Ad^*-equivariant for $\mathcal{K} = 0$ and has the cocycle (3.11) for $\mathcal{K} \neq 0$; see (3.10), (3.12). An example of such a momentum map is the momentum map (3.4) for the point vortex dynamics (3.3).

As in the analysis of the Hamiltonian meandering transition for Ad^*-equivariant momentum maps (see section 4.1), first the persistence of rotating waves to nearby momentum values is studied. Moreover, the behavior of the rotating waves in the limit of vanishing cocycle is analyzed in the following theorem.

Theorem 4.9. *Consider a Hamiltonian system with* SE(2)*-symmetry for which the momentum map* $\mathbf{J}(\cdot, \mathcal{K})$ *has a nonvanishing cocycle satisfying* (3.11), (3.12). *Then the following hold:*

(a) *For* $\mathcal{K} \neq 0$ *all relative equilibria are rotating waves. Any nondegenerate rotating wave* $\mathrm{SE}(2)x_0$ *persists to all nearby momentum values as a one-parameter family* $\mathrm{SE}(2)x_{\mathrm{RW}}(\nu)$, $\nu \in (\mathbf{g}^K_{\mu_0})^*$, $x_{\mathrm{RW}}(0) = x_0$.

(b) *The center of rotation* $c(\mathcal{K})$ *of any smooth family* $\mathrm{SE}(2)x_{\mathrm{RW}}(\mathcal{K})$, $\mathcal{K} \neq 0$, $\mathcal{K} \approx 0$, *of rotating waves with fixed linear momentum* $\mu^a \neq 0$ *and rotation frequency* $\omega^{\mathrm{rot}}(\mathcal{K})$, *such that* $\lim_{\mathcal{K}\to 0} \omega^{\mathrm{rot}}(\mathcal{K}) = \omega^{\mathrm{rot}}_0 \neq 0$, *tends to infinity, as* $\mathcal{K} \to 0$, *according to*

$$\|c(\mathcal{K})\| = \frac{\|\mu^a\|}{\mathcal{K}}.$$

(c) *Assume that there is a nondegenerate rotating wave* $\mathrm{SE}(2)x_0$ *at* $\mathcal{K} = 0$ *with momentum* $\mu_0 = (\mu^\phi_0, 0)$ *and rotation frequency* $\omega^{\mathrm{rot}}_0 \neq 0$. *Then this rotating wave can be continued to a rotating wave* $\mathrm{SE}(2)x_{\mathrm{RW}}(\nu^\phi, \mathcal{K})$ *for small* $\mathcal{K} \neq 0$, $\nu^\phi \neq 0$, *such that* $\mathbf{J}^\phi(x_{\mathrm{RW}}(\nu^\phi, \mathcal{K}), \mathcal{K}) = \mu^\phi_0 + \nu^\phi$ *and* $\mathbf{J}^a(x_{\mathrm{RW}}(\nu^\phi, \mathcal{K}), \mathcal{K}) = 0$.

Part (b) of this theorem implies that for $\mathcal{K} = 0$ rotating waves exist only for vanishing linear momentum; see Proposition 4.2.

Proof of Theorem 4.9. (a) If the momentum map for a Hamiltonian system with SE(2)-symmetry has a nonvanishing cocycle, then $\Gamma^\kappa_\mu \simeq \mathrm{SO}(2)$ for all $\mu \in \mathrm{se}(2)^*$, as shown in Remark 3.4. Hence $\mathbf{g}^K_\mu \simeq \mathrm{so}(2)$ for all $\mu \in \mathrm{se}(2)^*$, and by (3.16) all relative equilibria are

rotating waves. Moreover, in Example 3.9 it was shown that $\dot\nu \equiv 0$ in (3.21), $\nu \in (\mathbf{g}_\mu^K)^*$. For a nondegenerate rotating wave $\mathrm{SE}(2)x_0$ the matrix $\mathrm{D}_w^2 h(0)$ is invertible. Therefore, there is a path $w_{\mathrm{RW}}(\nu)$ of equilibria of the $\dot w$-equation near $w_{\mathrm{RW}}(0) = 0$. These equilibria correspond to rotating waves $\mathrm{SE}(2)x_{\mathrm{RW}}(\nu)$, $x_{\mathrm{RW}}(\nu) \simeq (\mathrm{id}, \nu, w_{\mathrm{RW}}(\nu))$, of (3.1) for all nearby momenta $\mathrm{SE}(2)\mathbf{J}(x_{\mathrm{RW}}(\nu), \mathcal{K}) = \mathrm{SE}(2)(\mu_0 + \nu)$, where $\mu_0 = \mathbf{J}(x_0, \mathcal{K})$.

(b) Let $\mathrm{SE}(2)x_{\mathrm{RW}}(\mathcal{K})$ be an rotating wave with momentum μ and drift velocity $\xi = \xi(\mathcal{K}) = (\xi^\phi, \xi^a)(\mathcal{K})$. Differentiating (3.15) and identifying $\mathbb{R}^2$ with $\mathbb{C}$ and $\mathrm{so}(2)$ with $\mathbb{R}$, one obtains that

$$(\xi^\phi, \xi^a) \in \mathbf{g}_\mu^K \quad \Leftrightarrow \quad \mathcal{K}\xi^a = \xi^\phi \mu^a. \tag{4.3}$$

Let $(\phi, a)x = e^{\mathrm{i}\phi}x + a$, $x \in \mathbb{R}^2 \simeq \mathbb{C}$. Then $x(t) = \exp(t(\xi^\phi, \xi^a))x_0$ satisfies the differential equation $\dot x(t) = \mathrm{i}\xi^\phi x(t) + \xi^a$. Solving this and setting $t = 1$, one gets

$$\exp(\xi^\phi, \xi^a)x_0 = e^{\mathrm{i}\xi^\phi}x_0 + \frac{1}{\mathrm{i}\xi^\phi}(e^{\mathrm{i}\xi^\phi} - 1)\xi^a.$$

Therefore, $\exp(\xi^\phi, \xi^a)$ is a rotation with center

$$c = \mathrm{i}\frac{\xi^a}{\xi^\phi} = \mathrm{i}\frac{\mu^a}{\mathcal{K}} = \frac{1}{\mathcal{K}}(-\mu_2^a, \mu_1^a)^T. \tag{4.4}$$

This proves (b).

(c) Let us reduce only by the $\mathrm{SO}(2) \times \{0\}$-symmetry. Then the system

$$\dot{\tilde\nu} = 0, \quad \dot{\tilde w} = \mathbb{J}_{\tilde{\mathcal{N}}_1} \mathrm{D}_{\tilde w}\tilde h(\tilde\nu, \tilde w, \mathcal{K})$$

is obtained on the slice $\tilde{\mathcal{N}}$ transverse to the $\mathrm{SO}(2)$-orbit $\mathrm{SO}(2)x_0$ at x_0, with $\tilde\nu = \nu^\phi$. By (3.8), (3.9) the matrix $\mathrm{ad}_{\xi_0} = -\mathrm{ad}_{\xi_0}^*$ has simple eigenvalues 0 and $\pm\mathrm{i}\omega_0^{\mathrm{rot}}$. By (2.4) and (3.20), at $\mathcal{K} = 0$ the linearization of the slice equation $\mathrm{D}f_{\tilde{\mathcal{N}}}(0,0)$ has one simple eigenvalue 0 corresponding to the equation $\dot{\tilde\nu} = 0$. Therefore, $\mathrm{D}_{\tilde w}^2\tilde h(0,0,0)$ is invertible, and so the rotating wave $\mathrm{SO}(2)x_0$ is nondegenerate. By Remark 3.7 the $\dot{\tilde w}$-equation is smooth in $\mathcal{K}$. So there is a smooth two-parameter family $\tilde w_{\mathrm{RW}}(\tilde\nu, \mathcal{K})$ of equilibria of the $\dot{\tilde w}$-equation. This gives a family $\mathrm{SE}(2)x_{\mathrm{RW}}(\nu^\phi, \mathcal{K})$, $x_{\mathrm{RW}}(\nu^\phi, \mathcal{K}) \simeq (\mathrm{id}, \nu^\phi, \tilde w_{\mathrm{RW}}(\nu^\phi, \mathcal{K}))$, of rotating waves of (3.1). By (3.18) their angular momentum is $\mathbf{J}^\phi(x_{\mathrm{RW}}(\nu^\phi, \mathcal{K}), \mathcal{K}) = \mu_0^\phi + \nu^\phi$. Since only a reduction by the $\mathrm{SO}(2)$-symmetry has been carried out, the rotating waves through $x_{\mathrm{RW}}(\nu^\phi, \mathcal{K})$ have drift velocities $\xi_{\mathrm{RW}}(\nu^\phi, \mathcal{K})$ with $\xi_{\mathrm{RW}}^a(\nu^\phi, \mathcal{K}) = 0$. Moreover, $\xi_{\mathrm{RW}}^\phi(\nu^\phi, \mathcal{K}) \approx \omega_0^{\mathrm{rot}} \neq 0$ for $\mathcal{K} \approx 0$. Therefore, (4.3) implies that $\mathbf{J}^a(x_{\mathrm{RW}}(\nu^\phi, \mathcal{K}), \mathcal{K}) = 0$. ■

The next proposition shows that in the case of a momentum map with cocycle (3.11) all RPOs are modulated rotating waves.

Proposition 4.10. *Consider a Hamiltonian system with* $\mathrm{SE}(2)$*-symmetry for which the momentum map* $\mathbf{J}(\cdot, \mathcal{K})$ *has a nonvanishing cocycle satisfying* (3.11), (3.12)*. Then the following hold:*

(a) *For* $\mathcal{K} \neq 0$ *all RPOs are modulated rotating waves.*

(b) *Any smooth family* $\mathcal{P}(\mathcal{K})$ *of proper modulated rotating waves of* (3.1)*, i.e., of RPOs with drift symmetry* $\gamma(\mathcal{K}) = (\phi(\mathcal{K}), a(\mathcal{K}))$*, where* $\phi(\mathcal{K}) \neq 0 \bmod 2\pi$ *for all* $\mathcal{K} \neq 0$*,* $\mathcal{K} \approx 0$*, with fixed linear momentum* $\mu^a \neq 0$ *has a center of rotation* $c(\mathcal{K})$ *diverging to* ∞ *for* $\mathcal{K} \to 0$ *according to* $\|c(\mathcal{K})\| = \frac{\|\mu^a\|}{\mathcal{K}}$.

Proof. (a) Similarly as in the proof of Theorem 4.9 (a), this follows from the fact that $\Gamma_\mu^\kappa \simeq \mathrm{SO}(2)$ for all $\mu \in \mathrm{se}(2)^*$, as shown in Remark 3.4. This, together with (3.17), implies that all RPOs are modulated rotating waves.

(b) This statement is proved similarly as Theorem 4.9 (b). By (3.15), the center of rotation $c(\mathcal{K}) = R_{\phi(\mathcal{K})}c(\mathcal{K}) + a(\mathcal{K})$ of the drift symmetry $\gamma(\mathcal{K}) = (\phi(\mathcal{K}), a(\mathcal{K}))$, $\phi(\mathcal{K}) \neq 0 \bmod 2\pi$, of the RPO is given by (4.4), and so $c(\mathcal{K}) \to \infty$ as $\mathcal{K} \to 0$ for $\mu^a \neq 0$. ∎

In the next theorem the transition from rotating waves to modulated rotating waves and modulated traveling waves is studied in the limit $\mathcal{K} \to 0$.

Theorem 4.11. *Let* $\mathrm{SE}(2)x_{\mathrm{RW}}(\mathcal{K})$ *be a nondegenerate rotating wave of a Hamiltonian system* (3.1) *which has a momentum map* $\mathbf{J}(\cdot, \mathcal{K})$ *with cocycle satisfying* (3.10), (3.11)*. Fix the momentum* $\mu_0 = \mathbf{J}(x_{\mathrm{RW}}(\mathcal{K}), \mathcal{K})$ *of* $x_{\mathrm{RW}}(\mathcal{K})$ *independent of* $\mathcal{K}$*. Then the following hold true.*

(a) *Fix* $\mathcal{K} \neq 0$*. Assume that the rotating wave* $\mathrm{SE}(2)x_{\mathrm{RW}}(\mathcal{K})$ *is elliptic and nonresonant in the sense of Definition 4.4, with normal frequencies* ω_j*,* $j = 1, \dots, \frac{1}{2}\dim\mathcal{M} - 2$*. Then there are* $(\frac{1}{2}\dim\mathcal{M} - 2)$ *many two-dimensional families* $\mathcal{P}_j(\nu, s, \mathcal{K})$ *of modulated rotating waves such that there are functions* $x_j(\nu, s, \mathcal{K}) \in \mathcal{P}_j(\nu, s, \mathcal{K})$ *which are smooth in* $s \geq 0$*,* $\mathcal{K}$*, and* $\nu \in (\mathbf{g}_{\mu_0}^K)^* \simeq \mathrm{so}(2)^*$ *with* $x_j(\nu, 0, \mathcal{K}) = x_{\mathrm{RW}}(\nu, \mathcal{K})$*. Here* $x_{\mathrm{RW}}(\nu, \mathcal{K})$ *lies on a rotating wave with momentum* $\mu_0 + \nu$*. The modulated rotating wave through* $x_j(\nu, s, \mathcal{K})$ *has relative period* $T_j(\nu, s, \mathcal{K})$ *with* $T_j(0, 0, \mathcal{K}) = 2\pi/|\omega_j|$*, energy* $H(x_j(\nu, s, \mathcal{K}), \mathcal{K}) = H(x_{\mathrm{RW}}(\nu, \mathcal{K}), \mathcal{K}) \pm s^2$ *(depending on the Krein signature of* ω_j*), and momentum* $\mathbf{J}(x_j(\nu, s, \mathcal{K}), \mathcal{K}) = \mu_0 + \nu$*,* $j = 1, \dots, \frac{1}{2}\dim\mathcal{M} - 2$.

(b) *Assume that there is a rotating wave* $\mathrm{SE}(2)x_0$ *at* $\mathcal{K} = 0$ *with rotation frequency* ω_0^{rot} *which is elliptic and nonresonant in the sense of Definition 4.4, and that* $\omega_0^{\mathrm{rot}}/\omega_j \notin \mathbb{Z}$ *for all eigenvalues* $\mathrm{i}\omega_j$ *of* $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}_w^2 h(0, 0, 0)$*. Then the* $(\frac{1}{2}\dim\mathcal{M} - 3)$ *families* $\mathcal{P}_j(\nu^\phi, s)$ *of modulated rotating waves near* $\mathrm{SE}(2)x_0$ *from Proposition 4.6 at* $\mathcal{K} = 0$ *can be continued to small* $\mathcal{K} \neq 0$ *and correspond to families* $\mathcal{P}_j(\nu, s, \mathcal{K})$ *from part* (a) *with* $\nu = \nu^\phi$.

(c) *Assume that the rotating wave* $\mathrm{SE}(2)x_0$ *at* $\mathcal{K} = 0$ *is elliptic and nonresonant and that* $\omega_j/\omega_0^{\mathrm{rot}} \notin \mathbb{Z}$ *for all eigenvalues* $\mathrm{i}\omega_j$ *of* $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}_w^2 h(0, 0, 0)$*. Then for* $\mathcal{K} \neq 0$*,* $\mathcal{K} \approx 0$*, one of the families* $\mathcal{P}(\nu, s, \mathcal{K})$ *from* (a) *corresponds to an eigenvalue* $\mathrm{i}\omega(\mathcal{K})$ *of* $\mathbb{J}_{\mathcal{N}_1(\mathcal{K})}\mathrm{D}_w^2 h(0, w_{\mathrm{RW}}(0, \mathcal{K}), \mathcal{K})$ *which depends smoothly on* $\mathcal{K}$ *such that* $\omega(0) = \omega_0^{\mathrm{rot}}$*. Here* ω_0^{rot} *is the rotation frequency of the rotating wave* $\mathrm{SE}(2)x_0$ *at* $\mathcal{K} = 0$*,* $\mathcal{N}_1(\mathcal{K})$ *is the symplectic normal space at the rotating wave through* $x_{\mathrm{RW}}(0, \mathcal{K}) \simeq (\mathrm{id}, 0, w_{\mathrm{RW}}(0, \mathcal{K})) \in \Gamma \times \mathcal{N}_0(\mathcal{K}) \oplus \mathcal{N}_1(\mathcal{K})$*,* $\mathcal{N}_0(\mathcal{K}) \simeq (\mathbf{g}_{\mu_0}^K)^*$*, and* $h(\nu, w, \mathcal{K})$ *is the Hamiltonian in symmetry-adapted coordinates at* $x_{\mathrm{RW}}(0, \mathcal{K})$ *for the cocycle parameter* $\mathcal{K}$*. As* $\mathcal{K} \to 0$*, this family converges to the family of modulated traveling waves from Theorem 4.3.*

Proof. (a) Fix $\mathcal{K} \neq 0$. Then, as shown in Example 3.9, $\mathcal{N}_0 = \mathcal{N}_0(\mathcal{K})$ is one-dimensional, and so $\mathcal{N}_1 = \mathcal{N}_1(\mathcal{K})$ has dimension $\dim\mathcal{M} - 4$. By the Lyapunov center theorem (see, e.g., [19]) applied to the ν-dependent $\dot{w}$-equation on $\mathcal{N}_1 = \mathcal{N}_1(\mathcal{K})$, there are $\frac{1}{2}\dim\mathcal{N}_1 = (\dim\mathcal{M} - 4)/2$ many families of nonlinear normal modes through $w_j(\nu, s, \mathcal{K})$. These give families of RPOs $\mathcal{P}_j(\nu, s, \mathcal{K})$ of (3.22) through $x_j(\nu, s, \mathcal{K}) \simeq (\mathrm{id}, \nu, w_j(\nu, s, \mathcal{K}))$ with momentum $\mathbf{J}(x_j(\nu, s, \mathcal{K})) = \mu_0 + \nu$; see (3.18). The statement about the energy of the RPOs is proved as in Proposition 4.6.

It was shown in Remark 3.4 that $\Gamma_\mu^\kappa \simeq \mathrm{SO}(2)$ for all $\mu \in \mathrm{se}(2)^*$. Hence (3.17) implies that all these RPOs are modulated rotating waves. This proves part (a).

(b) Let us reduce by SO(2)-symmetry only. The dynamics on the slice $\tilde{\mathcal{N}}(\mathcal{K})$ is then $\dot{\tilde{\nu}} \equiv 0$, $\tilde{\nu} = \nu^\phi \in \mathrm{so}(2)^*$, and $\dot{\tilde{w}} = \mathbb{J}_{\tilde{\mathcal{N}}_1(\mathcal{K})} \mathrm{D}_{\tilde{w}} \tilde{h}(\tilde{\nu}, \tilde{w}, \mathcal{K})$. Because of (3.8), (3.9), the matrix $\mathrm{ad}_{\xi_0} = -\mathrm{ad}_{\xi_0}^*$ has simple eigenvalues $0, \pm \mathrm{i}\omega_0^{\mathrm{rot}}$. Therefore, by (2.4) and (3.20), the equilibrium $\tilde{w} = 0$ of the $\dot{\tilde{w}}$-equation at $\tilde{\nu} = 0$, $\mathcal{K} = 0$, is elliptic; moreover, its linearization has, in addition to the eigenvalues $\pm \mathrm{i}\omega_j$, $j = 1, \dots, \frac{1}{2}\dim \mathcal{N}_1$, double eigenvalues $\pm \mathrm{i}\omega_0^{\mathrm{rot}}$. Let us apply the Lyapunov center theorem to the $\dot{\tilde{w}}$-equation. This gives relative normal modes $\tilde{w}_j(\tilde{\nu}, s, \mathcal{K})$ for all nonresonant normal frequencies. These correspond to all imaginary eigenvalues $\pm \mathrm{i}\omega_j$ with eigenvectors in the original symplectic normal space $\mathcal{N}_1$ for the full SE(2)-group action. Since by Remark 3.7 the $\dot{\tilde{w}}$-equation is smooth in $\mathcal{K}$, this gives $\frac{1}{2}\dim \mathcal{N}_1 = (\dim \mathcal{M} - 6)/2$ smooth families $\mathcal{P}_j(\nu^\phi, s, \mathcal{K})$ of RPOs through $x_j(\nu^\phi, s, \mathcal{K}) \simeq (\mathrm{id}, \tilde{\nu}, \tilde{w}_j(\nu^\phi, s, \mathcal{K}))$, $j = 1, \dots \frac{1}{2}\dim \mathcal{N}_1$, with momentum $\mathbf{J}(x_j(\nu^\phi, s, \mathcal{K})) = (\mu_0^\phi + \nu^\phi, 0)$; see (3.18). These RPOs are modulated rotating waves, since the system was reduced only by SO(2)-symmetry.

(c) Let $\mathcal{K} = 0$. Then (3.20) and (3.9) imply that the linearization of the $\dot{\nu}$-equation has eigenvalues $\pm \mathrm{i}\omega_0^{\mathrm{rot}}$ and 0, where the real eigenspace of $\pm \mathrm{i}\omega_0^{\mathrm{rot}}$ is given by $\{\nu = (0, \nu^a), \nu^a \in \mathbb{R}^2\} \subseteq \mathrm{se}(2)^*$. Compared to the slice equation (3.23) near rotating waves of momentum maps without cocycle, for $\mathcal{K} \neq 0$ the $\mathcal{N}_0(\mathcal{K})$ component of the slice $\mathcal{N}(\mathcal{K})$ is only one-dimensional instead of three-dimensional. Since $\mu_0^a = \mathbf{J}^a(x_0, 0) = 0$ and $\mu_0 = \mathbf{J}(x_{\mathrm{RW}}(\mathcal{K}), \mathcal{K})$ is fixed, Remark 3.9 implies that $\nu = \nu^\phi$ for $\mathcal{K} \neq 0$ and that then $\dot{\nu} \equiv 0$. For $\mathcal{K} \neq 0$ the eigenvalue $\pm \mathrm{i}\omega_0^{\mathrm{rot}}$ of $\mathrm{D}f_{\mathcal{N}}(0)$ perturbs into eigenvalues $\pm \mathrm{i}\omega(\mathcal{K})$ of $\mathrm{D}f_{\mathcal{N}(\mathcal{K})}(v_{\mathrm{RW}}(\mathcal{K}))$. Here $v_{\mathrm{RW}}(\mathcal{K})$ is the equilibrium of the slice equation at momentum μ_0 and cocycle parameter $\mathcal{K}$ corresponding to the relative equilibrium $x_{\mathrm{RW}}(\nu^\phi, \mathcal{K})$ at $\nu^\phi = 0$ from Theorem 4.9 (c). Since $\mathrm{D}f_{\mathcal{N}(\mathcal{K})}(v_{\mathrm{RW}}(\mathcal{K}))$ vanishes when restricted to $\mathcal{N}_0(\mathcal{K})$, it follows that $\pm \mathrm{i}\omega(\mathcal{K})$ are eigenvalues of $\mathrm{D}f_{\mathcal{N}_1(\mathcal{K})}(v_{\mathrm{RW}}(\mathcal{K}))$. By the Lyapunov center theorem and due to the nonresonance condition $\omega_j/\omega_0^{\mathrm{rot}} \notin \mathbb{Z}$, this gives one more family of nonlinear relative normal modes for $\mathcal{K} \neq 0$, $\mathcal{K} \approx 0$. It will now be shown that this additional family of RPOs converges to the family $\mathcal{P}(\nu^\phi, r^a)$ of modulated traveling waves from Theorem 4.3 as $\mathcal{K} \to 0$.

The proof of Theorem 4.3 can be extended to the case $\mathcal{K} \approx 0$, $\mathcal{K} \neq 0$. It is convenient to work in the slice coordinates $(\nu, w) \in \mathcal{N} = \mathcal{N}_0 \oplus \mathcal{N}_1$ at $\mathcal{K} = 0$, even when perturbations to cocycle parameters $\mathcal{K} \neq 0$ are considered. For any $\mathcal{K} \approx 0$ the dynamics on the slice $\mathcal{N}$ does not depend on γ and the energy $h(\nu, w, \mathcal{K})$ is still a conserved quantity. But note that the Poisson structure on $\mathcal{N}$ changes, and in particular $|\nu^a|^2$ is not a conserved quantity anymore. By (3.9), for $\mathcal{K} = 0$ the momentum group orbits $\mathrm{SE}(2)\mu$ are cylinders around the μ^ϕ-axis. For $\mathcal{K} \neq 0$, due to (3.10) and (3.11), they are paraboloids, centered along the μ^ϕ-axis. A Casimir, i.e., a function satisfying $C(\gamma \cdot_\kappa \mu) = C(\mu)$, where $\gamma \cdot_\kappa \mu$ is defined in (3.10), is given by $C(\mu^\phi, \mu^a) = \|\mu^a\|^2 + 2\mathcal{K}\mu^\phi$. Let $\mathbf{J}^a(\nu, w, \mathcal{K})$ and $\mathbf{J}^\phi(\nu, w, \mathcal{K})$ be the linear and angular momenta on the slice $\mathcal{N}$. Since the momentum map depends smoothly on $\mathcal{K}$, these maps are smooth in all variables. Then

$$C(\nu, w, \mathcal{K}) = \|\mathbf{J}^a(\nu, w, \mathcal{K})\|^2 + 2\mathcal{K}\mathbf{J}^\phi(\nu, w, \mathcal{K}) \tag{4.5}$$

is a conserved quantity for the slice equation. For $\mathcal{K} = 0$ one has $\mathbf{J}^a(\nu, w, \mathcal{K} = 0) = \nu^a$, $\mathbf{J}^\phi(\nu, w, \mathcal{K} = 0) = \nu^\phi$. Therefore, for $\mathcal{K} \approx 0$ the Casimir $C(\nu, w, \mathcal{K})$ is a small perturbation

of $C(\nu,w,\mathcal{K})|_{\mathcal{K}=0} = \|\nu^a\|^2$. The rotating waves through $x_{\mathrm{RW}}(\nu^\phi,\mathcal{K})$ from Theorem 4.9 (c) have energy $E_{\mathrm{RW}}(\nu^\phi,\mathcal{K}) = H(x_{\mathrm{RW}}(\nu^\phi,\mathcal{K}))$ with $\frac{\partial E_{\mathrm{RW}}}{\partial \nu^\phi}(0,0) = \omega_0^{\mathrm{rot}} \neq 0$. So they can be parametrized by $(E,\mathcal{K})$ instead of $(\nu^\phi,\mathcal{K})$ for $E \approx E_0 = H(x_0)$ and $\mathcal{K} \approx 0$. Let $x_{\mathrm{RW}}(E,\mathcal{K}) = (\mathrm{id}, \nu_{\mathrm{RW}}(E,\mathcal{K}), w_{\mathrm{RW}}(E,\mathcal{K})) \in \Gamma \times \mathcal{N}$ be the family of rotating waves near $x_0 \simeq (\mathrm{id},0,0)$. Then $\nu^a_{\mathrm{RW},1}(E,0) = 0$, $\nu^a_{\mathrm{RW},2}(E,0) = 0$ by Proposition 4.2. Since the slice equation conserves energy, the energy level sets $\mathcal{N}_{E,\mathcal{K}}$ of the slice $\mathcal{N}$ at cocycle parameter $\mathcal{K}$ are flow-invariant. Moreover, as $\mathrm{D}_{\nu^\phi} h(0) = \omega_0^{\mathrm{rot}}$, $\omega_0^{\mathrm{rot}} \neq 0$, they can be parametrized by

$$\begin{aligned}\mathcal{N}_{E,\mathcal{K}} &= \{(\nu,w) \in \mathcal{N}, \quad h(\nu,w,\mathcal{K}) = E\}\\ &= \{(\nu,w) \in \mathcal{N}, \quad \nu^\phi = \nu^\phi(E,\nu^a,w,\mathcal{K})\},\end{aligned}$$

where $E \approx E_0 = H(x_0)$, $\nu^a \approx 0$, $\mathcal{K} \approx 0$, $w \approx 0$. As in the proof of Theorem 4.3,

$$\begin{aligned}\mathcal{S}_{E,\mathcal{K}} = \{(\nu,w) \in \mathcal{N}, \quad \nu_2^a = \nu^a_{\mathrm{RW},2}(E,\mathcal{K}), \quad \nu_1^a > \nu^a_{\mathrm{RW},1}(E,\mathcal{K}),\\ \nu^\phi = \nu^\phi(E,\nu^a,w,\mathcal{K}), \quad w \in \mathcal{N}_1, \quad w \approx 0, \quad \nu^a \approx 0\}\end{aligned}$$

is a Poincaré section in $\mathcal{N}_{E,\mathcal{K}}$ for $E \approx E_0$, $\mathcal{K} \approx 0$. Denote $s = \nu_1^a - \nu^a_{\mathrm{RW},1}(E,\mathcal{K})$. Let us now look for fixed points of the Poincaré map $\Pi(E,s,w,\mathcal{K})$ which maps $\mathcal{S}_{E,\mathcal{K}}$ to itself.

Decompose $\Pi(E,s,w,\mathcal{K}) = (\Pi_0(E,s,w,\mathcal{K}), \Pi_1(E,s,w,\mathcal{K}))$, where $\Pi_1(\cdot)$ is the w-component ($w \in \mathcal{N}_1$) and $\Pi_0(\cdot)$ the s-component of Π. Due to the nonresonance condition, $\Pi_1(E,s,w,\mathcal{K}) = w$ can be solved for $\mathcal{K} \approx 0$, $s \approx 0$, $E \approx E_0 = H(x_0)$, by the implicit function theorem to obtain $w(E,s,\mathcal{K})$. Inserting this into Π_0, one obtains one scalar fixed point equation $s = \hat{\Pi}(E,s,\mathcal{K})$. This equation is satisfied due to the existence of the Casimir (4.5): Let $\hat{s} = \hat{\Pi}(E,s,\mathcal{K})$. Inserting $w = w(E,s,\mathcal{K})$ into (4.5), a function $C(E,s,\mathcal{K}) = s^2 + O(\mathcal{K})$ is obtained. Any equilibrium (ν,w) of the slice equation on $\mathcal{N}$ satisfies

$$\mathrm{D}\, C(\nu,w) \,||\, \mathrm{D}\, h(\nu,w)$$

or

$$\mathrm{D}(C|_{h=E}) = 0,$$

where $E = h(\nu,w)$. The equilibria corresponding to rotating waves of (3.1) are at $s = 0$, and therefore $\mathrm{D}_s C(E,0,\mathcal{K}) \equiv 0$. Moreover, $\mathrm{D}_s^2 C(E_0,0,0) = 2$, where $E_0 = H(x_0)$. So $s \to C(E,s,\mathcal{K})$ is monotonically increasing for $s \geq 0$, $E \approx E_0$, $\mathcal{K} \approx 0$, and one can solve for $s(E,C,\mathcal{K})$. Hence $\hat{s} = s$, and a family $(\nu,w)(E,C,\mathcal{K})$ of periodic orbits of the slice equation is obtained. This gives a family $\mathcal{P}(E,C,\mathcal{K})$ of RPOs of (3.22). Changing the parametrization of the RPOs $\mathcal{P}(E,C,\mathcal{K})$ from (E,C) back to (ν^ϕ,s), the notation of the theorem is recovered. Let $(\phi,a)(\nu^\phi,s,\mathcal{K})$ by the drift symmetry of the RPO $\mathcal{P}(\nu^\phi,s,\mathcal{K})$ at $x(\nu^\phi,s,\mathcal{K}) \simeq (\mathrm{id},(\nu,w)(\nu^\phi,s,\mathcal{K}))$. Note that for $\mathcal{K} = 0$, (3.18) gives $s = r^a$ and $\mathbf{J}(x(\nu^\phi,r^a,0)) = (\mu_0^\phi + \nu^\phi, r^a, 0)$. By (3.17) and (3.9) for $r^a \neq 0$ and $\mathcal{K} = 0$ all RPOs are modulated traveling waves. Therefore, $\phi(\nu^\phi,s,\mathcal{K}) \to 0$ as $\mathcal{K} \to 0$ and the RPOs $\mathcal{P}(\nu^\phi,s,\mathcal{K})$ become modulated traveling waves in the limit of vanishing cocycle. ■

5. Extensions to systems with other symmetry groups. In this section the Hamiltonian analogue of the meandering transition is discussed for systems with spherical symmetry and for systems with the Euclidean symmetry group of three-dimensional space. See Remarks 2.2 (b) and (c) for the corresponding dissipative case.

5.1. Hamiltonian meandering transition with spherical symmetry. In this section it is assumed that the Hamiltonian system (3.1) has spherical symmetry $\Gamma = \mathrm{SO}(3)$. Then persistence of rotating waves to modulated rotating waves at nearby momentum values is studied. Thereby the analogue of Remark 2.2 (b) is studied in the Hamiltonian context. The results can be applied to rotating point vortices on the sphere; see, e.g., [16].

For $\mu_0 = 0$ the momentum isotropy subalgebra is $\mathbf{g}_{\mu_0} = \mathrm{so}(3)$ and for $\mu_0 \neq 0$ (the typical case) it is $\mathbf{g}_{\mu_0} = \mathrm{so}(2)$. Let us first consider the generic case of a rotating wave with momentum $\mu_0 \neq 0$.

Theorem 5.1. *Let $\mathrm{SO}(3)x_0$ be a nondegenerate rotating wave with nonvanishing momentum μ_0 and drift velocity ξ_0. Align x_0 such that $\mu_0 = (0, 0, \mu_{0,3})$, $\xi_0 = (0, 0, \omega_0^{\mathrm{rot}})^T$. Then the following hold true:*

(a) *The rotating wave $\mathrm{SO}(3)x_0$ persists to every nearby momentum. Moreover, there is a one-dimensional family of rotating waves $\mathrm{SO}(3)x_{\mathrm{RW}}(\nu)$, $\nu \approx 0$, such that $x_{\mathrm{RW}}(\nu)$ is smooth and $x_{\mathrm{RW}}(0) = x_0$. The rotating wave through $x_{\mathrm{RW}}(\nu)$ has drift velocity $\xi_{\mathrm{RW}}(\nu)||e_3$, with $\xi_{\mathrm{RW}}(0) = \xi_0$, and momentum $\mathbf{J}(x_{\mathrm{RW}}(\nu)) = (0, 0, \mu_{0,3} + \nu)$.*

(b) *Let the rotating wave $\mathrm{SO}(3)x_0$ be elliptic and nonresonant in the sense of Definition 4.4 and denote its normal frequencies by ω_j. Then there are $(\frac{1}{2}\dim\mathcal{M} - 2)$ two-dimensional families $\mathcal{P}_j(\nu, s)$, $j = 1, \ldots \frac{1}{2}\dim\mathcal{M} - 2$, $\nu \approx 0$, $s \geq 0$, of modulated rotating waves nearby such that there are smooth functions $x_j(\nu, s) \in \mathcal{P}_j(\nu, s)$ with $x_j(\nu, 0) = x_{\mathrm{RW}}(\nu)$ (where $x_{\mathrm{RW}}(\nu)$ is from (a)). These modulated rotating waves have energy $H(x_j(\nu, s)) = H(x_{\mathrm{RW}}(\nu)) \pm s^2$ (depending on the Krein signature of ω_j), momentum $\mathbf{J}(x_j(\nu, s)) = (0, 0, \mu_{0,3} + \nu)$, relative period $T_j(\nu, s)$, such that $T_j(0, 0) = \frac{2\pi}{|\omega_j|}$, and average drift velocity $\xi_j(\nu, s)||e_3$ at $x_j(\nu, s)$, with $\xi_j(0, 0) = \xi_0$.*

So resonance drift cannot occur near rotating waves of SO(3)-symmetric Hamiltonian systems with nonvanishing angular momentum.

Proof of Theorem 5.1. If $\mu_0 \neq 0$, then $\mathcal{N}_0 \simeq \mathbf{g}_{\mu_0}^* \simeq \mathrm{so}_3(2)^*$ is one-dimensional. Here $\mathrm{so}_3(2)$ corresponds to infinitesimal rotations around the e_3-axis. So the $\dot{\nu}$-equation of (3.19) just becomes $\dot{\nu} = 0$.

(a) A nondegenerate rotating wave $\mathrm{SO}(3)x_0$ persists as equilibrium $w_{\mathrm{RW}}(\nu)$ of the $\dot{w}$-equation for $\nu \approx 0$. This gives a rotating wave of (3.1) through $x_{\mathrm{RW}}(\nu) \simeq (\mathrm{id}, \nu, w_{\mathrm{RW}}(\nu))$ with nonvanishing momentum $\mathbf{J}(x_{\mathrm{RW}}(\nu)) = (0, 0, \mu_{0,3} + \nu)$, $\nu \approx 0$; cf. (3.18). Due to SO(3)-equivariance the rotating wave persists to all nearby momenta.

(b) By the Lyapunov center theorem there are $\frac{1}{2}\dim(\mathcal{N}_1)$ families $w_j(\nu, s)$ of periodic orbits of the $\dot{w}$-equation, parametrized by ν and s. Here $\dim\mathcal{N}_1 = \dim\mathcal{M} - 4$. These give points $x_j(\nu, s) = (\mathrm{id}, \nu, w_j(\nu, s))$ on modulated rotating waves $\mathcal{P}_j(\nu, s)$ with $x_j(\nu, 0) = x_{\mathrm{RW}}(\nu)$ and with momentum $\mathbf{J}(x_j(\nu, s)) = \mu_0 + \nu = (0, 0, \mu_{0,3} + \nu_3)$; see (3.18). Let $\mathrm{SO}_3(2)$ be the group of rotations around e_3 with Lie algebra $\mathrm{so}_3(2)$. Since $\mathbf{J}(x_j(\nu, s)) \neq 0$ and $\mathbf{J}(x_j(\nu, s))||e_3$, by (3.17) the drift symmetry $R_j(\nu, s)$ of the RPO at $x_j(\nu, s)$ lies in $\mathrm{SO}_2(3)$, and so the average drift velocity $\xi_j(\nu, s)$ at $x_j(\nu, s)$ is in $\mathrm{so}_3(2)$. ∎

Next let us consider the case that the rotating wave $\mathrm{SO}(3)x_0$ has zero angular momentum $\mu_0 = 0$. In this case resonance drift typically occurs, as the following theorem shows.

Theorem 5.2. *Consider a nondegenerate rotating wave $\mathrm{SO}(3)x_0$ with momentum $\mu_0 = \mathbf{J}(x_0) = 0$ and nonvanishing drift velocity $\xi_0 \neq 0$. Choose x_0 such that $\xi_0 = (0, 0, \omega_0^{\mathrm{rot}})^T$, where ω_0^{rot} is the rotation frequency of the rotating wave at x_0. Then the following hold true.*

(a) *There is a one-parameter family* $\mathrm{SO}(3)x_{\mathrm{RW}}(\nu_3)$ *of rotating waves nearby,* $\nu_3 \approx 0$*, with momentum* $\mathbf{J}(x_{\mathrm{RW}}(\nu_3)) = (0,0,\nu_3)$ *and drift velocity* $\xi_{\mathrm{RW}}(\nu_3)||e_3$ *at* $x_{\mathrm{RW}}(\nu_3)$*, such that* $\xi_{\mathrm{RW}}(0) = \xi_0$*. Moreover, the rotating wave* $\mathrm{SO}(3)x_0$ *persists to all nearby momentum values.*

(b) *Assume that* $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}_w^2 h(0)$ *has no eigenvalues in* $\mathrm{i}\omega_0^{\mathrm{rot}}\mathbb{Z}$*. Then there is a two-parameter family* $\mathcal{P}_{\mathrm{MRW}}(\nu_2,\nu_3)$*,* $\nu_2 \geq 0$*,* $\nu_3 \approx 0$*, of modulated rotating waves of* (3.1) *such that* $x_{\mathrm{MRW}}(\nu_2,\nu_3) \in \mathcal{P}_{\mathrm{MRW}}(\nu_2,\nu_3)$ *is smooth in* (ν_2,ν_3) *and* $x_{\mathrm{MRW}}(0,\nu_3) = x_{\mathrm{RW}}(\nu_3)$*. The modulated rotating wave at* $x_{\mathrm{MRW}}(\nu_2,\nu_3)$ *has drift symmetry* $\gamma_{\mathrm{MRW}}(\nu_2,\nu_3)$*, relative period* $T(\nu_2,\nu_3)$*, and momentum* $\mathbf{J}(x_{\mathrm{MRW}}(\nu_2,\nu_3)) = (0,\nu_2,\nu_3)$*, and* $x_{\mathrm{MRW}}(0,0) = x_0$*,* $T(0,0) = 2\pi/|\omega_0^{\mathrm{rot}}|$*,* $\gamma_{\mathrm{MRW}}(0) = \mathrm{id}$*. This family contains a one-parameter family* $\mathcal{P}(\nu_2,0)$ *of modulated rotating waves which have an average drift velocity* $\xi(\nu_2,0)$ *at* $x_{\mathrm{MRW}}(\nu_2,0)$ *parallel to the* e_2*-axis.*

(c) *Assume that the rotating wave* $\mathrm{SO}(3)x_0$ *is elliptic and nonresonant in the sense of Definition* 4.4 *and that* $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}_w^2 h(0)$ *has no eigenvalues* $\mathrm{i}\omega_j$ *with* $\omega_0^{\mathrm{rot}}/\omega_j \in \mathbb{Z}$*. Then there are* $(\frac{1}{2}\dim\mathcal{M} - 3)$ *more two-parameter families* $\mathcal{P}_j(\nu_3,s)$*,* $j = 1,\dots\frac{1}{2}\dim\mathcal{M} - 3$*, of modulated rotating waves near the rotating wave and there are smooth functions* $x_j(\nu_3,s) \in \mathcal{P}_j(\nu_3,s)$ *with* $x_j(\nu_3,0) = x_{\mathrm{RW}}(\nu_3)$ *(where* $x_{\mathrm{RW}}(\nu_3)$ *is from part* (a)*). The modulated rotating wave* $\mathcal{P}_j(\nu_3,s)$ *has momentum* $\mathbf{J}(x_j(\nu_3,s)) = (0,0,\nu_3)$ *at* $x_j(\nu_3,s)$*, energy* $H(x_j(\nu_3,s)) = H(x_{\mathrm{RW}}(\nu_3)) \pm s^2$ *(depending on the Krein signature of* ω_j*), relative period* $T_j(\nu_3,s)$ *such that* $T_j(0,0) = 2\pi/|\omega_j|$*, and average drift velocity* $\xi_j(\nu_3,s)||e_3$*,* $s \geq 0$*, with* $\xi_j(0,0) = \xi_0$*,* $j = 1,\dots\frac{1}{2}\dim\mathcal{M} - 3$*.*

Proof. If the rotating wave $\mathrm{SO}(3)x_0$ has momentum $\mu_0 = 0$, then $\mathbf{g}_{\mu_0} = \mathrm{so}(3)$. In this case $\nu \in \mathrm{so}(3)^* \simeq \mathbb{R}^3$ and the $\dot{\nu}$-equation from (3.19) becomes

$$\dot{\nu} = \nu \times \mathrm{D}_\nu h(\nu, w). \tag{5.1}$$

(a) Since the $\dot{\nu}$ equation has nontrivial dynamics, let us reduce only by the symmetry group

$$\tilde{\Gamma} = \{\gamma \in \mathrm{SO}(3), \quad \mathrm{Ad}_\gamma \xi_0 = \xi_0\} = \mathrm{SO}_3(2),$$

which is the group of rotations around the e_3-axis. The corresponding slice is denoted by $\widetilde{\mathcal{N}} = \widetilde{\mathcal{N}}_0 \oplus \widetilde{\mathcal{N}}_1$. Then $\tilde{\nu} \in \widetilde{\mathcal{N}}_0$ is given by $\tilde{\nu} = \nu_3$ and $\dot{\tilde{\nu}} = 0$. Note that $\dim\widetilde{\mathcal{N}}_1 = \dim\mathcal{N}_1 + 4$. Let $\tilde{h}(\tilde{\nu},\tilde{w})$ be the Hamiltonian in the bundle coordinates $(\tilde{\gamma},\tilde{\nu},\tilde{w}) \in \tilde{\Gamma} \times \widetilde{\mathcal{N}}_0 \oplus \widetilde{\mathcal{N}}_1$. The matrix $\mathrm{ad}^*_{\xi_0}$ has eigenvalues $\pm\mathrm{i}\omega_0^{\mathrm{rot}}$ with real eigenspace $\{\nu = (\nu_1,\nu_2,0) \in \mathrm{so}(3)^*\}$ and a simple eigenvalue 0. Because of (2.4) and (3.20), the eigenvalues of the linearization $\mathbb{J}_{\widetilde{\mathcal{N}}_1}\mathrm{D}^2_{\tilde{w}}\tilde{h}(0)$ of the equilibrium $0 \in \widetilde{\mathcal{N}}_1$ corresponding to the rotating wave $\mathrm{SO}(3)x_0$ are given by the eigenvalues of $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}_w^2 h(0)$ and by the eigenvalues $\pm\mathrm{i}\omega_0^{\mathrm{rot}}$ of multiplicity two. Hence the rotating wave is nondegenerate when considered as a rotating wave of a Hamiltonian system with $\mathrm{SO}_3(2)$-symmetry. Therefore, $\dot{\tilde{w}} = 0$ for $\tilde{w}(\tilde{\nu})$ can be solved using the nondegeneracy condition. This gives rotating waves $\mathrm{SO}(3)x_{\mathrm{RW}}(\tilde{\nu})$, $x_{\mathrm{RW}}(\tilde{\nu}) \simeq (\mathrm{id},\tilde{\nu},\tilde{w}(\tilde{\nu}))$ for the original $\mathrm{SO}(3)$-equivariant Hamiltonian system (3.1). Since only a reduction by $\mathrm{SO}_3(2)$-symmetry was carried out, these rotating waves have drift velocities $\xi_{\mathrm{RW}}(\nu_3)||e_3$, where $\xi(0) = \xi_0 \neq 0$. Then (3.16) implies that also $\mathbf{J}(x(\nu_3))||e_3$. This proves part (a).

(b) The rotating waves through $x_{\mathrm{RW}}(\nu_3) \simeq (\gamma_{\mathrm{RW}}(\nu_3),\nu_{\mathrm{RW}}(\nu_3),w_{\mathrm{RW}}(\nu_3))$ from part (a) have, by (3.18), momentum

$$\mathbf{J}(x_{\mathrm{RW}}(\nu_3)) = (0,0,\nu_3) = \gamma_{\mathrm{RW}}(\nu_3)\nu_{\mathrm{RW}}(\nu_3)$$

and energy $E_{\mathrm{RW}}(\nu_3) = h(\nu_{\mathrm{RW}}(\nu_3), w_{\mathrm{RW}}(\nu_3))$, where

$$E'_{\mathrm{RW}}(0) = \mathrm{D}_\nu h(0)\nu'_{\mathrm{RW}}(0) = \omega_0^{\mathrm{rot}} \neq 0.$$

Therefore, they can be parametrized by energy E instead of ν_3. Let $x_{\mathrm{RW}}(E) \simeq (\gamma_{\mathrm{RW}}(E), \nu_{\mathrm{RW}}(E), w_{\mathrm{RW}}(E))$ again denote the corresponding path of rotating waves. Let $\omega^{\mathrm{rot}}(E)$ be the rotation frequency of the rotating wave $\mathrm{SO}(3)x_{\mathrm{RW}}(E)$. Since the slice equation conserves the energy $h(\nu, w)$ the energy level sets $\mathcal{N}_E$ of $\mathcal{N}$ are flow-invariant. Because of $\mathrm{D}_\nu h(0) = (0, 0, \omega_0^{\mathrm{rot}})$, $\omega_0^{\mathrm{rot}} \neq 0$, and $\mathrm{D}_w h(0) = 0$, they can be parametrized for $E \approx E_0 = h(0)$, similarly as in the proof of Theorem 4.3, by

$$\mathcal{N}_E = \{(\nu, w) \in \mathcal{N}, \quad h(\nu, w) = E\} = \{(\nu, w) \in \mathcal{N}, \quad \nu_3 = \nu_3(\nu_1, \nu_2, w, E)\}.$$

Then $\nu_3(\nu_{\mathrm{RW},1}(E), \nu_{\mathrm{RW},2}(E), w_{\mathrm{RW}}(E), E) = \nu_{\mathrm{RW},3}(E)$. Let us now consider the equilibrium $v_{\mathrm{RW}}(E) := (\nu_{\mathrm{RW}}(E), w_{\mathrm{RW}}(E))$ of the slice equation as periodic orbit with period $T^{\mathrm{rot}}(E) = \frac{2\pi}{|\omega^{\mathrm{rot}}(E)|}$. The matrix $\mathrm{ad}^*_{\xi_0}$ has a pair $\pm\mathrm{i}\omega_0^{\mathrm{rot}}$ of nonvanishing imaginary eigenvalues with real eigenspace spanned by the vectors $\{(\nu_1, \nu_2, 0),\ \nu_1, \nu_2 \in \mathbb{R}\} \subseteq \mathrm{so}(3)^*$. By (3.20), the linearization of the slice equation $\mathrm{D}_v f_{\mathcal{N}}(v_{\mathrm{RW}}(E))$ at $E = E_0$ also has this pair of eigenvalues which perturbs to the eigenvalues $\pm\mathrm{i}\omega^{\mathrm{rot}}(E)$ of $\mathrm{D}_v f_{\mathcal{N}}(v_{\mathrm{RW}}(E))$ for $E \approx E_0$. Consequently,

$$\begin{aligned}\mathcal{S}_E = \{(\nu, w) \in \mathcal{N}_E,\ \nu_1 = \nu_{\mathrm{RW},1}(E),\ \nu_2 > \nu_{\mathrm{RW},2}(E),\ \nu_2 \approx \nu_{\mathrm{RW},2}(E),\\ \nu_3 = \nu_3(\nu_{\mathrm{RW},1}(E), \nu_2, w, E),\ w \approx w_{\mathrm{RW}}(E)\}\end{aligned}$$

is a section transverse to the flow of the slice equation at $v_{\mathrm{RW}}(E)$ inside the energy level set $\mathcal{N}_E$ to the energy $E \approx E_0$. Let $s := \nu_2 - \nu_{\mathrm{RW},2}(E)$. The corresponding Poincaré map is denoted by $\Pi(\cdot, E) : \mathcal{S}_E \to \mathcal{S}_E$. Decompose $\Pi(\cdot, E) = (\Pi_0(\cdot, E), \Pi_1(\cdot, E))$, where Π_0 maps into the ray $s \geq 0$ and Π_1 into $\mathcal{N}_1$. By assumption there is no k : 1-resonance between ω_0^{rot} and any normal frequency on $\mathcal{N}_1$. Therefore, the equation $\Pi_1(s, w, E) = w$ can be solved for $w(s, E)$, such that $w(0, E) = w_{\mathrm{RW}}(E)$ for $E \approx E_0$. Plugging this into Π_0 a map $\tilde{\Pi}(\cdot, E)$ from the ray $s \geq 0$ into itself is obtained. The $\dot{\nu}$-equation (5.1) conserves the Casimir $C^R(\nu, w) = \|\nu\|_2^2$. Define $\nu_3(s, E) := \nu_3(\nu_{\mathrm{RW},1}(E), s+\nu_{\mathrm{RW},2}(E), w(s, E), E)$. Then $\hat{s} = \hat{\Pi}(s, E)$ satisfies $C(\hat{s}, E) = C(s, E)$, where

$$C(s, E) = (\nu_{\mathrm{RW},1}(E))^2 + (\nu_{\mathrm{RW},2}(E) + s)^2 + (\nu_3(s, E))^2.$$

The path of relative equilibria $\mathrm{SO}(3)x_{\mathrm{RW}}(E)$ corresponds to $(s, E) = (0, E)$. Note that

$$\mathrm{D}\, h(v_{\mathrm{RW}}(E)) \mid\mid \mathrm{D}\, C^R(v_{\mathrm{RW}}(E))$$

and that $\mathrm{D}\, C^R|_{h(v)=E} = 0$ at $v = v_{\mathrm{RW}}(E)$. As a result of this, $\mathrm{D}_s C(0, E) \equiv 0$. Moreover, from $\mathrm{D}_\nu h(0) = \omega_0^{\mathrm{rot}} e_3$ and $\mathrm{D}_w h(0) = 0$, it follows that $\mathrm{D}_s \nu_3(0, E_0) = 0$, and therefore $\mathrm{D}_s^2 C(0, E_0) = 2$. Hence $s \to C(s, E)$ is injective for $s \geq 0$, $s \approx 0$, for any fixed $E \approx E_0$. Consequently $\hat{s} = s$, and so $v(s, E) := (\nu(s, E), w(s, E))$, with $\nu(s, E) = (\nu_{\mathrm{RW},1}(E), s + \nu_{\mathrm{RW},2}(E), \nu_3(s, E))^T$, lies on a periodic orbit of the slice equation with period $T(s, E) \approx T(0, E) = \frac{2\pi}{|\omega^{\mathrm{rot}}(E)|}$.

Changing the parametrization back from E to ν_3, a two-parameter family $v(s, \nu_3) = (\nu(s, \nu_3), w(s, \nu_3))$ of periodic orbits of the slice equation with periods $T(s, \nu_3)$ is obtained, satisfying $T(0, \nu_3) = T^{\mathrm{rot}}(0, \nu_3)$. These give points $\hat{x}(s, \nu_3) \simeq (\gamma_{\mathrm{RW}}(\nu_3), v(s, \nu_3))$ on modulated rotating waves of the full Hamiltonian system (3.1) with momentum $\hat{\mu}(s, \nu_3) = \gamma_{\mathrm{RW}}(\nu_3)\nu(s, \nu_3)$;

see (3.18). At the rotating waves $\hat{\mu}(0,\nu_3) = \nu_3$ and hence $\mathrm{D}_{\nu_3}\hat{\mu}(0,\nu_3) \equiv e_3$. Moreover, $\gamma_{\mathrm{RW}}(0) = \mathrm{id}$ implies that $\mathrm{D}_s\hat{\mu}(0,0) \equiv e_2$. So smooth functions $\hat{\phi}_2(s,\nu_3)$, $\hat{\phi}_3(s,\nu_3)$ can be found such that $\hat{\gamma}(s,\nu_3) = \exp(\hat{\phi}_2(s,\nu_3)\xi_2 + \hat{\phi}_3(s,\nu_3)\xi_3)$ (with the notation from (2.9)) satisfies $(\hat{\gamma}(s,\nu_3)\hat{\mu}(s,\nu_3))_1 = 0$ and $\hat{\phi}_2(0,\nu_3) = 0$, $\hat{\phi}_3(0,\nu_3) = 0$. Let $\gamma(s,\nu_3) = \hat{\gamma}(s,\nu_3)\gamma_{\mathrm{RW}}(\nu_3)$. Then $x(s,\nu_3) \simeq (\gamma(s,\nu_3), v(s,\nu_3))$ lies on an RPO with momentum $\mu(s,\nu_3)$ such that $\mu_1(s,\nu_3) = 0$. Then s can be replaced by ν_2 and (ν_2,ν_3) can be tranformed such that $x(\nu_2,\nu_3)$ has momentum $\mu(\nu_2,\nu_3) = \mathbf{J}(x(\nu_2,\nu_3)) = (0,\nu_2,\nu_3)$, $\nu_2 \geq 0$.

The condition (3.17) implies that the drift symmetry $\gamma(\nu_2,\nu_3)$ at the RPO through $x(\nu_2,\nu_3)$ satisfies $\gamma(\nu_2,\nu_3)\mu(\nu_2,\nu_3) = \mu(\nu_2,\nu_3)$, where $\gamma(0) = \mathrm{id}$. Therefore, $\gamma(\nu_2,\nu_3)$ is a rotation around the vector $\mu(\nu_2,\nu_3)$ in the (x_2,x_3)-plane. Moreover, for $\nu_3 = 0$ the modulated rotating wave $\mathcal{P}(\nu_2,0)$ rotates around the e_2-axis with momentum $\mu(\nu_2) = (0,\nu_2,0)$ at $x(\nu_2,0)$.

(c) By assumption there is no $k:1$-resonance between any of the eigenvalues of $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}^2_w h(0)$ and between the eigenvalues of $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}^2_w h(0)$ and $\mathrm{i}\omega_0^{\mathrm{rot}}$. Hence for all normal frequencies on $\mathcal{N}_1$, part (c) follows from the Lyapunov center theorem applied on the space $\tilde{\mathcal{N}}_1$, after symmetry reduction by $\mathrm{SO}_3(2)$ as in part (a). ∎

5.2. Hamiltonian meandering transition with the Euclidean symmetry of three-dimensional space. In this section the Hamiltonian analogue of the resonance drift of Remark 2.2 (c) is studied. The symmetry group is again $\Gamma = \mathrm{SE}(3) = \mathrm{SO}(3) \ltimes \mathbb{R}^3$. Similarly as in (3.8), (3.9), the adjoint and coadjoint actions for $\Gamma = \mathrm{SE}(3)$ are

$$\begin{aligned} \mathrm{Ad}_{(R,a)}\xi &= (\, R\xi^r \,,\, R\xi^a - R\xi^r \times a \,)\,, \\ \mathrm{Ad}^*_{(R,a)^{-1}}\mu &= (\, R\mu^r + a \times R\mu^a \,,\, R\mu^a \,)\,, \end{aligned} \tag{5.2}$$

where $(R,a) \in \mathrm{SO}(3) \ltimes \mathbb{R}^3$; see, e.g., [15, 26]. So typically, when $\mu^a \neq 0$, then $\Gamma_\mu \simeq \mathrm{SO}(2) \times \mathbb{R}$. In this case resonance drift is not possible.

Proposition 5.3. *Let $\mathrm{SE}(3)x_0$ be a nondegenerate relative equilibrium with generic momentum value μ_0 satisfying $\mu_0^a \neq 0$ and with drift velocity ξ_0. Align x_0 such that $\xi_0^a || e_3$, $\xi_0^r || e_3$. Then the following hold true.*

(a) *There is a two-parameter family $\mathrm{SE}(3)x_{\mathrm{RE}}(\nu_3^r,\nu_3^a)$ of relative equilibria of (3.1) with $x_{\mathrm{RE}}(0,0) = x_0$. The relative equilibrium at $x_{\mathrm{RE}}(\nu_3^r,\nu_3^a)$ has angular momentum $\mathbf{J}^r(x_{\mathrm{RE}}(\nu_3^r,\nu_3^a)) = (0,0,\mu^r_{0,3} + \nu_3^r)$, linear momentum $\mathbf{J}^r(x_{\mathrm{RE}}(\nu_3^r,\nu_3^a)) = (0,0,\mu^a_{0,3} + \nu_3^a)$, and drift velocity $\xi_{\mathrm{RE}}(\nu_3^r,\nu_3^a)$, which satisfies $\xi^r_{\mathrm{RE}}(\nu_3^r,\nu_3^a) || e_3$, $\xi^a_{\mathrm{RE}}(\nu_3^r,\nu_3^a) || e_3$.*

(b) *Let the relative equilibrium $\mathrm{SE}(3)x_0$ be elliptic and nonresonant in the sense of Definition 4.4 and denote its normal frequencies by ω_j, $j = 1,\dots,\frac{1}{2}\dim\mathcal{M} - 4$. Then there are $(\frac{1}{2}\dim\mathcal{M} - 4)$ families of RPOs $\mathcal{P}_j(\nu_3^r,\nu_3^a,s)$, $s \geq 0$, and smooth functions $x_j(\nu_3^r,\nu_3^a,s) \in \mathcal{P}_j(\nu_3^r,\nu_3^a,s)$ such that $x_j(\nu_3^r,\nu_3^a,0) = x_{\mathrm{RE}}(\nu_3^r,\nu_3^a)$. The RPO at $x_j(\nu_3^r,\nu_3^a,s)$ has momentum*

$$\mathbf{J}^r(x_j(\nu_3^r,\nu_3^a,s)) = (0,0,\mu^r_{0,3} + \nu_3^r), \quad \mathbf{J}^a(x_j(\nu_3^r,\nu_3^a,s)) = (0,0,\mu^a_{0,3} + \nu_3^a),$$

energy $H(x_j(\nu_3^r,\nu_3^a,s)) = H_{\mathrm{RE}}(\nu_3^r,\nu_3^a)) \pm s^2$ (depending on the Krein signature of ω_j), relative period $T_j(\nu_3^r,\nu_3^a,s)$, such that $T_j(0,0,0) = 2\pi/|\omega_j|$, and average drift velocity $\xi_j(\nu_3^r,\nu_3^a,s)$ at $x_j(\nu_3^r,\nu_3^a,s)$, which satisfies $\xi_j^r(\nu_3^r,\nu_3^a,s) || e_3$, $\xi_j^a(\nu_3^r,\nu_3^a,s) || e_3$, $\xi_j(0,0) = \xi_0$.

Proof. (a) Note that, by (5.2), $\xi_0^a || e_3$, $\xi_0^r || e_3$ implies $\mu_0^a || e_3$, $\mu_0^r || e_3$. The Lie-group $\Gamma_{\mu_0} \simeq \mathrm{SO}_3(2) \times \mathbb{R}_3$ is abelian; therefore, $\dot{\nu} \equiv 0$ holds in the equations (3.19) near a relative equilibrium

$\mathrm{SE}(3)x_0$ with generic momentum value. By the nondegeneracy condition, $\dot{w} = 0$ can be solved for $w_{\mathrm{RE}}(\nu)$ to obtain relative equilibria through $x_{\mathrm{RE}}(\nu) \simeq (\mathrm{id}, \nu, w_{\mathrm{RE}}(\nu))$. The statement about the momentum of $x_{\mathrm{RE}}(\nu)$ follows from (3.18), and the statement about the velocity $\xi_{\mathrm{RE}}(\nu)$ from (3.16) and (5.2).

(b) The Lyapounov center theorem can be applied on the ν-dependent $\dot{w}$-equation. The statements about the momentum of the families of RPOs follows from (3.18). The fact that their drift symmetry lies in $\mathrm{SO}_3(2) \times \mathbb{R}_3$, and hence their average drift velocity in $\mathrm{so}_3(2) \times \mathbb{R}_3$, follows from (3.17) and the fact that $\Gamma_{\mu_0} \simeq \mathrm{SO}_3(2) \times \mathbb{R}_3$; see part (a). ∎

The situation is different if the relative equilibrium $\mathrm{SE}(3)x_0$ has a nongeneric momentum value $\mu_0 = \mathbf{J}(x_0)$. In what follows, it is shown that in this case resonance drift occurs generically.

If $\mu_0^a = 0$, then (5.2) implies that $\Gamma_{\mu_0} \simeq \mathrm{SO}(2) \ltimes \mathbb{R}^3$. Let us assume, without loss of generality, that $\mu_0^r || e_3$. Then the drift velocity $\xi_0 = (\xi_0^r, \xi_0^a)$ of the relative equilibrium at x_0 satisfies $\xi_0^r || e_3$. Choose x_0 in its $\mathrm{SE}(3)$ orbit such that also $\xi_0^a || e_3$. The momentum value μ_0 is nonsplit, and the $\dot{\nu}$-equation of (3.19) for $\nu = (\nu^r, \nu^a) \in \mathrm{so}(2)^* \oplus (\mathbb{R}^3)^*$ is nontrivial; see [26]. It can be easily checked that the functions

$$C^a(\mu) = \|\mu^a\|^2 \quad \text{and} \quad C^r(\mu) = \langle \mu^a, \mu^r \rangle \tag{5.3}$$

are invariant under the coadjoint action (5.2). These restrict to the functions $C^a(\nu) = \|\nu^a\|^2$ and $C^r(\nu) = \nu_3^a(\mu_3^r + \nu^r)$ on the slice $\mathcal{N}_0 \simeq \mathrm{so}(2)^* \oplus (\mathbb{R}^3)^*$ and are Casimirs, i.e., conserved quantities of the $\dot{\nu}$-equation.

In the following proposition persistence of a relative equilibrium with vanishing linear momentum is studied, as a prerequisite for the analysis of the Hamiltonian meandering transition.

Proposition 5.4. *Let $\mathrm{SE}(3)x_0$ be a nondegenerate relative equilibrium with momentum value $\mu_0 = (\mu_0^r, 0)$, $\mu_0^r || e_3$, and with drift velocity $\xi_0 = (\xi_0^r, \xi_0^a)$, where $\xi_0^r \neq 0$, $\xi_0^r || e_3$, $\xi_0^a || e_3$. Then there exists a two-dimensional family of relative equilibria $\mathrm{SE}(3)x_{\mathrm{RE}}(\nu^r, \nu_3^a)$ of (3.1) such that $x_{\mathrm{RE}}(\nu^r, \nu_3^a)$ is smooth in its parameters and $x_{\mathrm{RE}}(0) = x_0$, $\xi_{\mathrm{RE}}(0,0) = \xi_0$. The relative equilibrium through $x_{\mathrm{RE}}(\nu^r, \nu_3^a)$ has angular momentum $\mathbf{J}^r(x_{\mathrm{RE}}(\nu^r, \nu_3^a)) = \mu_0^r + \nu^r e_3$, linear momentum $\mathbf{J}^a(x_{\mathrm{RE}}(\nu^r, \nu_3^a)) = \nu_3^a e_3$, and drift velocity $\xi_{\mathrm{RE}}(\nu^r, \nu_3^a)$, where $\xi_{\mathrm{RE}}^r(\nu^r, \nu_3^a) = \omega^{\mathrm{rot}}(\nu^r, \nu_3^a) e_3$, $\xi_{\mathrm{RE}}^a(\nu^r, \nu_3^a) || e_3$.*

Proof. This proposition is an application of a persistence result for general noncompact symmetry groups; see [32, Example 5.3(a)]. But it can also be proved in an elementary way: Because of (2.4), (3.20), the linearization L_0 at x_0 has, by our nondegeneracy condition and the form of ad_{ξ_0} from (5.2), a four-dimensional kernel corresponding to two zero eigenvalues of ad_{ξ_0}. Therefore, let us reduce only by the abelian symmetry group $\tilde{\Gamma} = \mathrm{SO}_3(2) \times \mathbb{R}_3$ of rotations around and translations along the e_3-axis. This gives the reduced system

$$\dot{\tilde{\nu}} = 0, \quad \dot{\tilde{w}} = \mathbb{J}_{\widetilde{\mathcal{N}}_1} \mathrm{D}_{\tilde{w}} \tilde{h}(\tilde{\nu}, \tilde{w}) \tag{5.4}$$

on the slice $\widetilde{\mathcal{N}} = \widetilde{\mathcal{N}}_0 \oplus \widetilde{\mathcal{N}}_1$, where $\tilde{\nu} = (\nu^r, \nu_3^a) \in \widetilde{\mathcal{N}}_0 \simeq \mathrm{so}_3(2)^* \oplus \mathbb{R}_3^*$. Since $\mathbb{J}_{\widetilde{\mathcal{N}}_1} \mathrm{D}_{\tilde{w}}^2 \tilde{h}(0)$ is invertible, the equation $0 = \mathbb{J}_{\widetilde{\mathcal{N}}_1} \mathrm{D}_{\tilde{w}} h(\tilde{\nu}, \tilde{w})$ can be solved for $\tilde{w}(\tilde{\nu})$ by the implicit function theorem. This gives relative equilibria $\mathrm{SE}(3)x_{\mathrm{RE}}(\nu^r, \nu_3^a)$ of (3.1), where $x_{\mathrm{RE}}(\nu^r, \nu_3^a) \simeq (\mathrm{id}, \tilde{\nu}, \tilde{w}(\tilde{\nu}))$. The drift velocity $\xi_{\mathrm{RE}}(\nu^r, \nu_3^a)$ of the relative equilibrium at $x_{\mathrm{RE}}(\nu^r, \nu_3^a)$ lies in the Lie algebra of $\tilde{\Gamma}$

and therefore satisfies $\xi^r_{\mathrm{RE}}(\nu^r,\nu^a_3)=\omega^{\mathrm{rot}}_{\mathrm{RE}}(\nu^r,\nu^a_3)e_3$ and $\xi^a_{\mathrm{RE}}(\nu^r,\nu^a_3)||e_3$. The statement about the momentum of $x_{\mathrm{RE}}(\nu^r,\nu^a_3)$ follows from (3.18). ■

In this case resonance drift occurs as the following theorem shows.

Theorem 5.5. *Let, as before,* $\mathrm{SE}(3)x_0$ *be a nondegenerate relative equilibrium with momentum* $\mu_0=(\mu^r_0,0)$, $\mu^r_0||e_3$, $\mu^r_0\neq 0$, *and with nonvanishing rotational velocity vector* $\xi^r_0\neq 0$ *such that* $\xi^r_0=\omega^{\mathrm{rot}}_0e_3$, $\xi^a_0||e_3$. *Assume that* $\mathbb{J}_{\mathcal{N}_1}\mathrm{D}^2_w h(0)$ *has no eigenvalue in* $\mathrm{i}\omega^{\mathrm{rot}}_0\mathbb{Z}$. *Then the following hold:*

(a) *There is a three-dimensional manifold* $\mathcal{P}(\nu^r,\nu^a_2,\nu^a_3)$ *of RPOs such that* $x(\nu^r,\nu^a_2,\nu^a_3)\in\mathcal{P}(\nu^r,\nu^a_2,\nu^a_3)$ *is a smooth function of its parameters,* $\nu^a_2\geq 0$, $\nu^r\approx 0$, $\nu^a_3\approx 0$, *and* $x(\nu^r,0,\nu^a_3)=x_{\mathrm{RE}}(\nu^r,\nu^a_3)$. *The RPO through* $x(\nu^r,\nu^a_2,\nu^a_3)$ *has angular momentum* $\mathbf{J}^r(x(\nu^r,\nu^a_2,\nu^a_3))=\mu^r_0+\nu^re_3$, *linear momentum* $\mathbf{J}^a(x(\nu^r,\nu^a_2,\nu^a_3))=(0,\nu^a_2,\nu^a_3)$, *and relative period* $T(\nu^r,\nu^a_2,\nu^a_3)$ *with* $T(\nu^r,0,\nu^a_3)=2\pi/|\omega^{\mathrm{rot}}(\nu^r,\nu^a_3)|$. *Here* $\omega^{\mathrm{rot}}(\nu^r,\nu^a_3)$ *is the rotation frequency of the relative equilibrium* $\mathrm{SE}(3)x_{\mathrm{RE}}(\nu^r,\nu^a_3)$ *from Proposition* 5.4.

(b) *This family contains a two-dimensional submanifold* $x(\nu^r,\nu^a_2,0)$ *at* $\nu^a_3=0$ *which has an average rotational drift velocity* $\xi^r(\nu^r,\nu^a_2,0)$ *with* $\xi^r_3(\nu^r,\nu^a_2,0)=0$.

Note that the RPO through $x(\nu^r,\nu^a_2,0)$ rotates around and translates along a vector parallel to e_2, whereas the original relative equilibrium through x_0 rotates around and translates along the e_3 direction.

Proof of Theorem 5.5. Let $\nu^r\approx 0$, $\nu^a_3\approx 0$. Near the relative equilibria through $x_{\mathrm{RE}}(\nu^r,\nu^a_3)\simeq(\gamma_{\mathrm{RE}}(\nu^r,\nu^a_3),\nu_{\mathrm{RE}}(\nu^r,\nu^a_3),w_{\mathrm{RE}}(\nu^r,\nu^a_3))$ from Proposition 5.4, let us change coordinates on the slice $\mathcal{N}=\mathcal{N}_0\oplus\mathcal{N}_1$, $\mathcal{N}_0\simeq\mathrm{so}(2)^*\oplus(\mathbb{R}^3)^*$, from $v=(\nu^r,\nu^a_1,\nu^a_2,\nu^a_3,w)$ to $(E,C^r,\nu^a_1,\nu^a_2,w)$ as follows: first let

$$\nu^a_3(\nu^r,C^r)=\frac{C^r}{\nu^r+\mu^r_{0,3}}.$$

Here $C^r(\mu+\nu)=\langle\nu^a,\mu^r_0+\nu^re_3\rangle$ is the Casimir from (5.3) restricted to elements of the form $\mu_0+\nu$, where $\nu\in\mathcal{N}_0\simeq\mathrm{so}_3(2)^*\oplus(\mathbb{R}^3)^*$. Then, by the implicit function theorem, using that $\mathrm{D}_{\nu^r}h(0)=\omega^{\mathrm{rot}}_0\neq 0$, $\mathrm{D}_wh(0)=0$, $\mathrm{D}_{\nu^a}h(0)||e_3$, and $\mathrm{D}_{\nu^r}\nu^a_3(0,0)=0$, one obtains

$$\nu^r=\nu^r(E,C^r,\nu^a_1,\nu^a_2,w)$$

for $E\approx E_0=H(x_0)$, $C^r\approx 0$, where $E=h((\nu^r,\nu^a_1,\nu^a_2,\nu^a_3(\nu^r,C^r)),w)$. Solving $E=h(\nu_{\mathrm{RE}}(\nu^r,\nu^a_3(\nu^r,C^r)),w_{\mathrm{RE}}(\nu^r,\nu^a_3(\nu^r,C^r)))$ by the implicit function theorem for ν^r, the family of relative equilibria $\mathrm{SE}(3)x_{\mathrm{RE}}(E,C^r)$, $x_{\mathrm{RE}}(E,C^r)\simeq(\gamma_{\mathrm{RE}}(E,C^r),v_{\mathrm{RE}}(E,C^r))$, $v_{\mathrm{RE}}(E,C^r)=(\nu_{\mathrm{RE}}(E,C^r),w_{\mathrm{RE}}(E,C^r))$ is obtained, parametrized by the conserved quantities (E,C^r). Then, since $\omega^{\mathrm{rot}}_0\neq 0$ and $\xi^r_0=\omega^{\mathrm{rot}}_0e_3$, by (3.20) and (5.2), for $C^r\approx 0$, $E\approx E_0$, the section

$$\begin{aligned}\mathcal{S}_{E,C^r}=\{(\nu,w)\in\mathcal{N},\ &\nu^a_3=\nu^a_3(\nu^r,C^r),\ \nu^r=\nu^r(E,C^r,\nu^a_1,\nu^a_2,w),\\ &\nu^a_1=\nu^a_{\mathrm{RE},1}(E,C^r),\ \nu^a_2>\nu^a_{\mathrm{RE},2}(E,C^r),\\ &\nu^a_2\approx\nu^a_{\mathrm{RE},2}(E,C^r),\ w\approx w_{\mathrm{RE}}(E,C^r)\}\end{aligned}$$

is transversal to the flow of (3.1) in the flow-invariant manifold

$$\mathcal{N}_{E,C^r}:=\{(\nu,w)\in\mathcal{N},\ h(\nu,w)=E,\ C^r(\nu)=C^r\}.$$

Consider the Poincaré map $(s,w) \to \Pi(E,C^r,s,w)$ from $\mathcal{S}_{E,C^r}$ to itself, where $s = \nu_2^a - \nu_{\mathrm{RE},2}(E,C^r)$. At $(E,C^r,s,w) = (E_0,0,0,0)$ the Poincaré return time is $T(0) = \frac{2\pi}{|\omega_0^{\mathrm{rot}}|}$. Due to the nonresonance assumption the equation $\Pi_1(E,C^r,s,w) = w$ can be solved for $w(E,C^r,s)$ such that $w(E,C^r,0) = w_{\mathrm{RE}}(E,C^r)$. Here Π_1 is the $\mathcal{N}_1$ component of Π. What follows is a proof that $\hat{x}(E,C^r,s) = (\gamma_{\mathrm{RE}}(E,C^r), \nu(E,C^r,s), w(E,C^r,s))$ lies on an RPO. Here

$$
\begin{aligned}
\nu(E,C^r,s) &= (\nu^r(E,C^r,s), \nu^a(E,C^r,s)),\\
\nu^r(E,C^r,s) &= \nu^r(E,C^r,\nu^a_{\mathrm{RE},1}(E,C^r), s+\nu^a_{\mathrm{RE},2}(E,C^r), w(E,C^r,s)),\\
\nu_3^a(E,C^r,s) &= \nu_3^a(\nu^r(E,C^r,s),C^r),\\
\nu^a(E,C^r,s) &= (\nu^a_{\mathrm{RE},1}(E,C^r), s+\nu^a_{\mathrm{RE},2}(E,C^r), \nu_3^a(E,C^r,s)).
\end{aligned}
\tag{5.5}
$$

By construction, $\hat{x}(E,C^r,0) = x_{\mathrm{RE}}(E,C^r)$. Define

$$C^a(E,C^r,s) = \|\nu^a(E,C^r,s)\|^2.$$

The equilibria $v_{\mathrm{RE}}(E,C^r)$ of the slice equation are at $s=0$. The fact that $\mathrm{D}_{(\nu,w)}h$ is a linear combination of $\mathrm{D}_{(\nu,w)}C^a$ and $\mathrm{D}_{(\nu,w)}C^r$ at any equilibrium of the slice equation then implies that

$$\mathrm{D}_s C^a(E,C^r,0) = 0.$$

Moreover, $\mathrm{D}_s\nu_3^a(0,0,0) = 0$ since $\nu_3^a(\nu^r, C^r=0) \equiv 0$. This together with (5.5) gives

$$\mathrm{D}_s^2 C^a(E_0,0,0) = 2.$$

So for small positive s and fixed $E \approx E_0$, $C^r \approx 0$, the function $s \to C^a(E,C^r,s)$ is injective. This proves that there is a coordinate transformation from (E,C^r,s) to the conserved quantities (E,C^r,C^a). Therefore, $\hat{x}(E,C^r,s)$ lies on an RPO $\mathcal{P}(E,C^r,s)$ of (3.1), as claimed. Let us now change coordinates back from (E,C^r,s) to (ν^r,s,ν_3^a) and denote the corresponding function again by $\hat{x}(\nu^r,s,\nu_3^a) \simeq (\gamma_{\mathrm{RE}}(\nu^r,\nu_3^a), \nu(\nu^r,s,\nu_3^a), w(\nu^r,s,\nu_3^a)) \in \mathcal{P}(\nu^r,s,\nu_3^a)$.

It is now shown that a smooth function $\hat{\gamma}(\nu^r,s,\nu_3^a)$ can be found such that

$$x(\nu^r,s,\nu_3^a) \simeq (\gamma(\nu^r,s,\nu_3^a), v(\nu^r,s,\nu_3^a)),$$

where

$$\gamma(\nu^r,s,\nu_3^a) = \hat{\gamma}(\nu^r,s,\nu_3^a)\gamma_{\mathrm{RE}}(\nu^r,\nu_3^a) \quad \text{and} \quad v(\nu^r,s,\nu_3^a) = (\nu(\nu^r,s,\nu_3^a), w(\nu^r,s,\nu_3^a)),$$

satisfies

$$\mathbf{J}_1^a(x(\nu^r,s,\nu_3^a)) = 0, \quad \mathbf{J}_1^r(x(\nu^r,s,\nu_3^a)) = 0, \quad \mathbf{J}_2^r(x(\nu^r,s,\nu_3^a)) = 0, \tag{5.6}$$

and $\hat{\gamma}(\nu^r,0,\nu_3^a) = \mathrm{id}$. First note that this holds true at $s=0$ by Proposition 5.4. For $s \neq 0$, let

$$\hat{\gamma}(\nu^r,s,\nu_3^a) = (\hat{R}(\nu^r,s,\nu_3^a), \hat{a}(\nu^r,s,\nu_3^a)),$$

and $\hat{\mu}(\nu^r,s,\nu_3^a) = \mathbf{J}(\hat{x}(\nu^r,s,\nu_3^a))$. Then $\mathbf{J}^a(x_{\mathrm{RE}}(\nu^r,\nu_3^a)) = \nu_3^a e_3$ implies that $\mathrm{D}_{\nu_3^a}\hat{\mu}^a(0,0,0) = e_3$. From $\gamma_{\mathrm{RE}}(0,0) = \mathrm{id}$ one further gets $\mathrm{D}_s\hat{\mu}^a(0,0,0) = e_2$. Therefore, smooth functions $\hat{\phi}_2(\nu^r,s,\nu_3^a)$, $\hat{\phi}_3(\nu^r,s,\nu_3^a)$ can be found such that $\hat{\phi}_j(\nu^r,0,\nu_3^a) = 0$, $j=2,3$, and

$$(\hat{R}(\nu^r,s,\nu_3^a)\hat{\mu}^a(\nu^r,s,\nu_3^a))_1 = 0,$$

where, as in (2.9),

$$\hat{R}(\nu^r, s, \nu_3^a) = \exp(\hat{\phi}_2(\nu^r, s, \nu_3^a)\xi_2 + \hat{\phi}_3(\nu^r, s, \nu_3^a)\xi_3).$$

In this way the first equation of (5.6) is satisfied. Then ν_3^a and $\nu_2^a = s$ can be rescaled such that

$$\mathbf{J}^a(x(\nu^r, \nu_2^a, \nu_3^a)) = (0, \nu_2^a, \nu_3^a).$$

Let

$$\tilde{\mu}(\nu^r, \nu_2^a, \nu_3^a) = \mathrm{Ad}^*_{(\hat{R}(\nu^r, \nu_2^a, \nu_3^a), 0)^{-1}} \hat{\mu}(\nu^r, \nu_2^a, \nu_3^a).$$

Note that $\tilde{\mu}^a(\nu^r, \nu_2^a, \nu_3^a) = (0, \nu_2^a, \nu_3^a)^T$ has been achieved by choosing $\hat{R}(\nu^r, \nu_2^a, \nu_3^a)$. Moreover, (5.2) gives

$$\mathrm{Ad}_{(0,\hat{a})^{-1}} \tilde{\mu}^r = \tilde{\mu}^r + (\nu_3^a \hat{a}_2 - \nu_2^a \hat{a}_3, -\nu_3^a \hat{a}_1, \nu_2^a \hat{a}_1)^T.$$

Let $\hat{a}_2 \equiv 0$. When $\nu_2^a \neq 0$, then $\hat{a}_3 = \hat{a}_3(\nu^r, \nu_2^a, \nu_3^a)$ can be chosen such that

$$(\mathrm{Ad}^*_{(0,\hat{a}(\nu^r, \nu_2^a, \nu_3^a))^{-1}} \tilde{\mu}^r(\nu^r, \nu_2^a, \nu_3^a))_1 = 0,$$

thus satisfying the second equation of (5.6). If $\nu_2^a = 0$, i.e., at the relative equilibria, $\tilde{\mu}^r(\nu^r, 0, \nu_3^a) = \nu^r e_3$ anyway, and when $\nu_2^a \to 0$, then $\hat{a}_3 \to \mathrm{D}_{\nu_2^a} \tilde{\mu}_1^r(\nu^r, 0, \nu_3^a)$. Moreover, the equation

$$(\mathrm{Ad}^*_{(0,\hat{a}(\nu^r, \nu_2^a, \nu_3^a))^{-1}} \tilde{\mu}^r(\nu^r, \nu_2^a, \nu_3^a))_2 = 0,$$

and hence the last equation of (5.6), can be satisfied by choosing $\hat{a}_1(\nu^r, \nu_2^a, \nu_3^a)$ appropriately whenever $\nu_3^a \neq 0$. When $\nu_3^a = 0$, then $C^r = 0$ and $\nu_{\mathrm{RE}}(\nu^r, 0) = (\nu^r, 0)$. In particular, $\nu_{\mathrm{RE},1}^a(\nu^r, 0) = 0$, and so also $\nu_1^a(\nu^r, s, 0) = 0$ (see (5.5)). Therefore, $\hat{R}(\nu^r, s, 0) = \mathrm{id}$ and $\tilde{\mu}_2^r(\nu^r, \nu_2^a, 0) = 0$. Consequently, when $\nu_2^a \to 0$, then $\hat{a}_2 \to \mathrm{D}_{\nu_2^a} \tilde{\mu}_2^r(\nu^r, \nu_2^a, 0)$. Hence a smooth function $\hat{a}(\nu^r, \nu_2^a, \nu_3^a)$ has been found such that $\mu^r(\nu^r, \nu_2^a, \nu_3^a) := \mathrm{Ad}^*_{(0,\hat{a})^{-1}} \tilde{\mu}^r(\nu^r, \nu_2^a, \nu_3^a) \,||\, e_3$. For $\nu_2^a = 0$ the equality $\mu^r(\nu^r, 0, \nu_3^a) = \mu_0^r + \nu^r e_3$ holds. So coordinates can be changed such that $\mu^r(\nu^r, \nu_2^a, \nu_3^a) = \mu_0^r + \nu^r e_3$ for all $\nu_2^a \geq 0$, $\nu_2^a \approx 0$. This proves part (a) of the proposition.

For part (b), let $\gamma(\nu^r, \nu_2^a, \nu_3^a) = (R(\nu^r, \nu_2^a, \nu_3^a), a(\nu^r, \nu_2^a, \nu_3^a))$ be the drift symmetry of the RPO $\mathcal{P}(\nu^r, \nu_2^a, \nu_3^a)$ at $x(\nu^r, \nu_2^a, \nu_3^a)$, and write, as in (2.9), $R(\nu^r, \nu_2^a, \nu_3^a) = \exp(\sum_{i=1}^3 \phi_i \xi_i)$, where $\phi_i = \phi_i(\nu^r, \nu_2^a, \nu_3^a)$. Then $\phi_3(\nu^r, \nu_2^a, \nu_3^a) = 0$ at $\nu_3^a = 0$ needs to be satisfied. Equations (3.17) and (5.2) imply that $R(\nu^r, \nu_2^a, \nu_3^a)\mathbf{J}^a(x(\nu^r, \nu_2^a, \nu_3^a)) = \mathbf{J}^a(x(\nu^r, \nu_2^a, \nu_3^a))$, where $\mathbf{J}^a(x(\nu^r, \nu_2^a, \nu_3^a)) = (0, \nu_2^a, \nu_3^a)$. Therefore, $\sum_{i=1}^3 \phi_i \xi_i = \hat{\phi}(0, \nu_2^a, \nu_3^a)^T$ for some $\hat{\phi} \in \mathbb{R}$, where ξ_i is identified with $e_i \in \mathbb{R}^3$ and $\mathrm{so}(3)$ and $(\mathbb{R}^3)^*$ with $\mathbb{R}^3$. Because of this, $\phi_3 = 0$ for $\nu_3^a = 0$. ∎

In addition to the family of RPOs from the above theorem, there may be additional families of RPOs which rotate and translate about the same axis (without loss of generality the e_3-axis) as the relative equilibrium.

Proposition 5.6. *Let, as before,* $\mathrm{SE}(3)x_0$ *be a relative equilibrium of an* $\mathrm{SE}(3)$*-equivariant Hamiltonian system* (3.1)*, with momentum* $\mu_0 = (\mu_0^r, 0)$*,* $\mu_0^r \neq 0$*, and with nonvanishing rotational velocity vector* $\xi_0^r \neq 0$*. Choose* x_0 *such that* $\mu_0^r || e_3$*,* $\xi_0^r = \omega_0^{\mathrm{rot}} e_3$*,* $\xi_0^a || e_3$*. Assume that the relative equilibium is nonresonant and elliptic in the sense of Definition* 4.4 *and that* $\omega_0^{\mathrm{rot}}/\omega_j \notin \mathbb{Z}$ *for all eigenvalues* $\mathrm{i}\omega_j$ *of* $\mathbb{J}_{\mathcal{N}_1} \mathrm{D}_w^2 h(0)$*. Then there are three-dimensional manifolds* $\mathcal{P}_j(\nu^r, s, \nu_3^a)$ *of RPOs,* $j = 1, \ldots \frac{1}{2} \dim \mathcal{M} - 5$*, and smooth functions* $x_j(\nu^r, s, \nu_3^a) \in \mathcal{P}_j(\nu^r, s, \nu_3^a)$

such that $x_j(\nu^r, 0, \nu_3^a) = x_{\mathrm{RE}}(\nu^r, \nu_3^a)$ *(with* $x_{\mathrm{RE}}(\nu^r, \nu_3^a)$ *from Proposition 5.4). Moreover, the RPO through* $x_j(\nu^r, s, \nu_3^a)$ *has momentum* $(\mu_0^r + \nu^r e_3, \nu_3^a e_3)$*, energy* $H(x_j(\nu^r, s, \nu_3^a)) = H(x_{\mathrm{RE}}(\nu^r, \nu_3^a)) \pm s^2$ *(depending on the Krein signature of* ω_j*), relative period* $T_j(\nu^r, s, \nu_3^a)$*, where* $T_j(0,0,0) = 2\pi/|\omega_j|$*, and average drift velocity*

$$\xi_j(\nu^r, s, \nu_3^a) = (\xi_{j,3}^r(\nu^r, s, \nu_3^a)e_3, \xi_{j,3}^a(\nu^r, s, \nu_3^a)e_3), \xi_j(0,0,0) = \xi_0.$$

Proof. Let us, as in the proof of Proposition 5.4, reduce only by the symmetry group $\tilde{\Gamma} = \mathrm{SO}_3(2) \times \mathbb{R}_3$. The statement then follows by applying the Lyapunov center theorem on the $\dot{\tilde{w}}$-equation of (5.4). ∎

Conclusion and future directions. In this paper a Hamiltonian analogue of the well-known meandering transition from rotating waves to modulated rotating waves and modulated traveling waves in systems with Euclidean symmetries has been studied. This transition occurs, for example, in a finite-dimensional system of point vortices. Similar effects have been analyzed in systems with spherical symmetry and with the Euclidean symmetry of three-dimensional space. It remains a challenging open problem to extend these results to infinite-dimensional Hamiltonian systems such as PDE models of vortex dynamics.

REFERENCES

[1] H. Aref, *Addendum: "Three-vortex motion with zero total circulation,"* Z. Angew. Math. Phys., 40 (1989), pp. 495–500.

[2] V. I. Arnold and B. A. Khesin, *Topological Methods in Hydrodynamics*, Springer-Verlag, New York, Berlin, Heidelberg, 1998.

[3] D. Barkley, *Euclidean symmetry and the dynamics of rotating spiral waves*, Phys. Rev. Lett., 72 (1994), pp. 164–167, http://link.aps.org/abstract/PRL/v72/p164.

[4] D. Chan, *Hopf bifurcations from relative equilibria in spherical geometry*, J. Differential Equations, 226 (2006), pp. 118–134.

[5] D. Chan, *A Normal Form Approach to Nonresonant and Resonant Hopf Bifurcation from Relative Equilibria*, Ph.D. thesis, University of Surrey, Guildford, UK, 2007.

[6] D. Chan and I. Melbourne, *A geometric characterisation of resonance in Hopf bifurcation from relative equilibria*, Phys. D, 234 (2007), pp. 98–104.

[7] A. Comanici, *Transition from rotating waves to modulated rotating waves on the sphere*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 759–782.

[8] B. Fiedler, B. Sandstede, A. Scheel, and C. Wulff, *Bifurcation from relative equilibria to non-compact group actions: Skew products, meanders, and drifts*, Doc. Math., 1 (1996), pp. 479–505.

[9] V. Ginzburg and E. Lerman, *Existence of relative periodic orbits near relative equilibria*, Math. Res. Lett., 11 (2004), pp. 397–412.

[10] M. Golubitsky, V. LeBlanc, and I. Melbourne, *Meandering of the spiral tip—an alternative approach*, J. Nonlinear Sci., 7 (1997), pp. 557–586.

[11] V. Guillemin and S. Sternberg, *A normal form for the moment map*, in Differential Geometric Methods in Mathematical Physics, S. Sternberg, ed., D. Reidel, Dordrecht, The Netherlands, 1984.

[12] V. Guillemin and S. Sternberg, *Symplectic Techniques in Physics*, Cambridge University Press, Cambridge, UK, 1984.

[13] G. R. Kirchhoff, *Vorlesungen über Mathematische Physik. Mechanik*, Teubner, Leipzig, 1876.

[14] M. Krupa, *Bifurcations of relative equilibria*, SIAM J. Math. Anal., 21 (1990), pp. 1453–1486.

[15] N. E. Leonard and J. E. Marsden, *Stability and drift of underwater vehicle dynamics: Mechanical systems with rigid motion symmetry*, Phys. D, 105 (1997), pp. 130–162.

[16] C. Lim, J. A. Montaldi, and R. M. Roberts, *Relative equilibria of point vortices on the sphere*, Phys. D, 148 (2001), pp. 97–135.
[17] C.-M. Marle, *Modèle d'action Hamiltonienne d'un groupe de Lie sur une variété symplectique*, Rend. Sem. Mat. Univ. Politec. Torino, 43 (1985), pp. 227–251.
[18] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, Berlin, Heidelberg, 1994.
[19] K. R. Meyer and G. R. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Springer-Verlag, New York, 1992.
[20] J. Montaldi and T. Tokieda, *Openness of momentum maps and persistence of extremal relative equilibria*, Topology, 42 (2003), pp. 833–844.
[21] P. Newton, *The N-Vortex Problem*, Appl. Math. Sci. 145, Springer-Verlag, New York, 2001.
[22] J.-P. Ortega, *Relative normal modes for nonlinear Hamiltonian systems*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 665–704.
[23] J.-P. Ortega and T. Ratiu, *Relative equilibria near stable and unstable Hamiltonian relative equilibria*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 1407–1431.
[24] R. S. Palais, *On the existence of slices for actions of noncompact Lie groups*, Ann. of Math. (2), 73 (1961), pp. 295–323.
[25] G. W. Patrick, *Dynamics of perturbed relative equilibria of point vortices on the sphere or plane*, J. Nonlinear Sci., 10 (2000), pp. 401–415.
[26] R. M. Roberts, C. Wulff, and J. Lamb, *Hamiltonian systems near relative equilibria*, J. Differential Equations, 179 (2002), pp. 562–604.
[27] B. Sandstede, A. Scheel, and C. Wulff, *Bifurcations and dynamics of spiral waves*, J. Nonlinear Sci., 9 (1999), pp. 439–478.
[28] U. Scheerer and C. Wulff, *Reduced dynamics for momentum maps with cocycles*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 999–1004.
[29] J. L. Synge, *On the motion of three vortices*, Canadian J. Math., 1 (1949), pp. 257–270.
[30] A. T. Winfree, *Varieties of spiral wave behaviour: An experimentalist's approach to the theory of excitable media*, Chaos, 1 (1991), pp. 303–334.
[31] C. Wulff, *Transition from relative equilibria to relative periodic orbits*, Doc. Math., 5 (2000), pp. 227–274.
[32] C. Wulff, *Persistence of relative equilibria in Hamiltonian systems with noncompact symmetry*, Nonlinearity, 16 (2003), pp. 67–91.
[33] C. Wulff, *Persistence of Hamiltonian relative periodic orbits*, J. Geom. Phys., 48 (2003), pp. 309–338.
[34] C. Wulff and M. Roberts, *Hamiltonian systems near relative periodic orbits*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 1–43.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1247–1282
© 2008 Society for Industrial and Applied Mathematics

Traveling Waves and Synchrony in an Excitable Large-Scale Neuronal Network with Asymmetric Connections*

William C. Troy†

Abstract. We study (i) traveling wave solutions, (ii) the formation and spatial spread of synchronous oscillations, and (iii) the effects of variations of threshold in a system of integro-differential equations which describe the activity of large-scale networks of excitatory neurons on spatially extended domains. The independent variables are the activity level u of a population of excitatory neurons which have long range connections, and a recovery variable v. In the integral component of the equation for u the firing rate function is the Heaviside function, and the coupling function w is positive. Thus, there is no inhibition in the system. There is a critical value of the parameter β ($\beta_* > 0$) that appears in the equation for v, at which the eigenvalues $\mu^{\pm}$ of the linearization of the system around the rest state $(u, v) = (0, 0)$ change from real to complex. We focus on the range $\beta > \beta_*$, where $\mu^{\pm}$ are complex, and analyze properties of wave fronts and 1-pulse and 2-pulse waves when the connection function w is asymmetric. For wave fronts we demonstrate how an initial stimulus evolves into two solutions which propagate in opposite directions with different speeds and shapes. For 1-pulse waves our main theoretical result (Theorem 4.2) shows that there is a range of $\beta > \beta_*$ where two families of waves exist, each consisting of infinitely many solutions. The waves in these two families also propagate in opposite directions with different speeds and shapes. There is a critical value $\theta^* > 0$ such that if $\theta > \theta^*$, then 1-pulse waves can propagate only in one direction. In addition, there is a second critical β value, $\beta^* > \beta_*$, where bulk oscillations come into existence and the system becomes bistable. When $\beta \geq \beta_*$ we show how an initial stimulus evolves into a solution with large amplitude oscillations that spread out uniformly from the point of stimulus. The asymmetry in w causes the rate of spread of the "region of synchrony" to be more rapid to the right of the point of stimulus than to the left. When $\theta > \theta^*$ we construct a "unidirectional" circuit where synchronization in one region can trigger synchronization in a distant, second region. However, when synchronization is initially triggered in the second region, it cannot spread to the first region.

Key words. waves, integro-differential equation, nonlocal, excitatory

AMS subject classifications. 34B15, 34C23, 34C11

DOI. 10.1137/070709888

1. Introduction.

Functional behavior of the central nervous system includes such diverse phenomena as information processing from different receptor zones, sleep, and the control of vital autonomic functions [29, 30, 40]. These processes require cooperation between ensembles of cells organized into large-scale, spatially extended neuronal networks. The physical laws that govern the behavior of large-scale networks are different from those for a system consisting of small numbers of cells [15, 26, 34, 56]. In the study of spatially extended neuronal networks

*Received by the editors December 2, 2007; accepted for publication (in revised form) by T. Kaper July 10, 2008; published electronically October 24, 2008. This research was supported by NSF grant DMS0412370. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siads/7-4/70988.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (troy@math.pitt.edu).

considerable attention has been given to (i) traveling waves of activity, (ii) the formation and spatial spread of synchronous oscillations, and (iii) the effects of variations in the threshold of excitation. This includes both experimental [28, 3, 4, 5, 10, 18, 23, 31, 33, 36, 39, 37, 45, 46, 49, 53, 61, 57, 60] and theoretical [1, 12, 13, 14, 17, 2, 19, 20, 24, 25, 22, 32, 35, 38, 40, 42, 58, 59, 62] studies.

In this paper we investigate traveling waves, the spread of synchronous oscillations, and the effects of variations in threshold in the excitable, spatially extended model

$$\begin{aligned} \tfrac{\partial u(x,t)}{\partial t} &= -u(x,t) - v(x,t) + \textstyle\int_{-\infty}^{\infty} w(x-x')f(u(x',t)-\theta)dx' + \zeta(x,t), \\ \tfrac{\partial v(x,t)}{\partial t} &= \epsilon(\beta u(x,t) - v(x,t)). \end{aligned} \tag{1.1}$$

Systems of this form were introduced by Pinto and Ermentrout [42] to model the spread of excitation waves in slices of brain cortex in which synaptic inhibition is pharmacologically blocked [7, 10, 31, 61]. The variable u denotes the activity level of the population of excitatory neurons with long range connections; v represents a negative feedback recovery mechanism in which "the negative feedback could represent spike frequency adaptation, synaptic depression or some other process that limits excitation of the network" [42]. The function $\zeta(x,t)$ represents external input to the system. The parameters ϵ and β are positive and control the rate of change of v; θ is a positive constant which denotes the threshold level for u. We assume that the coupling function w is strictly positive. Thus, there is no inhibition in the system. We also assume that w is continuous, integrable, and either symmetric, i.e.,

$$w(x) = w(-x) \quad \forall x \in (-\infty, \infty), \tag{1.2}$$

or asymmetric, i.e.,

$$w(x) \neq w(-x) \quad \text{for some} \quad x \in (-\infty, \infty). \tag{1.3}$$

The firing rate function f is nonnegative and sigmoidal-shaped. In order to allow for comparison of our results with those of previous studies, we follow [2, 19, 42, 48, 58] and set

$$f(u-\theta) = H(u-\theta) = \begin{cases} 1 & \forall u \geq \theta, \\ 0 & \forall u < \theta, \end{cases} \tag{1.4}$$

and

$$w(x) = \frac{1}{2}e^{-|x|+\kappa x}, \quad 0 \leq \kappa \leq 1, \quad \text{and} \quad x \in \mathrm{R}. \tag{1.5}$$

We will make use of the observation that if $\kappa > 0$, then

$$w(x) > w(-x) \quad \forall x \in (-\infty, \infty). \tag{1.6}$$

In our numerical studies we employ two methods to initiate a solution such as a traveling wave. The first is to set the external stimulus to zero, i.e., $\zeta(x,0) \equiv 0$, and let $(u(x,0), v(x,0))$ be a perturbation from the rest state $(0,0)$ which has the typical form

$$(u(x,0), v(x,0)) \equiv (Ae^{-Bx^2}, 0), \quad x \in \mathrm{R},\ A > 0, \text{ and } B > 0. \tag{1.7}$$

Equivalently, we could set $(u(x,0),v(x,0)) = (0,0)$ and set $\zeta(x,0) = Ae^{-Bx^2}$.

Goals. We investigate the dynamics of (1.1) when w is asymmetric and the eigenvalues $\mu^{\pm}$ of the linearization of (1.1) around the rest state $(u,v) = (0,0)$ are *complex*. Section 2 shows how $\mu^{\pm}$ change from real to complex as β passes through the critical value $\beta_* = \frac{(\epsilon-1)^2}{4\epsilon}$ from below. As β increases from β_* the imaginary part of $\mu^{\pm}$ increases and solutions of the linearization oscillate with increasing frequency. In turn, this causes the dynamics of (1.1) to become more complicated. Thus, we study how the structure of families of traveling waves changes as β increases from β_*. There is a critical value $\beta^* > \beta_*$ where bulk oscillations appear and (1.1) becomes bistable. When $\beta \geq \beta^*$ we investigate the formation and spatial spread of synchronous oscillations. To put our investigation into perspective we summarize previous results in (A)–(C) below. Our specific aims are described in (D).

(A) *Symmetric couplings and real eigenvalues.* Pinto et al. [42, 43, 44] analyzed 1-pulse waves in parameter regimes where w is symmetric and $\mu^{\pm}$ are *real*. Richardson, Schiff, and Gluckman [48] make use of the results in [44] to study the effects of raising electric fields on 1-pulse waves in the mammalian cortex. The stability of solutions in the real eigenvalue setting has been studied in [12, 44, 50].

(B) *Asymmetric couplings and real eigenvalues.* Pinto and Troy [47] analyzed 1-pulse waves when w is asymmetric and $\mu^{\pm}$ are real. The asymmetry in w causes an initial stimulus of the form (1.7) to evolve into two 1-pulse waves which propagate in opposite directions with different amplitudes and speeds. Thus, traveling waves have a preferred direction of propagation when the coupling is asymmetric. In agreement with these theoretical results, we give experimental evidence which indicates that 1-pulse waves in the barrel cortex have a preferred direction of propagation.

(C) *Symmetric couplings and complex eigenvalues.* Troy and Shusterman [58] investigated (1.1) in parameter regimes where the coupling is the symmetric function $w(x) = \frac{1}{2}e^{-|x|}$ and $\beta > \beta_*$ where $\mu^{\pm}$ are *complex*. We analyzed one-dimensional wave fronts, single-pulse waves, and multipulse wave trains. We also explained why multipulse waves are expected to exist only in the complex eigenvalue regime. In two dimensions we analyzed the periodic formation of waves, as well as the mechanisms responsible for the formation of spiral waves. Spiral waves were recently discovered in tangential slices of rat brain tissue [31]. Thus, in the complex eigenvalue setting the dynamics of (1.1) are richer than in the real eigenvalue case. Furthermore, these dynamics closely resemble electrophysiological phenomena observed in clinical and experimental studies [40].

(D) *Specific aims: Asymmetric couplings and complex eigenvalues.* In this paper we extend the results described in (A)–(C) and investigate the dynamics of (1.1) when w is asymmetric and the eigenvalues $\mu^{\pm}$ are complex. Our goals are summarized in I–III below.

I. *Traveling waves.* Single-pulse and multipulse traveling activity waves have been observed in feline cortex [3, 4, 5], in the brain of freely moving mice [18], in tangential and coronal brain slice experiments [31, 61], and in seizure propagation across cortical regions [39, 60]. In sections 3–5 we analyze wave fronts, 1-pulse waves, and 2-pulse waves when $\beta > \beta_*$ and $w(x) = \frac{1}{2}e^{-|x|+\kappa x}$, where $\kappa \neq 0$. Wave front solutions cross the threshold level $u = \theta$ precisely once, whereas N-pulse waves cross threshold exactly 2N times. For a fixed initial stimulus we find that, as β increases, there is a natural evolution from wave fronts to 1-pulse waves, and subsequently to 2-pulse waves. Our main theoretical result (Theorem 4.2)

shows that there is a range of $\beta > \beta^*$ where two families of 1-pulse waves exist, each consisting of infinitely many coexisting solutions. The waves in these two families propagate in opposite directions with different speeds and shapes. Infinite families of solutions with such diverse properties do not exist when w is asymmetric and $\mu^{\pm}$ are real. Similar properties hold for wave fronts and 2-pulse waves. Because the eigenvalues are complex, technical difficulties make proofs more challenging than in the real eigenvalue case. These difficulties lead to open problems which are stated as we proceed.

II. *Synchronous oscillations.* There is a second critical value $\beta^* > \beta_*$ where spatially independent bulk oscillations come into existence (section 6). When $\beta \geq \beta^*$ the system (1.1) is bistable since these oscillations are stable and coexist with the stable rest state $(u, v) = (0, 0)$. An initial stimulus of the form (1.7) can evolve into a solution which exhibits large amplitude oscillations that spread out uniformly from the point of stimulus. The asymmetry in w causes the rate of spread of the "region of synchrony" to be more rapid to the right of the point of stimulus than to the left. Eventually, however, the solution oscillates uniformly over the entire spatial region. We also show how an initial stimulus can evolve into a stable 1-pulse wave which coexists with synchronous oscillations and the stable rest state. In section 6 we compare our theoretical results with clinical observations of epileptiform events.

III. *Variations in threshold.* In 1936 Hill [28] suggested that the value of threshold might change in response to the state of neuronal tissue. Following Hill's theoretical ideas, Coombes and Owen [13, 14] studied the effects of a state dependent threshold on bump-type solutions in a scalar Wilson–Cowan-type model in which w is of Mexican hat type; i.e., $w(x)$ is symmetric about $x = 0$ and changes sign on $(-\infty, \infty)$. Recently, Kowai, Lazar, and Metherate [35] have given experimental evidence which shows that the threshold of excitation can indeed change. In particular, when they expose axons of thalamocortical mouse neurons to nicotine, the threshold of excitation *decreases* and the firing rate of the neurons *doubles*. In [54] it is suggested that this causes "an increase in the amount of sensory information reaching the cortex," and that "this is a major reason that nicotine enhances cognitive functioning." It is also pointed out that in schizophrenia there is poor communication between the thalamus and the cortex, and therefore the high incidence of smoking in schizophrenics might be a method of self-medication. In our analysis of traveling waves and synchrony we investigate the effects of both increasing and decreasing the threshold θ, and also the strength κ of the asymmetric coupling. In agreement with [35] we find that the amplitudes and speeds of traveling waves *increase* dramatically as θ *decreases*. Because w is asymmetric there is a critical value θ^* such that waves can be transmitted in only one direction when $\theta > \theta^*$. In section 6 we investigate how variations in threshold can affect synchrony. As a first step toward understanding how communication between spatial regions can become inhibited when threshold is too high, we let $\theta > \theta^*$ and study how synchronization in one region can influence synchronization in a distant region. For this we construct a simple "unidirectional neuronal circuit" in which $\theta > \theta^*$ so that waves propagate only to the right. In this setting synchronization in one region causes the formation of a train of waves which propagate to the right and ultimately trigger a second, distant region to undergo synchronization. However, because $\theta > \theta^*$, synchronization in the second region is inhibited from emitting left propagating waves; hence the first region remains at rest.

Conclusions and directions for future research are given in section 7.

2. Traveling waves. Following [44, 58], we set $\zeta(x,t) = 0$ and look for traveling wave solutions of (1.1) of the form $(u,v) = (U(z), V(z))$, where $z = x + ct$. These satisfy

$$\begin{aligned} cU'(z) &= -U - V + \textstyle\int_{-\infty}^{\infty} w(z-z')H(U(z')-\theta)dz', \\ cV'(z) &= \epsilon(\beta U - V), \end{aligned} \tag{2.1}$$

where

$$w(z-z') = \frac{1}{2}e^{-|z-z'|+\kappa(z-z')}, \quad 0 \le \kappa < 1. \tag{2.2}$$

Our main focus is on the regime $\kappa \neq 0$, where the coupling is asymmetric. For simplicity we restrict our attention to the case $\kappa > 0$. It is easily verified that (2.1) is equivalent to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = c\frac{d}{dz}\int_{-\infty}^{\infty} w(z-z')H(U-\theta)dz' + \epsilon\int_{-\infty}^{\infty} w(z-z')H(U-\theta)dz'. \tag{2.3}$$

Linearizing (2.3) around the rest state $U = 0$ leads to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = 0. \tag{2.4}$$

The eigenvalues associated with (2.4) are

$$\mu^{\pm} = \frac{\lambda^{\pm}}{c} = \frac{-(\epsilon+1) \pm \mathrm{i}\sqrt{4\beta\epsilon - (\epsilon-1)^2}}{2c}. \tag{2.5}$$

We restrict our attention to the regime $0 < \epsilon < 1$. From this and (2.5) it follows that

$$\mu^{\pm} \text{ are real} \iff 0 < \beta \le \beta_* = \frac{(\epsilon-1)^2}{4\epsilon}. \tag{2.6}$$

Remarks. (i) When $\mu^{\pm}$ are real one expects to find only wave fronts or single-pulse waves [42, 43, 44, 48, 58]. Our interest is in the regime $\beta > \beta^*$, where $\mu^{\pm}$ are complex and the dynamics of (1.1) are richer. For example, when w is symmetric we previously found the coexistence of families of multipulse waves in one space dimension and periodic waves and rotating waves in two dimensions [58].

(ii) We have followed [44, 58] and let the independent variable have the form $z = x + ct$. Thus, when $c > 0$, solutions of (2.1)–(2.2) correspond to traveling waves of (1.1) which propagate "to the left" as t increases. If $z = x - ct$, then traveling wave solutions propagate "to the right."

Previous studies analyzed traveling waves when w is symmetric [2, 19, 20, 42, 44, 48, 58]. In this setting it can be assumed that $c > 0$, since a wave traveling with speed $c > 0$ can be transformed into a wave traveling in the opposite direction with the same shape, and with speed $-c < 0$. Here we study traveling waves when $\mu^{\pm}$ are complex and the w is asymmetric. Our numerical experiments in the next three sections show that an initial perturbation from rest can cause two waves to form which propagate in opposite directions with different speeds and shapes. We will make use of the quantities

$$\alpha = \mathrm{Re}(\mu^{\pm}) = \frac{-(\epsilon+1)}{2c} \quad \text{and} \quad \gamma = \mathrm{Im}(\mu^{\pm}) = \frac{\sqrt{4\epsilon\beta - (\epsilon-1)^2}}{2c}. \tag{2.7}$$

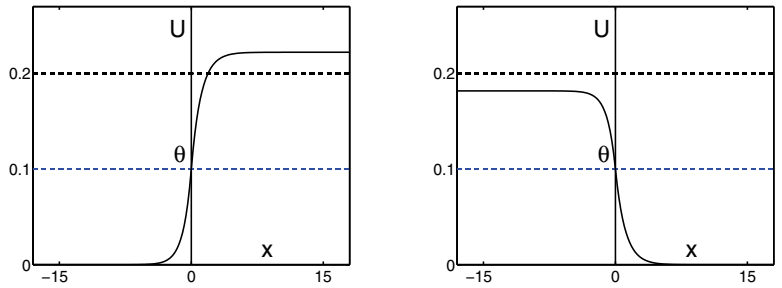


Figure 1. *Left panel: Stationary front solution of* (3.8)–(3.9)*, with* $U(\infty) = \frac{2\theta}{1-\kappa} = .23$ *when* $(\theta, \kappa) = (.1, .15)$ *and* $\beta = \beta_1 = \frac{0.5}{(\kappa+1)\theta} - 1 = 3.35$*. Right panel: Solution with* $U(-\infty) = \frac{2\theta}{1+\kappa} = .173$ *when* $\beta = \beta_2 = \frac{.5}{(1-\kappa)\theta} - 1 = 4.88$.

Our investigation indicates that stable traveling waves exist when

$$0 < \epsilon, \kappa < 1, \quad 0 < \theta < \min\left(\frac{2\epsilon}{(1-\kappa^2)(\epsilon+1)^2}, \frac{1}{4(\epsilon+1)}\right), \quad \beta > \beta_* = \frac{(\epsilon-1)^2}{4\epsilon}. \tag{2.8}$$

The first two sets of inequalities in (2.8) are mild restrictions which allow technical arguments to be completed, and the third inequality means that $\mu^{\pm}$ are complex. Throughout the paper we perform numerical experiments for parameters which satisfy (2.8). An important implication of (2.8) is that

$$\beta_* = \frac{(\epsilon-1)^2}{4\epsilon} < \beta_1 = \frac{1}{2(1+\kappa)\theta} - 1 < \beta_2 = \frac{1}{2(1-\kappa)\theta} - 1 \tag{2.9}$$

when $|\kappa|$ is small. In section 3 we will see that distinct branches of wave fronts come into existence at the critical values $\beta_1 = \frac{1}{2(1+\kappa)\theta} - 1$ and $\beta_2 = \frac{1}{2(1-\kappa)\theta} - 1$. Our analysis of 1-pulse waves in section 4 shows that infinitely many wave speeds are possible at $\beta = \beta_1$ and $\beta = \beta_2$. 2-pulse solutions are described in section 5. When $\mu^{\pm}$ are complex, underlying oscillatory terms lead to technical difficulties which make the completion of existence proofs especially challenging. These difficulties suggest open problems which are discussed as we proceed.

3. Wave fronts. We analyze wave front solutions when w is asymmetric. Although the discussion appears lengthy, it is necessary to give complete details in order to obtain a global understanding of the structure of solutions. Our study focuses on the following:

A. The construction of two families of stationary solutions with speed $c = 0$ (Figure 1).
B. Properties of wave fronts when $c > 0$. In this case a family of solutions bifurcates from a stationary solution (Figure 1, right panel) as c *increases* from $c = 0$.
C. Properties of wave fronts when $c < 0$. In this case a family of wave fronts with different speeds and shapes bifurcates from a second stationary solution (Figure 1, left panel) as c *decreases* from $c = 0$.
D. The effects of changing the threshold θ.
E. Open problems.

A. *Stationary solutions.* We set $c = 0$ and $w(x) = \frac{1}{2}e^{-|x-x'|+\kappa(x-x')}$ in (2.3) and investigate the existence of stationary wave front solutions of the form

$$U(x) = \frac{1}{2(\beta+1)} \int_{-\infty}^{\infty} e^{-|x-x'|+\kappa(x-x')} H\left(U(x') - \theta\right) dx'. \tag{3.1}$$

We find that there is a range of parameters for which two different solutions exist.

The first stationary solution. The first solution (Figure 1, left panel) satisfies

$$\begin{cases} U(x) \text{ and } U'(x) \text{ are continuous } \forall x \in (-\infty, \infty), \\ U(x) < \theta \ \forall x < 0,\ U(0) = \theta, \\ U(x) > \theta \ \forall x \in (0, \infty). \end{cases} \tag{3.2}$$

Without loss of generality we have assumed that $U(0) = \theta$ in (3.2) since (3.1) is translationally invariant. Then (3.2) reduces (3.1) to

$$U(x) = \frac{1}{2(\beta+1)} \int_{0}^{\infty} e^{-|x-x'|+\kappa(x-x')} dx'. \tag{3.3}$$

Solving (3.2)–(3.3) gives

$$U(x) = \begin{cases} \frac{0.5}{(\kappa+1)(\beta+1)} e^{(1+\kappa)x} & \forall x \le 0, \\ \frac{0.5}{\beta+1} \left(\frac{2}{1-\kappa^2} - \frac{e^{(\kappa-1)x}}{1-\kappa} \right), & x > 0, \end{cases} \tag{3.4}$$

where β, κ, θ satisfy

$$0 < \kappa < 1, \quad 0 < \theta < 1, \quad \beta = \beta_1 = \frac{0.5}{(\kappa+1)\theta} - 1. \tag{3.5}$$

This and (2.9) imply that

$$\beta_* < \beta_1 < \frac{1}{2\theta} - 1 \ \forall \kappa \in (0,1), \quad \beta_1 \to \frac{1}{2\theta} - 1 \text{ as } \kappa \to 0^+, \quad \beta_1 \to \frac{1}{4\theta} - 1 \text{ as } \kappa \to 1^-. \tag{3.6}$$

It follows from (3.4)–(3.5) that

$$\begin{cases} U'(x) > 0 \ \forall x \in (-\infty, \infty), \\ (U(x), U'(x)) \to \left(\frac{2\theta}{1-\kappa}, 0 \right) \text{ as } x \to \infty, \\ (U(x), U'(x)) \to (0,0) \text{ as } x \to -\infty. \end{cases} \tag{3.7}$$

The second stationary solution. The second solution (Figure 1, right panel) satisfies

$$\begin{cases} U(x) \text{ and } U'(x) \text{ are continuous } \forall x \in (-\infty, \infty), \\ U(x) > \theta \ \forall x \in (-\infty, 0),\ U(0) = \theta, \\ U(x) < \theta \ \forall x > 0. \end{cases} \tag{3.8}$$

Conditions (3.8) reduce (3.1) to

$$U(x) = \frac{1}{2(\beta+1)} \int_{-\infty}^{0} e^{-|x-x'|+\kappa(x-x')} dx'. \tag{3.9}$$

Solving (3.8)–(3.9) gives

$$U(x) = \begin{cases} \frac{.5}{\beta+1}\left(\frac{2}{1-\kappa^2} - \frac{e^{(1+\kappa)x}}{1+\kappa}\right), & x < 0, \\ \frac{.5}{(\beta+1)(1-\kappa)} e^{(\kappa-1)x} & \forall x \geq 0. \end{cases} \tag{3.10}$$

Conditions (3.8) hold when β, κ, θ, a satisfy

$$0 < \kappa < 1, \quad 0 < \theta < 1, \quad \beta = \beta_1 = \frac{.5}{(1-\kappa)\theta} - 1. \tag{3.11}$$

This implies that

$$\beta_1 > \frac{1}{2\theta} - 1 \ \forall \kappa \in (0,1), \quad \beta_1 \to \frac{1}{2\theta} - 1 \text{ as } \kappa \to 0^+, \quad \beta_1 \to +\infty \text{ as } \kappa \to 1^-. \tag{3.12}$$

Finally, it follows from (3.10)–(3.11) that U satisfies (3.8), and also

$$\begin{cases} U'(x) < 0 \ \forall x \in (-\infty, \infty), \\ (U(x), U'(x)) \to \left(\frac{2\theta}{1+\kappa}, 0\right) \text{ as } x \to -\infty, \\ (U(x), U'(x)) \to (0,0) \text{ as } x \to +\infty. \end{cases} \tag{3.13}$$

B. *Properties of wave fronts when $c > 0$.* In this and the next two sections we show how distinct branches of solutions bifurcate from the stationary solutions constructed above as c passes through $c = 0$. When $c > 0$ wave front solutions satisfy

$$\begin{cases} U(z) \text{ and } U'(z) \text{ are continuous } \forall z \in (-\infty, \infty), \\ U(z) < \theta \ \forall z < 0, \\ (U(z), U'(z)) \to (0,0) \text{ as } z \to -\infty, \\ U(0) = \theta, \quad U(z) > \theta \ \forall z > 0, \quad U'(z) \to 0 \text{ as } z \to \infty. \end{cases} \tag{3.14}$$

Again, without loss of generality we have assumed that $U(0) = \theta$ in (3.2) since (2.3) is translationally invariant. When conditions (3.14) hold, equation (2.3) reduces to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = \frac{c}{2}\frac{d}{dz}\int_0^\infty e^{-|z-z'|+\kappa(z-z')}dz' + \frac{\epsilon}{2}\int_0^\infty e^{-|z-z'|+\kappa(z-z')}dz'. \tag{3.15}$$

This further reduces to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = g(z), \tag{3.16}$$

where

$$g(z) = \begin{cases} .5(c + \frac{\epsilon}{1+\kappa})e^{(1+\kappa)z} & \forall z \leq 0, \\ .5(c - \frac{\epsilon}{1-\kappa})e^{-(1-\kappa)z} + \frac{\epsilon}{1-\kappa^2} & \text{if} \quad z > 0. \end{cases} \tag{3.17}$$

A combination of analysis and numerical experiments suggests that two branches of solutions coexist when $c > 0$ (Figure 2). To understand how these results are obtained, we investigate the following:

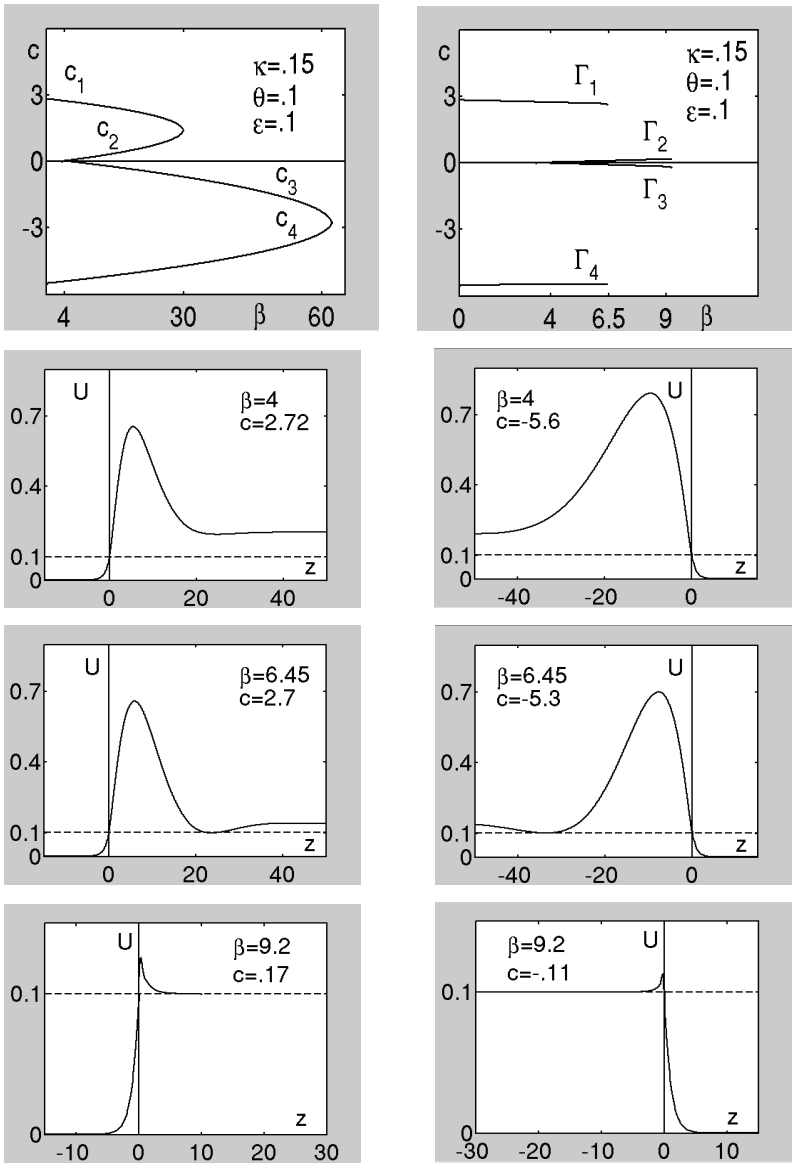


Figure 2. *Upper left:* c_1, c_2, c_3, *and* c_4 *versus* β *when* $(\epsilon, \theta, \kappa) = (.1, .1, .15)$. *Upper right:* Γ_1, Γ_2, Γ_3, *and* Γ_4 *in the* (β, c) *plane when* $(\epsilon, \theta, \kappa) = (.1, .1, .15)$. *Second row: Solutions at* $\beta = 4$ *on* Γ_1 *(left) and* Γ_4 *(right). Third row: Solutions at the right endpoints of* Γ_1 *(left) and* Γ_4 *(right), where* $\beta \approx 6.45$. *Fourth row: Solutions at the right endpoints of* Γ_2 *(left) and* Γ_3 *(right), where* $\beta = \frac{1}{(1-\kappa^2)\theta} - 1 \approx 9.2$.

(i) Analysis of solutions on $(-\infty, 0]$.
(ii) Analysis of solutions on $(0, \infty)$.
(iii) Numerical evidence for the existence of solutions.

(i) *Analysis of solutions on* $(-\infty, 0]$. On $(-\infty, 0]$ the solution of (3.16) is

$$U_0(z) = b_1 e^{\alpha z} \cos(\gamma z) + b_2 e^{\alpha z} \sin(\gamma z) + P_0(z), \tag{3.18}$$

where b_1 and b_2 are constants, and $P_0(z)$ is the particular solution

$$P_0(z) = \frac{.5(\epsilon + (1+\kappa)c)}{(1+\kappa)((1+\kappa)^2c^2 + (1+\kappa)(1+\epsilon)c + \epsilon(\beta+1))} e^{(1+\kappa)z} \quad \forall z \le 0. \tag{3.19}$$

The oscillatory terms $b_1 e^{\alpha z}\cos(\gamma z)$ and $b_2 e^{\alpha z}\sin(\gamma z)$ in (3.18) are due to $\mu^{\pm}$ being complex. Recall from (2.7) that $\alpha = \mathrm{Re}(\mu^{\pm}) < 0$ when $c > 0$. Thus, to satisfy the condition $U_0(-\infty) = U_0'(-\infty) = 0$, we conclude that $b_1 = b_2 = 0$, and (3.18)–(3.19) reduce to

$$U_0(z) = \frac{.5(\epsilon + (1+\kappa)c)}{(1+\kappa)((1+\kappa)^2c^2 + (1+\kappa)(1+\epsilon)c + \epsilon(\beta+1))} e^{(1+\kappa)z} \quad \forall z \le 0. \tag{3.20}$$

Substituting $U_0(0) = \theta$ into (3.20) gives the algebraic equation

$$\frac{.5(\epsilon + (1+\kappa)c)}{(1+\kappa)((1+\kappa)^2c^2 + (1+\kappa)(1+\epsilon)c + \epsilon(\beta+1))} = \theta. \tag{3.21}$$

It follows from (3.20)–(3.21) that

$$U_0(z) = \theta e^{(1+\kappa)z} < \theta \quad \forall z \le 0. \tag{3.22}$$

(ii) *Analysis of solutions on* $(0,\infty)$. When $z > 0$ the first step in analyzing solutions is to solve (3.21) for wave speed c. This gives the two positive values (Figure 2, upper left panel)

$$c_1 = \frac{.5 - \theta(1+\kappa)(\epsilon+1) + \sqrt{\left(.5 - \theta(1+\kappa)(1+\epsilon)\right)^2 - 4\epsilon\theta(1+\kappa)\left(\theta(1+\kappa)(\beta+1) - .5\right)}}{2\theta(1+\kappa)^2}, \tag{3.23}$$

$$c_2 = \frac{.5 - \theta(1+\kappa)(\epsilon+1) - \sqrt{\left(.5 - \theta(1+\kappa)(1+\epsilon)\right)^2 - 4\epsilon\theta(1+\kappa)\left(\theta(1+\kappa)(\beta+1) - .5\right)}}{2\theta(1+\kappa)^2}. \tag{3.24}$$

Recall from (2.8) that $0 < \epsilon < 1$ and $0 < \theta < \frac{1}{4(\epsilon+1)}$. This and (3.23)–(3.24) imply that

$$\begin{cases} c_1 > 0 \text{ if } 0 < \beta < \frac{1}{2\theta(1+\kappa)} - 1 + \frac{1}{4\epsilon\theta^2(1+\kappa)^2}\left(.5 - \theta(1+\kappa)(1+\epsilon)\right)^2, \\ c_2 = 0 \text{ if } \beta = \frac{1}{2\theta(1+\kappa)} - 1, \\ c_2 > 0 \text{ if } \frac{1}{2\theta(1+\kappa)} - 1 < \beta \le \frac{1}{2\theta(1+\kappa)} - 1 + \frac{1}{4\epsilon\theta^2(1+\kappa)^2}\left(.5 - \theta(1+\kappa)(1+\epsilon)\right)^2, \\ c_1 \text{ and } c_2 \text{ are complex if } \beta > \frac{1}{2\theta(1+\kappa)} - 1 + \frac{1}{4\epsilon\theta^2(1+\kappa)^2}\left(.5 - \theta(1+\kappa)(1+\epsilon)\right)^2. \end{cases} \tag{3.25}$$

When $z > 0$ the solution of (3.16) is

$$U_1 = k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z) + P_1(z), \tag{3.26}$$

where $\alpha < 0$ and $\gamma > 0$ are defined in (2.7), and $P_1(z)$ is the particular solution

$$P_1(z) = \frac{.5((1-\kappa)c - \epsilon)}{(1-\kappa)((1-\kappa)^2c^2 - (1-\kappa)(1+\epsilon)c + \epsilon(\beta+1))} e^{-(1-\kappa)z} + \frac{1}{(\beta+1)(1-\kappa^2)}, \tag{3.27}$$

where $c = c_1$ or $c = c_2$. To preserve continuity at $z = 0$ we require that $(U_1(0), U_1'(0)) = (U_0(0), U_0'(0))$. This and (3.26)–(3.27) show that k_1 and k_2 are uniquely defined by

$$k_1 = \theta - P_1(0) \quad \text{and} \quad k_2 = \frac{1}{\gamma}\left(\theta(1-\alpha) - P_1{}'(0) + \alpha P_1(0)\right). \tag{3.28}$$

(iii) *Numerical evidence for the existence of solutions.* To complete the proof that a solution satisfies all of the conditions in (3.14) for a traveling wave front we need only to show that

$$U_1(z) > \theta \ \ \forall z > 0 \quad \text{and} \quad \lim_{z\to\infty} U_1'(z) = 0. \tag{3.29}$$

To gain insight we let $\epsilon = .1$ and $\beta > \beta_* = 2.025$ and solve the initial value problem

$$\begin{aligned} u_t(x,t) &= -u - v + \tfrac{1}{2}\textstyle\int_{-120}^{120} e^{-|x-x'|+\kappa(x-x')} H(u(x',t)-\theta)dx', \\ v_t(x,t) &= \epsilon(\beta u - v), \\ u(x,0) &= e^{-x^2}, \quad \text{and} \quad v(x,0) = 0 \quad \forall x \in [-120, 120], \end{aligned} \tag{3.30}$$

where (θ, κ) satisfy (3.30). The limits $(-\infty, \infty)$ in the integral term have been replaced with $[-120, 120]$. Other choices for initial conditions give results similar to those described below. A second approach which leads to wave formation is to initially keep u and v at their resting levels, i.e., $u(x,0) = v(x,0) = 0$, and perturb the system with an external stimulus applied to the right side of the equation for u. To solve (3.30) we approximate the integral term with a Riemann sum, with step size $\Delta x = .1$, and use an explicit Euler time step $\Delta t = .05$. When $\beta > \beta_*$ the solution of (3.30) evolves into a traveling wave front (Figure 2), a single-pulse wave (Figures 4 and 6), or an N-pulse wave (Figure 7). Our numerical study of wave fronts suggests that (3.29) does not hold at every point along the curves c_1 and c_2 (Figure 2, upper left). However, (3.29) does hold along two connected subbranches Γ_1 and Γ_2 of these curves (Figure 2, upper right). Below we describe Γ_1 and Γ_2.

The lower branch Γ_2. When $c = c_2$ it follows from (3.26), (3.27), and (3.28) that

$$\begin{aligned} U_1(z) = {} & k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z) \\ & + \frac{.5((1-\kappa)c_2 - \epsilon)}{(1-\kappa)((1-\kappa)^2 c_2^2 - (1-\kappa)(1+\epsilon)c_2 + \epsilon(\beta+1))} e^{-(1-\kappa)z} + \frac{1}{(\beta+1)(1-\kappa^2)}. \end{aligned} \tag{3.31}$$

If $U_1(z) > \theta \ \forall z > 0$, then (3.31) implies that $U_1'(z) \to 0$ as $z \to \infty$, and therefore both conditions in (3.29) hold. Thus, it is sufficient to show that

$$U_1(z) > \theta \quad \forall z > 0. \tag{3.32}$$

It is difficult to prove (3.32) since the oscillatory component $k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z)$ of (3.31) can cause $U_1(z)$ to dip below the threshold level θ at some point in $(0, \infty)$. However, our numerical experiments indicate that there is a branch Γ_2 (Figure 2) of solutions satisfying

$$U(z) = \begin{cases} \theta e^{(1+\kappa)z} < \theta & \forall z < 0, \\ U_1(z) > \theta & \forall z > 0. \end{cases} \tag{3.33}$$

Along Γ_2 it follows from (3.23) and (3.25) that

$$c_2 \to 0^+ \text{ as } \beta \to \left(\frac{1}{2(1+\kappa)\theta} - 1\right)^+. \tag{3.34}$$

Thus, the left end of Γ_2 begins at $\beta = \frac{1}{2(1+\kappa)\theta} - 1$ where $c_2 = 0$ and U is the stationary solution defined by (3.4) (Figure 1, left panel). To determine the right end of Γ_2 we substitute the condition $U(\infty) \geq \theta$ into (3.33) and obtain $U(\infty) = \frac{1}{(1-\kappa^2)(\beta+1)} \geq \theta$. This implies that Γ_2 cannot extend past $\beta = \frac{1}{(1-\kappa^2)\theta} - 1$. Figure 2 (fourth row, left panel) shows the solution at the right endpoint of Γ_2 when $(\epsilon, \theta, \kappa) = (.1, .1, .15)$. For each $\beta \in [\frac{1}{2(1+\kappa)\theta} - 1, \frac{1}{(1-\kappa^2)\theta} - 1]$ our computations indicate that $U(z) > \theta \ \forall z > 0$; hence we conjecture that the interval of existence of Γ_2 is $[\frac{1}{2(1+\kappa)\theta} - 1, \frac{1}{(1-\kappa^2)\theta} - 1]$. Our study also suggests that all solutions on Γ_2 are unstable.

The upper branch Γ_1. Let Γ_1 denote the upper branch of wave fronts when $c = c_1$ (Figure 2). This branch extends to the left of $\beta = \beta_*$ down to $\beta = 0$. When $0 < \beta \leq \beta_*$ the eigenvalues $\mu^{\pm}$ are real, and solutions on Γ_1 are monotone for large z. When $\beta > \beta_*$ the eigenvalues $\mu^{\pm}$ are complex and solutions have the form

$$U(z) = \begin{cases} \theta e^{(1+\kappa)z} & \forall z \leq 0, \\ k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z) + \frac{.5((1-\kappa)c_1 - \epsilon)e^{-(1-\kappa)z}}{(1-\kappa)((1-\kappa)^1 c_1^2 - (1-\kappa)(1+\epsilon)c_1 + \epsilon(\beta+1))} & \\ \qquad + \frac{1}{(1-\kappa^2)(\beta+1)} & \forall z > 0, \end{cases} \tag{3.35}$$

where α, γ, k_1, and k_2 are evaluated at $c = c_2$. To complete the proof we need to show that $U(z) > \theta \ \forall z > 0$. This is difficult to prove, even in the real eigenvalue regime, since U can dip below θ at a positive z value. In Figure 2 (second row, left) we set $\beta = 4$ and see that $U(z)$ is a wave front since $U(z) > \theta \ \forall z > 0$. In the third row, left panel, we set $\beta = 6.45$ and see that $U(z)$ is not a wave front since it is tangent to $U = \theta$ at $z \approx 21$. When $\beta > 6.45$ the function U cannot be a wave front since it dips below θ at a positive z value. We conjecture that the interval of existence of Γ_1 is approximately $(0, 6.45)$ (Figure 2, first row, right). Our study suggests that solutions on Γ_1 are stable.

C. *Properties of wavefronts when* $c < 0$. When $c < 0$ we investigate solutions which satisfy

$$\begin{cases} U(z) \text{ and } U'(z) \text{ are continuous } \forall z \in (-\infty, \infty), \\ U(z) > \theta \ \forall z < 0, \quad U'(z) \to 0 \text{ as } z \to -\infty, \\ U(0) = \theta, \quad U(z) < \theta \ \forall z > 0, \\ (U(z), U'(z)) \to (0,0) \text{ as } z \to \infty. \end{cases} \tag{3.36}$$

When conditions (3.36) hold, (2.3) reduces to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = \frac{c}{2}\frac{d}{dz}\int_{-\infty}^{0} e^{-|z-z'|+\kappa(z-z')}dz' + \frac{\epsilon}{2}\int_{-\infty}^{0} e^{-|z-z'|+\kappa(z-z')}dz'. \tag{3.37}$$

This further reduces to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = g(z), \tag{3.38}$$

where

$$g(z) = \begin{cases} -.5(c + \frac{\epsilon}{1+\kappa})e^{(1+\kappa)z} + \frac{\epsilon}{1-\kappa^2} & \forall z \le 0, \\ -.5(c - \frac{\epsilon}{1-\kappa})e^{-(1-\kappa)z} & \text{if} \quad z > 0. \end{cases} \tag{3.39}$$

As in the case $c > 0$, we devote the remainder of this section to the following:

(iv) Analysis of solutions on $[0, \infty)$.
(v) Analysis of solutions on $(-\infty, 0)$.
(vi) Numerical evidence for the existence of solutions.

(iv) *Analysis of solutions on* $[0, \infty)$. Proceeding as above, we find that the solution of (3.38) is

$$U_2(z) = \frac{.5(\epsilon - (1-\kappa)c)}{(1-\kappa)((1-\kappa)^2c^2 + -(1-\kappa)(1+\epsilon)c + \epsilon(\beta+1))} e^{-(1-\kappa)z} \quad \forall z > 0. \tag{3.40}$$

Substituting the condition $U_2(0) = \theta$ into (3.40), we obtain

$$\frac{.5(\epsilon - (1-\kappa)c)}{(1-\kappa)((1-\kappa)^2c^2 - (1-\kappa)(1+\epsilon)c + \epsilon(\beta+1))} = \theta. \tag{3.41}$$

It follows from (3.40)–(3.41) that

$$U_2(z) = \theta e^{-(1-\kappa)z} < \theta \quad \forall z \ge 0. \tag{3.42}$$

Next, solving (3.41) for wave speed c gives the two negative values (Figure 2)

$$c_3 = \frac{\theta(1-\kappa)(\epsilon+1) - .5 + \sqrt{\left(.5 - \theta(1-\kappa)(1+\epsilon)\right)^2 - 4\epsilon\theta(1-\kappa)\left(\theta(1-\kappa)(\beta+1) - .5\right)}}{2\theta(1-\kappa)^2}, \tag{3.43}$$

$$c_4 = \frac{\theta(1-\kappa)(\epsilon+1) - .5 - \sqrt{\left(.5 - \theta(1-\kappa)(1+\epsilon)\right)^2 - 4\epsilon\theta(1-\kappa)\left(\theta(1-\kappa)(\beta+1) - .5\right)}}{2\theta(1-\kappa)^2}. \tag{3.44}$$

(v) *Analysis of solutions on* $(-\infty, 0)$. When $z < 0$ the solution of (3.16) is

$$U_3 = k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z) + P_2(z), \tag{3.45}$$

where $\alpha < 0$ and $\gamma > 0$ are defined in (2.7), and $P_2(z)$ is the particular solution

$$P_2(z) = \frac{-.5((1+\kappa)c + \epsilon)}{(1+\kappa)((1+\kappa)^2c^2 + (1+\kappa)(1+\epsilon)c + \epsilon(\beta+1))} e^{(1+\kappa)z} + \frac{1}{(\beta+1)(1-\kappa^2)}, \tag{3.46}$$

where $c = c_3$ or $c = c_4$. To preserve continuity at $z = 0$ we require that $(U_3(0), U_3'(0)) = (U_2(0), U_2'(0))$. This and (3.45)–(3.46) show that k_1 and k_2 are uniquely defined by

$$k_1 = \theta - P_2(0) \quad \text{and} \quad k_2 = \frac{1}{\gamma}\left(\theta(1-\alpha) - P_2'(0) + \alpha P_2(0)\right). \tag{3.47}$$

(vi) *Numerical evidence for the existence of solutions.* As in the case $c > 0$, to complete the proof that a solution satisfies all of the conditions in (3.14) for a wave front it suffices to show that

$$U_3(z) > \theta \quad \forall z < 0. \tag{3.48}$$

Our numerical experiments suggest that conditions (3.48) are not satisfied at every point along the curves c_3 and c_4 (Figure 2). However, (3.48) does hold along two subbranches Γ_3 and Γ_4 of these curves (Figure 2). Below we briefly describe properties of solutions along Γ_3 and Γ_4.

The branch Γ_3. When $c = c_3 < 0$ it follows from (3.45), (3.46), and (3.47) that

$$\begin{aligned} U_2(z) &= k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z) \\ &\quad - \tfrac{.5((1+\kappa)c_3+\epsilon)}{(1+\kappa)((1+\kappa)^2c_3^2+(1+\kappa)(1+\epsilon)c_3+\epsilon(\beta+1))}e^{(1+\kappa)z} + \tfrac{1}{(\beta+1)(1-\kappa^2)}. \end{aligned} \tag{3.49}$$

Our numerical experiments indicate that there is a branch Γ_3 (Figure 2) of solutions satisfying

$$U(z) = \begin{cases} \theta e^{-(1-\kappa)z} < \theta & \forall z \geq 0, \\ U_2(z) > \theta & \forall z < 0. \end{cases} \tag{3.50}$$

Along Γ_3 it follows from (3.43) that

$$c_3 \to 0^+ \text{ as } \beta \to \left(\frac{1}{2(1-\kappa)\theta} - 1\right)^+. \tag{3.51}$$

Thus, the left end of Γ_3 begins at $\beta = \frac{1}{2(1-\kappa)\theta} - 1$ where $c_3 = 0$ and U is the stationary solution defined by (3.4) (Figure 1, right panel). To determine the right end of Γ_3 we substitute the condition $U(\infty) \geq \theta$ into (3.50) and obtain $U(\infty) = \frac{1}{(1-\kappa^2)(\beta+1)} \geq \theta$. This implies that Γ_3 cannot extend past $\beta = \frac{1}{(1-\kappa^2)\theta} - 1$. Figure 2 (fourth row, right panel) shows the solution at the right endpoint of Γ_3 when $(\epsilon, \theta, \kappa) = (.1, .1, .15)$. For each $\beta \in [\frac{1}{2(1-\kappa)\theta} - 1, \frac{1}{(1-\kappa^2)\theta} - 1]$ our computations indicate that $U(z) > \theta$ $\forall z < 0$; hence we conjecture that the interval of existence of Γ_3 is $[\frac{1}{2(1-\kappa)\theta} - 1, \frac{1}{(1-\kappa^2)\theta} - 1]$. Our study also suggests that all solutions on Γ_3 are unstable.

Comparisons. To obtain Γ_2 we set $(\epsilon, \theta, \kappa) = (.1, .1, .15)$ and let β increase from the upper critical value $\beta = \frac{1}{2(1+\kappa)\theta} - 1$ where $c_2 = 0$ and the solution is the stationary front defined by (3.4). Γ_3 is found by letting β increase from the lower critical value $\beta = \frac{1}{2(1-\kappa)\theta} - 1$ where $c_3 = 0$ and the solution is the stationary front defined by (3.10). The eigenvalues $\mu^{\pm}$ are complex at $\beta = \frac{1}{2(1+\kappa)\theta} - 1$ and $\beta = \frac{1}{2(1-\kappa)\theta} - 1$ since $(\epsilon, \theta, \kappa)$ satisfy (2.8). It is interesting to contrast these bifurcation results with [2], where a similar phenomenon is found when the coupling is symmetric (i.e., $\kappa = 0$) and (ϵ, θ) are chosen so that $\mu^{\pm}$ are real. In that study (θ, β) are kept fixed and counterpropagating fronts bifurcate from the stationary solution as ϵ varies. It would be interesting to analytically determine if a similar phenomenon occurs here where w is asymmetric and $\mu^{\pm}$ are complex.

The branch Γ_4. We let Γ_4 denote the upper branch of wave fronts when $c = c_4 < 0$ (Figure 2). As in the case $c > 0$, this branch extends below $\beta = \beta_*$ down to $\beta = 0$. Solutions on Γ_4 have the form

$$U(z) = \begin{cases} \theta e^{-(1-\kappa)z} & \forall z > 0, \\ k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z) - \frac{.5((1+\kappa)c_4+\epsilon)e^{(1+\kappa)z}}{(1+\kappa)((1+\kappa)^2c_4^2+(1+\kappa)(1+\epsilon)c_4+\epsilon(\beta+1))} & \\ \qquad + \frac{1}{(1-\kappa^2)(\beta+1)} & \forall z \le 0, \end{cases} \tag{3.52}$$

where α, γ, k_1, and k_2 are evaluated at $c = c_4$. To complete the proof of existence we need to show that $U(z) > \theta\ \forall z < 0$. Again, this is difficult to prove since U can dip below θ at a negative value of z. In Figure 2 (second row, right panel) we set $\beta = 4$ and see that $U(z)$ is a wave front since $U(z) > \theta\ \forall z < 0$. In the third row, right panel, we set $\beta = 6.45$ and see that $U(z)$ is not a wave front since it is tangent to $U = \theta$ at $z \approx -21$. When $\beta > 6.45$ the function U cannot be a wave front since it dips below θ at a finite value of z. Thus, we conjecture that the interval of existence of Γ_4 is approximately $(0, 6.45)$. Our study suggests that all solutions on Γ_4 are stable.

D. *The effects of changing the threshold θ.* We study how variations in θ affect wave front formation when $\mu^{\pm}$ are complex and w is asymmetric. Figures 2 and 3 illustrate the differences in the behavior of solutions when $\theta = .1$ and $\theta = .2$, and lead to the following conjectures:

(i) There is an interval of β values, approximately $0 < \beta < 6.75$ when $(\epsilon, \kappa, \theta) = (.1, .15, .1)$, and a second interval $0 < \beta < 4.05$ when $(\epsilon, \kappa, \theta) = (.1, .15, .2)$, over which two stable wave fronts coexist, which propagate in opposite directions, with different speeds and amplitudes.

(ii) At fixed $(\epsilon, \kappa, \beta)$ the speeds and amplitudes *decrease* as θ increases.

(iii) At fixed (ϵ, κ) the values of β at the right endpoints of $\Gamma_1, \ldots, \Gamma_4$ decrease as θ increases.

E. *Open problems.* It remains an open problem to prove the existence of wave fronts along the branches $\Gamma_1, \ldots, \Gamma_4$ described above. Whether $\mu^{\pm}$ are real or complex, new methods are needed in order to overcome the technical difficulties in showing that solutions satisfy $U(z) = \theta$ only once on the interval $(-\infty, \infty)$. Although generalizations and insightful approximations have previously been given [2, 42], this property has not yet been verified for any set of parameters or couplings, even in the real eigenvalue regime. In [58] we considered parameter values where $\mu^{\pm}$ are *real* and developed a comparison method which addresses these issues when w is symmetric. It is hoped that extensions of our techniques will help complete existence proofs when w is asymmetric, and for a wider range of rate functions and parameter values. The proof of stability of solutions along Γ_1 and Γ_4, and the instability of solutions along Γ_2 and Γ_3, also remains an open problem. It is possible that this might be accomplished by extensions of Evans function methods developed in [12, 44, 50].

4. 1-pulse traveling waves.

We analyze 1-pulse traveling waves when w is asymmetric and $\mu^{\pm}$ are complex. Our study consists of the following:

A. The proof of nonexistence of stationary, 1-pulse solutions.

B. Positive and negative wave speeds: the statement and proof of Theorem 4.2 concerning the coexistence of two families of infinitely many 1-pulse waves.

C. The effects of changing the threshold and the identification of θ^*.

A. *Nonexistence of stationary solutions.* In the previous section we showed how two distinct families of wave fronts come into existence by means of a bifurcation from stationary solutions when $\kappa \neq 0$. In [58] we demonstrated how a branch of 1-pulse traveling waves

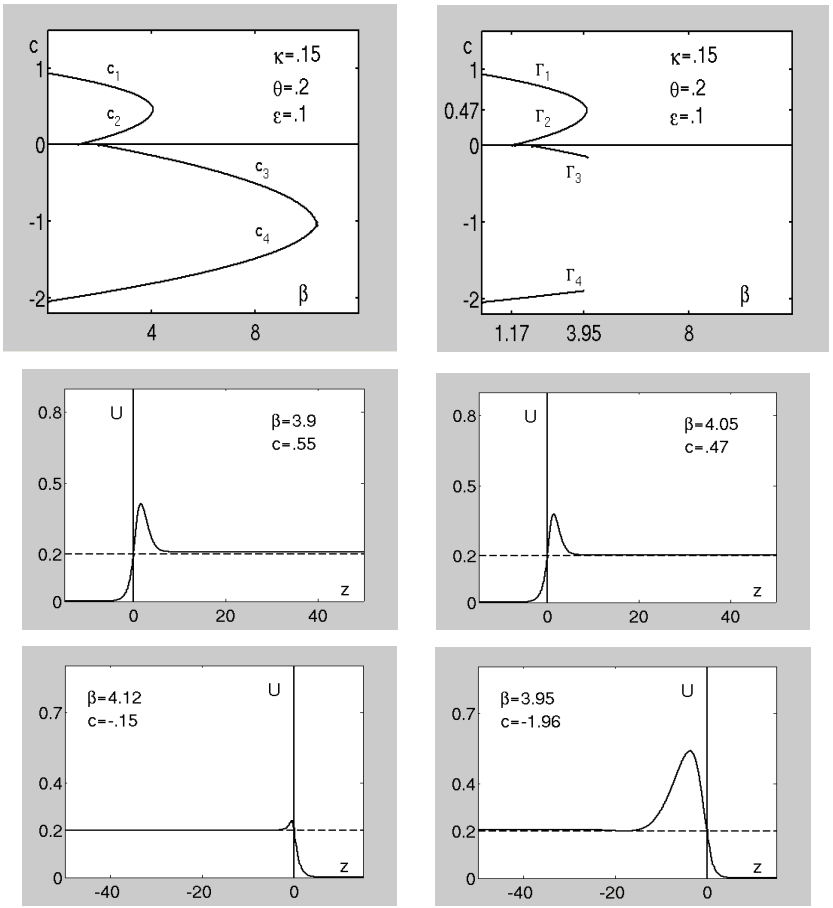


Figure 3. *Upper left:* c_1, c_2, c_3, *and* c_4 *versus* β *when* $(\epsilon, \theta, \kappa) = (.1, .2, .15)$. *Upper right:* Γ_1, Γ_2, Γ_3, *and* Γ_4 *in the* (β, c) *plane. Second row: Solutions at* $\beta = 3.9$ *on* Γ_1 *(left) and at* $\beta = 4.05$, *the right endpoint where* Γ_1 *meets* Γ_2 *(right). Third row: Solutions at the right endpoints of* Γ_3 *where* $\beta \approx 4.12$ *(left) and* Γ_4 *where* $\beta \approx 3.95$ *(right). It is conjectured that solutions on* Γ_1 *and* Γ_4 *stable, and those on* Γ_2 *and* Γ_3 *are unstable.*

bifurcates from a stationary solution when $\kappa = 0$ and the coupling is symmetric. Theorem 4.1 below shows that there is a fundamental difference when we consider asymmetric couplings. In particular, we prove that there is no stationary 1-pulse solution for the class of asymmetric couplings satisfying the general condition

$$w(x) - w(-x) > 0 \ \ \forall x \in (-\infty, \infty) \quad \text{or} \quad w(x) - w(-x) < 0 \ \ \forall x \in (-\infty, \infty). \tag{4.1}$$

The function $w(x) = \frac{1}{2} e^{-|x| + \kappa x}$ which we study in this paper falls within this class. Thus, for couplings satisfying condition (4.1), a family of 1-pulse waves cannot come into existence by means of a bifurcation from a stationary solution.

Theorem 4.1. *If* w *satisfies* (4.1), *then* (2.3) *does not have a stationary* 1*-pulse solution.*

Proof. If a stationary solution U exists, then $c = 0$ and (2.3) reduces to

$$U(x) = \frac{1}{\beta + 1} \int_{-\infty}^{\infty} w(x - x') dx'. \tag{4.2}$$

We assume, for contradiction, that $U(x)$ satisfies

$$\begin{cases} U(x) \text{ and } U'(x) \text{ are continuous } \forall x \in (-\infty, \infty), \\ U(x) < \theta \ \forall x < 0, \quad U(0) = \theta, \\ U(x) > \theta \ \forall x \in (0, a), \quad \text{for some } a > 0, \\ U(a) = \theta, \quad U(x) < \theta \ \forall x > a, \\ (U(x), U'(x)) \to 0 \text{ as } |x| \to \infty. \end{cases} \tag{4.3}$$

Without loss of generality we have assumed that $U(0) = \theta$ in (4.3) since (4.2) is translationally invariant. Conditions (4.3) reduce (4.2) to

$$U(x) = \frac{1}{\beta + 1} \int_0^a w(x - x')dx'. \tag{4.4}$$

The substitution $\eta = x - x'$ changes (4.4) into

$$U(x) = \frac{1}{\beta + 1} \int_{x-a}^{x} w(\eta)d\eta. \tag{4.5}$$

Setting $x = 0$ and $x = a$ in (4.5), we conclude from (4.3) that

$$U(0) = \frac{1}{\beta + 1} \int_{-a}^{0} w(\eta)d\eta = \theta \quad \text{and} \quad U(a) = \frac{1}{\beta + 1} \int_0^a w(\eta)d\eta = \theta. \tag{4.6}$$

From this it follows that $\int_{-a}^{0} w(\eta)d\eta = \int_0^a w(\eta)d\eta$; hence

$$\int_0^a [w(\eta) - w(-\eta)]d\eta = 0. \tag{4.7}$$

However, condition (4.1) implies that either

$$\int_0^a [w(\eta) - w(-\eta)]d\eta > 0 \quad \text{or} \quad \int_0^a [w(\eta) - w(-\eta)]d\eta < 0, \tag{4.8}$$

which contradicts (4.7). This completes the proof.

B. *Positive and negative wave speeds.* In this section we study properties of 1-pulse traveling waves for both positive and negative wave speeds.

Positive wave speeds. When $c > 0$ we investigate the existence of 1-pulse waves which satisfy

$$\begin{cases} U(z) \text{ and } U'(z) \text{ are continuous } \forall z \in (-\infty, \infty), \\ U(z) < \theta \ \forall z < 0, \quad U(0) = \theta, \\ U(z) > \theta \ \forall z \in (0, a), \quad \text{for some } a = a(c) > 0, \\ U(a) = \theta \quad \text{and} \quad U(z) < \theta \ \forall z \in (a, \infty), \\ (U(z), U'(z)) \to (0, 0) \text{ as } |z| \to \infty. \end{cases} \tag{4.9}$$

Again we have assumed that $U(0) = \theta$ since (2.3) is translationally invariant. When conditions (4.9) hold, (2.3) reduces to

$$c^2U'' + c(1 + \epsilon)U' + \epsilon(\beta + 1)U = c\frac{d}{dz} \int_0^a w(z - z')dz' + \epsilon \int_0^a w(z - z')dz'. \tag{4.10}$$

Because $w(x) = \frac{1}{2}e^{-|x|+\kappa x}$, (4.10) can be written in the equivalent form

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = f(z), \tag{4.11}$$

where

$$f(z) = \begin{cases} \frac{.5}{1+\kappa}\left(c(1+\kappa)+\epsilon\right)\left(1-e^{-(1+\kappa)a}\right)e^{(1+\kappa)z} & \forall z \le 0, \\ .5\left(c+\frac{\epsilon}{\kappa-1}\right)e^{(\kappa-1)z} - .5\left(c+\frac{\epsilon}{\kappa+1}\right)e^{(\kappa+1)(z-a)} + \frac{\epsilon}{1-\kappa^2} & \text{if} \quad 0 < z < a, \\ \frac{.5}{1-\kappa}(\epsilon - c(1-\kappa))(e^{(1-\kappa)a}-1)e^{-(1-\kappa)z} & \forall z \ge a. \end{cases} \tag{4.12}$$

Negative wave speeds. When $c < 0$, a 1-pulse traveling wave satisfies

$$\begin{cases} U(z) \text{ and } U'(z) \text{ are continuous } \forall z \in (-\infty, \infty), \\ U(z) < \theta \ \forall z \in (-\infty, a), \quad \text{for some } a = a(c) < 0,\ U(a) = 0, \\ U(z) > \theta \ \forall z \in (a, 0), \quad U(0) = \theta, \\ U(z) < \theta \ \forall z > 0, \\ (U(z), U'(z)) \to (0,0) \text{ as } |z| \to \infty. \end{cases} \tag{4.13}$$

When conditions (4.13) hold, (2.3) reduces to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = c\frac{d}{dz}\int_a^0 w(z-z')dz' + \epsilon\int_a^0 w(z-z')dz'. \tag{4.14}$$

Because $w(x) = \frac{1}{2}e^{-|x|+\kappa x}$, (4.14) can be written in the equivalent form

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = h(z), \tag{4.15}$$

where

$$h(z) = \begin{cases} \frac{.5}{1-\kappa}\left(c(1-\kappa)-\epsilon\right)\left(e^{(1-\kappa)a}-1\right)e^{-(1-\kappa)z} & \forall z \ge 0, \\ .5\left(c-\frac{\epsilon}{1-\kappa}\right)e^{-(1-\kappa)(z-a)} - .5\left(c+\frac{\epsilon}{\kappa+1}\right)e^{(\kappa+1)z} + \frac{\epsilon}{1-\kappa^2} & \text{if} \quad a < z < 0, \\ \frac{.5}{1+\kappa}(\epsilon + c(1+\kappa))(e^{-(1+\kappa)a}-1)e^{(1+\kappa)z} & \forall z \le a. \end{cases} \tag{4.16}$$

We prove the following theorem.

Theorem 4.2. *Let (ϵ, θ) satisfy (2.8). If $\kappa > 0$ is small and $\beta = \frac{1}{2(1-\kappa^2)\theta} - 1$, then the following hold.*

(i) *There are infinitely many $c \in (c_2, c_1)$ and $a(c) > 0$ and solutions U of (4.11)–(4.12) such that*

$$\begin{aligned} &U(z) = \theta e^{(1+\kappa)z}\ \forall z < 0, \quad U(0) = U(a(c)) = \theta, \quad \textit{and} \quad U'(0) = 1+\kappa, \\ &(U(z), U'(z)) \to (0,0) \quad \textit{as} \quad z \to \infty. \end{aligned} \tag{4.17}$$

(ii) *There are infinitely many $c \in (c_4, c_3)$ and $a(c) < 0$ and solutions U of (4.11)–(4.12) such that*

$$\begin{aligned} &U(z) = \theta e^{-(1-\kappa)z}\ \forall z > 0, \quad U(a(c)) = U(0) = \theta, \quad \textit{and} \quad U'(0) = \kappa - 1, \\ &(U(z), U'(z)) \to (0,0) \quad \textit{as} \quad z \to -\infty. \end{aligned} \tag{4.18}$$

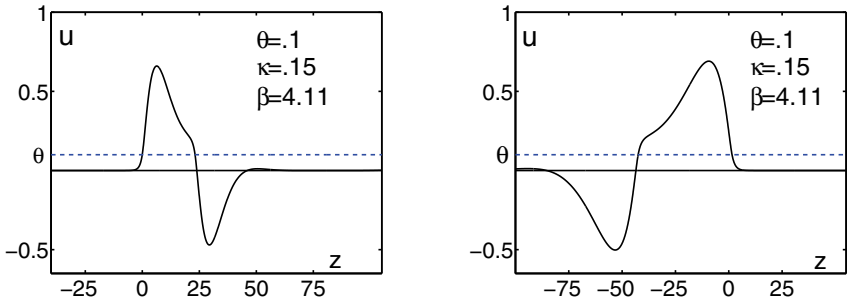


Figure 4. *1-pulse waves when* $\epsilon = \theta = .1$ *and* $\beta = \frac{1}{2\theta(1-\kappa^2)} - 1 \approx 4.11$ *(see Theorem 4.2). Left panel:* $c \approx 2.27$ *and* $a \approx 24$. *Right panel:* $c \approx -5.65$ *and* $a \approx -43$. *Clicking on the first panel displays the accompanying movie* (70988_01.mpg [1.4MB]).

Remarks. (i) The proof of Theorem 4.2 relies heavily on sine and cosine terms which arise due to $\mu^{\pm}$ being complex. Such terms are not present in the real eigenvalue case, and therefore we conjecture that Theorem 4.2 does not hold when $\mu^{\pm}$ are real.

(ii) The positive wave speed solutions described in Theorem 4.2 have monotone tails as $z \to -\infty$ and oscillatory tails as $z \to \infty$. By contrast, the speeds and amplitudes of the negative wave speed solutions are larger than those of the positive speed solutions (see Figure 4); they have oscillatory tails as $z \to -\infty$ and monotone tails as $z \to \infty$. The oscillatory tails are due to $\mu^{\pm}$ being complex.

(iii) To prove that the solutions in Theorem 4.2 are 1-pulse waves we must also prove that $z = 0$ and $z = a(c)$ are the *only* solutions of $U(z) = \theta$, and that $(U(z), U'(z)) \to (0,0)$ as $z \to \infty$. Because $\mu^{\pm}$ are complex, technical difficulties arise which make the verification of these properties a challenging problem which remains open.

Proof of Theorem 4.2. We prove part (i) for the case $c > 0$. The details for the case $c < 0$ are similar and are omitted for brevity. On the interval $(-\infty, 0)$ the system (4.11)–(4.12) reduces to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = \frac{.5}{1+\kappa}\left(c(1+\kappa)+\epsilon\right)(1-e^{-(1+\kappa)a})e^{(1+\kappa)z} \quad \forall z \le 0. \tag{4.19}$$

The general solution of (4.19) is

$$\begin{aligned} U_3(z) &= k_1 e^{\alpha z}\cos(\gamma z) + k_2 e^{\alpha z}\sin(\gamma z) \\ &\quad + \tfrac{.5((1+\kappa)c+\epsilon)(1-e^{-(1+\kappa)a})}{(1+\kappa)((1+\kappa)^2c^2+(1+\kappa)(1+\epsilon)c+\epsilon(\beta+1))}e^{(1+\kappa)z}. \end{aligned} \tag{4.20}$$

We need to show that there are values $c \in (c_2, c_1)$ and $a > 0$ such that

$$U_3(z) < \theta \ \ \forall z \in (-\infty, 0), \quad U_3(-\infty) = U_3'(-\infty) = 0, \quad \text{and} \quad U_3(0) = \theta. \tag{4.21}$$

Recall from (2.7) that $\alpha = \mathrm{Re}(\mu^{\pm}) < 0$. Thus, to satisfy $U_3(-\infty) = U_3'(-\infty) = 0$ we conclude that $k_1 = k_2 = 0$, and therefore

$$U_3(z) = \frac{.5((1+\kappa)c+\epsilon)(1-e^{-(1+\kappa)a})}{(1+\kappa)((1+\kappa)^2c^2+(1+\kappa)(1+\epsilon)c+\epsilon(\beta+1))}e^{(1+\kappa)z}. \tag{4.22}$$

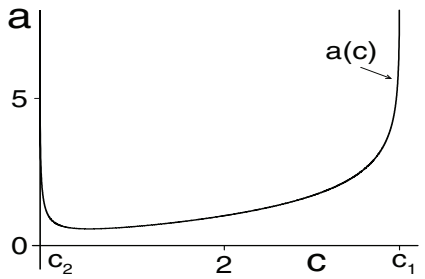


Figure 5. *Graph of $a(c)$ versus c when $(\epsilon,\theta,\kappa)=(.1,.1,.15)$ and $\beta=\frac{1}{2(1-\kappa^2)\theta}-1\approx 4.11$.*

Substituting the continuity requirement $U_3(0)=\theta$ into (4.22) gives

$$\frac{.5((1+\kappa)c+\epsilon)(1-e^{-(1+\kappa)a})}{(1+\kappa)((1+\kappa)^2c^2+(1+\kappa)(1+\epsilon)c+\epsilon(\beta+1))}=\theta. \tag{4.23}$$

Combining (4.22) and (4.23) gives $U_3(z)=\theta e^{(1+\kappa)z}\ \forall z\le 0$.

The interval $(0,a)$. On the interval $(0,a)$ the system (4.11)–(4.12) reduces to

$$c^2U''+c(1+\epsilon)U'+\epsilon(\beta+1)U=.5\left(c+\frac{\epsilon}{\kappa-1}\right)e^{(\kappa-1)z}-.5\left(c+\frac{\epsilon}{\kappa+1}\right)e^{(\kappa+1)(z-a)}+\frac{\epsilon}{1-\kappa^2}. \tag{4.24}$$

The general solution of (4.24) is

$$U_4(z)=m_1e^{\alpha z}\cos(\gamma z)+m_2e^{\alpha z}\sin(\gamma z)+P_4(z), \tag{4.25}$$

where α and γ are defined in (2.7), and P_4 is the particular solution

$$\begin{aligned}P_4(z)=&\tfrac{.5(c(1-\kappa)-\epsilon)e^{(\kappa-1)z}}{(1-\kappa)((1-\kappa)^2c^2-(1+\epsilon)(1-\kappa)c+\epsilon(\beta+1))}\\&-\tfrac{.5(\epsilon+c(1+\kappa))e^{(\kappa+1)(z-a)}}{(1+\kappa)((1+\kappa)^2c^2+(1+\epsilon)c(1+\kappa)+\epsilon(\beta+1))}+\tfrac{1}{(\beta+1)(1-\kappa^2)}.\end{aligned} \tag{4.26}$$

Continuity at $z=0$ requires that $(U_4(0),U_4'(0))=(U_3(0),U_3'(0))=(\theta,(1+\kappa)\theta)$. Combining this with (4.25) and (4.26) shows that the coefficients m_1 and m_2 in (4.25) are defined by

$$m_1=\theta-P_4(0)\quad\text{and}\quad m_2=\frac{1}{\gamma}\left(\theta(1+\kappa-\alpha)-P_4'(0)+\alpha P_4(0)\right). \tag{4.27}$$

Next, solve (4.23) for $e^{-(1+\kappa)a}$ and get

$$e^{-(1+\kappa)a(c)}=\frac{.5(\epsilon+c(1+\kappa))-\theta(1+\kappa)\left((1+\kappa)^2c^2+(1+\epsilon)c(1+\kappa)+\epsilon(\beta+1)\right)}{.5(\epsilon+c(1+\kappa))}. \tag{4.28}$$

It follows from (3.21) that the right side of (4.28) is zero at $c=c_1$ and $c=c_2$ and that (Figure 5)

$$a(c)>0\ \ \forall c\in(c_2,c_1)\quad\text{and}\quad\lim_{c\to c_2^+}a(c)=\lim_{c\to c_1^-}a(c)=\infty. \tag{4.29}$$

We will use (4.29) to help prove that there exist values $c \in (c_2, c_1)$ and $a = a(c) > 0$ such that

$$U_4(a(c)) = \theta. \tag{4.30}$$

Substituting (4.25) into (4.30) gives

$$m_1 e^{\alpha a(c)} \cos(\gamma a(c)) + m_2 e^{\alpha a(c)} \sin(\gamma a(c)) + P_4(a(c)) = \theta. \tag{4.31}$$

Thus, it remains to show that there are infinitely many $c \in (c_2, c_1)$ such that the function

$$g(c) = m_1 e^{\alpha a(c)} \cos(\gamma a(c)) + m_2 e^{\alpha a(c)} \sin(\gamma a(c)) + P_4(a(c)) - \theta \tag{4.32}$$

satisfies $g(c) = 0$. The first step is to write (4.28) as

$$\frac{.5(\epsilon + c(1+\kappa))}{(1+\kappa)^2 c^2 + (1+\epsilon)(1+\kappa)c + \epsilon(\beta+1)} = \frac{.5(\epsilon + c(1+\kappa))e^{-(1+\kappa)a}}{(1+\kappa)^2 c^2 + (1+\epsilon)(1+\kappa)c + \epsilon(\beta+1)} + \theta(1+\kappa). \tag{4.33}$$

Substituting (4.33) into (4.26) and using the hypothesis $\beta = \frac{1}{2(1-\kappa^2)\theta} - 1$, we obtain

$$\begin{aligned} P_4(a(c)) - \theta = {} & e^{-(1-\kappa)a(c)} \left(\tfrac{.5(c(1-\kappa)-\epsilon)}{(1-\kappa)(1-\kappa)^2 c^2 - (1+\epsilon)(1-\kappa)c + \epsilon(\beta+1)} \right) \\ & - e^{-(1+\kappa)a(c)} \left(\tfrac{.5(\epsilon + c(1+\kappa))}{(1+\kappa)(1+\kappa)^2 c^2 + (1+\epsilon)(1+\kappa)c + \epsilon(\beta+1)} \right). \end{aligned} \tag{4.34}$$

Next, substitute (4.34) into (4.32) and get

$$g(c) = e^{\alpha a(c)} g_1(c), \tag{4.35}$$

$$\begin{aligned} g_1 = {} & m_1 \cos(\gamma a(c)) + m_2 \sin(\gamma a(c)) + e^{-(1+\alpha-\kappa)a(c)} \left(\tfrac{.5(c(1-\kappa)-\epsilon)}{(1-\kappa)(1-\kappa)^2 c^2 - (1+\epsilon)(1-\kappa)c + \epsilon(\beta+1)} \right) \\ & - e^{-(1+\alpha+\kappa)a(c)} \left(\tfrac{.5(\epsilon + c(1+\kappa))}{(1+\kappa)(1+\kappa)^2 c^2 + (1+\epsilon)(1+\kappa)c + \epsilon(\beta+1)} \right). \end{aligned} \tag{4.36}$$

Because $e^{\alpha a(c)} > 0$, it suffices to show that $g_1(c)$ changes sign infinitely often on (c_2, c_1). To analyze $g_1(c)$ we need to determine the limiting behavior, as $c \to c_1^-$, of the terms on the right side of (4.36). For this we need the five basic estimates developed below.

(i) *The behavior of $e^{-(1+\alpha-\kappa)a(c)}$ and $e^{-(1+\alpha+\kappa)a(c)}$ as $c \to c_1^-$.* From (2.7) it follows that

$$\lim_{c \to c_1^-} (1 + \alpha - \kappa) = \frac{2c_1(1-\kappa) - 1 - \epsilon}{2c_1}. \tag{4.37}$$

It follows from (3.23), and the hypothesis $\beta = \frac{1}{2(1-\kappa^2)\theta} - 1$, that

$$c_1 = \frac{.5 - \theta(1+\kappa)(\epsilon+1) + \sqrt{(.5 - \theta(1+\kappa)(1+\epsilon))^2 - \frac{2\epsilon\theta\kappa(1+\kappa)}{1-\kappa}}}{2\theta(1+\kappa)^2}. \tag{4.38}$$

We substitute (4.38) into (4.37) and conclude from an algebraic manipulation, and the restriction $0 < \theta < \frac{1}{4(\epsilon+1)}$ given in (2.8), that

$$\lim_{(c,\kappa) \to (c_1^-, 0)} (1 + \alpha - \kappa) = \frac{1 - 3\theta(1+\epsilon)}{1 - 2\theta(1+\epsilon)} > 0. \tag{4.39}$$

From (4.39) and continuity it follows that if $\kappa > 0$ is small, then

$$\lim_{c\to c_1^-}(1+\alpha-\kappa) > 0. \tag{4.40}$$

Thus, if $\kappa > 0$ is small, then (4.29) and (4.40) imply that

$$\lim_{c\to c_1^-} e^{-(1+\alpha-\kappa)a(c)} = \lim_{c\to c_1^-} e^{-(1+\alpha+\kappa)a(c)} = 0. \tag{4.41}$$

(ii) *The behavior of $\gamma a(c)$ as $c \to c_1^-$.* Recall that $\beta = \frac{1}{2(1-\kappa^2)\theta} - 1$. This and (2.7)–(2.8) imply that

$$\gamma = \frac{(\epsilon+1)}{2c\sqrt{(1-\kappa^2)\theta}}\sqrt{\frac{2\epsilon}{(\epsilon+1)^2} - (1-\kappa^2)\theta} > 0 \quad \forall c \in [c_2, c_1] \text{ and } \forall \kappa \in [0,1). \tag{4.42}$$

From this and (4.29) we conclude that $\gamma a(c)$ is continuous in c and that

$$\gamma a(c) > 0 \ \forall c \in (c_2, c_1) \quad \text{and} \quad \lim_{c\to c_1^-} \gamma a(c) = \infty \ \forall \kappa \in [0,1). \tag{4.43}$$

(iii) *The behavior of $(1-\kappa)^2c^2 - (1+\epsilon)(1-\kappa)c + \epsilon(\beta+1)$ as $c \to c_1^-$.* We show that

$$\lim_{c\to c_1^-}\left((1-\kappa)^2c^2 - (1+\epsilon)(1-\kappa)c + \epsilon(\beta+1)\right) > 0 \quad \forall \kappa \in [0,1). \tag{4.44}$$

It follows from (2.8) that the discriminant of the limiting term

$$(1-\kappa)^2c_1^2 - (1+\epsilon)(1-\kappa)c_1 + \epsilon(\beta+1) \tag{4.45}$$

satisfies

$$\text{discriminant} = (1-\kappa)^2\left((1+\epsilon)^2 - 4\epsilon(\beta+1)\right) < 0 \quad \forall \kappa \in [0,1),$$

since $\epsilon > 0$ and $\beta > \frac{(\epsilon-1)^2}{4\epsilon}$. This and continuity imply that (4.44) holds.

(iv) *The behavior of m_1 as $c \to c_1^-$.* We show that m_1 is bounded away from 0 as $c \to c_1^-$ when $\kappa > 0$ is small. Since $a(c) \to \infty$ as $c \to c_1^-$, it follows from (4.26) and the definition of m_1 given in (4.27) that m_1 is continuous in κ and c, and

$$\lim_{c\to c_2^-} m_1 = \theta - \frac{.5(c_1(1-\kappa)-\epsilon)}{(1-\kappa)\left((1-\kappa)^2c_1^2 - (1+\epsilon)(1-\kappa)c_1 + \epsilon(\beta+1)\right)} - \frac{1}{(\beta+1)(1-\kappa^2)}. \tag{4.46}$$

We solve (3.21) for c^2, substitute the resultant expression into (4.46), and obtain the limiting value

$$\lim_{(\kappa,c)\to(0,c_1^-)} m_1 = \frac{\theta(\epsilon - 2\theta(1+\epsilon)c_1)}{(.5 - 2\theta(1+\epsilon))c_1 + .5\epsilon} - \frac{1}{\beta+1}, \tag{4.47}$$

where, by (3.23), c_1 has the limiting value

$$c_1 = \frac{.5 - \theta(1+\epsilon)}{\theta} > 0. \tag{4.48}$$

Substituting (4.48) into the numerator of the first term in (4.47) gives the limiting value

$$\lim_{(\kappa,c)\to(0,c_1^-)} m_1 = \frac{\theta(-1+2\theta(1+\epsilon)^2)}{(.5-2\theta(1+\epsilon))c_1+.5\epsilon} - \frac{1}{\beta+1}. \tag{4.49}$$

Recall from (2.8) that $0<\epsilon<1$ and $0<\theta<\frac{1}{4(\epsilon+1)}$. This implies that $-1+2\theta(1+\epsilon)^2 < .5(\epsilon-1)<0$ and $.5-2\theta(1+\epsilon)>0$. Thus, the first term in (4.49) is negative, and therefore

$$\lim_{(\kappa,c)\to(0,c_1^-)} m_1 < -\frac{1}{\beta+1} = -2\theta. \tag{4.50}$$

Finally, it follows from (4.50) and continuity that if $\kappa>0$ is small, then

$$\lim_{c\to c_1^-} m_1 < -2\theta. \tag{4.51}$$

(v) *The behavior of $g_1(c)$.* We now use the estimates given above to determine the behavior of $g_1(c)$ when $\kappa>0$ is small. It follows from (4.43) that, for small $\kappa>0$, there is an increasing sequence $\{c_n\}$ such that

$$c_n \to c_2^- \text{ as } n\to\infty \quad \text{and} \quad \gamma a(c_n) = 2n\pi \text{ for large } n. \tag{4.52}$$

Let $c=c_n$ in (4.36). Then $\sin(\gamma a(c_n))=0$, $\cos(\gamma a(c_n))=1$, and (4.36) reduces to

$$\begin{aligned} g_1(c_n) = m_1 &+ e^{-(1+\alpha-\kappa)a(c_n)}\left(\tfrac{.5(c_n(1-\kappa)-\epsilon)}{(1-\kappa)(1-\kappa)^2c_n^2-(1+\epsilon)(1-\kappa)c_n+\epsilon(\beta+1)}\right) \\ &- e^{-(1+\alpha+\kappa)a(c_n)}\left(\tfrac{.5(\epsilon+c_n(1+\kappa))}{(1+\kappa)(1+\kappa)^2c_n^2+(1+\epsilon)(1+\kappa)c_n+\epsilon(\beta+1)}\right). \end{aligned} \tag{4.53}$$

Combining the estimates in (4.41), (4.44), (4.51), and (4.52), we conclude from (4.53) that

$$g_1(c_n) < -\theta \text{ for } n >> 1. \tag{4.54}$$

Likewise, for small $\kappa>0$, there is an increasing sequence $\{c^n\}$ such that

$$c^n \to c_2^- \text{ as } n\to\infty \quad \text{and} \quad \gamma a(c^n) = (2n+1)\pi \text{ for } n >> 1. \tag{4.55}$$

Let $c=c^n$ in (4.36). Then $\sin(\gamma a(c^n))=0$ and $\cos(\gamma a(c^n))=-1$. From this, (4.41), (4.44), (4.51), and (4.55) it follows that

$$g_1(c^n) > \theta \text{ for large } n >> 1. \tag{4.56}$$

It follows from (4.54) and (4.56) and continuity that $g_1(c)$, and therefore $g(c)$, have infinitely many zeros on (c_1,c_2) when $\kappa>0$ is small. Thus, we have proved that there are infinitely many $c\in(c_1,c_2)$ and $a(c)>0$, and corresponding solutions U of (2.3), such that

$$U(z) = \begin{cases} \theta e^{(1+\kappa)z} & \forall z\le 0, \\ U_4(z), & 0<z<a(c), \end{cases} \tag{4.57}$$

where $U_4(z)$ is defined in (4.25)–(4.26) and satisfies

$$U_4(0) = U_4(a(c)) = \theta \quad \text{and} \quad U_4'(0) = 1 + \kappa. \tag{4.58}$$

This completes the first part of the proof of (4.17). It remains to show that the solution satisfies $(U(z), U'(z)) \to (0,0)$ as $z \to \infty$. This is addressed below.

The interval (a, ∞). On the interval (a, ∞) the system (4.11)–(4.12) reduces to

$$c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U = \frac{.5}{1-\kappa}(\epsilon - c(1-\kappa))(e^{(1-\kappa)a} - 1)e^{-(1-\kappa)z} \quad \forall z > a. \tag{4.59}$$

The general solution of (4.59) is

$$U_5(z) = q_1 e^{\alpha z}\cos(\gamma z) + q_2 e^{\alpha z}\sin(\gamma z) + P_5(z), \tag{4.60}$$

where α and γ are defined in (2.7), and P_5 is the particular solution

$$P_5(z) = \tfrac{(\epsilon - c(1-\kappa))(e^{(1-\kappa)a}-1)e^{-(1-\kappa)z}}{(1-\kappa)((1-\kappa)^2c^2 - (1+\epsilon)(1-\kappa)c + \epsilon(\beta+1))}. \tag{4.61}$$

The coefficients q_1 and q_2 in (4.60) are uniquely defined by the continuity conditions

$$U_5(a(c)) = U_4(a(c)) = \theta \quad \text{and} \quad U_5'(a(c)) = U_4'(a(c)). \tag{4.62}$$

Finally, it follows from (4.60)–(4.61) and the fact that $\alpha < 0$ that

$$(U_5(z), U_5'(z)) \to (0,0) \text{ as } z \to \infty. \tag{4.63}$$

This completes the proof of part (i) of Theorem 4.2.

C. *The effects of changing threshold and the identification of* θ^*. To understand how solutions change as θ varies we proceed as in section 3 and compare the behavior of solutions when $\theta = .1$ and $\theta = .2$. For these values our numerical experiments suggest that all conditions for a 1-pulse wave are satisfied along a branch Λ_1 of solutions when $c > 0$, and also along a second branch Λ_2 when $c < 0$. Solutions along Λ_1 and Λ_2 are stable and propagate in opposite directions with different speeds and amplitudes. Figure 6 (top row) illustrates these branches when $(\epsilon, \kappa) = (.1, .15)$.

(i) *The case* $\theta = .1$. Figure 6 (second row, first panel) shows two solutions which coexist at the right endpoints of Λ_1 and Λ_2, where $\beta \approx 12.75$.

(ii) *The case* $\theta = .2$. The middle panel of the second row shows two solutions which coexist when $\beta = 3.35$. The right endpoint of Λ_1 occurs at $\beta \approx 3.5$, where the positive wave speed solution ceases to exist. When $\beta > 3.5$ the negative wave speed solution continues to exist until $\beta \approx 7.5$, where the branch Λ_2 comes to an end. The third panel illustrates the solution at $\beta = 7.5$.

Our experiments lead to the following conjectures whose proofs remain open problems:

- At fixed $(\epsilon, \kappa, \beta)$ the amplitudes and speeds of 1-pulse solutions *decrease* as θ *increases*.
- At fixed (ϵ, κ) the right endpoint of Λ_1 decreases more rapidly than the right endpoint of Λ_2 as θ increases. Thus, there is an interval I of β values such that if $\beta \in$ I is fixed, then a critical value $\theta^* = \theta^*(\epsilon, \kappa, \beta)$ exists such that there are two stable solutions when $0 < \theta \leq \theta^*$ and only one solution when $\theta > \theta^*$.

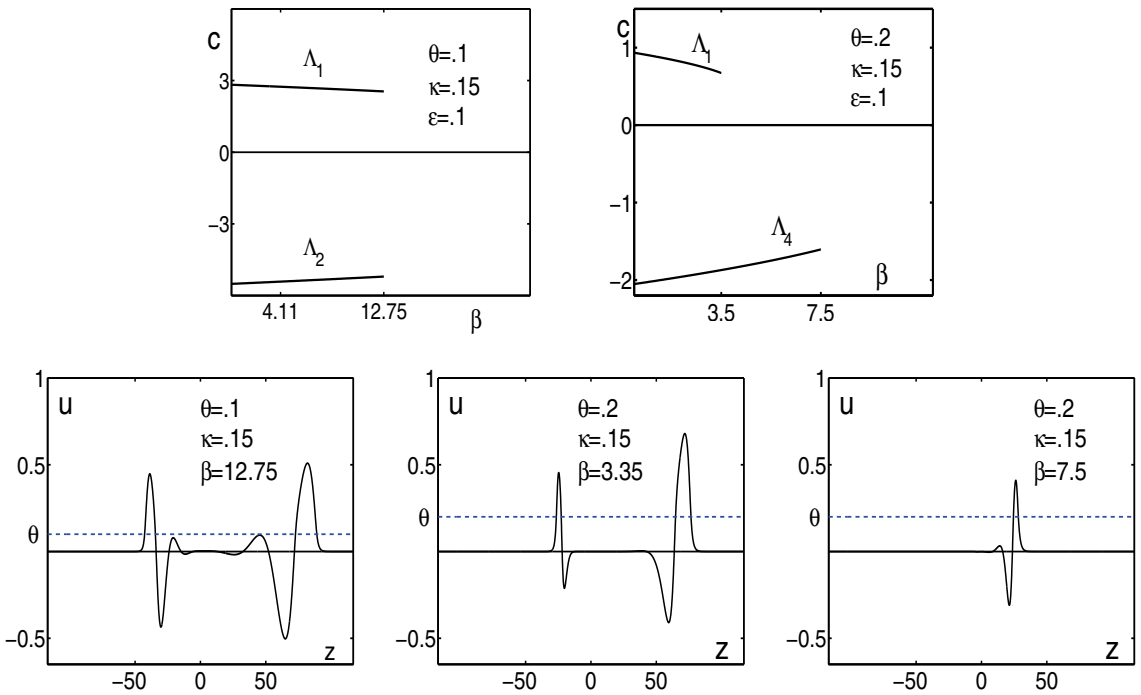


Figure 6. *First row: Bifurcation curves Λ_1 and Λ_2 in the (β, c) plane for families of 1-pulse waves when $(\epsilon, \kappa, \theta) = (.1, .15, .1)$ (left) and $(\epsilon, \kappa, \theta) = (.1, .15, .2)$ (right). Second row: Solutions at specific points along Λ_1 and Λ_2. Please click on each figure in the second row to display the accompanying movie* (70988_02.mpg [2.64MB], 70988_03.mpg [2.1MB], and 70988_04.mpg [1.41MB]).

Experimental implications. By controlling the electric field in disinhibited slices of mammalian cortex, Richardson, Schiff, and Gluckman [48] can determine properties of 1-pulse waves for different values of threshold. Our results show that, when threshold has a low value, an appropriate stimulus causes waves to form which propagate in opposite directions with different amplitudes and speeds. The methods in [48] might allow one to determine if similar properties hold for low electric field values and also whether there is a critical value of the field where one of the waves disappears. This, together with an analysis of the speeds and amplitudes of waves as a function of electric field strength, could provide a plausible method to obtain a measure of the asymmetry in the connectivity between neuronal groups.

5. 2-pulse waves. To understand how 2-pulse waves form when $\beta > \beta_*$, we analyze the linearization of (2.3) around the rest state $U = 0$:

$$c^2 H'' + c(1+\epsilon)H' + \epsilon(\beta+1)H = 0. \tag{5.1}$$

When $\mu^{\pm}$ are complex the general solution of (5.1) is

$$H(z) = b_1 e^{\alpha z} \sin(\gamma z) + b_2 e^{\alpha z} \cos(\gamma z), \tag{5.2}$$

where

$$\alpha = \mathrm{Re}(\mu^{\pm}) = \frac{-(\epsilon+1)}{2c} < 0 \quad \text{and} \quad \gamma = \mathrm{Im}(\mu^{\pm}) = \frac{\sqrt{4\epsilon\beta - (\epsilon-1)^2}}{2c} > 0. \tag{5.3}$$

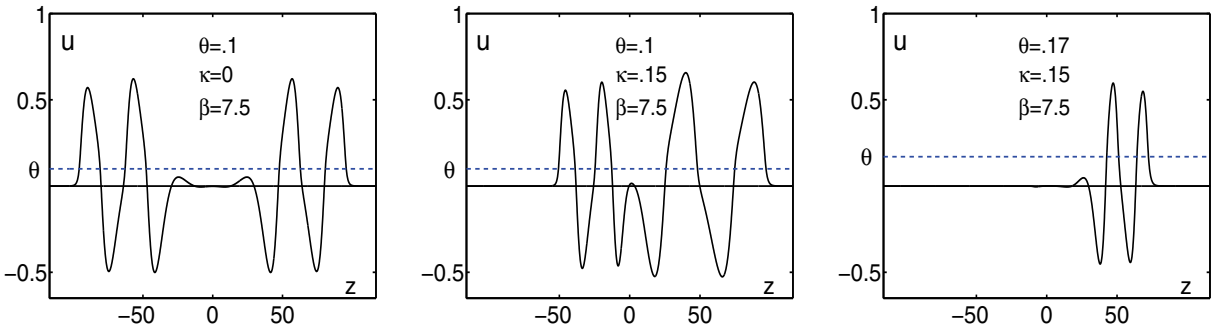


Figure 7. *2-pulse waves when* $\beta = 7.5$ *and* $(\epsilon, \theta, \kappa) = (.1, .1, 0)$ *(left),* $(\epsilon, \theta, \kappa) = (.1, .1, .15)$ *(middle), and* $(\epsilon, \theta, \kappa) = (.1, .17, .15)$ *(right). In the right panel the wave propagates only to the right since* $\theta = .17 > \theta^*$ *when* $(\epsilon, \kappa) = (.1, .15)$. *Click on each panel to display the accompanying movie* (70988_05.mpg [3.51MB], 70988_06.mpg [2.94MB], and 70988_07.mpg [2.73MB]).

It follows from (5.2)–(5.3) that the frequency of oscillation of $H(z)$ increases as β increases from β_*. In turn, this causes solutions of (1.1) to become more oscillatory as β increases, making it increasingly likely that an initial perturbation will evolve into a 2-pulse traveling wave. It follows from (2.3) that a 2-pulse traveling wave satisfies

$$\begin{aligned} c^2U'' + c(1+\epsilon)U' + \epsilon(\beta+1)U &= c\tfrac{d}{dz}\textstyle\int_0^a w(z-z')dz' + \epsilon\int_0^a w(z-z')dz' \\ &\quad + c\tfrac{d}{dz}\textstyle\int_b^d w(z-z')dz' + \epsilon\int_b^d w(z-z')dz', \end{aligned} \tag{5.4}$$

where

$$\begin{cases} U(0) = U(a) = U(b) = U(d) = \theta \text{ for some } d > b > a > 0, \\ U(z) \neq \theta \text{ if } z \notin \{0, a, b, d\}, \\ (U(z), U'(z)) \to (0,0) \text{ as } |z| \to \infty. \end{cases} \tag{5.5}$$

In Figure 7 we consider the representative parameter set $\epsilon = .1$ and $\beta = 7.5$ and illustrate three different types of 2-pulse waves. For this we solve the initial value problem

$$\begin{aligned} &u_t(x,t) = -u - v + \tfrac{1}{2}\textstyle\int_{-120}^{120} e^{-|x-x'|+\kappa(x-x')}H(u(x',t)-\theta)dx', \\ &v_t(x,t) = \epsilon(\beta u - v), \\ &u(x,0) = .6e^{-x^2}, \quad \text{and} \quad v(x,0) = 0 \quad \forall x \in [-120, 120]. \end{aligned} \tag{5.6}$$

In the left panel we let $\theta = .1$ and set $\kappa = 0$ so that w is symmetric. The initial stimulus splits into two 2-pulse waves which propagate outward from $x = 0$. The wave propagating to the left has the same speed and shape as the wave propagating to the right. In the second panel we keep $\theta = .1$ and set $\kappa = .15$ so that w is asymmetric. The initial stimulus again splits into two waves. The wave traveling to the left is slower than the one traveling to the right, and its shape is also different. In the third panel we keep $\kappa = .15$ and raise θ to $\theta = .17$. Once again, the initial perturbation splits into two waves which begin to propagate outward from the $x = 0$. However, because $\theta = .17 > \theta^*$, the "weaker" wave propagating to the left cannot be sustained, and it quickly shrinks and collapses onto the steady state $(u, v) = (0, 0)$. The

wave propagating to the right persists indefinitely. We note that (i) other choices for initial conditions give the same results, and (ii) similar properties were observed for 1-pulse waves (see Figures 4 and 6).

For general $N \geq 2$ a simple extension of (5.4)–(5.5) gives the criteria satisfied by N-pulse traveling waves. In particular, it is necessary to prove that the solution intersects $U = \theta$ exactly 2N times. However, as with 1-pulse waves, the nonlocal terms in the equation lead to technical difficulties in proving this key property, and therefore existence proofs remain a challenging open problem. Finally, we note that our numerical study suggests that the traveling waves in Figure 7 are all stable as solutions of (1.1). It would be of interest to develop Evans function methods to prove this conjecture.

6. Synchronous oscillations. In this section we investigate the formation and spread of uniformly synchronous oscillations when β increases past a second critical value $\beta^* > \beta_*$ where (1.1) becomes bistable. In particular, we show how β^* arises and study the following when $\beta \geq \beta^*$:

A. Bistability: The coexistence of bulk oscillations and a stable rest state.
B. Synchrony: The formation of uniformly synchronous oscillations which spread outward from a point of stimulus.
C. The coexistence of synchronous oscillations and 1-pulse waves.
D. How synchronization in one region can trigger synchronization in another.

A. *Bistability: The coexistence of bulk oscillations and a stable rest state.* Our goal here is to show that there is a critical value β^* where spatially independent bulk oscillations come into existence. Such solutions depend only on the variable t and satisfy

$$\begin{aligned} \tfrac{du}{dt} &= -u - v + f(u-\theta)\textstyle\int_{-\infty}^{\infty} w(\eta)d\eta, \\ \tfrac{dv}{dt} &= \epsilon(\beta u - v). \end{aligned} \tag{6.1}$$

In [58] we considered symmetric couplings and showed that periodic solutions of (6.1) come into existence at a critical β^*. Figure 8 illustrates how this happens for the parameter set $(\epsilon, \theta, \kappa) = (.1, .1, 0)$. In the first row we let $\beta = 12.6$ and see that the solution with initial condition $(u(0), v(0)) = (.4, 0)$ returns to the rest state $(u, v) = (0, 0)$ as $t \to \infty$. The same behavior occurs when $0 < \beta < 12.6$, and for all other initial conditions. Thus, the rest state is globally stable when $0 < \beta \leq 12.6$. At $\beta = \beta^* \approx 12.61$ a periodic solution comes into existence (second row, right panel). Its trajectory forms a closed loop in the (u, v) plane and intersects the $u' = 0$ nullcline at the "threshold point" $(u, v) = (\theta, -\theta)$ (left panel). This property causes the periodic solution to be semistable. That is, solutions of (6.1) whose initial conditions lie outside its trajectory approach a translate of the periodic solution as $t \to \infty$, and solutions with initial conditions inside the trajectory satisfy $(u, v) \to (0, 0)$ as $t \to \infty$. As β increases from β^* a family of stable periodic orbits bifurcates from the periodic solution at β^*. The third row shows such a solution at $\beta = 17$. Our study shows that this mechanism also occurs when $\kappa \neq 0$, and over a wide range of θ, which includes the critical value θ^*. For small $\theta > 0$ a phase plane approach can be used to prove the existence of periodic solutions. General proofs remain open.

B. *Synchrony.* When $\beta \geq \beta^*$ our numerical experiments on the full system (1.1) show that uniformly synchronous oscillations can form and spread outward from a point of stimulus. In

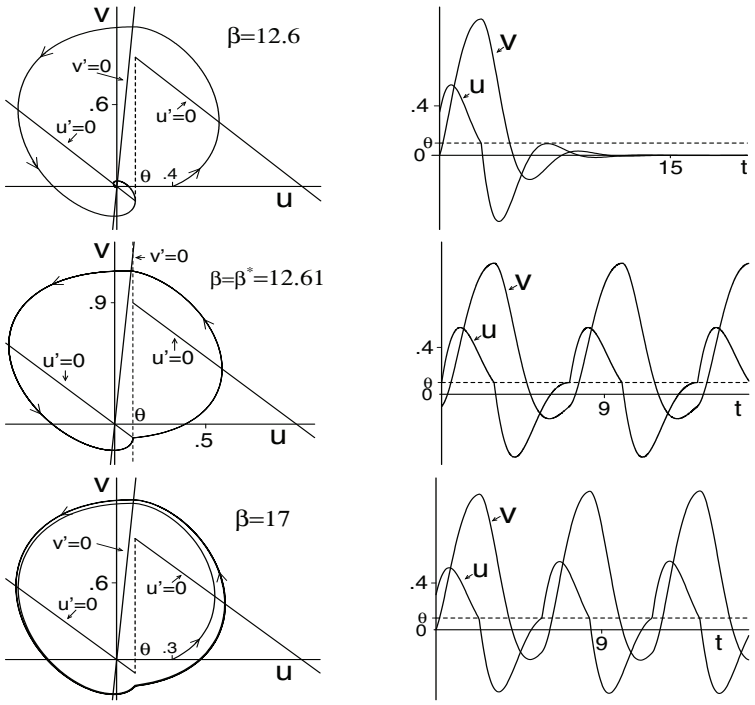


Figure 8. *Bulk oscillations: Solutions of* (6.1) *graphed in the* (u, v) *phase plane (left column), and* u *and* v *as functions of* t *(right column). Parameter values:* $(\epsilon, \theta, \kappa) = (.1, .1, 0)$.

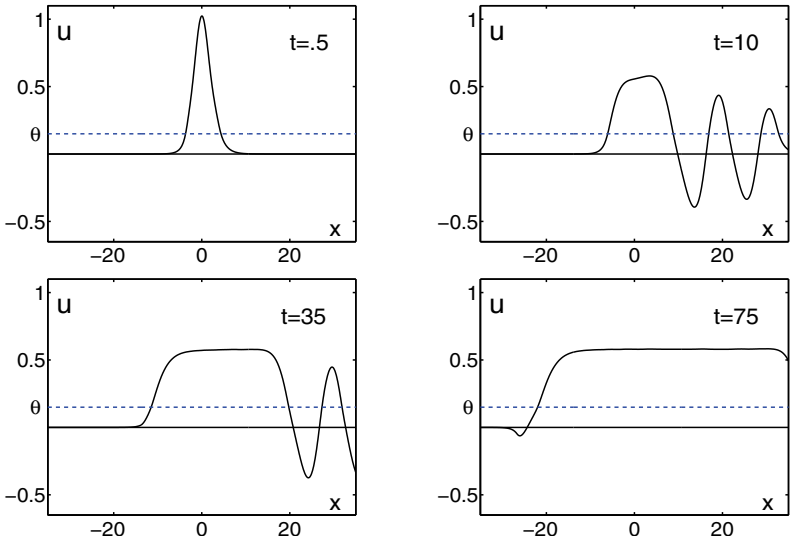


Figure 9. *The formation and spread of uniformly synchronous oscillations in response to the initial stimulus* $(u(x,0), v(x,0)) = (.6e^{-x^2}, 0)$ *when* $(\epsilon, \theta, \kappa, \beta) = (.1, .15, .15, 17)$. *Clicking on the first image displays the accompanying movie* (70988_08.mpg [3.17MB]).

Figure 9 we set $(\beta, \epsilon, \theta, \kappa) = (17, .1, .15, .15)$ and consider the initial condition

$$(u(x,0),v(x,0)) = (.6e^{-x^2},0), \quad x \in \mathrm{R}. \tag{6.2}$$

The stimulus (6.2) is sufficiently strong so that the bulk oscillations affect the solution and cause it to begin oscillating at the center of the stimulus. The oscillations are spatially uniform and in phase over an ever expanding "region of synchrony." Our study suggests the following:

(i) During each oscillation the region of synchrony expands outward by an amount proportional to the speed of a 1-pulse wave.

(ii) During each oscillation the expanding region sheds a traveling wave. Since $\theta > \theta^*$ for our choice of parameters, these waves can propagate only to the right.

Because of the asymmetry in w, the expansion of the region of synchrony is most rapid in the direction to the right of the point of initial stimulus and less rapid to the left. Eventually, however, the solution oscillates uniformly over the entire spatial region, and with the same period as the bulk oscillation. There are similarities between the theoretical spread of a region of synchrony and clinical observations of epileptiform events. Milton and Jung [40, pp. 346–347] point out that "during a seizure there is a propagation of synchrony over the cortical surface" (see Figures 5.7 and 5.8 in [40]) and that optical imaging shows wave-like properties of epileptic propagation. In electocorticographic studies, Towel et al. [57] show how spatially uniform synchronous oscillations develop in a region behind the leading edge of a seizure as it propagates across the cortex (see Figures 6.2, 6.5, and 6.11 in [57]). Milton [39, pp. 18–19] notes that a seizure spreads relatively slowly when compared with spike propagation rates. Our numerical study indicates that the region of synchrony also spreads slowly, at a rate which is only a fraction of the speed of a traveling wave. An important challenge for future research is to extend the methods in [58] and derive a theoretical formula for the rate of expansion which takes into account asymmetric couplings as well as variations in the threshold θ.

C. *The coexistence of synchronous oscillations and 1-pulse waves.* Wright and Sergejew [60] have demonstated the presence of traveling waves in EEG studies of seizure propagation. In [58] we found that a similar phenomenon occurs theoretically; i.e., 1-pulse traveling waves can form in the same parameter regime as synchronous oscillations when $\kappa = 0$ and $\theta > 0$ is small. Here we examine the robustness of this property when $\kappa \neq 0$ and for larger values of θ. For this we consider the representative parameter set $(\beta, \epsilon, \theta, \kappa) = (17, .1, .15, .15)$. At these values the system is bistable, and synchronous oscillations form and spread outward in response to a sufficiently strong initial stimulus. However, synchrony is not always the outcome. For example, Figure 10 illustrates how the solution with the smaller amplitude initial perturbation

$$(u(x,0),v(x,0)) = (.18e^{-x^2},0), \quad x \in \mathrm{R}, \tag{6.3}$$

evolves into a 1-pulse wave. It propagates only to the right since $\kappa = .15$ and $\theta = .15 > \theta^*$. Our study suggests that synchronous oscillations and stable 1-pulse waves coexist for a broad range of parameters. Proofs remain an open problem.

D. *How synchronization in one region can trigger synchronization in another.* We investigate how synchronous oscillations in one region can spread and initiate synchronization in distant regions. This aspect of our study is motivated by two diverse settings.

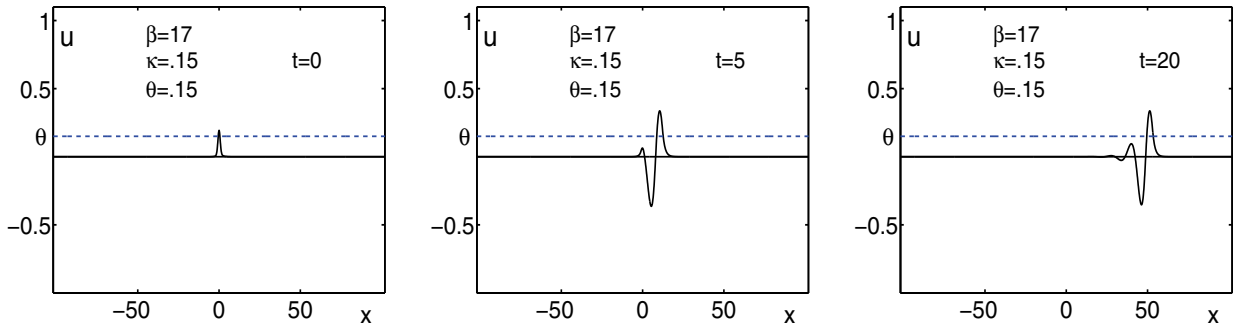


Figure 10. *A small amplitude stimulus* $(u(x,0),v(x,0)) = (.18e^{-x^2},0)$ *evolves into a* 1*-pulse wave when* $(\beta,\epsilon,\theta,\kappa) = (17,.1,.15,.15)$. *Its speed* $c \approx |c_4| = .257$ *is computed from* (3.43). *Clicking on the first panel displays the accompanying movie* (70988_09.mpg [1.73MB]).

(i) In section 1 we described recent experiment results which show that exposure to nicotine reduces the threshold of excitation in thalamocortical mouse neurons and that this dramatically increases their firing rates [35, 54]. It is suggested that in the brains of schizophrenics, the poor communication between thalamus and cortex might be improved by such lowering of threshold.

(ii) Chkhenkeli and Milton [8] describe how seizures in one region of the brain can trigger seizure onset in another. In particular, they give evidence which shows how rhythmic oscillations in the amygdala can trigger a seizure in the hippocampus (see Figure 3.4 in [8]). They also show how a seizure which starts in the hippocampus does not necessarily spread to the amygdala (see Figure 3.4 in [8]).

Our goal here is to understand how similar phenomena occur in our model when $\kappa \neq 0$ and the threshold θ varies. In particular, we find that when θ increases past the critical value θ^*, synchronization in one region can trigger the onset of synchronization in another, but the reverse is not true. To see how this happens we let β vary as a function of x and define

$$\beta(x) = \begin{cases} 17 & \forall x \in [-100,-35), \\ 10 & \forall x \in [-35,35), \\ 17 & \forall x \in [35,100]. \end{cases} \tag{6.4}$$

As in B and C above, we set $(\epsilon,\theta,\kappa) = (.1,.15,.15)$. Since $\beta = 17$ in $[-100,35)$ and $[35,100]$ the results of part B show that spatially uniform synchronous oscillations can develop in these intervals. However, since $\beta = 10 < \beta^*$ when $x \in [-35,35)$, synchronization cannot occur in this interval. Thus, the interval $[-35,35)$ forms a "buffer" region separating the two outer intervals where synchronization can occur. Furthermore, $\theta > \theta^*$ for our choice of parameters; hence waves can propagate only to the right in any of the three subintervals. To understand the behavior of this "unidirectional neuronal circuit," we performed three simple experiments.

I. In the first row of Figure 11 a small initial stimulus centered at $x = -50$ evolves into a 1-pulse wave propagating to the right (left panel). When the wave enters the interval $[-35,35)$ its amplitude and speed increase since β has the lower value $\beta = 10$ (middle panel). As the wave crosses into the interval $[35,100]$ it is not sufficient to trigger synchronization. Instead, its amplitude and speed *decrease* and the wave passes through the interval $[35,100]$ as time t

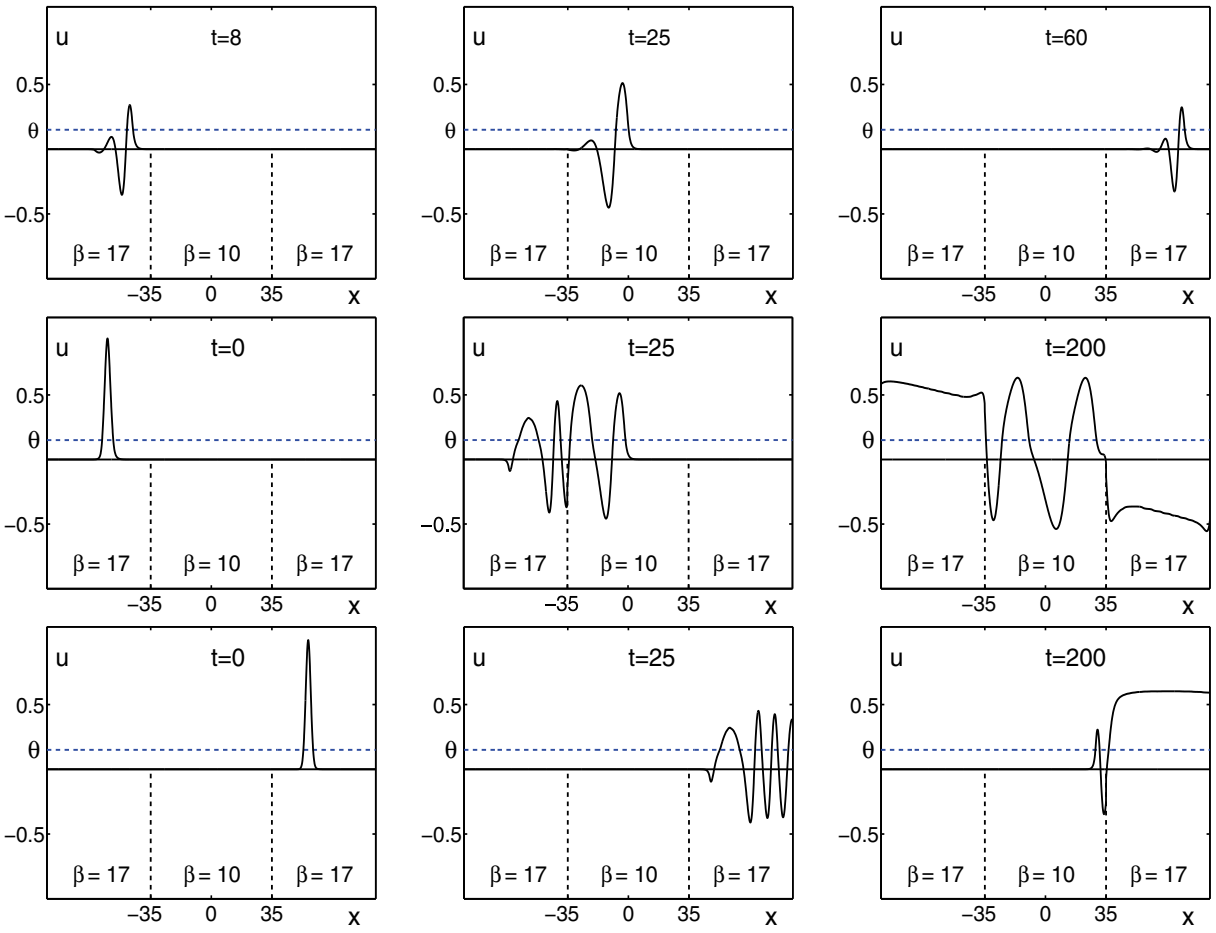


Figure 11. *First row: A small amplitude stimulus* $(u(x,0), v(x,0)) = (.18e^{-(x+50)^2}, 0)$ *evolves into a* 1*-pulse wave. The wave propagates through the entire circuit without triggering any part of it to undergo synchronization. Parameters are* $(\epsilon, \theta, \kappa) = (.1, .15, .15)$. *Second row: A larger amplitude stimulus* $(u(x,0), v(x,0)) = (e^{-(x+50)^2}, 0)$ *triggers synchronization in the interval* $[-100, -35)$. *The train of waves emitted by the synchronizing region propagates through the circuit and initiates synchronization in the interval* $[35, 100]$. *Third row: A stimulus* $(u(x,0), v(x,0)) = (e^{-(x-50)^2}, 0)$ *triggers synchronization in the interval* $[35, 100]$. *Since waves cannot propagate to the left, there can be no initiation of synchronization in the left end interval* $[-100, -35)$. *Click on the first figure in each row to display the accompanying movie* (70988_10.mpg [1.64MB], 70988_11.mpg [16.1MB], and 70988_12.mpg [4.02MB]).

increases (right panel). Thus, a small amplitude stimulus centered in the interval $[-100, -35)$ evolves into a 1-pulse wave which propagates through the entire circuit without triggering synchronization.

II. In the second row a stimulus centered at $x = -50$ (left panel) is sufficiently strong to trigger the formation and spread of synchronous oscillations in the interval $[-100, -35)$. During each cycle the region of synchrony expands by a small amount and a wave is emitted. Thus, a train of waves is created. These waves propagate to the right and enter $[-35, 35)$, where their amplitudes and speeds increase (middle panel). As the waves pass the point

$x = 35$ they are sufficient to initiate synchronization in the interval $[35, 100]$ (right panel). Thus, synchronization in the region $[-100, -35)$ causes the formation of a train of waves which eventually trigger synchronization in the region $[35, 100]$.

Remark. We have found that at least three waves must pass the point $x = 35$ in order to initiate synchronization in the region $[35, 100]$. To test this we found that the initial stimulus $(u(x, 0), v(x, 0)) = (10e^{-(x+20)^2}, 0)$ evolves into a 2-pulse wave which propagates to the right and passes right on through the interval $[35, 100]$ without initiating synchronization.

III. In the third row we test whether synchronous oscillations in the right end interval $[35, 100]$ can ultimately trigger synchronous oscillations in the left end interval $[-100, 35)$. An initial stimulus centered at $x = 50$ (left panel) initiates synchronous oscillations which spread uniformly over the entire interval $[35, 100]$ (second and third panels). However, our choice of parameters does not allow for left propagating waves; hence there is no initiation of synchronous oscillations in $[-100, -35)$ where the solution remains in the rest state $u = v = 0$.

Remarks. The results described above in I–III still hold when β has different values in the outer intervals $[-100, 35)$ and $[35, 100]$. In this setting, when synchronous oscillations are initiated in these intervals, their frequencies do not necessarily entrain. Analytical proofs of the numerical experiments described in parts A–D remain a challenging open problem.

7. Conclusions. In this paper we analyzed the dynamic behavior of a system of integro-differential equations that models the activity of excitatory neurons on large-scale, spatially extended domains. The independent variables represent the activity level of a population of excitatory neurons with long range connections (u) and recovery (v). We considered positive, asymmetric coupling functions (w) and a Heaviside firing rate (f).

There is a critical value of the parameter β ($\beta_* > 0$) that appears in the equation for v, at which the eigenvalues $\mu^{\pm}$ of the linearization of the system around the rest state $(u, v) = (0, 0)$ change from real to complex. If $0 < \beta \leq \beta_*$, then $\mu^{\pm}$ are real and both wave fronts and 1-pulse traveling waves can exist. In [58] we explained why multipulse waves are not expected in the real eigenvalue case. By contrast, when $\beta > \beta_*$ and $\mu^{\pm}$ are complex, the range of behavior is much richer. For example, our analysis provides evidence for the coexistence of at least two distinct families of stable wave fronts. Because w is asymmetric, these solutions propagate in opposite directions with different speeds and shapes. We have also found a range of $\beta > \beta_*$ where two families of 1-pulse traveling wave solutions exist (Theorem 4.2). Each family consists of infinitely many coexisting solutions, and solutions in the two families propagate in opposite directions with different speeds and shapes. In addition, we study the effects of variations of threshold θ on the dynamics of the system. As θ increases, the speeds and amplitudes of the waves in each family decrease until a critical value $\theta^* > 0$ is reached where solutions in the first family disappear. That is, left propagating 1-pulse waves cease to exist when $\theta > \theta^*$. In addition there is a range $\theta > \theta^*$ where 2-pulse waves can propagate only in one direction. This phenomenon does not occur when the coupling is a symmetric function. To our knowledge this is the first description of such unidirectional wave propagation in this class of nonlocal model.

There is a a second critical value $\beta^* > \beta_*$ where (1.1) becomes bistable and a family of spatially independent bulk oscillations comes into existence. These solutions influence the global dynamics of (1.1). For example, when $\beta \geq \beta^*$ a strong initial stimulus evolves into uniformly synchronous oscillations which spread outward from the point of stimulus. However,

a weak stimulus does not trigger synchronization and merely evolves into a 1-pulse traveling wave. We also study how variations in θ affect the spread of synchrony. In particular, we let $\kappa > 0$ be fixed and allow $\beta > \beta^*$ to be a function of the spatial coordinate x. We then raise θ to a value above the critical value θ^* so that waves propagate only to the right. In the resulting unidirectional circuit we show how synchronization in one region can trigger synchronization in a second, distant region. However, when synchronization is triggered in the second region, it cannot spread to the first region and the first region remains at rest.

In all cases formidable technical difficulties preclude the completion of the final step of existence proofs. It remains an open problem to extend our methods so that existence proofs can be completed for a wider range of parameters, and also for more general coupling and firing rate functions.

Our numerical experiments were performed when the firing rate is the Heaviside function. To test the robustness of our results we have considered more general, sigmoidal-shaped firing rates of the form

$$f(u) = \frac{1}{1 + Ke^{-r(u-\theta)}}, \quad K > 0,\ r > 0, \tag{7.1}$$

and

$$f(u) = Ke^{-\frac{r}{(u-\theta)^2}} H(u-\theta), \quad K > 0,\ r > 0. \tag{7.2}$$

With f given by (7.1) or (7.2), our numerical results continue to hold when M is of moderate size and R is large (e.g., $K \approx 1$ and $R \geq 50$). It remains an open problem to determine the maximal range of parameters, firing rate, and coupling functions over which the numerical results are valid.

Our theoretical results might have important implications for experimental and clinical neurophysiology. In particular, our finding that the dynamics of (1.1) undergo qualitative transitions when $\mu^{\pm}$ become complex, or θ exceeds θ^*, offers a plausible explanation of trailing-end instabilities and wave speed variations observed in cortical experiments [35, 45, 46]. Further explanation of observed variability in cortical waves might be provided by our findings that the asymmetry in w leads to the coexistence of entire families of traveling waves which propagate in opposite directions with different shapes and speeds. The unpredictable variation in trailing ends and wave speeds could be caused by solutions switching from one member of the family to another. A possible biophysical mechanism of such switching may involve a variable neurohormonal concentration affecting neuronal recovery and strength of intercellular connections [31].

Our observation that bifurcations of the system behavior occur at the critical values $\beta = \beta_*$ or $\theta = \theta^*$ also has important practical correlates. It predicts that by pushing the system above or below one of these values one can qualitatively change the system behavior and obtain a broad range of dynamical phenomena. One experimental example of such macrobehavior is an evoked response, which might persist long after the stimulus [46]. Understanding the cellular mechanisms responsible for such important functional changes in neuronal networks requires further study.

Acknowledgments. The author thanks Brent Doiron and Vladimir Shusterman for useful discussions during the preparation of the manuscript.

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral inhibition type neural fields*, Biol. Cybernet., 27 (1977), pp. 77–87.
[2] P. C. Bressloff and S. E. Folias, *Front bifurcations in an excitatory neural network*, SIAM J. Appl. Math., 65 (2004), pp. 131–151.
[3] B. D. Burns, *Some properties of the cat's isolated cerebral cortex*, J. Physiol., 111 (1950), pp. 50–68.
[4] B. D. Burns, *Some properties of the isolated cerebral cortex in the unanaesthetized cat*, J. Physiol., 112 (1951), pp. 156–175.
[5] B. D. Burns and B. Grafstein, *The function and structure of some neurones in the cat's cerebral cortex*, J. Physiol., 118 (1952), pp. 412–433.
[6] G. Buzsaki and A. Draguhn, *Neuronal oscillations in cortical networks*, Science, 304 (2004), pp. 1926–1929.
[7] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Periodicity and directionality in the propagation of epileptiform discharges across neocortex*, J. Neurophys., 60 (1988), pp. 1695–1713.
[8] S. A. Chkhenkeli and J. Milton, *Dynamic epileptic systems versus epileptic foci*, in Epilepsy as a Dynamic Disease, Biological and Medical Physics Series, J. Milton and P. Jungs, eds., Springer-Verlag, New York, 2003, Chapter 3.
[9] P. H. Chu, J. Milton, and J. D. Cowan, *Connectivity and the dynamics of integrate and fire neural networks*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1992), pp. 237–243.
[10] B. W. Connors and Y. Amati, *Generation of epileptiform discharge by local circuits of neocortex*, in Epilepsy: Models, Mechanisms and Concepts, P. A. Schwartkroin, ed., Cambridge University Press, Cambridge, UK, 1993, pp. 388–423.
[11] S. Coombes, *Waves and bumps in neural field theories*, Biol. Cybernet., 93 (2005), pp. 91–108.
[12] S. Coombes and M. R. Owen, *Evans functions for integral neural field equations with Heaviside firing rate function*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 574–600.
[13] S. Coombes and M. R. Owen, *Bumps, breathers, and waves in a neural network with spike frequency adaptation*, Phys. Rev. Lett., 94 (2005), 148102.
[14] S. Coombes and M. R. Owen, *Exotic dynamics in a firing rate model of neuronal tissue with threshold accommodation*, in Fluids and Waves: Recent Trends in Applied Analysis, Contemp. Math. 440, F. Bothelo, T. Hagan, and J. Jamison, eds., AMS, Providence, RI, 2007, pp. 123–144.
[15] J. D. Drover and B. Ermentrout, *Nonlinear coupling near a degenerate Hopf (Bautin) bifurcation*, SIAM J. Appl. Math., 63 (2003), pp. 1627–1647.
[16] J. S. Ebersol and J. Milton, *The electroencephalogram (EEG): A measure of neural synchrony*, Chapter 5 of Epilepsy as a Dynamic Disease, Biological and Medical Physics Series, Springer-Verlag, New York, 2003.
[17] G. B. Ermentrout and J. B. McLeod, *Existence and uniqueness of traveling waves for a neural network*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 461–478.
[18] I. Ferezou, S. Bolea, and C. Petersen, *Visualizing the cortical representation of whisker touch: Voltage sensitive dye imaging in freely moving mice*, Neuron, 50 (2006), pp. 617–629.
[19] S. Folias and P. Bressloff, *Breathers in two dimensional excitable neural media*, Phys. Rev. Lett., 95 (2004), 208107.
[20] S. E. Folias and P. C. Bressloff, *Breathing pulses in an excitatory neural network*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 378–407.
[21] J. Glanz, *Mastering the nonlinear brain*, Science, 277 (5333) (1997), pp. 1758–1760.
[22] D. Golomb, *Models of neuronal transient synchrony during propagation of activity through neocortical circuitry*, J. Neurophys., 79 (1998), pp. 1–12, 1335–1348.
[23] D. Golomb and Y. Amati, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophys., 78 (1997), pp. 1199–1211, 1335–1348.

[24] Y. Guo and C. C. Chow, *Existence and stability of standing pulses in neural networks:* I. *Existence*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 217–248.
[25] Y. Guo and C. C. Chow, *Existence and stability of standing pulses in neural networks:* II. *Stability*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 249–281.
[26] B. Gutkin, D. Pinto, and B. Ermentrout, *Mathematical neuroscience: From neurons to circuits to systems*, J. Physiol. Paris, 97 (2003), pp. 209–219.
[27] S. P. Hastings, *Single and multiple pulse waves for the FitzHugh–Nagumo equations*, SIAM J. Appl. Math., 42 (1982), pp. 247–260.
[28] A. V. Hill, *Excitation and accommodation in nerve*, Proc. Roy. Soc. London Ser. B, 119 (1936), pp. 305–355.
[29] J. A. Hobson, *Dreaming: An Introduction to the Science of Sleep*, Oxford University Press, Oxford, UK, 2004.
[30] J. A. Hobson and R. W. McCarley, *The brain as a dream state generator. An activation-synthesis of the dream process*, Amer. J. Psychiatry, 134 (1977), pp. 1335–1348.
[31] X. Huang, W. C. Troy, Q. Yang, H. Ma, C. Laing, S. Schiff, and J. Y. Wu, *Spiral waves in disinhibited mammalian cortex*, J. Neurosci., 24 (2004), pp. 9897–9902.
[32] M. A. P. Idiart and L. F. Abbott, *Propagation of excitation in neural network models*, Network, 4 (1993), pp. 285–294.
[33] D. Kleinfield, K. R. Delaney, M. S. Fee, J. A. Flores, D. W. Tank, and A. Galperin, *Dynamics of propagating waves in the olfactory network of a terrestrial mollusk: An electrical and optical study*, J. Neurophys., 72 (1994), pp. 1402–1419.
[34] N. Kopell and B. Ermentrout, *Chemical and electrical synapses perform complementary roles in the synchronization of interneuronal networks*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 15482–15487.
[35] H. Kowai, R. Lazar, and R. Metherate, *Nicotine control of axon excitability regulates thalamocortical transmission*, Nature Neuroscience, 10 (2007), pp. 1168–1175.
[36] Y. W. Lam, L. B. Cohen, M. Wachowiak, and M. R. Zochowski, *Odors elicit three different oscillations in the turtle olfactory bulb*, J. Neurosci., 20 (2000), pp. 749–762.
[37] R. Miles, R. D. Traub, and R. K. Wong, *Spread of synchronous firing in longitudinal slices from the CA3 region of the hippocampus*, J. Neurophys., 60 (1988), pp. 1481–1496.
[38] J. Milton, T. Mundel, U. an der Heiden, J. Sprire, and J. Cowan, *Traveling activity waves*, in Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA, 1994, pp. 994–996.
[39] J. Milton, *Insights into seizure propagation from axonal conductance times*, in Epilepsy as a Dynamic Disease, Biological and Medical Physics Series, J. Milton and P. Jung, eds., Springer-Verlag, New York, 2003, Chapter 2.
[40] J. Milton and P. Jung, *The electroencephalogram (EEG): A measure of and neural synchrony*, in Epilepsy as a Dynamic Disease, Biological and Medical Physics Series, J. Milton and P. Jung, eds., Springer-Verlag, New York, 2003, Chapter 5.
[41] C. Peskin, *Mathematical Aspects of Heart Physiology*, Courant Institute of Mathematical Sciences, New York University, New York, 1975.
[42] D. J. Pinto and B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* I. *Traveling fronts and pulses*, SIAM J. Appl. Math., 62 (2001), pp. 206–225.
[43] D. J. Pinto and B. Ermentrout, *Spatially structured activity in synaptically coupled neuronal networks:* II. *Lateral inhibition and standing pulses*, SIAM J. Appl. Math., 62 (2001), pp. 226–243.
[44] D. J. Pinto, R. K. Jackson, and C. E. Wayne, *Existence and stability of traveling pulses in a continuous neuronal network*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 954–984.
[45] D. J. Pinto, S. A. Patrick, H. W. Huang, and B. Connors, *Initiation, propagation and termination of epileptiform activity in rodent neocortex in vitro involve distinct mechanisms*, J. Neurosci., 25 (2005), pp. 8131–8140.
[46] J. C. Prechtl, L. B. Cohen, B. Pasaram, P. P. Mitra, and D. Kleinfeld, *Visual stimuli induce waves of electrical activity in turtle cortex*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 7621–7626.
[47] D. Pinto and W. C. Troy, *The Effects of Asymmetric Coupling on Traveling Waves in Neocortex*, in preparation, 2007.
[48] K. Richardson, S. J. Schiff, and B. J. Gluckman, *Control of traveling waves in mammalian cortex*, Phys. Rev. Lett., 94 (2005), 028103.

[49] A. Rosenbluth and W. B. Cannon, *Cortical responses to electrical stimulation*, Amer. J. Physiol., 135 (1942), pp. 690–741.
[50] B. Sandstede, *Evans functions and nonlinear stability of traveling waves in neuronal network models*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 17 (2007), pp. 2693–2704.
[51] B. Schechter, *How the brain gets its rhythm*, Science, 274 (5286) (1996), p. 339.
[52] J. Schofflen, R. Oostenveld, and P. Fries, *Neuronal coherence as a mechanism of effective corticospinal interaction*, Science, 308 (2003), pp. 111–113.
[53] I. A. Shevlev, E. N. Tsicalov, A. M. Gorbach, K. P. Budko, and G. A. Sharaev, *Temperature tomography of the brain cortex: Thermoencephalography*, J. Neurosci. Methods, 46 (1992), pp. 49–57.
[54] Science News, 172 (2000), pp. 148–149.
[55] V. Shusterman and W. C. Troy, *From baseline to epileptiform activity: A path to synchronized rhythmicity in large-scale neural networks*, Phys. Rev. E, 77 (2008), 061911.
[56] S. Strogatz, *SYNC*, Hyperion Books, New York, 2003.
[57] V. I. Towel, F. Ahmad, M. Kohrman, H. Hecox, and S. Chkhenkeli, *Electrocortigraphic coherence patterns of epileptic seizures*, in Epilepsy as a Dynamic Disease, Biological and Medical Physics Series, J. Milton and P. Jungs, eds., Springer-Verlag, New York, 2003, Chapter 6.
[58] W. C. Troy and V. Shusterman, *Patterns and features of families of traveling waves in large-scale neuronal networks*, SIAM J. Appl. Dyn. Syst., 6 (2007), pp. 263–292.
[59] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.
[60] J. J. Wright and A. Sergejew, *Radial coherence, wave velocity and damping of electrocortical waves*, Electroenceph. Clin. Neurophysiol., 79 (1991), pp. 403–412.
[61] J. Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neurosci., 19 (1999), pp. 5005–5015.
[62] L. Zhang, *On stability of traveling wave solutions in synaptically coupled neuronal networks*, Differential Integral Equations, 16 (2003), pp. 513–536.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 7, No. 4, pp. 1283–1322
© 2008 Society for Industrial and Applied Mathematics

TC-HAT ($\widehat{\text{TC}}$): A Novel Toolbox for the Continuation of Periodic Trajectories in Hybrid Dynamical Systems*

Phanikrishna Thota† and Harry Dankowicz‡

Abstract. This paper describes the underlying formulation and functionality of the newly developed software program $\widehat{\text{TC}}$ ("TC-HAT"), to perform bifurcation analysis of systems in which continuous-in-time dynamics are interrupted by discrete-in-time events, often referred to as *hybrid dynamical systems*. Boundary-value-problem formulations corresponding to single- and two-parameter continuations of periodic trajectories and selected associated codimension-one bifurcations in such systems are presented. Finally, the capabilities of the program are illustrated by performing bifurcation analysis of a few example hybrid dynamical systems.

Key words. hybrid dynamical systems, bifurcation analysis, continuation software, TC-HAT

AMS subject classifications. 34B10, 34B08, 65P30, 65L50, 65L10, 37N30

DOI. 10.1137/070703028

1. Introduction.

A combination of theoretical and computational tools for bifurcation analysis of dynamical systems offers distinct advantages to brute-force forward-time simulation [29]. Such a combination enables prediction of behavior and response without the need for a vast collection of simulations based at distinct initial conditions. More importantly, it may offer an understanding of, and an underlying explanation for, changes in behavior and response that are not available through simple simulation. Such analysis may further establish the existence of structure in the response of a dynamical system that would not be accessible to forward-time simulation. This may, in turn, enable a critical evaluation of the validity of the output of computer simulations.

A comprehensive bifurcation analysis of a dynamical system seeks to establish the existence of characteristic classes of responses, such as equilibria or periodic responses. In each case, this involves locating and tracking families of such responses under variations in system parameters in a process known as *continuation* [2, 10, 12, 17, 18, 19, 28, 34, 37, 38, 40, 44]. The study of the robustness of particular system responses further emphasizes parameter values where such families merge or terminate or where the stability characteristics of the corresponding responses change. Here, characteristic normal forms may be used to establish universal unfoldings of the associated *bifurcation structure* of response families [26]. In turn, these unfoldings provide guidance for further continuation.

*Received by the editors September 9, 2007; accepted for publication (in revised form) by W. Beyn June 3, 2008; published electronically October 31, 2008. This material is based upon work supported by the National Science Foundation under grants 0237370 and 0635469.
http://www.siam.org/journals/siads/7-4/70302.html
†Department of Engineering Mathematics, University of Bristol, Bristol, BS8 1TR, UK (Phani.Thota@bris.ac.uk).
‡Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (danko@uiuc.edu).

Hybrid dynamical systems describe a class of mathematical models in which the continuous evolution of continuous and discrete state variables is punctuated by discontinuous changes in these variables and in the description of their future evolution. In mechanical engineering applications, hybrid dynamical systems occur in typical models of impacts (corresponding to instantaneous jumps in system velocities) or dry friction (including distinct conditions of sustained stick or slip) [31, 33, 35, 54]. In models of electrical circuitry, hybrid systems represent idealized versions of nonlinear circuit elements, such as diodes and transistors. Similarly, switched feedback strategies occur commonly in control applications. Finally, models of biochemical systems, including those describing biomolecular reactions, often include chemical switches and other triggered events, for example, the mitotic halving of the cell mass [51].

Hybrid dynamical systems exhibit bifurcations of the same fundamental nature as those observed in smooth systems, including saddle-node, Hopf, and period-doubling bifurcations. In contrast to smooth systems, however, hybrid dynamical systems also exhibit *discontinuity-induced bifurcations* associated with changes in the hybrid time history of the reference response. For example, *grazing bifurcations*, associated with the onset of zero-relative-velocity contact in systems with impact-like discontinuities, are known to be associated with strong instabilities and complicated sequences of postgrazing bifurcations.

Analytical tools for the study of discontinuity-induced bifurcations of periodic trajectories include the discontinuity-mapping technique pioneered by Nordmark and collaborators [9, 14, 41, 42, 50]. This method introduces a unique correction to the prebifurcation description of the local dynamics and an effective normal-form description of the postbifurcation behavior. In particular, the technique allows one to resolve the degree of singularity associated with the discontinuity-induced bifurcation and to establish the existence of nearby families of periodic trajectories with distinct hybrid time histories.

A number of computational tools are available for bifurcation analysis of characteristic classes of response, such as equilibria, periodic trajectories, homo- or heteroclinic trajectories between equilibria and/or periodic trajectories, quasiperiodic trajectories on invariant tori, and stable and unstable manifolds. These include general algebraic and two-point boundary-value solvers for ordinary differential equations, such as AUTO (and specialized drivers, such as HOMCONT [4, 6] and SLIDECONT [11]), MATCONT [13], and SYMPERCON [53]; boundary-value solvers for delay differential equations, such as DDE-BIFTOOL [20] and PDDE-CONT [48]; tools for large-scale systems, such as LOCA [45]; and implementations in MATLAB [5, 27].

The purpose of this paper is to establish a working definition of hybrid dynamical systems that is amenable to bifurcation analysis and to the implementation of continuation algorithms for periodic responses in such systems. In particular, selected boundary-value-problem formulations are proposed that enable single- and two-parameter continuations of periodic responses and some associated bifurcations. The formulation and its implementation in the Fortran-based software application $\widehat{\text{TC}}$ ("TC-HAT") provides a semiautomated tool for computational bifurcation analysis of periodic responses that complements the functionality present in the packages mentioned above.

The manuscript is organized as follows. Section 2 formulates a working definition of a hybrid dynamical system in terms of its essential components and its solutions. Several fundamental boundary-value-problem formulations are given in section 3. Section 4 reviews the general orthogonal collocation scheme employed in AUTO with particular emphasis on the

modifications necessary to accommodate hybrid dynamics and presents some specific comments on the implementation and usage of $\widehat{\text{TC}}$. A series of model examples is presented in section 5 to illustrate the formalism. Finally, a concluding discussion in section 6 highlights a number of desirable additions to the $\widehat{\text{TC}}$ functionality that would enable a more comprehensive study of the solution structure of hybrid dynamical systems. This section also summarizes drawbacks in the present implementation that may affect computational accuracy and that would be successfully addressed through a reimplementation of the boundary-value-problem formulations in a stand-alone hybrid system continuation application.

2. Hybrid dynamical systems.

2.1. General setup. In this paper (cf. Thota [49] and Kang et al. [32]), a *hybrid dynamical system* assumes the existence of a state space $\mathbb{X}$ of dimension n and an associated smooth vector-valued function $\mathbf{f}_{\mathfrak{m}} : \mathbb{X} \to \mathbb{X}$ known as the *vector field*, indexed by a *mode variable* $\mathfrak{m}$ in some finite set of modes $\mathfrak{M}$. Moreover, denote by $\mathfrak{E}$ a finite collection of *events*, and associate to each element $\mathfrak{e} \in \mathfrak{E}$ a pair $\pi_{\mathfrak{e}} = (\mathfrak{m}, \mathfrak{m}')$ of elements of $\mathfrak{M}$, a smooth *event function* $h_{\mathfrak{e}} : \mathbb{X} \to \mathbb{R}$, and a smooth *jump function* $\mathbf{g}_{\mathfrak{e}} : \mathbb{X} \to \mathbb{X}$. Then, a pair of modes $(\mathfrak{m}_{\text{in}}, \mathfrak{m}_{\text{out}})$ and a pair of states $(\mathbf{x}_{\text{in}}, \mathbf{x}_{\text{out}})$ are said to be *connected* by the event $\mathfrak{e}$ if

$$\pi_{\mathfrak{e}} = (\mathfrak{m}_{\text{in}}, \mathfrak{m}_{\text{out}}), \tag{2.1}$$

$\mathbf{x}_{\text{in}}$ is a point on the *event surface*

$$\{\mathbf{x} \mid h_{\mathfrak{e}}(\mathbf{x}) = 0,\ \partial_{\mathbf{x}} h_{\mathfrak{e}}(\mathbf{x}) \cdot \mathbf{f}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}) \leq 0\},$$

i.e.,

$$h_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = 0, \tag{2.2}$$

and

$$\mathbf{g}_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = \mathbf{x}_{\text{out}} \tag{2.3}$$

(cf. Figure 1).

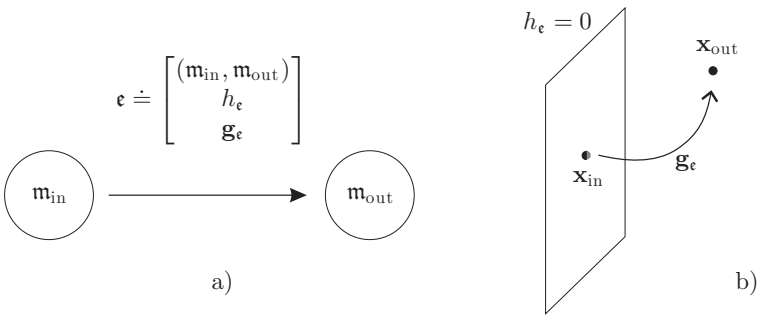


Figure 1. *The event $\mathfrak{e}$ results in a change in mode from $\mathfrak{m}_{\text{in}}$ to $\mathfrak{m}_{\text{out}}$ and a jump in state from $\mathbf{x}_{\text{in}}$ to $\mathbf{x}_{\text{out}}$.*

A *solution* of (or *trajectory* of) the corresponding dynamical system on a finite interval of time $[t_0, t_N]$ is a sequence $\xi = \{\mathbf{x}_j : [t_{j-1}, t_j] \to \mathbb{X}\}_{j=1}^{N}$ of N smooth curves, an associated

sequence of modes $\{\mathfrak{m}_j\}_{j=1}^{N}$, and an associated sequence of events $\{\mathfrak{e}_j\}_{j=1}^{N-1}$, such that the corresponding tangent vector at $\mathbf{x}_j(t)$ equals $\mathbf{f}_{\mathfrak{m}_j}(\mathbf{x}_j(t))$, i.e., in the case of $\mathbb{X} = \mathbb{R}^n$,

$$\frac{d}{dt}\mathbf{x}_j(t) = \mathbf{f}_{\mathfrak{m}_j}(\mathbf{x}_j(t)), \tag{2.4}$$

and such that, for $j \in [1, N-1]$, the pair of modes $(\mathfrak{m}_j, \mathfrak{m}_{j+1})$ and the pair of states $(\mathbf{x}_j(t_j), \mathbf{x}_{j+1}(t_j))$ are connected by the event $\mathfrak{e}_j$, i.e.,

$$h_{\mathfrak{e}_j}(\mathbf{x}_j(t_j)) = 0 \tag{2.5}$$

and

$$\mathbf{g}_{\mathfrak{e}_j}(\mathbf{x}_j(t_j)) = \mathbf{x}_{j+1}(t_j). \tag{2.6}$$

The concatenation $\mathbf{\Sigma} = \{\mathfrak{m}_1, \mathfrak{e}_1, \mathfrak{m}_2, \mathfrak{e}_2, \ldots, \mathfrak{m}_{N-1}, \mathfrak{e}_{N-1}, \mathfrak{m}_N\}$ will be called the solution's *signature* (cf. [36]). It is clear that this definition encapsulates the case of a smooth dynamical system, for which $\mathfrak{M}$ consists of a single element and $\mathfrak{E}$ is empty. A solution of a hybrid dynamical system will be said to be *simple* in the case that the signature has only a single element and *hybrid* otherwise.

2.2. The hybrid flow. Denote by $\mathbf{\Phi}_{\mathfrak{m}}$ the flow function corresponding to the vector field $\mathbf{f}_{\mathfrak{m}}$. Then, for $j \in [1, N-1]$ it follows that

$$\mathbf{\Phi}_{\mathfrak{m}_j}(\mathbf{x}_j(t_{j-1}), t_j - t_{j-1}) = \mathbf{x}_j(t_j), \tag{2.7}$$

where $T_j \overset{\text{def}}{=} t_j - t_{j-1}$ is the *time-of-flight* of the jth segment. As in smooth systems, the Jacobian $\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, t)$ is obtained as the solution to the variational initial-value problem

$$\frac{d}{dt}\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, t) = \partial_{\mathbf{x}}\mathbf{f}_{\mathfrak{m}}(\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, t)) \cdot \partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, t), \tag{2.8}$$

$$\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, 0) = Id, \tag{2.9}$$

where Id denotes the $n \times n$ identity matrix. Moreover, differentiation of $\mathbf{f}_{\mathfrak{m}}(\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, t))$ with respect to time shows that

$$\mathbf{f}_{\mathfrak{m}}(\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, t)) = \partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}}(\mathbf{x}, t) \cdot \mathbf{f}_{\mathfrak{m}}(\mathbf{x}) \tag{2.10}$$

and, in particular, that

$$\mathbf{f}_{\mathfrak{m}_j}(\mathbf{x}_j(t_j)) = \partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}_j}(\mathbf{x}_j(t_{j-1}), t_j - t_{j-1}) \cdot \mathbf{f}_{\mathfrak{m}_j}(\mathbf{x}_j(t_{j-1})). \tag{2.11}$$

An event $\mathfrak{e}$ with $\pi_{\mathfrak{e}} = (\mathfrak{m}_{\text{in}}, \mathfrak{m}_{\text{out}})$ is *transversal* if

$$\partial_{\mathbf{x}}h_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) \cdot \mathbf{f}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}_{\text{in}}) < 0. \tag{2.12}$$

In this case, the function

$$F(\mathbf{x}, t) = h_{\mathfrak{e}}(\mathbf{\Phi}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}, t)) \tag{2.13}$$

satisfies the conditions

$$F(\mathbf{x}_{\text{in}}, 0) = h_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = 0 \tag{2.14}$$

and

$$\partial_t F(\mathbf{x}_{\text{in}}, 0) = \partial_{\mathbf{x}} h_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) \cdot \mathbf{f}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}_{\text{in}}) < 0. \tag{2.15}$$

From the implicit function theorem it follows that there exists a unique smooth function $\tau_{\mathfrak{e}}(\mathbf{x})$ for $\mathbf{x} \approx \mathbf{x}_{\text{in}}$, such that

$$\tau_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = 0, \tag{2.16}$$

$$h_{\mathfrak{e}}(\mathbf{\Phi}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}, \tau_{\mathfrak{e}}(\mathbf{x}))) \equiv 0, \tag{2.17}$$

and

$$\partial_{\mathbf{x}} \tau_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = -\frac{\partial_{\mathbf{x}} h_{\mathfrak{e}}(\mathbf{x}_{\text{in}})}{\partial_{\mathbf{x}} h_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) \cdot \mathbf{f}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}_{\text{in}})}. \tag{2.18}$$

For $\mathbf{x} \approx \mathbf{x}_{\text{in}}$ the *discontinuity mapping*

$$\mathbf{D}_{\mathfrak{e}}(\mathbf{x}) = \mathbf{\Phi}_{\mathfrak{m}_{\text{out}}}(\mathbf{g}_{\mathfrak{e}}(\mathbf{\Phi}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}, \tau_{\mathfrak{e}}(\mathbf{x}))), -\tau_{\mathfrak{e}}(\mathbf{x})) \tag{2.19}$$

satisfies the conditions

$$\mathbf{D}_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = \mathbf{g}_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = \mathbf{x}_{\text{out}} \tag{2.20}$$

and

$$\partial_{\mathbf{x}} \mathbf{D}_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) = \partial_{\mathbf{x}} \mathbf{g}_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) + \frac{(\mathbf{f}_{\mathfrak{m}_{\text{out}}}(\mathbf{x}_{\text{out}}) - \partial_{\mathbf{x}} \mathbf{g}_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) \cdot \mathbf{f}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}_{\text{in}})) \cdot \partial_{\mathbf{x}} h_{\mathfrak{e}}(\mathbf{x}_{\text{in}})}{\partial_{\mathbf{x}} h_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) \cdot \mathbf{f}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}_{\text{in}})} \tag{2.21}$$

(cf. [1, 39]). In particular,

$$\partial_{\mathbf{x}} \mathbf{D}_{\mathfrak{e}}(\mathbf{x}_{\text{in}}) \cdot \mathbf{f}_{\mathfrak{m}_{\text{in}}}(\mathbf{x}_{\text{in}}) = \mathbf{f}_{\mathfrak{m}_{\text{out}}}(\mathbf{x}_{\text{out}}). \tag{2.22}$$

Now suppose that ξ denotes a solution sequence of length N with $T_j \neq 0$ for all $1 \leq j \leq N$ and an associated signature $\mathbf{\Sigma}$ with all transversal events. Then, for every $\mathbf{x} \approx \mathbf{x}_1(t_0)$, $\tilde{t}_0 \approx t_0$, and $\tilde{t}_N \approx t_N$, the above analysis establishes the existence of a sequence $\tilde{\xi} = \left\{\tilde{\mathbf{x}}_j : \left[\tilde{t}_{j-1}, \tilde{t}_j\right] \to \mathbb{X}\right\}_{j=1}^{N}$ of N smooth curves with $\tilde{\mathbf{x}}_1(\tilde{t}_0) = \mathbf{x}$, such that conditions (2.4)–(2.6) hold with $\mathbf{x}$ and t replaced by $\tilde{\mathbf{x}}$ and $\tilde{t}$, respectively. In particular, it follows that $\tilde{\mathbf{x}}_j(\tilde{t}_j) = \mathbf{\Phi}_{\mathfrak{m}_j}(\tilde{\mathbf{x}}_j(\tilde{t}_{j-1}), \tilde{t}_j - \tilde{t}_{j-1})$. Associate with ξ and the signature $\mathbf{\Sigma}$ the *hybrid flow*

$$\mathbf{\Phi}_{\xi,\mathbf{\Sigma}} \overset{\text{def}}{=} \mathbf{\Phi}_{\mathfrak{m}_N}(\cdot, T_N) \circ \mathbf{D}_{\mathfrak{e}_{N-1}} \circ \mathbf{\Phi}_{\mathfrak{m}_{N-1}}(\cdot, T_{N-1}) \circ \mathbf{D}_{\mathfrak{e}_{N-2}} \circ \cdots \circ \mathbf{\Phi}_{\mathfrak{m}_2}(\cdot, T_2) \circ \mathbf{D}_{\mathfrak{e}_1} \circ \mathbf{\Phi}_{\mathfrak{m}_1}(\cdot, T_1) \tag{2.23}$$

defined for $\mathbf{x} \approx \mathbf{x}_1(t_0)$ and such that

$$\tilde{\mathbf{x}}_N\left(\tilde{t}_0 + \sum_{i=1}^{N} T_i\right) = \mathbf{\Phi}_{\xi,\mathbf{\Sigma}}(\tilde{\mathbf{x}}_1(\tilde{t}_0)). \tag{2.24}$$

To linear order it follows that

$$\tilde{\mathbf{x}}_N\left(\tilde{t}_0+\sum_{i=1}^{N}T_i\right)-\mathbf{x}_N\left(t_0+\sum_{i=1}^{N}T_i\right)=\partial_{\mathbf{x}}\boldsymbol{\Phi}_{\xi,\boldsymbol{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)\cdot\left(\tilde{\mathbf{x}}_1\left(\tilde{t}_0\right)-\mathbf{x}_1\left(t_0\right)\right), \tag{2.25}$$

where

$$\partial_{\mathbf{x}}\boldsymbol{\Phi}_{\xi,\boldsymbol{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)=\partial_{\mathbf{x}}\boldsymbol{\Phi}_{\mathfrak{m}_N}\left(\mathbf{x}_N\left(t_{N-1}\right),T_N\right)\cdot\prod_{j=N-1}^{1}\partial_{\mathbf{x}}\mathbf{D}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j\right)\right)\cdot\partial_{\mathbf{x}}\boldsymbol{\Phi}_{\mathfrak{m}_j}\left(\mathbf{x}_j\left(t_{j-1}\right),T_j\right). \tag{2.26}$$

Equations (2.11) and (2.22) imply that

$$\partial_{\mathbf{x}}\boldsymbol{\Phi}_{\xi,\boldsymbol{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)\cdot\mathbf{f}_{\mathfrak{m}_1}\left(\mathbf{x}_1\left(t_0\right)\right)=\mathbf{f}_{\mathfrak{m}_N}\left(\mathbf{x}_N\left(t_N\right)\right), \tag{2.27}$$

i.e., that deviations in the initial condition along the initial vector field result in deviations of the terminal point along the final vector field.

The above discussion pertains to the a posteriori characterization of a sequence of curves and an associated signature as a solution to a hybrid dynamical system. The question of how to generate such a solution a priori requires a definition of the forward dynamics of a hybrid dynamical system. For this purpose, associate with each mode $\mathfrak{m}\in\mathfrak{M}$ an *event map* $\iota_{\mathfrak{m}}:\mathbb{X}\to\mathfrak{E}$, such that $\pi_{\iota_{\mathfrak{m}}(\mathbf{x})}=(\mathfrak{m},\cdot)$. Then, given a state vector $\mathbf{x}_j\left(t_{j-1}\right)$ and an associated mode $\mathfrak{m}_j$ at $t=t_{j-1}$, apply the flow $\boldsymbol{\Phi}_{\mathfrak{m}_j}$ until the earliest time $t=t_j$ that condition (2.5) is satisfied for some event function $h_{\mathfrak{e}}$, for which $\pi_{\mathfrak{e}}=(\mathfrak{m}_j,\cdot)$. Proceed to apply the event map $\iota_{\mathfrak{m}_j}$ to yield $\mathfrak{e}_j=\iota_{\mathfrak{m}_j}\left(\mathbf{x}_j\left(t_j\right)\right)$, $\mathbf{x}_{j+1}\left(t_j\right)=\mathbf{g}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j\right)\right)$, and $\mathfrak{m}_{j+1}$, where $\pi_{\mathfrak{e}_j}=(\mathfrak{m}_j,\mathfrak{m}_{j+1})$. Append the curve segment $\mathbf{x}_j\left(t\right)$ for $t\in[t_{j-1},t_j]$ and $\mathfrak{m}_j$ and $\mathfrak{e}_j$ to the solution sequence ξ and the signature $\boldsymbol{\Sigma}$, respectively. Repeat this construction as many times as desired. Degenerate situations may occur if two event functions are reached simultaneously, in which case priority must be given on an ad hoc, domain-specific basis.

It is clear that there may not exist a solution with an initial condition $\mathbf{x}_0$ and a prescribed signature. This, however, is of no concern to the above construction of the sequence $\tilde{\xi}$, since this presupposes an existing trajectory with signature $\boldsymbol{\Sigma}$ and transversal intersections of the corresponding event surfaces. On the other hand, it is certainly possible that trajectories may exist with a given signature that could not occur during forward simulation of the hybrid dynamical system. To ensure consistency between the a posteriori characterization of trajectories and the a priori generation of such trajectories using the transition function, care and domain-specific knowledge is required.

2.3. The hybrid Poincaré map. Suppose again that the event $\mathfrak{e}$ with $\pi_{\mathfrak{e}}=(\mathfrak{m}_{\text{in}},\mathfrak{m}_{\text{out}})$ is *transversal*, such that $h_{\mathfrak{e}}=0$ is a local Poincaré section for trajectory segments based at nearby initial conditions and governed by the $\mathbf{f}_{\mathfrak{m}_{\text{in}}}$ vector field. Then, for $\mathbf{x}\approx\mathbf{x}_{\text{in}}$, the *projection*

$$\mathbf{P}_{\mathfrak{e}}\left(\mathbf{x}\right)=\boldsymbol{\Phi}_{\mathfrak{m}_{\text{in}}}\left(\mathbf{x},\tau_{\mathfrak{e}}\left(\mathbf{x}\right)\right) \tag{2.28}$$

satisfies the conditions

$$h_{\mathfrak{e}}\left(\mathbf{P}_{\mathfrak{e}}\left(\mathbf{x}\right)\right)=0, \tag{2.29}$$

$$\mathbf{P}_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)=\mathbf{x}_{\text{in}}, \tag{2.30}$$

and

$$\partial_{\mathbf{x}}\mathbf{P}_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)=Id-\frac{\mathbf{f}_{\mathfrak{m}_{\text{in}}}\left(\mathbf{x}_{\text{in}}\right)\cdot\partial_{\mathbf{x}}h_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)}{\partial_{\mathbf{x}}h_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)\cdot\mathbf{f}_{\mathfrak{m}_{\text{in}}}\left(\mathbf{x}_{\text{in}}\right)}. \tag{2.31}$$

In particular,

$$\partial_{\mathbf{x}}\mathbf{P}_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)\cdot\mathbf{f}_{\mathfrak{m}_{\text{in}}}\left(\mathbf{x}_{\text{in}}\right)=\mathbf{0}. \tag{2.32}$$

Consider again the sequence $\xi=\{\mathbf{x}_j:[t_{j-1},t_j]\to\mathbb{X}\}_{j=1}^{N}$ of N smooth curves with associated signature $\mathbf{\Sigma}$ and all transversal events. Associate with ξ and $\mathbf{\Sigma}$ the *hybrid Poincaré map*

$$\mathbf{P}_{\xi,\mathbf{\Sigma}}\overset{\text{def}}{=}\mathbf{g}_{\mathfrak{e}_{N-1}}\circ\mathbf{P}_{\mathfrak{e}_{N-1}}\circ\mathbf{\Phi}_{\mathfrak{m}_{N-1}}\left(\cdot,T_{N-1}\right)\circ\mathbf{g}_{\mathfrak{e}_{N-2}}\circ\mathbf{P}_{\mathfrak{e}_{N-2}}\circ\mathbf{\Phi}_{\mathfrak{m}_{N-2}}\left(\cdot,T_{N-2}\right)\circ\cdots\circ\mathbf{g}_{\mathfrak{e}_1}\circ\mathbf{P}_{\mathfrak{e}_1}\circ\mathbf{\Phi}_{\mathfrak{m}_1}\left(\cdot,T_1\right) \tag{2.33}$$

defined for $\mathbf{x}\approx\mathbf{x}_1\left(t_0\right)$ and such that

$$\tilde{\mathbf{x}}_N\left(\tilde{t}_{N-1}\right)=\mathbf{P}_{\xi,\mathbf{\Sigma}}\left(\tilde{\mathbf{x}}_1\left(\tilde{t}_0\right)\right) \tag{2.34}$$

for a nearby sequence $\tilde{\xi}=\left\{\tilde{\mathbf{x}}_j:\left[\tilde{t}_{j-1},\tilde{t}_j\right]\to\mathbb{X}\right\}_{j=1}^{N}$ of N smooth curves based at the point $\tilde{\mathbf{x}}_1\left(\tilde{t}_0\right)\approx\mathbf{x}_1\left(t_0\right)$. To linear order it follows that

$$\tilde{\mathbf{x}}_N\left(\tilde{t}_{N-1}\right)-\mathbf{x}_N\left(t_{N-1}\right)=\partial_{\mathbf{x}}\mathbf{P}_{\xi,\mathbf{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)\cdot\left(\tilde{\mathbf{x}}_1\left(\tilde{t}_0\right)-\mathbf{x}_1\left(t_0\right)\right),$$

where

$$\partial_{\mathbf{x}}\mathbf{P}_{\xi,\mathbf{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)=\prod_{j=N-1}^{1}\partial_{\mathbf{x}}\mathbf{g}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j\right)\right)\cdot\partial_{\mathbf{x}}\mathbf{P}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j\right)\right)\cdot\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}_j}\left(\mathbf{x}_j\left(t_{j-1}\right),T_j\right). \tag{2.35}$$

Equations (2.11) and (2.32) imply that

$$\partial_{\mathbf{x}}\mathbf{P}_{\xi,\mathbf{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)\cdot\mathbf{f}_{\mathfrak{m}_1}\left(\mathbf{x}_1\left(t_0\right)\right)=\mathbf{0}, \tag{2.36}$$

i.e., that deviations in the initial condition along the initial vector field result in zero deviations of the outgoing state subsequent to the last event.

2.4. Periodic trajectories. A trajectory ξ of a hybrid dynamical system on a finite interval $[t_0,t_0+T]$ with signature $\mathbf{\Sigma}$ of length $2N-1$ will be said to be *periodic* with period T if $\mathbf{x}_1\left(t_0\right)=\mathbf{x}_N\left(t_0+T\right)$ and $\mathfrak{m}_N=\mathfrak{m}_1$. The same need to check for consistency with forward simulation as described in the general case naturally applies here. In particular, it is necessary to establish that the terminal state on the Nth segment does not trigger a nontrivial event.

For a periodic trajectory, the Jacobian $\partial_{\mathbf{x}}\mathbf{\Phi}_{\xi,\mathbf{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)$ is known as the *monodromy matrix* and describes the local stability properties of the trajectory. In particular, its eigenvalues are the celebrated *Floquet multipliers*, all of which (excluding the trivial eigenvalue at 1 with eigenvector equal to the initial vector field; cf. (2.27)) must lie within the unit circle in the complex plane to guarantee asymptotic orbital stability.

From the general statement

$$\mathbf{A}\cdot\mathbf{B}\cdot\mathbf{v}=\lambda\mathbf{v}\Rightarrow\mathbf{B}\cdot\mathbf{A}\cdot\mathbf{B}\cdot\mathbf{v}=\lambda\mathbf{B}\cdot\mathbf{v}, \tag{2.37}$$

it follows that if $\mathbf{v}$ is an eigenvector of $\partial_{\mathbf{x}}\mathbf{\Phi}_{\xi,\mathbf{\Sigma}}\left(\mathbf{x}_1\left(t_0\right)\right)$ with eigenvalue λ, then

$$\left(\prod_{j=N-1}^{1}\partial_{\mathbf{x}}\mathbf{D}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j\right)\right)\cdot\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}_j}\left(\mathbf{x}_j\left(t_{j-1}\right),T_j\right)\right)\cdot\mathbf{v} \tag{2.38}$$

is an eigenvector of the matrix

$$\left(\prod_{j=N-1}^{1}\partial_{\mathbf{x}}\mathbf{D}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j\right)\right)\cdot\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}_j}\left(\mathbf{x}_j\left(t_{j-1}\right),T_j\right)\right)\cdot\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}_1}\left(\mathbf{x}_N\left(t_{N-1}\right),T_N\right) \tag{2.39}$$

with the same eigenvalue. This suggests redefining a *hybrid* periodic trajectory with N segments as the collection of all cyclic permutations of the sequence $\xi_{\circlearrowleft}=\{\mathbf{x}_j:[t_{j-1},t_j]\to\mathbb{X}\}_{j=1}^{N-1}$ of $N-1$ smooth curves obtained by prepending the final segment to the initial segment and relabeling the time partition. Specifically, associate with $\xi_{\circlearrowleft}$ the sequence of modes $\{\mathfrak{m}_j\}_{j=1}^{N-1}$ and the sequence of transitions $\{\mathfrak{e}_j\}_{j=1}^{N-1}$ such that, in addition to the conditions prescribed above for a general trajectory,

$$\pi_{\mathfrak{e}_{N-1}}=\left(\mathfrak{m}_{N-1},\mathfrak{m}_1\right), \tag{2.40}$$

the $(N-1)$th segment terminates at an intersection with the event surface

$$\left\{\mathbf{x}\mid h_{\mathfrak{e}_{N-1}}\left(\mathbf{x}\right)=0,\,\partial_{\mathbf{x}}h_{\mathfrak{e}_{N-1}}\left(\mathbf{x}\right)\cdot\mathbf{f}_{\mathfrak{m}_{N-1}}\left(\mathbf{x}\right)\leq 0\right\},$$

i.e.,

$$h_{\mathfrak{e}_{N-1}}\left(\mathbf{x}_{N-1}\left(t_{N-1}\right)\right)=0, \tag{2.41}$$

and the connectivity between the $(N-1)$th and 1st segments is given by the function $\mathbf{g}_{\mathfrak{e}_{N-1}}$, i.e.,

$$\mathbf{g}_{\mathfrak{e}_{N-1}}\left(\mathbf{x}_{N-1}\left(t_{N-1}\right)\right)=\mathbf{x}_1\left(t_0\right). \tag{2.42}$$

In this case, the *cyclic signature* equals the collection of all cyclic permutations of the sequence $\mathbf{\Sigma}_{\circlearrowleft}=\{\mathbf{I}_j\}_{j=1}^{N-1}$ consisting of pairs $\mathbf{I}_j=(\mathfrak{m}_j,\mathfrak{e}_j)$ or, equivalently, of triplets $\mathbf{I}_j=\left(\mathbf{f}_{\mathfrak{m}_j},h_{\mathfrak{e}_j},\mathbf{g}_{\mathfrak{e}_j}\right)$.

From this construction, it follows that the Floquet multipliers are the eigenvalues of the matrix

$$\partial_{\mathbf{x}}\mathbf{\Phi}_{\xi_{\circlearrowleft},\mathbf{\Sigma}_{\circlearrowleft}}\overset{\text{def}}{=}\prod_{j=N-1}^{1}\partial_{\mathbf{x}}\mathbf{D}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j\right)\right)\cdot\partial_{\mathbf{x}}\mathbf{\Phi}_{\mathfrak{m}_j}\left(\mathbf{x}_j\left(t_{j-1}\right),T_j\right). \tag{2.43}$$

For any event $\mathfrak{e}$ with $\pi_{\mathfrak{e}}=(\mathfrak{m}_{\text{in}},\mathfrak{m}_{\text{out}})$ it is straightforward to show that

$$\left(\partial_{\mathbf{x}}\mathbf{D}_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)-\partial_{\mathbf{x}}\mathbf{g}_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)\cdot\partial_{\mathbf{x}}\mathbf{P}_{\mathfrak{e}}\left(\mathbf{x}_{\text{in}}\right)\right)\cdot\mathbf{v}\parallel\mathbf{f}_{\mathfrak{m}_{\text{out}}}\left(\mathbf{x}_{\text{out}}\right) \tag{2.44}$$

for all $\mathbf{v}$. Suppose, in particular, that $\mathbf{v}$ is an eigenvector of $\partial_{\mathbf{x}}\mathbf{\Phi}_{\xi_\circlearrowleft,\mathbf{\Sigma}_\circlearrowleft}$ with eigenvalue λ and transversal to $\mathbf{f}_{\mathfrak{m}_1}(\mathbf{x}_1(t_0))$. Then, there exists a scalar α such that

$$\partial_{\mathbf{x}}\mathbf{P}_{\xi_\circlearrowleft,\mathbf{\Sigma}_\circlearrowleft}(\mathbf{x}_1(t_0))\cdot\mathbf{v}=\partial_{\mathbf{x}}\mathbf{\Phi}_{\xi_\circlearrowleft,\mathbf{\Sigma}_\circlearrowleft}\cdot\mathbf{v}+\alpha\mathbf{f}_{\mathfrak{m}_1}(\mathbf{x}_1(t_0)), \tag{2.45}$$

and, consequently, $\lambda\mathbf{v}+\alpha\mathbf{f}_{\mathfrak{m}_1}(\mathbf{x}_1(t_0))$ is an eigenvector of $\partial_{\mathbf{x}}\mathbf{P}_{\xi_\circlearrowleft,\mathbf{\Sigma}_\circlearrowleft}(\mathbf{x}_1(t_0))$ with the same eigenvalue. In addition, from (2.36) it follows that $\mathbf{f}_{\mathfrak{m}_1}(\mathbf{x}_1(t_0))$ is an eigenvector of $\partial_{\mathbf{x}}\mathbf{P}_{\xi_\circlearrowleft,\mathbf{\Sigma}_\circlearrowleft}(\mathbf{x}_1(t_0))$ with eigenvalue 0. In the cyclic formulation, the Jacobian $\partial_{\mathbf{x}}\mathbf{P}_{\xi_\circlearrowleft,\mathbf{\Sigma}_\circlearrowleft}(\mathbf{x}_1(t_0))$ thus yields an equivalent description of the local stability properties of the trajectory.

2.5. An example oscillator. To illustrate the formalism introduced above, consider a mechanical system consisting of an oscillating mass m, termed the *impactor*, that is suspended within a frame and whose motion relative to the frame is harmonically excited by an external force (cf. Figure 2) with some known angular frequency ω and amplitude A. Here, the clearance between the frame and the impactor is designed so that collisions may occur with sufficient displacement of the impactor relative to the frame. Interactions between the frame and the impactor transmitted through the suspension are modeled with a combination of a linear elastic element with stiffness k and a linear dissipative element with damping coefficient c. Finally, collisions between the frame and the impactor are modeled as instantaneous impacts that, through the imposition of conservation of momentum and Newton's law of restitution, result in discontinuous-in-time changes in the impactor's velocity relative to the frame.

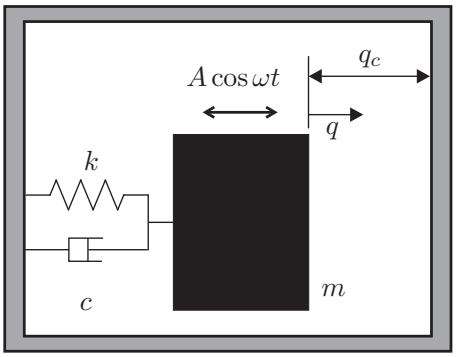


Figure 2. *The lateral motion of the impactor of mass m is harmonically excited with amplitude A and angular frequency ω. Instantaneous changes occur in the impactor's velocity $\dot{q}$ when q reaches q_c from below.*

The dynamics of the impactor may be formulated as a hybrid dynamical system in the following way. Denote by q the displacement of the impactor relative to the frame. The impactor motion is then governed by the linear differential equation

$$m\ddot{q}+c\dot{q}+kq=A\cos\omega t \tag{2.46}$$

as long as $q\leq q_c$ and such that if

$$\lim_{t\to t_c-}q(t)=q_c,\quad \lim_{t\to t_c-}\dot{q}(t)\geq 0 \tag{2.47}$$

for some time $t=t_c$, then

$$\lim_{t\to t_c+}q(t)=q_c,\quad \lim_{t\to t_c+}\dot{q}(t)=-e\lim_{t\to t_c-}\dot{q}(t), \tag{2.48}$$

where e is the coefficient of restitution. We omit from consideration situations in which the impactor remains in contact with the frame throughout a solution segment or solution signatures with infinite length.

Let

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} q \\ \dot{q} \\ \omega t \bmod 2\pi \end{pmatrix} \in \mathbb{R}^2 \times \mathbb{S}^1 \tag{2.49}$$

represent the state of the impactor. The smooth motion of the impactor is then governed by the vector field

$$\mathbf{f}_{\text{smooth}}(\mathbf{x}) = \begin{pmatrix} x_2 \\ \frac{1}{m}\left(A\cos x_3 - cx_2 - kx_1\right) \\ \omega \end{pmatrix} \tag{2.50}$$

and impacts between the impactor and the frame occur when

$$h_{\text{impact}}(\mathbf{x}) \stackrel{\text{def}}{=} q_c - x_1 = 0, \tag{2.51}$$

resulting in a discontinuous jump in state given by the state jump function

$$\mathbf{g}_{\text{impact}}(\mathbf{x}) = \begin{pmatrix} x_1 \\ -ex_2 \\ x_3 \end{pmatrix}. \tag{2.52}$$

Moreover, a discontinuous jump in the phase coordinate x_3 occurs when

$$h_{\text{phase}}(\mathbf{x}) \stackrel{\text{def}}{=} 2\pi - x_3 = 0 \tag{2.53}$$

and corresponds to the state jump function

$$\mathbf{g}_{\text{phase}}(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2 \\ x_3 - 2\pi \end{pmatrix}. \tag{2.54}$$

Finally, for purposes of detection of grazing events with the event surface corresponding to h_{impact}, consider the event function

$$h_{\text{turning}}(\mathbf{x}) \stackrel{\text{def}}{=} x_2 \tag{2.55}$$

and the associated state jump function

$$\mathbf{g}_{\text{identity}}(\mathbf{x}) = \mathbf{x} \tag{2.56}$$

(cf. Figure 3).

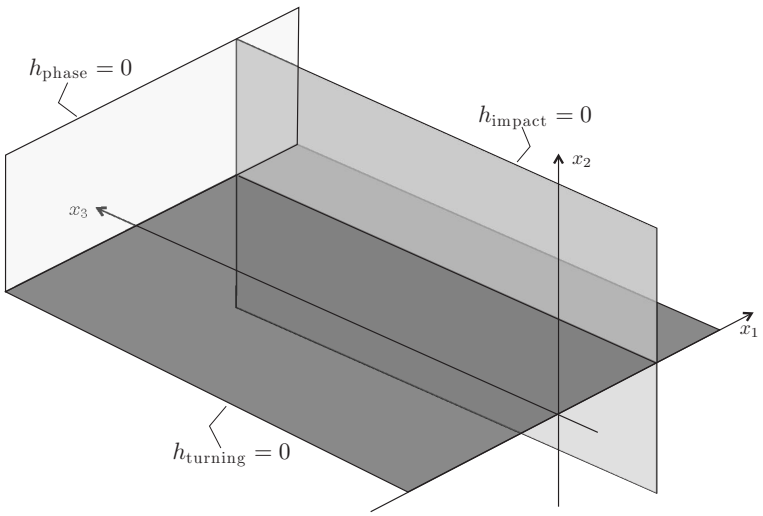


Figure 3. *A state-space schematic of the event surfaces describing the dynamics of the linear impact oscillator.*

Let $\mathfrak{M} = \{\mathfrak{smooth}\}$ and $\mathfrak{E} = \{\mathfrak{impact}, \mathfrak{phase}, \mathfrak{turning}\}$, such that $\mathbf{f}_{\mathfrak{smooth}} = \mathbf{f}_{\text{smooth}}$,

$$\mathfrak{impact} \doteq \begin{bmatrix} (\mathfrak{smooth}, \mathfrak{smooth}) \\ h_{\text{impact}} \\ \mathbf{g}_{\text{impact}} \end{bmatrix}, \tag{2.57}$$

$$\mathfrak{phase} \doteq \begin{bmatrix} (\mathfrak{smooth}, \mathfrak{smooth}) \\ h_{\text{phase}} \\ \mathbf{g}_{\text{phase}} \end{bmatrix}, \tag{2.58}$$

and

$$\mathfrak{turning} \doteq \begin{bmatrix} (\mathfrak{smooth}, \mathfrak{smooth}) \\ h_{\text{turning}} \\ \mathbf{g}_{\text{identity}} \end{bmatrix}, \tag{2.59}$$

corresponding to the *event graph* in Figure 4 showing the relationship between a mode and a given vector field as well as the event function and jump function associated with a given event. For example, $\mathfrak{e} = \mathfrak{impact}$ connects a solution segment governed by the vector field $\mathbf{f}_{\text{smooth}}$ and terminating at $\mathbf{x}_{\text{in}}$ on the event surface corresponding to h_{impact}, to the next solution segment based at $\mathbf{x}_{\text{out}}$ and again governed by the vector field $\mathbf{f}_{\text{smooth}}$, such that $\mathbf{x}_{\text{out}} = \mathbf{g}_{\text{impact}}(\mathbf{x}_{\text{in}})$. In the case of periodic trajectories, the dynamics of the hybrid system is captured by three distinct values of the index vector, namely,

$$\mathbf{I}_1 = (\mathfrak{smooth}, \mathfrak{impact}), \tag{2.60}$$

$$\mathbf{I}_2 = (\mathfrak{smooth}, \mathfrak{phase}), \tag{2.61}$$

$$\mathbf{I}_3 = (\mathfrak{smooth}, \mathfrak{turning}). \tag{2.62}$$

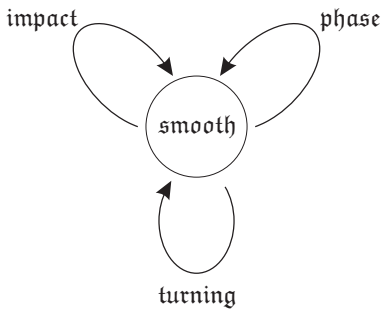


Figure 4. *Event graph corresponding to the linear impact oscillator.*

In particular, a solution will be termed *impacting* if its signature contains $\mathbf{I}_1$ and *nonimpacting* otherwise.

Consider, as an alternative, the state vector

$$\tilde{\mathbf{x}} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \end{pmatrix} \in \mathbb{R}^4 \tag{2.63}$$

and the corresponding smooth vector field

$$\tilde{\mathbf{f}}_{\text{smooth}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \tilde{x}_2 \\ \frac{1}{m}\left(A\tilde{x}_4 - c\tilde{x}_2 - k\tilde{x}_1\right) \\ \tilde{x}_3 + \omega\tilde{x}_4 - \tilde{x}_3\left(\tilde{x}_3^2 + \tilde{x}_4^2\right) \\ \tilde{x}_4 - \omega\tilde{x}_3 - \tilde{x}_4\left(\tilde{x}_3^2 + \tilde{x}_4^2\right) \end{pmatrix}, \tag{2.64}$$

event functions

$$\tilde{h}_{\text{impact}}(\tilde{\mathbf{x}}) = q_c - \tilde{x}_1, \tag{2.65}$$

$$\tilde{h}_{\text{phase}}(\tilde{\mathbf{x}}) = -\tilde{x}_3, \tag{2.66}$$

$$\tilde{h}_{\text{turning}}(\tilde{\mathbf{x}}) = \tilde{x}_2, \tag{2.67}$$

and state jump functions

$$\tilde{\mathbf{g}}_{\text{impact}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \tilde{x}_1 \\ -e\tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \end{pmatrix}, \tag{2.68}$$

$$\tilde{\mathbf{g}}_{\text{identity}}(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}. \tag{2.69}$$

It is straightforward to show that the surface $\tilde{x}_3^2 + \tilde{x}_4^2 - 1 = 0$ is invariant and globally attractive under the flow of the vector field $\tilde{\mathbf{f}}_{\text{smooth}}$. In particular, it can be easily seen that $\tilde{x}_3 = \sin(\omega t + \theta_0)$ and $\tilde{x}_4 = \cos(\omega t + \theta_0)$ on this surface.

Let $\mathfrak{M} = \{\mathfrak{smooth}\}$ and $\mathfrak{E} = \{\mathfrak{impact}, \mathfrak{phase}, \mathfrak{turning}\}$, such that $\tilde{\mathbf{f}}_{\mathfrak{smooth}} = \tilde{\mathbf{f}}_{\text{smooth}}$,

$$\mathfrak{impact} \doteq \begin{bmatrix} (\mathfrak{smooth}, \mathfrak{smooth}) \\ \tilde{h}_{\text{impact}} \\ \tilde{\mathbf{g}}_{\text{impact}} \end{bmatrix}, \tag{2.70}$$

$$\mathfrak{phase} \doteq \begin{bmatrix} (\mathfrak{smooth}, \mathfrak{smooth}) \\ \tilde{h}_{\text{phase}} \\ \tilde{\mathbf{g}}_{\text{identity}} \end{bmatrix}, \tag{2.71}$$

and

$$\mathfrak{turning} \doteq \begin{bmatrix} (\mathfrak{smooth}, \mathfrak{smooth}) \\ \tilde{h}_{\text{turning}} \\ \tilde{\mathbf{g}}_{\text{identity}} \end{bmatrix} \tag{2.72}$$

(cf. the event graph in Figure 4). In the case of periodic trajectories, the dynamics of the hybrid system is again captured by three distinct values of the index vector, namely,

$$\mathbf{I}_1 = (\mathfrak{smooth}, \mathfrak{impact}), \tag{2.73}$$

$$\mathbf{I}_2 = (\mathfrak{smooth}, \mathfrak{phase}), \tag{2.74}$$

$$\mathbf{I}_3 = (\mathfrak{smooth}, \mathfrak{turning}). \tag{2.75}$$

There is a one-to-one relationship between solutions to this hybrid dynamical system on $\tilde{x}_3^2 + \tilde{x}_4^2 - 1 = 0$ with $\theta_0 = 0$ and solutions to the original dynamical system. Solution segments for which $\tilde{x}_3^2 + \tilde{x}_4^2 - 1 \neq 0$, however, have no physical meaning. The introduction of the state variables $\tilde{x}_3$ and $\tilde{x}_4$ mimics the approach enabling the continuation of periodic trajectories of harmonically forced smooth dynamical systems in AUTO 97 effectively embedding $\mathbb{S}^1$ as a normally hyperbolic, attracting, invariant manifold of a smooth dynamical system in $\mathbb{R}^2$. Within the hybrid dynamical system formulation considered here, however, it is possible to consider an intrinsic parametrization in terms of the phase x_3, avoiding the need to artificially enlarge state space. This is particularly useful in the case of nonharmonic periodic or quasiperiodic forcing.

3. Boundary-value problems.

3.1. Periodic trajectories. The discussion below is focused on the task of finding a trajectory of a hybrid dynamical system with a prescribed signature satisfying the auxiliary boundary conditions

$$\mathbf{g}\left(\mathbf{x}_1\left(t_0\right), \mathbf{x}_N\left(t_N\right)\right) = \mathbf{0} \tag{3.1}$$

for some function $\mathbf{g}$ and any number of additional equations (typically generalized integral equations) that guarantee well-posedness.

As an example, in the case of a *periodic trajectory* of a hybrid dynamical system with a prescribed signature of length $2N - 1$, the auxiliary boundary condition corresponds to the connectivity condition

$$\mathbf{x}_1\left(t_0\right) - \mathbf{x}_N\left(t_N\right) = \mathbf{0}. \tag{3.2}$$

Together with an additional scalar condition, for example, the celebrated phase integral condition

$$\sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} \mathbf{x}\left(t\right)^T \cdot \dot{\mathbf{x}}^*\left(t\right) dt = 0 \tag{3.3}$$

for some reference trajectory $\mathbf{x}^*\left(t\right)$, this yields a well-posed formulation for locating such a periodic trajectory (away from singularities). No such auxiliary boundary or integral conditions are necessary in the cyclic formulation in section 2.4.

The boundary-value formulation imposes only conditions of equality, e.g., the termination of trajectory segments on the zero-level surfaces of the corresponding event functions. It does not, however, automatically ensure that the event function is locally decreasing along the trajectory segment at the termination point. Solutions to the boundary-value formulation must therefore be postprocessed to agree with this condition.

3.2. Grazing/sliding bifurcations. Suppose that, for some value of a free parameter, a trajectory has been found that achieves approximate *grazing contact* at a point $\mathbf{x}_*$ along the jth segment with an event surface corresponding to the event function $h_{\mathcal{D}}$, such that

$$h_{\mathcal{P}}\left(\mathbf{x}_*\right) = 0 \tag{3.4}$$

and (without loss of generality)

$$\partial_{\mathbf{x}} h_{\mathcal{P}}\left(\mathbf{x}_*\right) \cdot \mathbf{f}_{\mathbf{I}_j}\left(\mathbf{x}_*\right) \tag{3.5}$$

is distinctly negative, where

$$h_{\mathcal{P}}\left(\mathbf{x}\right) \stackrel{\text{def}}{=} \partial_{\mathbf{x}} h_{\mathcal{D}}\left(\mathbf{x}\right) \cdot \mathbf{f}_{\mathbf{I}_j}\left(\mathbf{x}\right). \tag{3.6}$$

Then replace the jth segment with mode $\mathfrak{m}_j$, and the jth event $\mathfrak{e}_j$ with two segments with modes $\mathfrak{m}_{j'} = \mathfrak{m}_{j''} = \mathfrak{m}_j$ and terminating events $\mathfrak{e}_{j'}$ and $\mathfrak{e}_{j''}$, such that $\pi_{\mathfrak{e}_{j'}} = (\mathfrak{m}_j, \mathfrak{m}_j)$, $\pi_{\mathfrak{e}_{j''}} = (\mathfrak{m}_j, \mathfrak{m}_{j+1})$, $h_{\mathfrak{e}_{j'}} = h_{\mathcal{P}}$, $h_{\mathfrak{e}_{j''}} = h_{\mathfrak{e}_j}$, $\mathbf{g}_{\mathfrak{e}_{j''}} = \mathbf{g}_{\mathfrak{e}_j}$, and $\mathbf{g}_{\mathfrak{e}_{j'}}$ is the identity. The auxiliary boundary condition

$$h_{\mathcal{D}}\left(\mathbf{x}_j\left(t_j\right)\right) = 0 \tag{3.7}$$

serves to locate the parameter value and the trajectory corresponding to actual grazing contact.

The imposition of an additional boundary condition may also be used to detect codimension-one bifurcations in Filippov systems in which a trajectory segment terminates on the boundary of the sliding region [14]. Here, the event surface, termed the *switching manifold*, corresponding to the event function $h_{\text{switching}}$ locally separates two distinct regions of state space with vector fields $\mathbf{f}_>$ when $h_{\text{switching}} > 0$ and $\mathbf{f}_<$ when $h_{\text{switching}} < 0$. The *sliding region* is the subset on the switching manifold where

$$\partial_{\mathbf{x}} h_{\text{switching}} \cdot \mathbf{f}_< - \partial_{\mathbf{x}} h_{\text{switching}} \cdot \mathbf{f}_> > 0 \tag{3.8}$$

and

$$-1 \leq h_{\text{sliding}} \overset{\text{def}}{=} -\frac{\partial_{\mathbf{x}} h_{\text{switching}} \cdot \mathbf{f}_{>} + \partial_{\mathbf{x}} h_{\text{switching}} \cdot \mathbf{f}_{<}}{\partial_{\mathbf{x}} h_{\text{switching}} \cdot \mathbf{f}_{<} - \partial_{\mathbf{x}} h_{\text{switching}} \cdot \mathbf{f}_{>}} \leq 1. \tag{3.9}$$

The parameter values corresponding to the termination of the jth trajectory segment on the boundary of the sliding region may then be obtained using either of the auxiliary boundary conditions

$$h_{\text{sliding}}\left(\mathbf{x}_j\left(t_j\right)\right) = 1 \tag{3.10}$$

and

$$h_{\text{sliding}}\left(\mathbf{x}_j\left(t_j\right)\right) = -1. \tag{3.11}$$

3.3. Limit points. *Limit points* in the single-parameter continuation of boundary-value problems occur when the component of the tangent vector to the corresponding solution branch in the direction of the bifurcation parameter changes sign. Such limit points may be located by appending to the existing equations, the equations for the tangent vector to the trajectory branch, and imposing the condition that this component vanish. This functionality is standard in AUTO 97.

Specifically, let $\mu(s)$ and $\xi(s)$ represent the bifurcation parameter and a one-parameter family of solutions of a hybrid system with identical signature, all transversal intersections, and $t_j(s) - t_{j-1}(s) \neq 0$ for $1 \leq j \leq N$, such that $\mu'(0) = 0$. In particular, for $j \in [1, N-1]$,

$$\partial_t \mathbf{x}_j(t,s) - \mathbf{f}_{\mathfrak{m}_j}\left(\mathbf{x}_j(t,s), \mu(s)\right) = \mathbf{0}$$

for $t \in [t_{j-1}(s), t_j(s)]$ and

$$h_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j(s), s\right), \mu(s)\right) = 0, \tag{3.12}$$

$$\mathbf{x}_{j+1}\left(t_j(s), s\right) - \mathbf{g}_{\mathfrak{e}_j}\left(\mathbf{x}_j\left(t_j(s), s\right), \mu(s)\right) = \mathbf{0}. \tag{3.13}$$

Differentiation with respect to s and letting $s = 0$ then yield

$$\partial_t \partial_s \mathbf{x}_j^{lp}(t) - \partial_{\mathbf{x}} \mathbf{f}_{\mathfrak{m}_j}\left(\mathbf{x}_j^{lp}(t), \mu^{lp}\right) \cdot \partial_s \mathbf{x}_j^{lp}(t) = \mathbf{0}$$

and

$$\partial_{\mathbf{x}} h_{\mathfrak{e}_j}\left(\mathbf{x}_j^{lp}\left(t_j^{lp}\right), \mu^{lp}\right) \cdot \left[\partial_t \mathbf{x}_j^{lp}\left(t_j^{lp}\right) t_j^{lp\prime} + \partial_s \mathbf{x}_j^{lp}\left(t_j^{lp}\right)\right] = 0, \tag{3.14}$$

$$\partial_t \mathbf{x}_{j+1}^{lp}\left(t_j^{lp}\right) t_j^{lp\prime} + \partial_s \mathbf{x}_{j+1}^{lp}\left(t_j^{lp}\right) - \partial_{\mathbf{x}} \mathbf{g}_{\mathfrak{e}_j}\left(\mathbf{x}_j^{lp}\left(t_j^{lp}\right), \mu^{lp}\right) \cdot \left[\partial_t \mathbf{x}_j^{lp}\left(t_j^{lp}\right) t_j^{lp\prime} + \partial_s \mathbf{x}_j^{lp}\left(t_j^{lp}\right)\right] = \mathbf{0}, \tag{3.15}$$

where the superscript ^{lp} refers to limit point values at $s = 0$. Solving the second equation for $t_j^{lp\prime}$ and substituting into the third equation then yields

$$\partial_s \mathbf{x}_{j+1}^{lp}\left(t_j^{lp}\right) - \partial_{\mathbf{x}} \mathbf{D}_{\mathfrak{e}_j}\left(\mathbf{x}_j^{lp}\left(t_j^{lp}\right), \mu^{lp}\right) \cdot \partial_s \mathbf{x}_j^{lp}\left(t_j^{lp}\right) = \mathbf{0}. \tag{3.16}$$

It follows that, on each segment, $\partial_s \mathbf{x}_j^{lp}(t)$ is a solution to the variational equation that connects to the solution on the subsequent segment by premultiplication with the Jacobian $\partial_{\mathbf{x}} \mathbf{D}_{\mathfrak{e}_j}\left(\mathbf{x}_j^{lp}\left(t_j^{lp}\right), \mu^{lp}\right)$ of the corresponding discontinuity mapping.

In the case of periodic trajectories of a hybrid dynamical system with a prescribed signature of length $2N-1$, limit points correspond to *saddle-node bifurcations*, for which $\partial_{\mathbf{x}} \mathbf{\Phi}_{\xi_\circlearrowleft, \mathbf{\Sigma}_\circlearrowleft}$ has an eigenvector with eigenvalue 1 distinct from the initial vector field. To locate a saddle-node bifurcation point, consider an augmented hybrid dynamical system with state space $\tilde{\mathbb{X}} = \mathbb{X} \times \mathbb{X}$, mode set $\tilde{\mathfrak{M}} = \mathfrak{M}$, and event set $\tilde{\mathfrak{E}} = \mathfrak{E}$. Let the associated vector fields $\tilde{\mathbf{f}}_{\mathfrak{m}} : \tilde{\mathbb{X}} \to \tilde{\mathbb{X}}$ be given by

$$\tilde{\mathbf{f}}_{\mathfrak{m}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \mathbf{f}_{\mathfrak{m}}(\mathbf{u}) \\ \mathbf{f}_{\mathfrak{m},\mathbf{x}}(\mathbf{u}) \cdot \mathbf{v} + \beta \mathbf{f}_{\mathfrak{m}}(\mathbf{u}) \end{pmatrix}, \quad \tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}, \tag{3.17}$$

where β is an auxiliary free parameter whose value is subsequently found to equal 0. Moreover, let $\tilde{h}_{\mathfrak{e}}(\tilde{\mathbf{x}}) = h_{\mathfrak{e}}(\mathbf{u})$ and

$$\tilde{\mathbf{g}}_{\mathfrak{e}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \mathbf{g}_{\mathfrak{e}}(\mathbf{u}) \\ \partial_{\mathbf{x}} \mathbf{D}_{\mathfrak{e}}(\mathbf{u}) \cdot \mathbf{v} \end{pmatrix}. \tag{3.18}$$

Together with the auxiliary boundary conditions

$$\mathbf{u}_1(t_0) - \mathbf{u}_m(t_m) = \mathbf{0}, \tag{3.19}$$

$$\mathbf{v}_1(t_0) - \mathbf{v}_m(t_m) = \mathbf{0} \tag{3.20}$$

and the integral conditions

$$\sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} \|\mathbf{v}_j(t)\|^2 \, dt = 1, \tag{3.21}$$

$$\sum_{j=1}^{N} \int_{t_{j-1}}^{t_j} \mathbf{v}_j(t)^T \cdot \mathbf{f}_{\mathfrak{m}_j}(\mathbf{u}_j(t)) \, dt = 0, \tag{3.22}$$

this yields a well-posed formulation for locating the periodic trajectory and the corresponding eigenvector of $\partial_{\mathbf{x}} \mathbf{\Phi}_{\xi_\circlearrowleft, \mathbf{\Sigma}_\circlearrowleft}$. Here, the second integral condition ensures that the eigenvector is distinct from the initial vector field.

As will be discussed further below, the formulation above mimics the implementation of $\widehat{\text{TC}}$ and exploits existing features of the boundary-value problem solver for smooth dynamical systems in AUTO 97. An alternative, and somewhat simpler, formulation is afforded using the cyclic formulation and imposing the condition that $\partial_{\mathbf{x}} \mathbf{P}_{\xi_\circlearrowleft, \mathbf{\Sigma}_\circlearrowleft}(\mathbf{x}_1(t_0))$ has an eigenvector with eigenvalue 1. For this purpose, consider again an augmented hybrid dynamical system with state space $\tilde{\mathbb{X}} = \mathbb{X} \times \mathbb{X}$, mode set $\tilde{\mathfrak{M}} = \mathfrak{M}$, and event set $\tilde{\mathfrak{E}} = \mathfrak{E}$. Let the associated vector fields $\tilde{\mathbf{f}}_{\mathfrak{m}} : \tilde{\mathbb{X}} \to \tilde{\mathbb{X}}$ be given by

$$\tilde{\mathbf{f}}_{\mathfrak{m}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \mathbf{f}_{\mathfrak{m}}(\mathbf{u}) \\ \mathbf{f}_{\mathfrak{m},\mathbf{x}}(\mathbf{u}) \cdot \mathbf{v} \end{pmatrix}, \quad \tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}. \tag{3.23}$$

Moreover, let $\tilde{h}_{\mathfrak{e}}(\tilde{\mathbf{x}}) = h_{\mathfrak{e}}(\mathbf{u})$ and

$$\tilde{\mathbf{g}}_{\mathfrak{e}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \mathbf{g}_{\mathfrak{e}}(\mathbf{u}) \\ \partial_{\mathbf{x}}\mathbf{g}_{\mathfrak{e}}(\mathbf{u}) \cdot \partial_{\mathbf{x}}\mathbf{P}_{\mathfrak{e}}(\mathbf{u}) \cdot \mathbf{v} \end{pmatrix}. \tag{3.24}$$

Then, the cyclic formulation and the integral condition

$$\sum_{j=1}^{N-1} \int_{t_{j-1}}^{t_j} \|\mathbf{v}_j(t)\|^2 \, dt = 1 \tag{3.25}$$

serve to locate the parameter values and the trajectory corresponding to the saddle-node bifurcation point. As $\mathbf{f}_{\mathfrak{m}_1}(\mathbf{x}_1(t_0))$ corresponds to an eigenvalue 0 of $\partial_{\mathbf{x}}\mathbf{P}_{\xi_{\circlearrowleft},\boldsymbol{\Sigma}_{\circlearrowleft}}(\mathbf{x}_1(t_0))$, there is no longer a need for the additional integral condition or the auxiliary parameter β.

3.4. Period-doubling bifurcations. Finally, consider the task of finding a periodic trajectory of a hybrid dynamical system with a prescribed signature in the presence of two free parameters, such that $\partial_{\mathbf{x}}\boldsymbol{\Phi}_{\xi_{\circlearrowleft},\boldsymbol{\Sigma}_{\circlearrowleft}}$ has an eigenvector with eigenvalue -1 corresponding to a *period-doubling bifurcation*. For this purpose, consider the augmented hybrid dynamical system with state space $\tilde{\mathbb{X}} = \mathbb{X} \times \mathbb{X}$, mode set $\tilde{\mathfrak{M}} = \mathfrak{M}$, event set $\tilde{\mathfrak{E}} = \mathfrak{E}$, and associated vector fields $\tilde{\mathbf{f}}_{\mathfrak{m}} : \tilde{\mathbb{X}} \to \tilde{\mathbb{X}}$, where

$$\tilde{\mathbf{f}}_{\mathfrak{m}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \mathbf{f}_{\mathfrak{m}}(\mathbf{u}) \\ \mathbf{f}_{\mathfrak{m},\mathbf{x}}(\mathbf{u}) \cdot \mathbf{v} \end{pmatrix}, \quad \tilde{\mathbf{x}} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}. \tag{3.26}$$

Moreover, let $\tilde{h}_{\mathfrak{e}}(\tilde{\mathbf{x}}) = h_{\mathfrak{e}}(\mathbf{u})$ and

$$\tilde{\mathbf{g}}_{\mathfrak{e}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \mathbf{g}_{\mathfrak{e}}(\mathbf{u}) \\ \partial_{\mathbf{x}}\mathbf{D}_{\mathfrak{e}}(\mathbf{u}) \cdot \mathbf{v} \end{pmatrix}. \tag{3.27}$$

Then, the auxiliary boundary conditions

$$\mathbf{u}_1(t_0) - \mathbf{u}_m(t_m) = \mathbf{0}, \tag{3.28}$$

$$\mathbf{v}_1(t_0) + \mathbf{v}_m(t_m) = \mathbf{0} \tag{3.29}$$

and the integral condition

$$\sum_{j=1}^{m} \int_{t_{j-1}}^{t_j} \|\mathbf{v}_j(t)\|^2 \, dt = 1 \tag{3.30}$$

serve to locate the parameter values and the trajectory corresponding to the period-doubling bifurcation point.

Alternatively, let

$$\tilde{\mathbf{g}}_{\mathfrak{e}}(\tilde{\mathbf{x}}) = \begin{pmatrix} \mathbf{g}_{\mathfrak{e}}(\mathbf{u}) \\ \varkappa\partial_{\mathbf{x}}\mathbf{g}_{\mathfrak{e}}(\mathbf{u}) \cdot \partial_{\mathbf{x}}\mathbf{P}_{\mathfrak{e}}(\mathbf{u}) \cdot \mathbf{v} \end{pmatrix}, \tag{3.31}$$

where $\varkappa = -1$ when $\mathfrak{e} = \mathfrak{e}_{N-1}$ and $\varkappa = 1$ otherwise. In this case, the cyclic formulation and the integral condition

$$\sum_{j=1}^{N-1} \int_{t_{j-1}}^{t_j} \|\mathbf{v}_j(t)\|^2 \, dt = 1 \tag{3.32}$$

serve to locate the parameter values and the trajectory, for which $\partial_{\mathbf{x}} \mathbf{P}_{\xi_\circlearrowleft, \mathbf{\Sigma}_\circlearrowleft}(\mathbf{x}_1(t_0))$ has an eigenvector with eigenvalue -1.

4. $\widehat{\text{TC}}$ ("TC-HAT").

4.1. Functionality. $\widehat{\text{TC}}$ ("TC-HAT") is a novel Fortran-based software application developed by the authors that encompasses a set of basic tools for the bifurcation analysis of periodic trajectories of hybrid dynamical systems. In this regard, $\widehat{\text{TC}}$ was inspired by, and closely resembles in general code implementation, the existing software application SLIDECONT [11], developed by Fabio Dercole and Yuri Kuznetsov for the study of hybrid dynamical systems with sliding dynamics but no state-space jumps. As detailed below, $\widehat{\text{TC}}$ replaces and improves on some of the functionality of SLIDECONT but suffers, in other respects, from the same computational limitations regarding problem discretization. $\widehat{\text{TC}}$ (just like SLIDECONT) functions as a driver to a modified version of AUTO 97 [16], a Fortran-based software application for the bifurcation analysis of smooth dynamical systems. In particular, $\widehat{\text{TC}}$ exploits AUTO 97's general boundary-value-problem formulation to locate and continue periodic trajectories of hybrid dynamical systems and a selected set of codimension-one bifurcation points under variations in system parameters.

In its current form, $\widehat{\text{TC}}$ can perform the following specific tasks:

1. Single-parameter continuation of multisegment periodic trajectories of a hybrid dynamical system with a given signature.
2. Two-parameter continuation of multisegment periodic trajectories of a hybrid dynamical system with a given signature and with grazing incidence at the terminal point of the first segment with some event surface.
3. Two-parameter continuation of multisegment periodic trajectories of a hybrid dynamical system with a given signature and with the terminal point of the first segment intersecting a switching manifold in a Filippov system on the boundary of the sliding region.
4. Two-parameter continuation of multisegment periodic trajectories of a hybrid dynamical system with a given signature and corresponding to saddle-node or period-doubling bifurcation points.

SLIDECONT is able only to partially perform the above tasks for periodic trajectories with at most three distinct segments. In particular, SLIDECONT is not able to handle nontrivial state jump functions or characterize the Lyapunov stability (Floquet multipliers) of periodic trajectories. In the standard implementation of AUTO 97, the Lyapunov stability properties of a periodic orbit are determined through a computation of the eigenvalues of the Jacobian $\partial_{\mathbf{x}} \mathbf{\Phi}(\mathbf{x}_0, T)$, where $\mathbf{x}_0$ is a point on the periodic orbit and T is the period. Although SLIDECONT relies on a multisegment formulation similar to that described for $\widehat{\text{TC}}$ (see below),

it does not account for the corrections to the flow Jacobian imposed by the piecewise nature of the solution trajectory as in (2.21). In contrast, $\widehat{\text{TC}}$ ships with a modified version of AUTO 97 that includes these corrections and, consequently, is able to accurately characterize the linearized stability properties of periodic trajectories of hybrid dynamical systems.

The ability to correctly compute the eigenvalues of $\partial_{\mathbf{x}}\mathbf{\Phi}_{\xi_\circlearrowleft,\mathbf{\Sigma}_\circlearrowleft}$ implies that $\widehat{\text{TC}}$ (with the help of AUTO 97) can detect bifurcations associated with the crossing of the unit circle of one or several eigenvalues, for example, saddle-node and period-doubling bifurcations. These form the starting points for the two-parameter continuations of task 4 which are implemented as new boundary-value formulations in the modified version of AUTO 97. Note that SLIDECONT is able to continue saddle-node bifurcations (for hybrid trajectories of up to three segments and without jumps) as per the generic formulation for limit points outlined above.

Similarly, $\widehat{\text{TC}}$ (and SLIDECONT for hybrid trajectories of up to three segments and without jumps) may be employed to detect the grazing contact of a trajectory segment with a properly identified event surface so as to enable subsequent two-parameter continuation as in task 2. In this case, it is typically necessary to cyclically reorder the trajectory segments in the grazing trajectory as well as to modify its signature so that grazing contact is imposed on the first trajectory segment and so that all termination points correspond to transversal intersections. In the continuation of periodic trajectories that graze an event surface corresponding to a mechanical-impact-like discontinuity, $\widehat{\text{TC}}$ is also able to detect and locate a selected set of characteristic codimension-two grazing bifurcations that serve as organizing centers for a variety of codimension-one bifurcation curves such as saddle-node, period-doubling, and grazing bifurcations (cf. Thota, Zhao, and Dankowicz [50]).

Finally, $\widehat{\text{TC}}$ (and SLIDECONT for hybrid trajectories of up to three segments and without jumps) may be employed to detect the crossing of the terminal point of a trajectory segment terminating on the switching manifold in a Filippov system with the boundary of the corresponding sliding region. Two-parameter continuation as in task 3 is then possible subsequent to reordering the trajectory segments so that the additional boundary condition is imposed on the first segment.

4.2. Discretization. Piecewise polynomial collocation methods provide an accurate and highly adaptive procedure for computing approximate solutions of boundary-value problems involving differential equations. In this method, approximants of the form of piecewise polynomials of some predetermined order are sought that satisfy the given differential equation at a discrete set of points in the interval of definition, the *collocation points*. The robustness of this method has made it an indisputable candidate in solving some of the difficult problems in differential equations [16, 22] (see also [3, 21]).

Following [16], denote by $\mathbf{x}(t)$ a solution on the interval $[0,1]$ of the differential equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}) \tag{4.1}$$

for some vector field $\mathbf{f}$. Introduce the partition

$$0 = t_0 < \cdots < t_{j-1} < \cdots < t_{j-i/m} = t_j - \frac{i}{m}\Delta_j < \cdots < t_j < \cdots < t_N = 1, \tag{4.2}$$

where

$$\Delta_j = t_j - t_{j-1}, \tag{4.3}$$

and suppose that the sequence $\mathbf{x}_0, \ldots, \mathbf{x}_{j-1}, \ldots, \mathbf{x}_{j-i/m}, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_N$ approximates the values of $\mathbf{x}(t)$ on this partition for some integers N and m, $j = 1, \ldots, N$, and $i = 0, \ldots, m$. On each interval $[t_{j-1}, t_j]$, define the Lagrange polynomials

$$l_{j,i}(t) = \prod_{k=0,\, k\neq i}^{m} \frac{t - t_{j-k/m}}{t_{j-i/m} - t_{j-k/m}}, \quad i = 0, \ldots, m \text{ and } j = 1, \ldots, N, \tag{4.4}$$

such that $l_{j,i}\left(t_{j-i/m}\right) = 1$ and $l_{j,i}\left(t_{j-k/m}\right) = 0$ for $k \neq i$. Then, the piecewise polynomial function $\mathbf{p}(t)$, such that

$$\mathbf{p}(t) = \sum_{i=0}^{m} l_{j,i}(t)\mathbf{x}_{j-i/m} \tag{4.5}$$

for $t \in [t_{j-1}, t_j]$, interpolates the values $\mathbf{x}_{j-i/m}$ for $j = 1, \ldots, N$ and $i = 0, \ldots, m$. Now, consider the mth order Legendre polynomial on the interval $[0, 1]$ and denote its roots by z_i, $i = 1, \ldots, m$. For each interval $[t_{j-1}, t_j]$, define $z_{j,i}$ as

$$z_{j,i} = t_{j-1} + z_i \Delta_j. \tag{4.6}$$

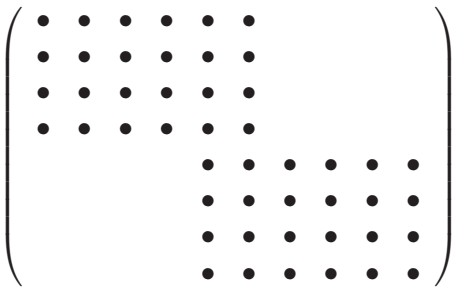
Then, an approximation to the solution to the original differential equation is obtained by seeking numerical values for the $mN + 1$ discrete approximants $\mathbf{x}_{j-i/m}$ for $j = 1, \ldots, N$, $i = 1, \ldots, m$, and $\mathbf{x}_N$ so that $\mathbf{p}$ satisfies the system of mN vector-valued equations

$$\mathbf{p}'(z_{j,i}) - \mathbf{f}(\mathbf{p}(z_{j,i})) = \mathbf{0} \tag{4.7}$$

for $j = 1, \ldots, N$ and $i = 1, \ldots, m$ and so that the associated boundary conditions are satisfied by $\mathbf{x}_0$ and $\mathbf{x}_N$.

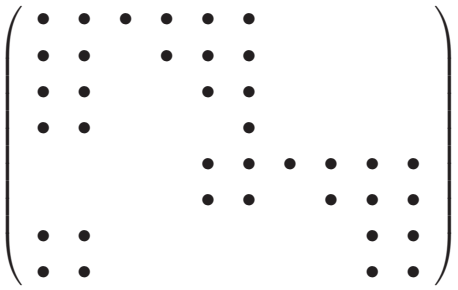
Suppose, for example, that $n = m = N = 2$. In this case, the linearization of (4.7) takes the form

(4.8)



where •'s denote nontrivial elements and zero elements have been omitted. Condensation of parameters and a subsequent step of nested dissection applied to an uncoupled subset of entries (see [16]) then yield a reduced matrix of the form

(4.9)



In particular, denote by P_0 and P_1 the 2×2 block matrices at the lower left and lower right, respectively, of this reduced matrix. It follows that, when evaluated on the converged solution,

$$P_0 \cdot \Delta \mathbf{x}_0 + P_1 \cdot \Delta \mathbf{x}_2 = 0, \tag{4.10}$$

i.e.,

$$\Delta \mathbf{x}_2 = -\left(P_1\right)^{-1} \cdot P_0 \cdot \Delta \mathbf{x}_0. \tag{4.11}$$

Hence, the matrix $-\left(P_1\right)^{-1} \cdot P_0$ gives the lowest-order relationship between perturbations in the initial point $\mathbf{x}_0$ and deviations in the corresponding terminal point $\mathbf{x}_2$.

For a multisegment trajectory of a hybrid dynamical system with solution sequence ξ and signature $\mathbf{\Sigma}$, the above formalism applies to each solution segment following the application of the time transformation

$$t \to t_{j-1} + tT_j \Rightarrow \mathbf{f}_{\mathfrak{m}_j} \to \frac{1}{T_j} \mathbf{f}_{\mathfrak{m}_j} \tag{4.12}$$

and the introduction of internal boundary conditions associated with the event sequence contained in $\mathbf{\Sigma}$. As described in previous sections, these couple the terminal point on the $(j-1)$st segment with the initial point on the jth segment and constrain the numerical values of the unknown times-of-flight T_j. As the coupling between distinct segments is imposed only on the boundary points, each segment may be treated independently from every other segment when formulating the piecewise polynomial approximant and the associated discretized differential equations. It follows that the sequence of matrix manipulations described previously can be applied for each segment independently of each other segment. Thus, the Jacobian $\partial_{\mathbf{x}} \mathbf{\Phi}_{\mathbf{I}_j} \left(\mathbf{x}_j \left(t_{j-1}\right), t_j - t_{j-1}\right)$ of the flow function that describes the sensitivity of the terminal point $\mathbf{x}_j \left(t_j\right)$ of the jth segment to changes in the initial point $\mathbf{x}_j \left(t_{j-1}\right)$ may again be obtained from the corresponding product $-\left(P_1\right)^{-1} \cdot P_0$.

A severely constrained implementation of the discretization scheme for a multisegment trajectory is afforded by the application of the same time partition to each solution segment. This fails to accommodate segment-specific error control and meshing algorithms, for example, in the case of solution segments with distinct curve characteristics. Instead, it corresponds to replacing the multipoint boundary-value problem with a regular two-point boundary-value problem for a single-segment trajectory of an augmented dynamical system with vector field

$$\mathbf{f} = \begin{pmatrix} \frac{1}{T_1} \mathbf{f}_{\mathfrak{m}_1} \\ \vdots \\ \frac{1}{T_N} \mathbf{f}_{\mathfrak{m}_N} \end{pmatrix} \tag{4.13}$$

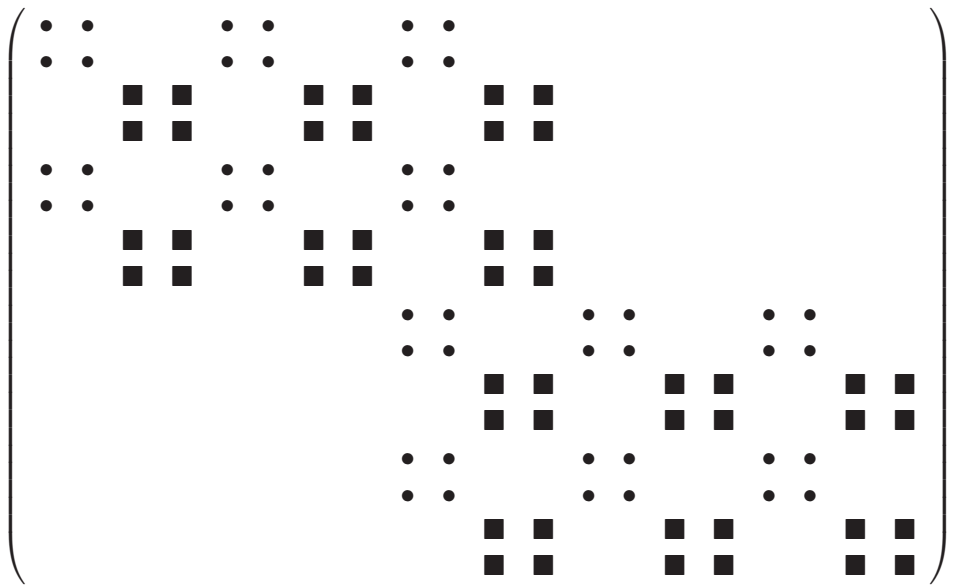
and solution

$$\mathbf{x}(t) = \begin{pmatrix} \mathbf{x}_1(t_0 + tT_1) \\ \vdots \\ \mathbf{x}_N(t_{N-1} + tT_N) \end{pmatrix}. \tag{4.14}$$

This also corresponds to the current implementation of $\widehat{\text{TC}}$ and SLIDECONT as these both rely on the two-point boundary-value solver contained in AUTO 97.

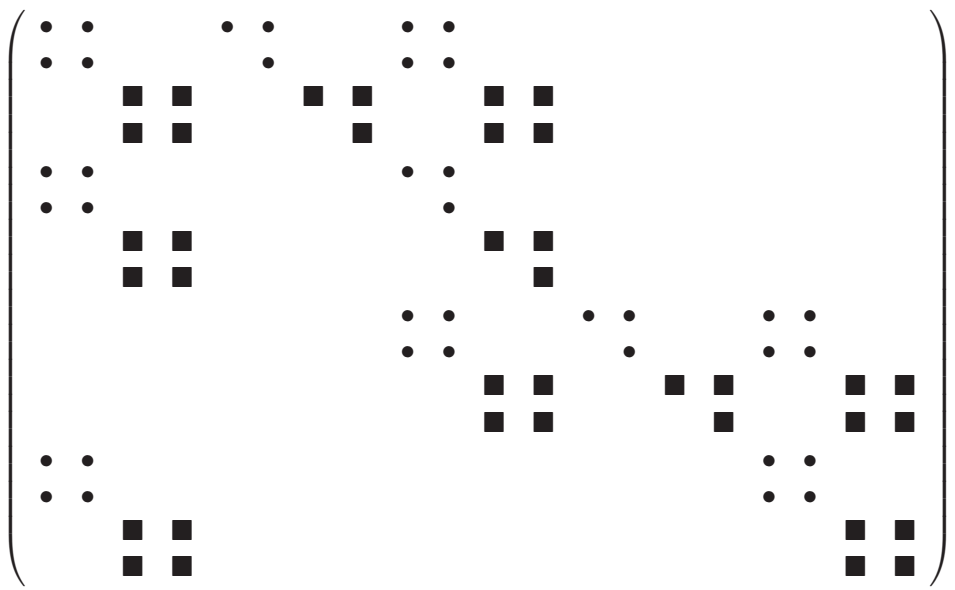
Suppose again that $n = m = N = 2$. Then, the linearization of (4.7) in the case of a two-segment trajectory takes the form

(4.15)



where •'s and ■'s denote nontrivial elements for the first and second segments, respectively, and zero elements have been omitted. Condensation of parameters and nested dissection as implemented in AUTO 97 then yields a reduced matrix of the form

(4.16)



In particular, denote by P_0 and P_1 the 4×4 block-diagonal matrices at the lower left and lower right, respectively, of this reduced matrix. It again follows that the matrix $-(P_1)^{-1}\cdot P_0$ gives the lowest-order relationship between perturbations in the initial point $\mathbf{x}_0$ and deviations in the corresponding terminal point $\mathbf{x}_2$ independently for each segment.

4.3. Usage. This section focuses on those aspects of $\widehat{\text{TC}}$ that distinguish it from AUTO 97. The section thus assumes some prior experience with the latter program. For further details on $\widehat{\text{TC}}$, see [49].

Essential information regarding the hybrid dynamical system is provided to $\widehat{\text{TC}}$ by the user in the <name>.f file.[1] This file contains the vector fields $\mathbf{f_I}$, the event functions $h_\mathbf{I}$, and the state jump functions $\mathbf{g_I}$ and, typically, their first (and, possibly, second) derivatives with respect to state variables and parameters. In addition, user-specific test functions may be included for monitoring during continuation. These may include event functions describing event surfaces with which the detection and two-parameter continuation of grazing incidence is desired or event functions describing the boundary of a sliding region on a switching manifold in a Filippov system for detection and continuation of selected sliding bifurcations.

Event functions corresponding to an identity state jump function may be introduced at liberty along a given trajectory of a hybrid dynamical system. This can be used to represent a trajectory by the value of some state variable on an event surface (cf. a Poincaré section) or to enforce a more accurate detection of grazing incidence (the latter necessarily occurring at a local extremum of the corresponding event function). In some instances, such as those where grazing occurs with an event surface at a point corresponding to a nontrivial jump in state space, an alternative signature may be intentionally employed to enable accurate detection and continuation of a grazing bifurcation curve (see below).

As with AUTO 97, continuation of periodic trajectories in $\widehat{\text{TC}}$ may be initialized with an approximate solution obtained by alternative means (for example, forward simulation) or from a previous run of $\widehat{\text{TC}}$. In contrast to the continuation of periodic trajectories in smooth systems, however, discontinuity-induced codimension-one bifurcations in hybrid dynamical systems are typically accompanied by a change in segment structure and length of signature. This necessitates an understanding of the trajectory branching associated with a given discontinuity-induced bifurcation and a subsequent reinitialization with a modified segment structure and signature. This may also be desirable prior to detecting a discontinuity-induced bifurcation as suggested above.

In the event that the initial solution trajectory is provided by the user, the time histories of the state variables for one complete time period of the periodic trajectory are contained in the <name>.dat file. As described in the previous section, each time interval $[t_{j-1}, t_j]$ is discretized by a segment-independent partition of the interval $[0, 1]$ (contained in the first column of the <name>.dat file) scaled to the length of the segment. The time history of the ith state variable along the jth segment is then contained in the $(1 + n\,(j-1) + i)$th column, where n is the state-space dimension of the hybrid dynamical system.

In the event that the initial solution trajectory is obtained from a previous continuation, the corresponding data is contained in a q.<name> file and labeled so as to enable further

[1] Here, <name> corresponds to the user-specified name of a file.

continuation. In the event that resegmentation is necessary, for example, when a discontinuity-induced bifurcation is associated with the birth of a branch of periodic trajectories with an additional segment, data in the q.<name> file may be extracted manually and modified externally as discussed in the previous paragraph.

To uniquely identify the continuation task to be undertaken, a number of numerical flags and continuation-specific parameter values must be provided by the user in the gc.<name> file. These include the state-space dimension of the hybrid dynamical system, the desired signature, the system parameters that will be allowed to vary during continuation (this list includes the segment times-of-flight), a segment-dependent list of event functions for which the detection of zero-crossings is desirable, and additional boundary conditions associated with continuation of grazing or sliding bifurcation curves.

5. Numerical examples.

5.1. A linear impact oscillator. To illustrate the methodology and functionality of $\widehat{\text{TC}}$ consider the first hybrid dynamical system discussed in section 2.5. Figure 5 shows a nonimpacting periodic trajectory of the hybrid dynamical system for $m = 1$, $c = 0.1$, $k = 1$, $q_c = 1$, $e = 0.8$, $\omega = 2.0$, and $A = 3.0$ with cyclic signature $\{\mathbf{I}_3, \mathbf{I}_2\}$. A segment of the corresponding branch of periodic trajectories under variations in ω with identical signature is shown in the left panel of Figure 6. Here, trajectories for which $\max x_1 > 1$ are inconsistent with the forward dynamics conditions and must be discarded. For $\omega = 1.998$ one finds a periodic trajectory that achieves grazing incidence with the event surface corresponding to h_{impact} at a point $\mathbf{x}^* = \begin{pmatrix} 1.0 & 0 & 3.075 \end{pmatrix}^T$. A two-parameter continuation of the corresponding grazing bifurcation curve under simultaneous variations in ω and A is shown in the right panel of Figure 6.

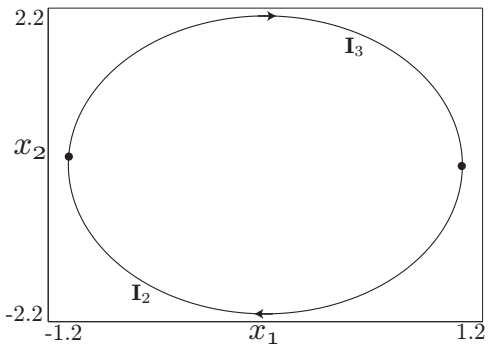


Figure 5. *A nonimpacting periodic trajectory of the linear oscillator for $\omega = 2.0$ and $A = 3.0$. Here and in the later figures, a dot on a periodic trajectory indicates the terminal point corresponding to a segment of that trajectory.*

Using the method of discontinuity mappings [8], it is possible to show that a unique branch of period-one impacting trajectories with a single impact per period emanates from the grazing curve under variations in A for all values of ω. To map out such a branch requires replacing the periodic trajectory with the equivalent trajectory with cyclic signature $\{\mathbf{I}_1, \mathbf{I}_2\}$. The result of such continuation for two distinct values of ω are shown in Figures 7 and 8. In each case, the

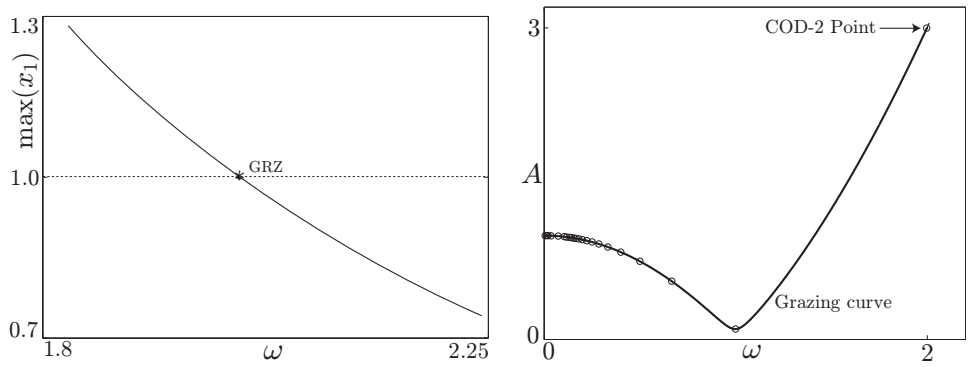


Figure 6. *Left panel: Diagram indicating the continuation of a nonimpacting periodic trajectory corresponding to the linear impact oscillator. Here,* $\widehat{\text{TC}}$ *detects the parameter value corresponding to a grazing incidence with the event surface* $h_{\text{impact}} = 0$ *that can be used as a starting solution to obtain a grazing curve. Right panel: Grazing curve in the* $(A$–$\omega)$ *space obtained using* $\widehat{\text{TC}}$*. The o's on the grazing curve correspond to the codimension-two bifurcation points, in particular to impact oscillators and detected by* $\widehat{\text{TC}}$*, that form organizing centers for a variety of codimension-one bifurcation curves* [50].

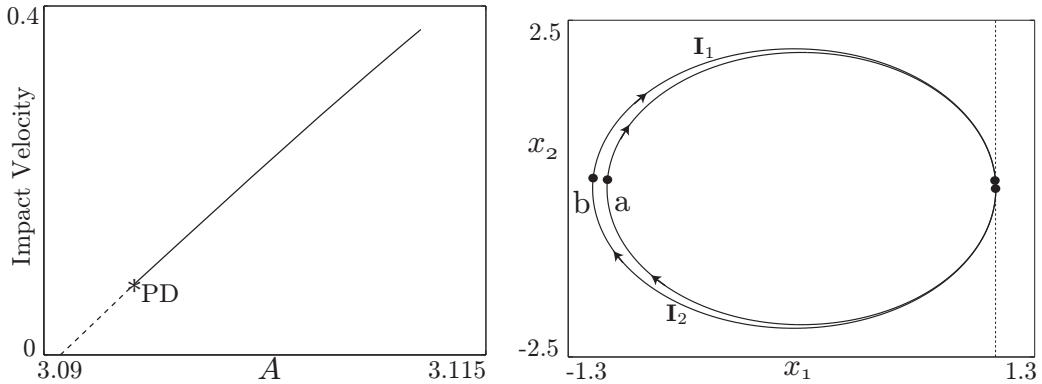


Figure 7. *Left panel: Bifurcation diagram indicating the continuation of an impacting periodic orbit as a function of* A *with a grazing periodic orbit as a starting solution for* $\omega = 2.0209$ *and* $A = 3.0909$*. This impacting periodic trajectory experiences a period-doubling bifurcation at* $A = 3.0951$ *resulting in a stable impacting trajectory. Right panel: Impacting periodic trajectories corresponding to the grazing incidence* (a) *and period-doubling bifurcation* (b) *points from the left panel.*

right panel shows two selected impacting periodic trajectories corresponding to the grazing and period-doubling bifurcation points and the grazing and saddle-node bifurcation points, respectively.

Figure 9 shows the results of two-parameter continuation of the saddle-node and period-doubling bifurcation curves under simultaneous variations in ω and A along with a segment of the grazing bifurcation curve. The former curves terminate at a point ($\omega = 1.9975$ and $A = 2.9966$) of quadratic contact with the grazing curve corresponding to a codimension-two grazing bifurcation point [23, 24, 30, 50].

In the case of the single-parameter continuation of impacting trajectories with cyclic signature $\{\mathbf{I}_1, \mathbf{I}_2\}$ from the grazing bifurcation curve considered above, the direction of continuation

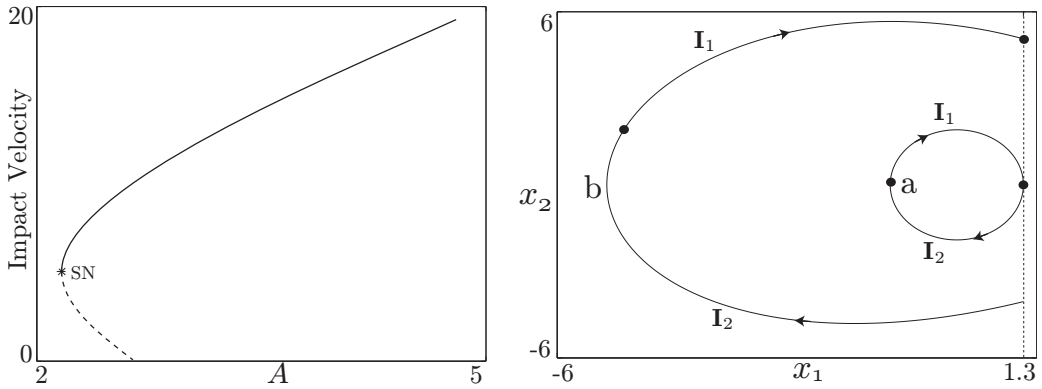


Figure 8. *Left panel: Bifurcation diagram indicating the continuation of an impacting periodic trajectory as a function of A with a grazing periodic trajectory as a starting solution for $\omega = 1.9086$ and $A = 2.6498$. This impacting periodic trajectory experiences a saddle-node bifurcation at $A = 2.1662$ resulting in a stable impacting trajectory. Right panel: Impacting periodic trajectories corresponding to the grazing incidence* (a) *and saddle-node bifurcation points* (b) *from the left panel.*

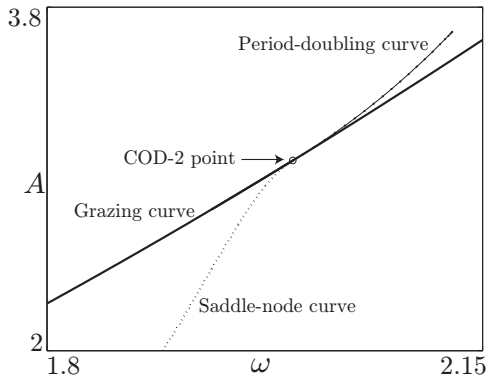


Figure 9. *Diagram depicting the two-parameter continuation of the grazing, saddle-node, and period-doubling bifurcation curves corresponding to impacting periodic trajectories. The saddle-node and period-doubling curves terminate at a joint tangential intersection with the grazing bifurcation curve.*

decides the validity of the solution trajectory obtained. As an example, continuation may result in convergence to a periodic trajectory similar to that shown in Figure 10. While this is a valid solution to the associated boundary-value problem, it is inconsistent with the requirement that event functions be locally decreasing along the corresponding trajectory segments at the corresponding termination points and should thus be discarded.

5.2. A Filippov system with impacts. As a second example, consider a mechanical system consisting of an oscillating mass m pressed against a rough horizontal surface with a corresponding maximum friction force of magnitude F_f [47]. Suppose that the motion of the mass relative to its environment is influenced by a combination of a linear elastic element with stiffness k and instantaneous impacts with coefficient of restitution e with a massive impactor with prescribed displacement time history $q_c(t)$ relative to the position of the mass corresponding to the unstretched length of the spring (cf. Figure 11).

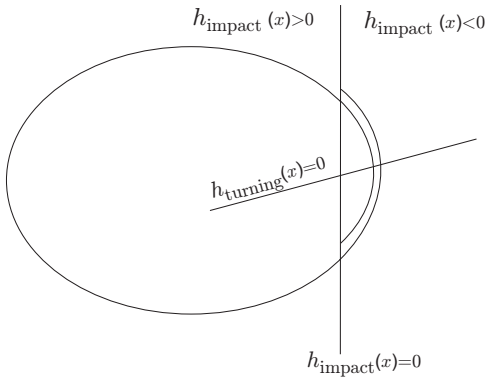


Figure 10. *Valid solution to the boundary value problem with no physical significance.*

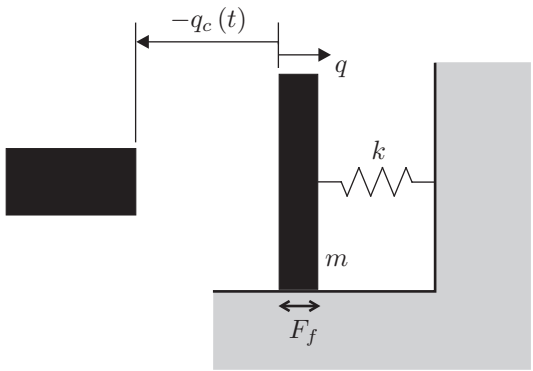


Figure 11. *The lateral motion of the oscillator of mass m is excited by impacts with a massive impactor with prescribed displacement time history $q_c(t)$.*

The dynamics of the oscillator may be formulated as a hybrid dynamical system in the following way. Denote by q the signed displacement of the oscillator relative to the position of unstretched length of the spring. The oscillator motion is then governed by the linear differential equation

$$m\ddot{q} + kq = F \tag{5.1}$$

as long as $q - q_c \geq 0$, where

$$F = \begin{cases} -F_f \frac{\dot{q}}{|\dot{q}|} & \text{for } \dot{q} \neq 0, \\ kq & \text{for } \dot{q} = 0 \text{ and } |kq| \leq F_f \end{cases} \tag{5.2}$$

and is otherwise defined so as to guarantee left-continuity with respect to time along a corresponding time history. Moreover, if

$$\lim_{t \to t_c -} q(t) = q_c(t_c), \quad \lim_{t \to t_c -} \dot{q}(t) \leq \dot{q}_c(t_c) \tag{5.3}$$

for some time $t = t_c$, then

$$\lim_{t \to t_c +} q(t) = q_c(t_c), \quad \lim_{t \to t_c +} \dot{q}(t) = -e \lim_{t \to t_c -} \dot{q}(t) + (1+e)\dot{q}_c(t_c), \tag{5.4}$$

where e is the coefficient of restitution. We again omit from consideration situations in which the oscillator remains in contact with the impactor throughout a solution segment or corresponds to signatures of infinite length.

Suppose that

$$q_c(t) = -b + a\sin\omega t \tag{5.5}$$

for $a, b > 0$, and let

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} q \\ \dot{q} \\ \omega t \ mod\ 2\pi \end{pmatrix} \in \mathbb{R}^2 \times \mathbb{S}^1 \tag{5.6}$$

represent the state of the oscillator. The smooth motion of the oscillator is then governed by the vector fields

$$\mathbf{f}_{\text{positive slip}}(\mathbf{x}) = \begin{pmatrix} x_2 \\ \frac{1}{m}(-F_f - kx_1) \\ \omega \end{pmatrix}, \tag{5.7}$$

$$\mathbf{f}_{\text{negative slip}}(\mathbf{x}) = \begin{pmatrix} x_2 \\ \frac{1}{m}(F_f - kx_1) \\ \omega \end{pmatrix}, \tag{5.8}$$

and

$$\mathbf{f}_{\text{stick}}(\mathbf{x}) = \begin{pmatrix} 0 \\ 0 \\ \omega \end{pmatrix}. \tag{5.9}$$

Impacts between the impactor and the frame occur when

$$h_{\text{impact}}(\mathbf{x}) \stackrel{\text{def}}{=} x_1 - q_c(x_3) = 0, \tag{5.10}$$

resulting in a discontinuous jump in state given by the state jump function

$$\mathbf{g}_{\text{impact}}(\mathbf{x}) = \begin{pmatrix} x_1 \\ -ex_2 + (1+e)\,\omega q_c'(x_3) \\ x_3 \end{pmatrix}. \tag{5.11}$$

Moreover, a discontinuous jump in the phase coordinate x_3 occurs when

$$h_{\text{phase}}(\mathbf{x}) \stackrel{\text{def}}{=} 2\pi - x_3 = 0 \tag{5.12}$$

and corresponds to the state jump function

$$\mathbf{g}_{\text{phase}}(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix}. \tag{5.13}$$

Discontinuous changes in the vector field are associated with intersections of system trajectories with the event surface

$$h_{\text{stick}\pm}(\mathbf{x}) \stackrel{\text{def}}{=} \pm x_2 = 0. \tag{5.14}$$

Finally, for purposes of detection of grazing events with the event surface corresponding to h_{impact}, consider the event function

$$h_{\text{turning}}(\mathbf{x}) \stackrel{\text{def}}{=} x_2 - \omega q_c'(x_3) \tag{5.15}$$

and the associated state jump function

$$\mathbf{g}_{\text{identity}}(\mathbf{x}) = \mathbf{x} \tag{5.16}$$

(cf. Figure 12).

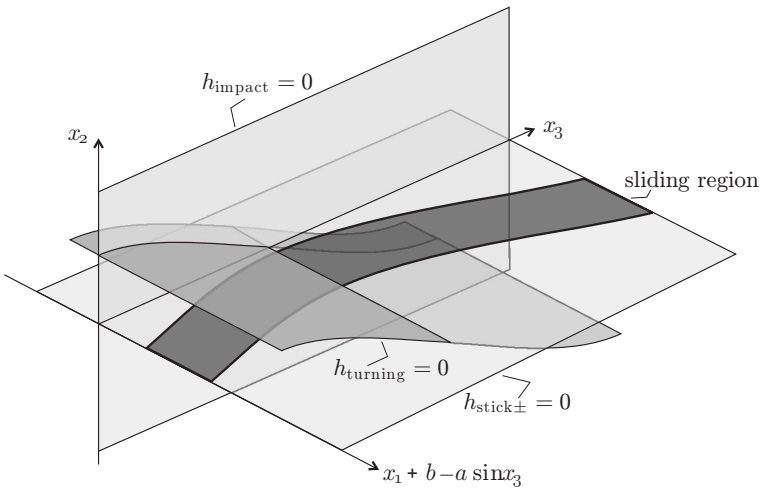


Figure 12. *A state-space schematic of the event surfaces describing the dynamics of the Filippov oscillator with impacts.*

A periodic trajectory of this hybrid dynamical system may be characterized in terms of a sequence of triplets of the form $(\mathbf{f}, h, \mathbf{g})$ corresponding to a solution segment governed by the vector field $\mathbf{f}$, terminating on the event surface corresponding to h, and connected to the next solution segment by the state jump function $\mathbf{g}$ as per the following list:

$$\mathbf{I}_1 = (\mathbf{f}_{\text{positive slip}}, h_{\text{impact}}, \mathbf{g}_{\text{impact}}), \tag{5.17}$$

$$\mathbf{I}_2 = (\mathbf{f}_{\text{positive slip}}, h_{\text{phase}}, \mathbf{g}_{\text{phase}}), \tag{5.18}$$

$$\mathbf{I}_3 = (\mathbf{f}_{\text{positive slip}}, h_{\text{stick}+}, \mathbf{g}_{\text{identity}}), \tag{5.19}$$

$$\mathbf{I}_4 = (\mathbf{f}_{\text{positive slip}}, h_{\text{turning}}, \mathbf{g}_{\text{identity}}), \tag{5.20}$$

$$\mathbf{I}_5 = (\mathbf{f}_{\text{negative slip}}, h_{\text{impact}}, \mathbf{g}_{\text{impact}}), \tag{5.21}$$

$$\mathbf{I}_6 = (\mathbf{f}_{\text{negative slip}}, h_{\text{phase}}, \mathbf{g}_{\text{phase}}), \tag{5.22}$$

$$\mathbf{I}_7 = (\mathbf{f}_{\text{negative slip}}, h_{\text{stick}-}, \mathbf{g}_{\text{identity}}), \tag{5.23}$$

$$\mathbf{I}_8 = (\mathbf{f}_{\text{negative slip}}, h_{\text{turning}}, \mathbf{g}_{\text{identity}}), \tag{5.24}$$

$$\mathbf{I}_9 = (\mathbf{f}_{\text{stick}}, h_{\text{impact}}, \mathbf{g}_{\text{impact}}), \tag{5.25}$$

$$\mathbf{I}_{10} = (\mathbf{f}_{\text{stick}}, h_{\text{phase}}, \mathbf{g}_{\text{phase}}), \tag{5.26}$$

$$\mathbf{I}_{11} = (\mathbf{f}_{\text{stick}}, h_{\text{turning}}, \mathbf{g}_{\text{identity}}) \tag{5.27}$$

(cf. Figure 13). In particular, a solution will be termed *impacting* if its signature contains $\mathbf{I}_1$, $\mathbf{I}_5$, and/or $\mathbf{I}_9$ and *nonimpacting* otherwise. Similarly, a solution will be termed *slipping* if its signature does not contain $\mathbf{I}_9$, $\mathbf{I}_{10}$, or $\mathbf{I}_{11}$, *sticking* if its signature contains only $\mathbf{I}_9$, $\mathbf{I}_{10}$, and/or $\mathbf{I}_{11}$, and a *stick-slip oscillation* otherwise.

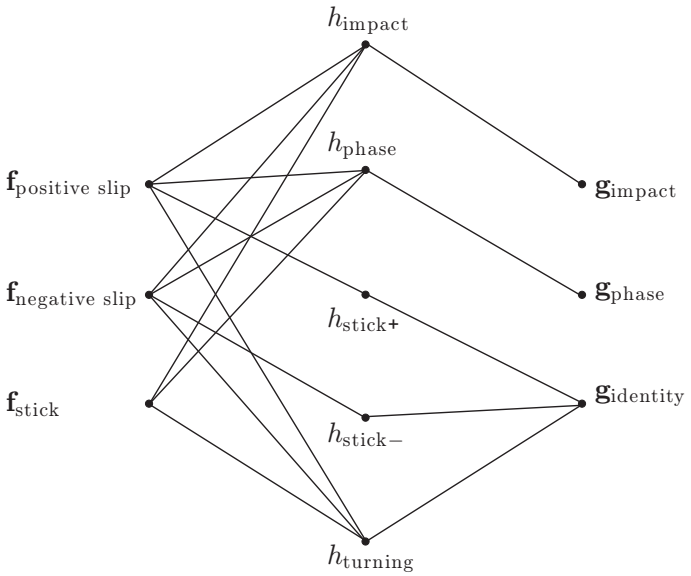


Figure 13. *Each branch of this graph based at an element in the leftmost column corresponds to the characterization of a solution segment of the Filippov system with impacts in terms of a triplet* $(\mathbf{f}, h, \mathbf{g})$.

This hybrid dynamical system is an example of a Filippov system with switching manifold given by the zero-level surface of $h_{\text{stick}+}$ (or, equivalently, $h_{\text{stick}-}$). Here, the sliding region is given by

$$-1 \leq h_{\text{sliding}}(\mathbf{x}) = \frac{kx_1}{F_f} \leq 1. \tag{5.28}$$

The vector field $\mathbf{f}_{\text{stick}}$ then corresponds to the *sliding vector field* (as per Utkin's equivalent control method [52]) along the sliding region.

The left panel of Figure 14 shows a periodic trajectory of the hybrid dynamical system for $m = 1$, $a = 1$, $\omega = 1$, $e = 0.9$, $k = 6$, $F_f = 0.7961$, and $b = 0.8471$ with cyclic signature $\{\mathbf{I}_9, \mathbf{I}_3, \mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}\}$. The corresponding branch of periodic trajectories under single-parameter variations in k is shown in the right panel of Figure 14. In particular, one end of the branch is seen to terminate at a point $k = 8.0544$, where the third segment of the periodic trajectory

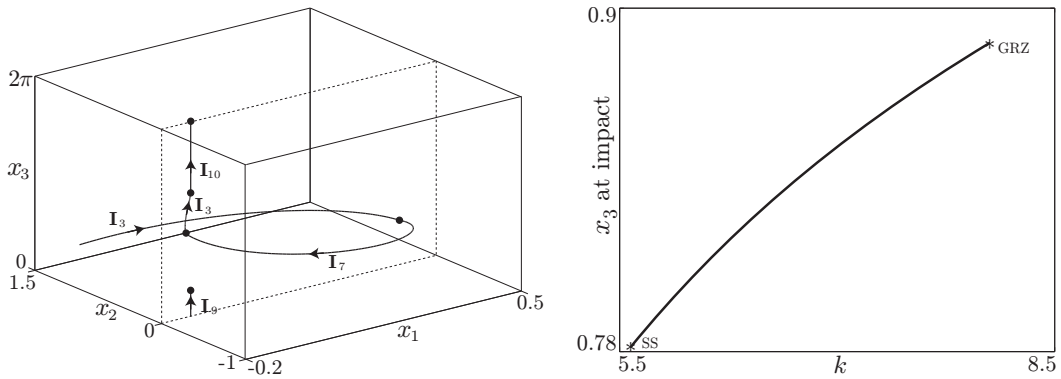


Figure 14. *Left panel: Periodic trajectory of the hybrid dynamical system for* $r = 0.9$, $k = 6$, $F_f = 0.7961$, *and* $b = 0.8471$. *Right panel: Single parameter continuation using the periodic trajectory in the left panel as an initial condition and varying* k.

achieves grazing incidence with the event surface corresponding to h_{impact}. Similarly, the other end of the branch terminates at $k = 5.5731$ as the third segment of the periodic trajectory terminates on the left boundary of the sliding region. The left and right panels of Figure 15 show the corresponding periodic trajectories.

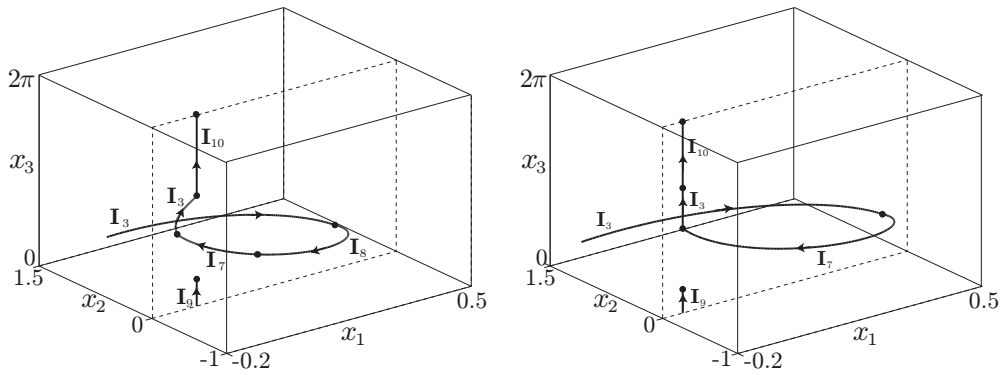


Figure 15. *Left panel: Periodic trajectory of the given hybrid dynamical system for* $k = 8.0544$ *that achieves grazing incidence with the event surface corresponding to* h_{impact}. *Right panel: Periodic trajectory of the given hybrid dynamical system for* $k = 5.5731$ *that intersects the switching manifold on the boundary of the sliding region.*

The dash-dotted curve in Figure 16 shows a two-parameter continuation of the grazing bifurcation curve under simultaneous variations in k and b based at an equivalent resegmentation with cyclic signature $\{\mathbf{I}_8, \mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ of the trajectory shown in the left panel of Figure 15. Resegmentation is here achieved by splitting the $\mathbf{I}_7$ segment into an $\mathbf{I}_8$ segment that terminates on the $h_{\text{turning}} = 0$ event surface followed by a truncated $\mathbf{I}_7$ segment and subsequently reordering the signature so that the additional boundary condition $h_{\text{impact}} = 0$ is applied to the terminal point of the first segment.

Two-parameter continuation of the sliding bifurcation curve under simultaneous variations in k and b based at the trajectory shown in the right panel of Figure 15 requires, at the very

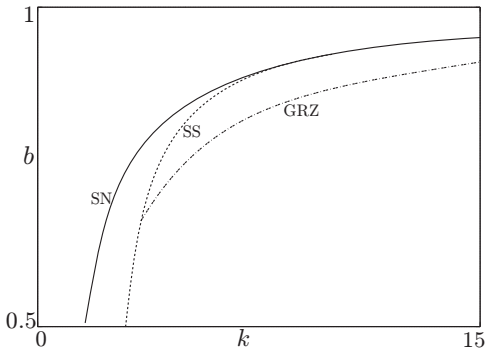


Figure 16. *Two parameter continuations of the grazing, sliding, and saddle-node bifurcation curves in the* (k, b) *parameter space.*

least, a reordering $\{\mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ of the cyclic signature so that the additional boundary condition $h_{\text{sliding}} = -1$ may be applied to the terminal point of the first segment. Some care is necessary, however, to ensure a nonsingular boundary-value problem. Specifically, in the limit as the $\mathbf{I}_7$ segment of the original periodic trajectory terminates on the left boundary of the sliding region, it follows from (5.7) and (5.9) that the subsequent $\mathbf{I}_3$ segment converges to a portion of the equilibrium trajectory $x_1 = -\frac{F_f}{k}$ and $x_2 = 0$ of the $\mathbf{f}_{\text{positive slip}}$ and $\mathbf{f}_{\text{stick}}$ vector fields, and terminates after a fixed time $\pi\sqrt{\frac{m}{k}}$ at a *nontransversal* event on $x_2 = 0$. In this limit, the cyclic signature $\{\mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ thus corresponds to a singular boundary-value problem and cannot be used for continuation of the sliding bifurcation.

A nonsingular boundary-value formulation may be obtained by replacing the terminal condition on the singular $\mathbf{I}_3$ segment with the condition that the segment's time-of-flight equal $\pi\sqrt{\frac{m}{k}}$. In this formulation, the correction matrix equation (2.21) corresponding to the singular $\mathbf{I}_3$ segment is given by the identity matrix. This permits two-parameter continuation of the sliding bifurcation curve under simultaneous variations in k and b (the dashed curve in Figure 16) and is consistent with subsequent one-parameter continuation under variations in k away from the sliding bifurcation curve using the original $\{\mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ boundary-value formulation.

Alternatively, a nonsingular boundary-value formulation may be obtained by replacing the terminal condition on the singular $\mathbf{I}_3$ segment with the condition that k be constant. This permits one-parameter continuation under variations in the segment's time-of-flight while retaining k among the unknowns of the boundary-value problem. For example, the singular $\mathbf{I}_3$ segment's time-of-flight may be reduced to zero while equivalently extending the subsequent $\mathbf{I}_{10}$ segment.

An alternative nonsingular boundary formulation may now be obtained by replacing the terminal condition on the singular $\mathbf{I}_3$ segment with the condition that the segment's time-of-flight equal 0. This again permits two-parameter continuation of the sliding bifurcation curve under simultaneous variations in k and b but is not consistent with subsequent one-parameter continuation away from the sliding bifurcation curve.

Finally, an equivalent resegmentation with cyclic signature $\{\mathbf{I}_7, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ may be obtained by replacing the consecutive $\mathbf{I}_3$ and $\mathbf{I}_{10}$ segments with a single extended $\mathbf{I}_{10}$ segment. This

is equivalent to eliminating the zero-length $\mathbf{I}_3$ segment obtained using the first alternative boundary-value formulation. This again permits two-parameter continuation of the sliding bifurcation curve under simultaneous variations in k and b and is consistent with subsequent one-parameter variation without the need for further resegmentation.

Indeed, a branch of periodic trajectories with cyclic signature $\{\mathbf{I}_7, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ can be shown to emanate from the sliding bifurcation point at $k = 5.5731$. A saddle-node bifurcation is found along this branch at $k = 4.9517$. The solid curve in Figure 16 shows the corresponding saddle-node bifurcation curve. It is curious to note that the grazing bifurcation curve appears to terminate on the sliding bifurcation curve at a point of transversal intersection, whereas the sliding bifurcation curve terminates on the saddle-node curve at a point of tangential contact. In the former case, the termination point is likely artificial, as a continuation with modified signature $\{\mathbf{I}_8, \mathbf{I}_7, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ should exist beyond this point. In contrast, the latter case is an example of a codimension-two sliding bifurcation (cf. [41]).

5.3. A heuristic cell-cycle model. As a final example, consider the biochemical reactions governing the activity of a set of biomolecules during distinct stages in the cell cycle of eukaryotes [51]. Here, the initiation of characteristic transitions during the cell cycle are reduced to variations in the concentrations of specific cyclin-enzyme dimers that catalyze the phosphorylation of key cell proteins. The activity of such cyclin/Cdk (cyclin-dependent kinase) dimers in the cell nucleus is controlled by a nonlinear feedback loop that regulates the availability of cyclin and various inhibitory agents. Specifically, the replication of DNA is triggered by increased levels of cyclin synthesis and reduced levels of cyclin degradation. Similarly, anaphase-promoting complexes (APCs) that enable the segregation of the chromatids during mitosis increase the level of cyclin degradation during telophase, returning the cell to its resting state.

Denote by x_1, x_2, and x_3 the concentrations of cyclin/Cdk dimers, active Cdh1/APCs, and the Cdh1/APC activator Cdc14, respectively. In this heuristic model, cyclin/Cdk dimers activate Cdc20/APCs while inhibiting Cdh1/APCs. Active Cdc20/APCs destroy an inhibitor of Cdc14, which in turn counters the inhibitory action of cyclin/Cdk on Cdh1/APCs. Finally, active Cdh1/APCs destroy cyclin. Denote by x_4 the cell mass, and suppose that during the cell's growth phase its rate of growth is independent of the concentrations of the biomolecules considered here. Cell division is assumed to be effectively instantaneous and triggered by the concentration of cyclin/Cdk dimers falling below a critical level x_1^*.

The dynamics of the cell-cycle model may be formulated as a hybrid dynamical system in the following way. Smooth variations in the state variables are governed by the vector field

$$\mathbf{f}_{\text{growth}}(\mathbf{x}) = \begin{pmatrix} k_1 - (k_2' + k_2'' x_2)\, x_1 \\ \frac{(k_3' + k_3'' x_3)(1 - x_2)}{J_3 + 1 - x_2} - \frac{k_4 x_1 x_2 x_4}{J_4 + x_2} \\ k_5' + k_5'' \frac{(x_1 x_4)^n}{J_5^n + (x_1 x_4)^n} - k_6 x_3 \\ \mu x_4 \left(1 - \frac{x_4}{x_4^*}\right) \end{pmatrix}, \tag{5.29}$$

where k_1, k_2', k_2'', k_3', k_3'', k_4, k_5', k_5'', and k_6 are rate constants, J_3, J_4, and J_5 are Michaelis–Menten constants (see [51]), n is some integer, μ is the linear growth rate in the limit of

$x_4 \ll x_4^*$, and x_4^* is the limiting cell mass in the absence of division events. Mitosis occurs when

$$h_{\text{mitosis}}(\mathbf{x}) = x_1 - x_1^*,$$

resulting in a discontinuous jump in state given by the state jump function

$$\mathbf{g}_{\text{mitosis}}(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4/2 \end{pmatrix}. \tag{5.30}$$

Let $\mathfrak{M} = \{\mathfrak{growth}\}$ and $\mathfrak{E} = \{\mathfrak{mitosis}\}$ such that $\mathbf{f}_{\mathfrak{growth}} = \mathbf{f}_{\text{growth}}$ and

$$\mathfrak{mitosis} \doteq \begin{bmatrix} (\mathfrak{growth}, \mathfrak{growth}) \\ h_{\text{mitosis}} \\ \mathbf{g}_{\text{mitosis}} \end{bmatrix} \tag{5.31}$$

and such that periodic trajectories have signatures consisting of one or several copies of the pair

$$\mathbf{I}_1 = (\mathfrak{growth}, \mathfrak{mitosis}). \tag{5.32}$$

The left panel of Figure 17 shows a periodic trajectory when $k_1 = 0.04$, $k_2' = 0.04$, $k_2'' = 1.0$, $k_3' = 1.0$, $k_3'' = 10.0$, $k_4 = 35.0$, $k_5' = 0.005$, $k_5'' = 0.2$, $k_6 = 0.1$, $J_3 = 0.04$, $J_4 = 0.04$, $J_5 = 0.3$, $n = 4$, $\mu = 0.01$, $m^* = 10.0$, and $x_1^* = 0.1$. A single-parameter continuation of this periodic trajectory with the cyclic signature $\{\mathbf{I}_1\}$ under variations in k_3'' is shown in the right panel of Figure 17. The corresponding branch of stable limit cycle terminates is limited by two saddle-node bifurcations at $k_3'' = 7.0213$ and $k_3'' = 45.1157$, respectively. A two-parameter continuation of the leftmost of these saddle-node bifurcations under simultaneous variations in k_3'' and k_5'' is shown in Figure 18.

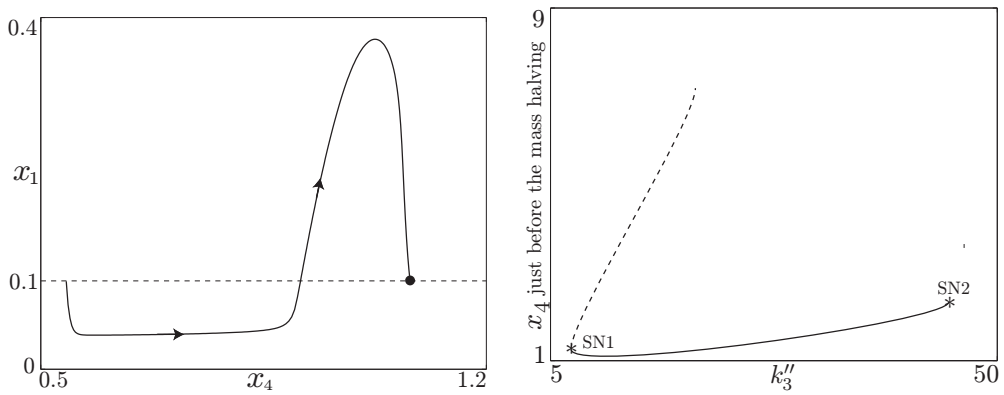


Figure 17. *Left panel: Periodic trajectory of the hybrid dynamical system for $k_3'' = 10$. Right panel: Single parameter continuation with varying k_3'' of the periodic trajectory shown in the left panel. Loci of saddle-node bifurcations SN1 and SN2 are evident in the figure.*

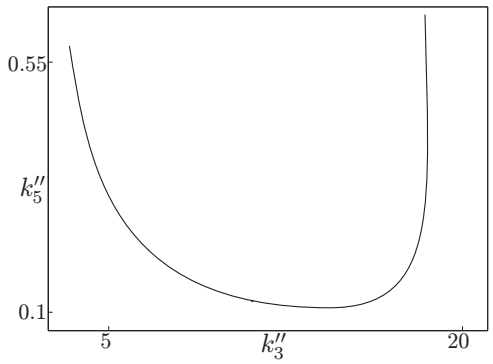


Figure 18. *Two-parameter continuation of the locus of the saddle-node bifurcation point, SN*1*, found in the single-parameter continuation shown in the right panel of Figure* 17*.*

6. Discussion. The preponderance of physical and biological systems that are modeled with hybrid dynamical systems makes the task of providing user-friendly computational tools for bifurcation analysis pressing. The existing package SLIDECONT was a first step in this direction with emphasis on Filippov systems in the plane and with the ability to provide a comprehensive bifurcation analysis of equilibria and selected bifurcation analysis of periodic trajectories. The boundary-value formulation introduced here extends SLIDECONT's formulation to enable successful continuation of multisegment periodic trajectories in general hybrid dynamical systems and selected associated codimension-one bifurcations. It should be straightforward to arrive at similar formulations in the case of other codimension-one bifurcations associated with the eigenvalue spectrum, for example, torus bifurcations.

It should also be possible to modify the more general boundary-value formulation to accommodate efforts to locate and continue multisegment homo- and heteroclinic trajectories between equilibria and/or limit cycles in hybrid dynamical systems (cf. [7, 15, 25, 46]). A related challenge is the task of locating and continuing periodic trajectories of a hybrid dynamical system consisting of a countable set of segments with (at least) one point of accumulation of the switching times t_j. As in the case of homo- or heteroclinic trajectories, this necessitates omitting from the boundary-value formulation an infinite set of equations for $|t|$ larger than some critical T in the homo- or heteroclinic case and for t in some finite interval in the case of finite-time accumulation. In the latter case, however, rather than relying on a linear approximation of the stable and unstable manifolds, it becomes necessary to provide a connectivity condition between the accumulation point and the final nonignored event [43].

A more challenging task is that of *branch switching* between branches of distinct signature, in which domain-specific knowledge (as captured by the index jump function) is coupled with a local unfolding of singularities in parameter space corresponding to the intersection of distinct branches. Instances of such switching were discussed in the examples above (see also [32]) and relied in all cases on an understanding of the types of signature changes that could be achieved in a manner consistent with the conditions of forward simulation and with a continuous change in the segment structure. This poses a particular challenge in the case of singularities associated with the onset of grazing contact in systems with mechanical impact-like discontinuities. Here, infinitely many branches of periodic trajectories may emanate

continuously from the singularity. Methods of unfolding, such as the discontinuity-mapping technique [42], could then be used to provide tangent directions to the corresponding trajectory branches at the singularity.

The implementation of $\widehat{\text{TC}}$ as a driver for a modified version of AUTO 97 enables a range of automated continuation tasks without the need to recreate the necessary source code infrastructure. As AUTO 97 is designed to handle only two-point boundary-value problems, however, this necessitates a numerically ill-advised segment-independent time partition that is unable to adapt the partition of each segment to local properties. An obvious implication is the sudden growth of problem dimension that occurs, for example, with the addition of a solution segment with zero time-of-flight. As AUTO 97 implements direct matrix-equation solvers, this growth in dimension affects both CPU time as well as solution accuracy. There is also no direct attempt made to exploit the block structure of the linearized equations. More generally, the partition scheme will likely be dominated by features on the geometrically most complex segment, requiring unnecessarily fine discretization for other segments. These shortcomings are likely to become pronounced for solutions with long signatures. Collectively, these observations warrant a redesign of $\widehat{\text{TC}}$ that accommodates segment-specific meshing algorithms and that exploits the (sparse) structure of the linearized equations.

An exhaustive quantitative comparison of the numerical performance of the implementation of $\widehat{\text{TC}}$ within AUTO 97 is beyond the scope of this manuscript. A preliminary demonstration of the effect of the boundary-value formulation is afforded, e.g., by investigating the accuracy with which the switching-sliding bifurcation point is detected under single-parameter variations of k in the Filippov system with impacts discussed in section 5.2 using the different equivalent segmentations of the critical periodic trajectory. It follows directly from explicit expressions for the flows corresponding to the distinct vector fields $\mathbf{f}_{\text{stick}}$, $\mathbf{f}_{\text{positive slip}}$, and $\mathbf{f}_{\text{negative slip}}$ that periodic trajectories with cyclic signature $\{\mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ correspond to fixed points $\tilde{x}^*$ on the interval $[-1, 1]$ of the map

$$\tilde{x} \longmapsto \sqrt{(1+\tilde{x})^2 + \Omega^2\left(\tilde{a}^2 - \left(\tilde{b}+\tilde{x}\right)^2\right)} - 5, \tag{6.1}$$

where

$$\tilde{x} = \frac{kq_0}{F_f}, \quad \tilde{a} = \frac{ka}{F_f}, \quad \tilde{b} = \frac{kb}{F_f}, \quad \Omega = (1+e)\,\omega\sqrt{\frac{m}{k}}, \tag{6.2}$$

and q_0 is the initial position of the oscillator at the beginning of the $\mathbf{I}_9$ segment. In particular,

$$\tilde{x}^* = \frac{-4 - \tilde{b}\Omega^2 + \sqrt{16 - 24\Omega^2 + 8\tilde{b}\Omega^2 + \tilde{a}^2\Omega^4}}{\Omega^2} \tag{6.3}$$

corresponds to the branch shown in the right panel of Figure 14. The switching-sliding bifurcation then corresponds to

$$\Omega^2 = \frac{16}{\tilde{a}^2 - \left(\tilde{b}-1\right)^2}, \tag{6.4}$$

and the corresponding nonzero Floquet multiplier is given by

$$\frac{4\left(1-\tilde{b}\right)}{\tilde{a}^2-\left(\tilde{b}-1\right)^2}. \tag{6.5}$$

Similarly, periodic trajectories with cyclic signature $\{\mathbf{I}_7, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ correspond to fixed points $\tilde{x}^*$ on the interval $[-1, 1]$ of the map

$$\tilde{x} \longmapsto 3-\sqrt{(1+\tilde{x})^2+\Omega^2\left(\tilde{a}^2-\left(\tilde{b}+\tilde{x}\right)^2\right)}. \tag{6.6}$$

In particular,

$$\tilde{x}^* = \frac{4-\tilde{b}\Omega^2-\sqrt{16-8\Omega^2-8\tilde{b}\Omega^2+\tilde{a}^2\Omega^4}}{\Omega^2} \tag{6.7}$$

corresponds to the branch of four-segment trajectories emanating from the switching-sliding bifurcation curve and the corresponding nonzero Floquet multiplier is now given by

$$\frac{4\left(\tilde{b}-1\right)}{\tilde{a}^2-\left(\tilde{b}-1\right)^2}. \tag{6.8}$$

Table 1 shows the absolute difference between estimated values (k_5, λ_5) and (k_4, λ_4), respectively, of the value of k and the nonzero Floquet multiplier λ at the switching-sliding bifurcation point detected using $\widehat{\text{TC}}$ during single-parameter variation of periodic trajectories with cyclic signatures $\{\mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ and $\{\mathbf{I}_7, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ and those predicted from (6.4)–(6.5). Here, the AUTO 97 constants *NTST*, *NCOL*, *EPSL*, *EPSU*, and *EPSS* refer to the number of discretization intervals per segment, the number of collocation points per discretization interval, the convergence tolerance for free parameters and for solutions, and the continuation step size tolerance when locating special solutions, respectively.

Table 1

Absolute difference between estimated values (k_5, λ_5) and (k_4, λ_4), respectively, of the value of k and the nonzero Floquet multiplier λ at the switching-sliding bifurcation point detected using $\widehat{\text{TC}}$ during single-parameter variation of periodic trajectories with cyclic signatures $\{\mathbf{I}_7, \mathbf{I}_3, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ and $\{\mathbf{I}_7, \mathbf{I}_{10}, \mathbf{I}_9, \mathbf{I}_3\}$ and those predicted from (6.4)–(6.5). Here, $EPSL = EPSU = 10^{-6}$, $EPSS = 10^{-4}$, and $NCOL = 4$.

NTST	$\lvert k_5 - k\rvert$	$\lvert k_4 - k\rvert$	$\lvert \lambda_5 - \lambda\rvert$	$\lvert \lambda_4 - \lambda\rvert$
200	1.7×10^{-4}	2.0×10^{-5}	1.0×10^{-6}	1.3×10^{-6}
100	6.3×10^{-4}	8.0×10^{-5}	3.5×10^{-6}	6.6×10^{-6}
50	2.4×10^{-3}	2.9×10^{-4}	1.3×10^{-5}	2.6×10^{-5}
30	6.5×10^{-3}	7.8×10^{-4}	3.6×10^{-5}	6.7×10^{-5}
10	5.1×10^{-2}	5.8×10^{-3}	3.0×10^{-4}	5.2×10^{-4}
5	2.0×10^{-1}	2.3×10^{-2}	1.4×10^{-3}	2.3×10^{-3}

Acknowledgments. The authors express their thanks to Eusebius Doedel, Fabio Dercole, and Yuri Kuznetsov for access to and advice regarding programming of AUTO 97 and SLIDECONT. Thanks also to Alan Champneys for challenging us to clearly establish a general-purpose hybrid system formalism that is not necessarily restricted to periodic trajectories.

REFERENCES

[1] J. Adolfsson, H. Dankowicz, and A. B. Nordmark, *3D passive walkers: Stability analysis in the presence of discontinuities*, Nonlinear Dynam., 24 (2001), pp. 205–229.

[2] E. L. Allgower and K. Georg, *Numerical Continuation Methods: An Introduction*, Springer-Verlag, New York, 1990.

[3] D. A. W. Barton, B. Krauskopf, and R. E. Wilson, *Collocation schemes for periodic solutions of neutral delay differential equations*, J. Differ. Equations Appl., 12 (2006), pp. 1087–1101.

[4] W. J. Beyn, A. Champneys, E. Doedel, W. Govaerts, Y. A. Kuznetsov, and B. Sandstede, *Numerical continuation and computation of normal forms*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, G. Iooss, and N. Kopell, eds., North–Holland, Amsterdam, 2002, pp. 149–219.

[5] D. S. Bindel, J. W. Demmel, M. J. Friedman, W. J. Govaerts, and Y. A. Kuznetsov, *Bifurcation analysis of large equilibrium systems in MATLAB*, in Proceedings of the International Conference on Computational Science (ICCS), Lecture Notes in Comput. Sci. 3514, Springer-Verlag, Berlin, 2005, pp. 50–57.

[6] A. R. Champneys, Y. A. Kuznetsov, and B. Sandstede, *A numerical toolbox for homoclinic bifurcation analysis*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 867–887.

[7] A. R. Champneys and Y. A. Kuznetsov, *Numerical detection and continuation of codimension-two homoclinic orbits*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 785–822.

[8] H. Dankowicz and A. B. Nordmark, *On the origin and bifurcations of stick-slip oscillations*, Phys. D, 136 (1999), pp. 280–302.

[9] H. Dankowicz and X. Zhao, *Local analysis of co-dimension-one and co-dimension-two grazing bifurcations in impact microactuators*, Phys. D, 202 (2005), pp. 238–257.

[10] M. Dellnitz, *Computational bifurcation of periodic solutions in systems with symmetry*, IMA J. Numer. Anal., 12 (1992), pp. 429–455.

[11] F. Dercole and Y. A. Kuznetsov, *SLIDECONT: An AUTO* 97 *driver for bifurcation analysis of Filippov systems*, ACM Trans. Math. Software, 31 (2005), pp. 95–119.

[12] P. Deuflhard, *Computation of periodic solutions of nonlinear ODE's*, BIT, 24 (1984), pp. 456–466.

[13] A. Dhooge, W. Govaerts, and Y. A. Kuznetsov, *MATCONT: A MATLAB package for numerical bifurcation analysis of ODEs*, ACM Trans. Math. Software, 29 (2003), pp. 141–164.

[14] M. di Bernardo, P. Kowalczyk, and A. Nordmark, *Bifurcations of dynamical systems with sliding: Derivation of normal-form mappings*, Phys. D, 170 (2002), pp. 175–205.

[15] E. J. Doedel and M. J. Friedman, *Numerical computation of heteroclinic orbits*, J. Comput. Appl. Math., 26 (1989), pp. 155–170.

[16] E. J. Doedel, H. B. Keller, and J. P. Kernevez, *Numerical analysis and control of bifurcation problems, part* II: *Bifurcation in infinite dimensions*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 745–772.

[17] E. J. Doedel, R. C. Paffenroth, H. B. Keller, D. J. Dichmann, J. Galan-Vioque, and A. Vanderbauwhede, *Computation of periodic solutions of conservative systems with application to the 3-body problem*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 13 (2003), pp. 1353–1381.

[18] E. J. Doedel, W. Govaerts, and Yu. A. Kuznetsov, *Computation of periodic solution bifurcations in ODEs using bordered systems*, SIAM J. Numer. Anal., 41 (2003), pp. 401–435.

[19] E. J. Doedel, W. Govaerts, Y. A. Kuznetsov, and A. Dhooge, *Numerical continuation of branch points of equilibria and periodic orbits*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 841–860.

[20] K. Engelborghs, T. Luzyanina, and D. Roose, *Numerical bifurcation analysis of delay differential equations using* DDE-BIFTOOL, ACM Trans. Math. Software, 28 (2002), pp. 1–21.

[21] K. Engelborghs, T. Luzyanina, K. J. In't Hout, and D. Roose, *Collocation methods for the computation of periodic solutions of delay differential equations*, SIAM J. Sci. Comput., 22 (2000), pp. 1593–1609.
[22] K. Engelborghs and E. J. Doedel, *Stability of piecewise polynomial collocation for computing periodic solutions of delay differential equations*, Numer. Math., 91 (2002), pp. 627–648.
[23] S. Foale, *Analytical determination of bifurcations in an impact oscillator*, Proc. R. Soc. Lond. Ser. A, 347 (1994), pp. 373–364.
[24] S. Foale and R. Bishop, *Bifurcations in impacting systems*, Nonlinear Dynam., 6 (1994), pp. 285–299.
[25] M. Friedman, W. Govaerts, Y. A. Kuznetsov, and B. Sautois, *Continuation of homoclinic orbits in MATLAB*, in Proceedings of the International Conference on Computational Science (ICCS), Lecture Notes in Comput. Sci. 3514, Springer-Verlag, Berlin, 2005, pp. 263–270.
[26] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Springer-Verlag, New York, 1985.
[27] W. Govaerts, Y. A. Kuznetsov, and A. Dhooge, *Numerical continuation of bifurcations of limit cycles in MATLAB*, SIAM J. Sci. Comput., 27 (2005), pp. 231–252.
[28] W. Govaerts, *Numerical bifurcation analysis for ODEs*, J. Comput. Appl. Math., 125 (2000), pp. 57–68.
[29] J. Guckenheimer, *Computer simulation and beyond—for the 21st century*, Notices Amer. Math. Soc., 45 (1998), pp. 1120–1123.
[30] A. P. Ivanov, *Bifurcations in impact systems*, Chaos Solitons Fractals, 7 (1996), pp. 1615–1634.
[31] J. D. Jansen, *Nonlinear rotor dynamics as applied to oilwell drillstring vibrations*, J. Sound Vibration, 147 (1991), pp. 115–135.
[32] W. Kang, P. Thota, B. Wilcox, and H. Dankowicz, *Bifurcation analysis of a microactuator using a new toolbox for continuation of hybrid system trajectories*, ASME J. Comput. Nonlinear Dyn., to appear.
[33] K. Karagiannis and F. Pfeiffer, *Theoretical and experimental investigations of gear-rattling*, Nonlinear Dynam., 2 (1991), pp. 367–387.
[34] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Problems*, Springer-Verlag, New York, 1987.
[35] R. I. Leine, D. H. van Campen, and W. J. G. Keultjes, *Stick-slip whirl interaction in drillstring dynamics*, J. Vibration Acoustics Transactions of the ASME, 124 (2002), pp. 209–220.
[36] A. C. J. Luo and B. C. Gegg, *On the mechanism of stick and nonstick, periodic motions in a periodically forced, linear oscillator with dry friction*, J. Vibration Acoustics Transactions of the ASME, 128 (2006), pp. 399–418.
[37] K. Lust, D. Roose, A. Spence, and A. R. Champneys, *An adaptive Newton–Picard algorithm with subspace iteration for computing periodic solutions*, SIAM J. Sci. Comput., 19 (1998), pp. 1188–1209.
[38] T. Luzyanina, K. Engelborghs, and D. Roose, *Numerical bifurcation analysis of differential equations with state-dependent delay*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 737–753.
[39] P. C. Müller, *Calculation of Lyapunov exponents for dynamics systems with discontinuities*, Chaos Solitons Fractals, 5 (1995), pp. 1671–1681.
[40] F. J. Munoz-Almaraz, E. Freire, J. Galan, E. Doedel, and A. Vanderbauwhede, *Continuation of periodic orbits in conservative and Hamiltonian systems*, Phys. D, 181 (2003), pp. 1–38.
[41] A. B. Nordmark and P. Kowalczyk, *A codimension-two scenario of sliding solutions in grazing-sliding bifurcations*, Nonlinearity, 19 (2006), pp. 1–26.
[42] A. B. Nordmark, *Existence of periodic orbits in grazing bifurcations of impacting mechanical oscillators*, Nonlinearity, 14 (2001), pp. 1517–1542.
[43] A. B. Nordmark, *Personal communication.*
[44] D. Roose, K. Lust, A. Champneys, and A. Spence, *A Newton-Picard shooting method for computing periodic solutions of large-scale dynamical systems*, Chaos Solitons Fractals, 5 (1995), pp. 1913–1925.
[45] A. G. Salinger, E. A. Burroughs, R. P. Pawlowski, E. T. Phipps, and L. A. Romero, *Bifurcation tracking algorithms and software for large scale applications*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 1015–1032.
[46] G. Samaey, K. Engelborghs, and D. Roose, *Numerical computation of connecting orbits in delay differential equations*, Numer. Algorithms, 30 (2002), pp. 335–352.
[47] F. Svahn and H. Dankowicz, *Energy transfer in vibratory systems with friction exhibiting low-velocity*

collisions, J. Vib. Control, 14 (2008), pp. 255–284.
[48] R. Szalai, G. Stépán, and S. J. Hogan, *Continuation of bifurcations in periodic delay-differential equations using characteristic matrices*, SIAM J. Sci. Comput., 28 (2006), pp. 1301–1317.
[49] P. Thota, *Analytical and Computational Tools for the Study of Grazing Bifurcations of Periodic Orbits and Invariant Tori*, Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA, 2007.
[50] P. Thota, X. Zhao, and H. Dankowicz, *Codimension-two grazing bifurcation points in single-degree-of-freedom impact oscillators*, J. Computational Nonlinear Dynam., 1 (2006), pp. 328–335.
[51] J. J. Tyson and B. Novák, *Cell cycle controls*, in Computational Cell Biology, C. P. Fall, E. S. Marland, J. M. Wagner, and J. J. Tyson, eds., Springer Science+Business Media, New York, 2002, pp. 261–284.
[52] V. I. Utkin, *Sliding Modes in Control Optimization*, Springer-Verlag, Berlin, 1992.
[53] C. Wulff and A. Schebesch, *Numerical continuation of symmetric periodic orbits*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 435–475.
[54] X. Zhao, C. K. Reddy, and A. H. Nayfeh, *Nonlinear dynamics of an electrically driven impact microactuator*, Nonlinear Dynam., 40 (2005), pp. 227–239.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1323–1354
© 2008 Society for Industrial and Applied Mathematics

Realization of Critical Eigenvalues for Scalar and Symmetric Linear Delay-Differential Equations*

P.-L. Buono† and V. G. LeBlanc‡

Abstract. This paper studies the link between the number of critical eigenvalues and the number of delays in certain classes of delay-differential equations. There are two main results. The first states that for k purely imaginary numbers which are linearly independent over the rationals, there exists a scalar delay-differential equation depending on k fixed delays whose spectrum contains those k purely imaginary numbers. The second result is a generalization of the first result for delay-differential equations which admit a characteristic equation consisting of a product of s factors of scalar type. In the second result, the k eigenvalues can be distributed among the different factors. Since the characteristic equation of scalar equations contain only exponential terms, the proof exploits a toroidal structure which comes from the arguments of the exponential terms in the characteristic equation. Our second result is applied to delay coupled $\mathbf{D}_n$-symmetric cell systems with one-dimensional cells. In particular, we provide a general characterization of delay coupled $\mathbf{D}_n$-symmetric systems with an arbitrary number of delays and cell dimension.

Key words. bifurcation, delay-differential equations, symmetry, realizability

AMS subject classifications. 37G10, 37G40, 34K06, 34K18

DOI. 10.1137/08071363X

1. Introduction and background.

Delay-differential equations (DDEs) have been used as mathematical models for phenomena in population dynamics [21], physiology [12, 3], physics [23], climate modeling [26, 28], and engineering [27], among others. DDEs behave like abstract ordinary differential equations (ODEs) on an infinite-dimensional (Banach) phase space and many results which are known for ODEs on finite-dimensional spaces have analogues in the context of DDEs. Many scalar DDE models have been developed over the years, such as for Cheynes–Stokes respiration [12] and the regulation of hematopoiesis [12], the delayed Nicholson blowflies equation [18] in population dynamics, a two-delay model of an experiment on Parkinsonian tremor [2], and many more.

The bifurcation analysis of DDEs is done essentially in the same way as that of ODEs, although the technical details differ. Consider the neighborhood of an equilibrium solution of a nonlinear DDE; then the analysis of the linearization at the equilibrium point leads to stable, unstable, and center invariant subspaces where only the stable subspace is infinite-dimensional. There exist local invariant manifolds (stable, unstable, and center manifolds) tangent to the

*Received by the editors January 17, 2008; accepted for publication (in revised form) by B. Krauskopf June 18, 2008; published electronically October 31, 2008. This research is partly supported by the Natural Sciences and Engineering Research Council of Canada in the form of a Discovery Grant (PLB, VGL).
http://www.siam.org/journals/siads/7-4/71363.html
†Faculty of Science, University of Ontario Institute of Technology, Oshawa, ONT L1H 7K4, Canada (luciano.buono@uoit.ca).
‡Department of Mathematics and Statistics, University of Ottawa, Ottawa, ONT K1N 6N5, Canada (vleblanc@uottawa.ca).

corresponding invariant subspaces of the linearized equations about the equilibrium point on which the flow near the equilibrium is exponentially attracting (stable manifold), exponentially repelling (unstable manifold), or nonhyperbolic (center manifold). Now, bifurcations near equilibria are determined by the flow on the center manifold, and the dimension of this manifold is determined by the number of eigenvalues of the linearization on the imaginary axis.

The first result of our paper is Theorem 2.1 and goes as follows. Consider n nonzero imaginary numbers $i\omega_1, \ldots, i\omega_n$, where the imaginary parts $\omega_1, \ldots, \omega_n$ are positive and not rationally dependent. We show that there exists a scalar linear DDE depending on n discrete delays written

$$\dot{x} = \sum_{j=1}^{n} a_j x(t - \tau_j), \tag{1}$$

where $x \in \mathbb{R}$, $a_j \in \mathbb{R}$, and $\tau_j \in [0, \tau]$ for all $j = 1, \ldots, n$ such that the characteristic equation of (1), given by

$$\lambda - \sum_{j=1}^{n} a_j e^{-\lambda \tau_j} = 0, \tag{2}$$

has eigenvalues $\pm i\omega_1, \ldots, \pm i\omega_n$. This result generalizes explicit computations done in the case of one and two delays; see [20, 9, 1]. The proof is done by embedding the problem as a mapping which is solved by the implicit function theorem at a carefully chosen point. From the implicit function theorem, we are able to define a smooth mapping whose transversal intersection with a dense curve on an n-dimensional torus provides solutions. The incommensurability of the n frequencies enables us to define the dense curve on the n-torus. This type of argument using a dense curve on an n-dimensional torus was used in Choi and LeBlanc [7].

This result falls within the category of so-called realization theorems, for instance, the realization theorem of linear ODEs by linear DDEs obtained by Faria and Magalhães [11]. They show that for any finite-dimensional matrix B, a necessary and sufficient condition for the existence of a bounded linear operator $\mathcal{L}_0$ from $C([-\tau, 0], \mathbb{R}^n)$ into $\mathbb{R}^n$ with infinitesimal generator having spectrum containing the spectrum of B is that n be larger than or equal to the largest number of Jordan blocks associated with each eigenvalue of B. Other results in this direction are concerned with the realization of finite jets of ODEs on a finite-dimensional center manifold by DDEs; see [11, 7]. To our knowledge, the realization theorems in this paper are the first general results linking the number of critical eigenvalues of linear DDEs with the number of discrete delays.

The next significant result is an openness theorem, that is, the realization of n imaginary numbers (not necessarily rationally independent) as eigenvalues of a linear scalar DDE is valid in a neighborhood of any set of n rationally independent imaginary numbers. The proof of this theorem also relies on the implicit function theorem.

We then turn our attention to the context of symmetric systems of DDEs. Several examples of symmetric systems of DDEs [16, 24] have characteristic equations which decompose in factors, some of which have the same form as the characteristic equation (2). The decomposition of the characteristic equation is induced by the isotypic decomposition of the space,

and we present a general derivation of this decomposition. We show that isotypic components consisting of a unique one-dimensional complex irreducible representation contribute a factor of the form (2) in the characteristic equation, and so Theorem 2.1 can be applied directly to each such factor separately.

We present a generalization of Theorem 2.1 to the case where several factors of the characteristic equation have purely imaginary eigenvalues simultaneously. Theorem 2.4 shows that a set of n rationally independent purely imaginary complex numbers can be realized from several factors of the characteristic equation of a DDE with n delays given that two nondegeneracy conditions on the characteristic equation are satisfied. The statement of the theorem is independent of any symmetric structure, and the proof is a generalization of the proof of Theorem 2.1.

We illustrate the above result on $\mathbf{D}_n$-symmetric rings of n delay equations with delayed coupling. Hopf bifurcation from such symmetric networks has been studied by several authors [6, 10, 15, 16, 17, 24, 25, 29, 30]. We study only the case in which n is odd, since, for n even, one of the nondegeneracy conditions of Theorem 2.4 is not always satisfied, as we illustrate in a $\mathbf{D}_4$ example.

In order to apply Theorem 2.4 to this context, we derive an explicit form of the coupling matrix in terms of the connections in the graph representation of the ring for cells of any dimension and arbitrary numbers of connections and delays. This is a generalization of the networks considered in the articles listed in the previous paragraph. We specialize to the case of one-dimensional cells and show how Theorem 2.4 applies to $\mathbf{D}_n$-symmetric coupled cell systems with n odd.

The paper is organized as follows. The first section contains brief preliminary remarks, and then we state and prove our main result (Theorem 2.1) and the openness result (Theorem 2.2). Then we introduce the context leading to Theorem 2.4 and state this result. Section 3 is devoted to Γ-symmetric systems of DDEs, and the section begins with a general discussion. Section 3.1 presents a characterization of $\mathbf{D}_n$-symmetric rings of delay coupled cells with an arbitrary number of delays, and the characteristic equation in the case of one-dimensional cells is derived. Theorem 2.4 is applied to $\mathbf{D}_n$-symmetric rings of one-dimensional cells with n odd. Section 4 contains the proof of Theorem 2.4. We conclude with a discussion of open problems along the lines of those presented in this paper.

2. Realization theorems. We now discuss some aspects of the spectral theory of linear scalar DDEs. In fact, we just introduce the basic facts, in a nonabstract setting, needed for the statement of our first main theorem. For a complete treatment, see Diekmann et al. [9] or Hale and Verduyn-Lunel [20].

Consider the scalar DDE

$$\dot{x}(t) = \sum_{j=1}^{n} a_j x(t-\tau_j), \tag{3}$$

where $a_j \in \mathbb{R}$ and $\tau_j \in [0,\tau]$ for all $j = 1,\dots,n$ and $\tau > 0$. The characteristic equation for (3) can be obtained by substituting $x(t) = Ce^{\lambda t}$, where C is a constant, into the equation. Thus,

$$\lambda C e^{\lambda t} = \sum_{j=1}^{n} a_j C e^{\lambda(t-\tau_j)} = \sum_{j=1}^{n} a_j C e^{-\lambda\tau_j} e^{\lambda t},$$

and by rearranging the terms we obtain

$$\left(\lambda - \sum_{j=1}^{n} a_j e^{-\lambda\tau_j}\right) x(t) = 0.$$

So, $x(t)$ is a nonzero solution of (3) if and only if

$$\Delta(\lambda) := \lambda - \sum_{j=1}^{n} a_j e^{-\lambda\tau_j} = 0.$$

The complex number λ is an eigenvalue of equation (3) if it is a solution of the characteristic equation $\Delta(\lambda) = 0$.

The question we address in this paper is related to the number of imaginary eigenvalues (with incommensurable frequencies) which can satisfy $\Delta(\lambda) = 0$. The case $n = 1$ with one nonzero delay is a straightforward calculation and $\Delta(\lambda) = 0$ for only one nonzero imaginary eigenvalue λ; see [20]. The case $n = 2$ with $\tau_1 = 0$ and $\tau_2 \in (0, \tau]$ in (3) can be found in [9]. There, it is shown that $\Delta(\lambda) = 0$ can have at most two nonzero imaginary eigenvalues. The case $n = 2$ with $\tau_1, \tau_2 > 0$ is done in [1], where it is shown that $\Delta(\lambda) = 0$ can have at most two nonzero imaginary eigenvalues. We are now ready to state our first result.

Theorem 2.1. *Suppose* $\omega_1 > 0$, $\omega_2 > 0$, $\ldots$, $\omega_n > 0$ *are linearly independent over the rationals. Then there exist* $\tau_1 > 0$, $\tau_2 > 0$, $\ldots$, $\tau_n > 0$, $a_1 \in \mathbb{R}$, $a_2 \in \mathbb{R}$, $\ldots$, $a_n \in \mathbb{R}$ *such that the linear DDE*

$$\dot{x}(t) = a_1\, x(t-\tau_1) + a_2\, x(t-\tau_2) + \cdots + a_n\, x(t-\tau_n) \tag{4}$$

has solutions $x_j^{\pm}(t) = e^{\pm\, i\omega_j t}$ *for all* $j = 1, \ldots, n$.

Proof. A necessary and sufficient condition for the conclusion of the theorem to hold is that the algebraic system of $2n$ equations

$$\begin{aligned} \sum_{k=1}^{n} a_k\, e^{-i\omega_j\tau_k} &= i\omega_j, \qquad j = 1, \ldots, n, \\ \sum_{k=1}^{n} a_k\, e^{i\omega_j\tau_k} &= -i\omega_j, \quad j = 1, \ldots, n, \end{aligned} \tag{5}$$

has a solution in the $2n$ unknowns $(\tau_1, \tau_2, \ldots, \tau_n, a_1, a_2, \ldots, a_n)$. Although (5) is in complex form, since the second equation in (5) is just the complex conjugate of the first equation in (5), system (5) is equivalent to a system of $2n$ *real* equations. This fact is taken for granted throughout what follows, even though we continue to use complex notation.

It is useful to use the following matrix notation for (5):

$$\begin{pmatrix} P(\tau;\omega) \\ P(-\tau;\omega) \end{pmatrix} A^T = \begin{pmatrix} i\omega^T \\ -i\omega^T \end{pmatrix}, \tag{6}$$

where $\tau = (\tau_1, \ldots, \tau_n)$, $\omega = (\omega_1, \ldots, \omega_n)$, $A = (a_1, \ldots, a_n)$, superscript T denotes transpose, and $P(\tau;\omega) = P(\tau_1, \ldots, \tau_n; \omega_1, \ldots, \omega_n)$ is the $n \times n$ matrix whose entry at row j column k is

$$[\, P(\tau;\omega)\,]_{jk} = e^{-i\omega_j\tau_k}.$$

Note that $\overline{P(\tau;\omega)} = P(-\tau;\omega)$.

Instead of attempting to solve (6) directly, we adopt an approach based on the following fact. For $j,k = 1,\dots,n$, consider the exponents $\omega_j\tau_k$ in $P(\tau;\omega)$ taken modulo 2π. Since the ω_j are rationally independent, for $\tau_k \geq 0$, the vector $\tau_k\omega$ mod 2π generates a dense orbit, denoted by $\mathcal{O}_k$, on an n-torus $\mathbf{T}^n$, where $\mathbf{T} = \mathbb{R}/2\pi\mathbb{Z}$. If (τ, A) is a solution to (6), then τ produces a point on $V := (\mathbf{T}^n)^n$ via the dense orbits.

Thus, we embed the problem of finding solutions of (6) into the problem of finding solutions of a mapping F defined on V and which is an extension of (6). The idea is that an explicit solution of $F = 0$ is easily obtained, and we use the implicit function theorem to find a submanifold of solutions to $F = 0$. We then show that the structure of the dense orbits $\mathcal{O}_k$ on this submanifold yields an infinite number of solutions to $F = 0$ and therefore to (6).

Choose coordinates on V as follows:

$$V := \{\, \Phi = (\Phi^1,\dots,\Phi^n) \mid \Phi^j = (\varphi_1^j,\dots,\varphi_n^j) \in \mathbf{T}^n,\ j = 1,\dots,n \,\},$$

and consider the following mapping associated to (6):

$$F : V \times \mathbb{R}^n \longmapsto \mathbb{R}^{2n}$$

defined by

$$F(\Phi, A;\omega) = \begin{pmatrix} \widetilde{P}(\Phi) \\ \widetilde{P}(-\Phi) \end{pmatrix} A^T - i \begin{pmatrix} \omega^T \\ -\omega^T \end{pmatrix}, \tag{7}$$

where A and ω are as previously defined and $\widetilde{P}(\Phi)$ is the $n \times n$ matrix whose entry at row j column k is

$$\Big[\, \widetilde{P}(\Phi) \,\Big]_{jk} = \Big[\, \widetilde{P}(\Phi^1,\dots,\Phi^n) \,\Big]_{jk} = e^{-i\varphi_j^k}.$$

Letting $\Psi = \Phi^n$ is a convenient notation to use when applying the implicit function theorem, i.e.,

$$\Psi = (\psi_1,\dots,\psi_n) = \Phi^n = (\varphi_1^n,\dots,\varphi_n^n).$$

We write $V = V_\Phi \times V_\Psi$, where $V_\Phi \cong (\mathbf{T}^n)^{n-1}$ and $V_\Psi \cong \mathbf{T}^n$ so that

$$F : V_\Phi \times V_\Psi \times \mathbb{R}^n \longmapsto \mathbb{R}^{2n}$$

is written as $F(\Phi, \Psi, A;\omega)$ in (7) (we have relabeled $\Phi = (\Phi^1,\dots,\Phi^{n-1})$ to designate coordinates for $V_\Phi \cong (\mathbf{T}^n)^{n-1}$).

We now find an explicit solution to $F = 0$. If $\{e_1,\dots,e_n\}$ denotes the canonical basis of vectors in $\mathbb{R}^n$, we define the vectors $v_1,\dots,v_n$ by $v_1 = \sum_{k=1}^n e_k$, and for $j = 2,\dots,n$,

$$v_j = v_1 - \sum_{\ell=0}^{j-2} 2\, e_{n-\ell}.$$

By construction, the set $\{v_1,\dots,v_n\}$ is linearly independent, and so the $n \times n$ matrix $\mathcal{I}$, whose jth column is the vector v_j^T, is invertible.

Consider the following point in $V_\Phi \times V_\Psi$:

$$(\widehat{\Phi}, \widehat{\Psi}) = -\frac{\pi}{2}\left((v_1, \ldots, v_{n-1}), v_n\right);$$

then it is easy to compute that

$$\widetilde{P}(\widehat{\Phi}, \widehat{\Psi}) = i\mathcal{I}, \tag{8}$$

where $\widetilde{P}$ is as in (7). If we define

$$\widehat{A}^T \equiv (\hat{a}_1, \ldots, \hat{a}_n)^T = \mathcal{I}^{-1}\omega^T,$$

then

$$F(\widehat{\Phi}, \widehat{\Psi}, \widehat{A}; \omega) = 0.$$

Because the ω_j are rationally independent, it follows that the components $\hat{a}_k$ of $\widehat{A}$ are all nonzero.

We now show that we can use the implicit function theorem at the point $(\widehat{\Phi}, \widehat{\Psi})$. Define the $n \times n$ invertible matrix $\mathcal{U}_j$ to be the diagonal matrix whose kth diagonal element is the kth component of the vector v_j (in particular, $\mathcal{U}_1$ is the identity matrix). Note also that $\mathcal{U}_j^{-1} = \mathcal{U}_j$, $j = 1, \ldots, n$. We easily compute the derivative

$$J \equiv D_{(\Psi,A)}F(\widehat{\Phi}, \widehat{\Psi}, \widehat{A}; \omega) = \begin{pmatrix} \hat{a}_n\mathcal{U}_n & i\mathcal{I} \\ \hat{a}_n\mathcal{U}_n & -i\mathcal{I} \end{pmatrix},$$

which is invertible, and its inverse is easily computed as

$$J^{-1} = \begin{pmatrix} \frac{1}{2\hat{a}_n}\mathcal{U}_n & \frac{1}{2\hat{a}_n}\mathcal{U}_n \\ -\frac{i}{2}\mathcal{I}^{-1} & \frac{i}{2}\mathcal{I}^{-1} \end{pmatrix}.$$

By the implicit function theorem, there exist a neighborhood N of $\widehat{\Phi}$ in V_Φ and a unique smooth function

$$\begin{aligned} G : N &\longmapsto V_\Psi \times \mathbb{R}^n, \\ \Phi &\longmapsto G(\Phi) = (G_\Psi(\Phi), G_A(\Phi)) \end{aligned} \tag{9}$$

such that

$$G(\widehat{\Phi}) = (\widehat{\Psi}, \widehat{A})$$

and

$$F(\Phi, G(\Phi); \omega) \equiv 0 \quad \forall\, \Phi \in N. \tag{10}$$

Recall that $\mathcal{O}_k$ is the dense orbit generated by $\tau_k\omega$ mod 2π on the n-torus $\mathbf{T}^n$. Let $\mathcal{O}_\Phi \subset V_\Phi$ be the direct product of the dense orbits $\mathcal{O}_k$ for $k = 1, \ldots, n-1$ and $\mathcal{O}_\Psi$ be the dense orbit in V_Ψ. From (9), if $\Phi \in \mathcal{O}_\Phi$ and $\Psi = G_\Psi(\Phi) \in \mathcal{O}_\Psi$, then $A = G_A(\Phi)$ yields a solution to the original system of equations (6). Thus, to complete the proof, it remains to show that there exists a point $\Phi \in \mathcal{O}_\Phi$ which is mapped by G_Ψ to a point $\Psi \in \mathcal{O}_\Psi$.

We begin by showing that G_Ψ is regular at $\widehat{\Phi}$. An easy calculation shows that

$$K \equiv D_\Phi F(\widehat{\Phi}, \widehat{\Psi}, \widehat{A}; \omega) = \begin{pmatrix} \hat{a}_1 \mathcal{U}_1 & \hat{a}_2 \mathcal{U}_2 & \cdots & \hat{a}_{n-1} \mathcal{U}_{n-1} \\ \hat{a}_1 \mathcal{U}_1 & \hat{a}_2 \mathcal{U}_2 & \cdots & \hat{a}_{n-1} \mathcal{U}_{n-1} \end{pmatrix},$$

and implicit differentiation of (10) yields that

$$\begin{aligned} DG(\widehat{\Phi}) &= \begin{pmatrix} DG_\Psi(\widehat{\Phi}) \\ DG_A(\widehat{\Phi}) \end{pmatrix} = -J^{-1}K \\ &= \begin{pmatrix} -\frac{\hat{a}_1}{\hat{a}_n} \mathcal{U}_n \mathcal{U}_1 & -\frac{\hat{a}_2}{\hat{a}_n} \mathcal{U}_n \mathcal{U}_2 & \cdots & -\frac{\hat{a}_{n-1}}{\hat{a}_n} \mathcal{U}_n \mathcal{U}_{n-1} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}, \end{aligned} \tag{11}$$

where $\mathbf{0}$ denotes the $n \times n$ zero matrix. Consequently,

$$DG_\Psi(\widehat{\Phi}) = \begin{pmatrix} -\frac{\hat{a}_1}{\hat{a}_n} \mathcal{U}_n \mathcal{U}_1 & -\frac{\hat{a}_2}{\hat{a}_n} \mathcal{U}_n \mathcal{U}_2 & \cdots & -\frac{\hat{a}_{n-1}}{\hat{a}_n} \mathcal{U}_n \mathcal{U}_{n-1} \end{pmatrix}, \tag{12}$$

and it follows that the mapping

$$G_\Psi : N \longrightarrow V_\Psi$$

is regular at $\widehat{\Phi}$.

The density of $\mathcal{O}_k$ in $\mathbf{T}^n$ for $k = 1, \ldots, n-1$ implies that for every $\epsilon > 0$ there exists $(\tau_{1,\epsilon}, \ldots, \tau_{n-1,\epsilon})$ such that

$$\Phi^*_\epsilon = (\tau_{1,\epsilon}\omega, \ldots, \tau_{n-1,\epsilon}\omega) \bmod 2\pi$$

is in an ϵ-neighborhood of $\widehat{\Phi}$, and we define a small $(n-1)$-dimensional surface in V_Φ based at Φ^*_ϵ by

$$S^h_{\Phi^*_\epsilon} = \{(\tau_1\omega, \ldots, \tau_{n-1}\omega) \bmod 2\pi \mid \tau_j \in (\tau_{j,\epsilon} - h, \tau_{j,\epsilon} + h)\}$$

with ϵ, h small enough so that $S^h_{\Phi^*_\epsilon} \subset N$. Note that this surface is generated by small nonempty open intervals of $\mathcal{O}_k$ for $k = 1, \ldots, n-1$.

We now show that the image of $S^h_{\Phi^*_\epsilon}$ by G_Ψ has a nontrivial transversal intersection with $\mathcal{O}_\Psi$. To do this, we consider the function

$$\mathcal{T} : N \longrightarrow \mathbb{R}$$

defined by

$$\mathcal{T}(\Phi) = \det \begin{pmatrix} DG_\Psi(\Phi) \cdot W_1^T & DG_\Psi(\Phi) \cdot W_2^T & \cdots & DG_\Psi(\Phi) \cdot W_{n-1}^T & \omega^T \end{pmatrix}, \tag{13}$$

where

$$W_j := \frac{d}{dx_j}(x_1\omega, \ldots, x_j\omega, \ldots, x_{n-1}\omega)$$

for $j = 1, \dots, n-1$ are $n-1$ linearly independent vectors in $(\mathbb{R}^n)^{n-1}$. Obviously, $\mathcal{T}$ is continuous, and

$$\begin{aligned}
\mathcal{T}(\widehat{\Phi}) &= \det\left(-\tfrac{\hat{a}_1}{\hat{a}_n}\mathcal{U}_n\mathcal{U}_1\omega^T \quad -\tfrac{\hat{a}_2}{\hat{a}_n}\mathcal{U}_n\mathcal{U}_2\omega^T \quad \cdots \quad -\tfrac{\hat{a}_{n-1}}{\hat{a}_n}\mathcal{U}_n\mathcal{U}_{n-1}\omega^T \quad \mathcal{U}_n\mathcal{U}_n\omega^T \right) \\
&= \frac{(-1)^{n-1}}{\hat{a}_n^{n-1}} \det\mathcal{U}_n \det\left(\hat{a}_1\mathcal{U}_1\omega^T \quad \hat{a}_2\mathcal{U}_2\omega^T \quad \cdots \quad \hat{a}_{n-1}\mathcal{U}_{n-1}\omega^T \quad \mathcal{U}_n\omega^T \right) \\
&= \frac{(\omega_1\,\omega_2\,\cdots\,\omega_n)\,(\hat{a}_1\,\hat{a}_2\,\cdots\,\hat{a}_{n-1})}{\hat{a}_n^{n-1}} \det\mathcal{I} \\
&\neq 0.
\end{aligned}$$

It follows that there is a neighborhood $N' \subseteq N$ in which $\mathcal{T} \neq 0$. So, by choosing ϵ, h small enough such that $S^h_{\Phi^*_\epsilon} \subset N'$, the image of $S^h_{\Phi^*_\epsilon}$ by G_Ψ is transverse to $\mathcal{O}_\Psi$. The density of the orbit $\mathcal{O}_\Psi$ in V_Ψ guarantees that there are infinitely many intersections with $G_\Psi(\mathcal{S}^h_{\Phi^*_\epsilon})$ near the point $\widehat{\Psi} = G_\Psi(\widehat{\Phi})$. ■

The next theorem shows that the previous realization result holds for open sets near solutions found in Theorem 2.1.

Theorem 2.2. *Suppose $\omega_1 > 0$, $\omega_2 > 0$, …, $\omega_n > 0$ are linearly independent over the rationals. There exist a neigborhood $\mathcal{N}$ of $\omega = (\omega_1, \dots, \omega_n)$ in $\mathbb{R}^n$ and a smooth mapping*

$$\begin{aligned}
H: \quad \mathcal{N} &\longrightarrow \mathbb{R}^n \times \mathbb{R}^n, \\
\omega &\longmapsto H(\omega) = (\tau(\omega), A(\omega)) = ((\tau_1(\omega), \dots, \tau_n(\omega)), (a_1(\omega), \dots, a_n(\omega)))
\end{aligned}$$

such that

$$\begin{aligned}
\sum_{k=1}^{n} a_k(\omega)\, e^{-i\omega_j \tau_k(\omega)} &= i\omega_j, \qquad j = 1, \dots, n, \\
\sum_{k=1}^{n} a_k(\omega)\, e^{i\omega_j \tau_k(\omega)} &= -i\omega_j, \quad j = 1, \dots, n,
\end{aligned} \tag{14}$$

for all $\omega \in \mathcal{N}$.

Proof. We consider the system $F = 0$ given by (7). We have already shown in Theorem 2.1 that, for fixed ω linearly independent over the rationals, there exist infinitely many solutions to $F = 0$. We again use an implicit function theorem argument combined with the density of irrational torus flows.

Consider the mapping

$$\begin{aligned}
Q \quad &: \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n \times \mathbb{R}^n, \\
&(\tau, A, \omega) \longmapsto Q(\tau, A, \omega) = F((\tau_1\omega, \dots, \tau_{n-1}\omega), \tau_n\omega, A; \omega),
\end{aligned} \tag{15}$$

where F is as in (7). Therefore,

(16)

$$D_\tau Q(\tau, A, \omega) = D_{((\Phi^1, \dots, \Phi^{n-1}), \Psi)} F((\tau_1\omega, \dots, \tau_{n-1}\omega), \tau_n\omega, A; \omega) \cdot \begin{pmatrix} \omega^T & 0 & 0 & \cdots & 0 \\ 0 & \omega^T & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \omega^T \end{pmatrix},$$

where each 0 in the matrix above is an n-dimensional zero column vector; and

$$D_A Q(\tau, A, \omega) = D_A F((\tau_1\omega, \dots, \tau_{n-1}\omega), \tau_n\omega, A; \omega). \tag{17}$$

Thus, we wish to show that the $2n \times 2n$ matrix

$$\begin{pmatrix} D_\tau Q(\tau, A, \omega) & D_A Q(\tau, A, \omega) \end{pmatrix} \tag{18}$$

is invertible at the solutions to (5) we have found in Theorem 2.1.

For positive integers p and q, let $\mathrm{Mat}_{p,q}$ denote the space of $p \times q$ matrices. Consider the following mappings associated to (16), (17), and (18):

$$\mathcal{R}_1 : V_\Phi \times V_\Psi \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathrm{Mat}_{2n,n}$$

defined by

$$\mathcal{R}_1(\Phi, \Psi, A, \omega) = D_{(\Phi,\Psi)} F(\Phi, \Psi, A; \omega) \cdot \begin{pmatrix} \omega^T & 0 & 0 & \cdots & 0 \\ 0 & \omega^T & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \omega^T \end{pmatrix},$$

$$\mathcal{R}_2 : V_\Psi \times V_\Psi \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathrm{Mat}_{2n,n}$$

defined by

$$\mathcal{R}_2(\Phi, \Psi, A, \omega) = D_A F(\Phi, \Psi, A; \omega),$$

and

$$\mathcal{R} : V_\Phi \times V_\Psi \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathrm{Mat}_{2n,2n}$$

defined by

$$\mathcal{R}(\Phi, \Psi, A, \omega) = \begin{pmatrix} \mathcal{R}_1(\Phi, \Psi, A, \omega) & \mathcal{R}_2(\Phi, \Psi, A, \omega) \end{pmatrix}.$$

Now, a simple computation (similar to those done in the proof of Theorem 2.1) shows that

$$\mathcal{R}\left(-\frac{\pi}{2}(v_1, \dots, v_n), A, \omega\right) = \begin{pmatrix} \mathcal{Z} & i\mathcal{I} \\ \mathcal{Z} & -i\mathcal{I} \end{pmatrix},$$

where

$$\mathcal{Z} = \begin{pmatrix} a_1 \mathcal{U}_1 \omega^T & a_2 \mathcal{U}_2 \omega^T & \cdots & a_n \mathcal{U}_n \omega^T \end{pmatrix}.$$

If none of the a_j vanish, then the $n \times n$ matrix $\mathcal{Z}$ is invertible, since its determinant is

$$\det \mathcal{Z} = \prod_{j=1}^{n} a_j \omega_j \det \mathcal{I} \neq 0.$$

Thus,

$$\mathcal{R}\left(-\frac{\pi}{2}(v_1, \dots, v_n), A, \omega\right)^{-1} = \begin{pmatrix} \frac{1}{2}\mathcal{Z}^{-1} & \frac{1}{2}\mathcal{Z}^{-1} \\ -\frac{i}{2}\mathcal{I}^{-1} & \frac{i}{2}\mathcal{I}^{-1} \end{pmatrix}.$$

By continuity, there is thus a neighborhood $\mathcal{N}$ of the point $-\frac{\pi}{2}(v_1, \dots, v_n)$ in $V_\Phi \times V_\Psi$ in which $\mathcal{R}$ is invertible. By Theorem 2.1, there are infinitely many solutions of $Q = 0$ (see (15)) in $\mathcal{N}$, and the Jacobian matrix (18) is thus invertible at these solutions. We get the conclusion of Theorem 2.2 by the implicit function theorem. ∎

2.1. Example: $\mathbf{D}_3$-symmetric system. Theorem 2.1 is written in the context of scalar DDEs. However, in this section, we look at an example of a $\mathbf{D}_3$-symmetric system of DDEs where Theorem 2.1 can be applied and then proceed to explain the generalization of this theorem, which has applications to symmetric systems of DDEs.

Example 2.3. Let $\Gamma = \mathbf{D}_3$, the group generated by κ and γ, act on $\mathbb{R}^3$ as follows:

$$\kappa.(x_1, x_2, x_3) = (x_1, x_3, x_2), \qquad \gamma.(x_1, x_2, x_3) = (x_3, x_1, x_2).$$

Consider a linear $\mathbf{D}_3$-symmetric coupled cell system with delayed coupling where each cell is one-dimensional and has the form

$$\begin{aligned} \dot{x}_1 &= a_1 x_1(t-\tau_1) + a_2[x_2(t-\tau_2) + x_3(t-\tau_2)], \\ \dot{x}_2 &= a_1 x_2(t-\tau_1) + a_2[x_3(t-\tau_2) + x_1(t-\tau_2)], \\ \dot{x}_3 &= a_1 x_3(t-\tau_1) + a_2[x_1(t-\tau_2) + x_2(t-\tau_2)], \end{aligned} \tag{19}$$

where $x_i \in \mathbb{R}$ for $i = 1, 2, 3$ and $a_1, a_2, a_3 \in \mathbb{R}$. The characteristic equation of system (19) is obtained by substituting $(x_1, x_2, x_3) = (w_1 e^{\lambda t}, w_2 e^{\lambda t}, w_3 e^{\lambda t})$ into the equations. We obtain after simplification

$$\begin{aligned} \lambda w_1 &= a_1 e^{-\lambda\tau_1} w_1 + a_2 e^{-\lambda\tau_2}[w_2 + w_3], \\ \lambda w_2 &= a_1 e^{-\lambda\tau_1} w_2 + a_2 e^{-\lambda\tau_2}[w_3 + w_1], \\ \lambda w_3 &= a_1 e^{-\lambda\tau_1} w_3 + a_2 e^{-\lambda\tau_2}[w_1 + w_2], \end{aligned}$$

and rearranging the terms we have

$$\left[(\lambda - a_1 e^{-\lambda\tau_1})I - a_2 e^{-\lambda\tau_2} \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}\right] \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = 0, \tag{20}$$

where I is the 3×3 identity matrix. Letting $\alpha = \lambda - a_1 e^{-\lambda\tau_1}$ and $\beta = -a_2 e^{-\lambda\tau_2}$ equation (20) becomes

$$\begin{pmatrix} \alpha & \beta & \beta \\ \beta & \alpha & \beta \\ \beta & \beta & \alpha \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = 0.$$

Let

$$\Delta(\lambda) = \begin{pmatrix} \alpha & \beta & \beta \\ \beta & \alpha & \beta \\ \beta & \beta & \alpha \end{pmatrix}.$$

We complexify $\mathbb{R}^3$ and look at the isotypic decomposition of $\mathbb{C}^3$ by the action of $\mathbf{D}_3$:

$$\mathbb{C}^3 = V_0 \oplus V_1 \oplus V_2,$$

where V_0 is the trivial representation of $\mathbf{D}_3$ and V_1, V_2 are the standard irreducible representations of $\mathbf{D}_3$ (all representations are one-dimensional complex). A basis for V_0 is $u_0 = (v, v, v)^t$,

a basis for V_1 is $u_1 = (v, e^{2\pi i/3}v, e^{4\pi i/3}v)^t$, and a basis for V_2 is $u_2 = (v, e^{4\pi i/3}v, e^{2\pi i/3}v)^t$. Therefore,

$$\Delta(\lambda)u_0 = (\alpha + 2\beta)u_0$$

and

$$\Delta(\lambda)u_1 = (\alpha - \beta)u_1, \qquad \Delta(\lambda)u_2 = (\alpha - \beta)u_2$$

since $e^{4\pi i/3} = \overline{e^{2\pi i/3}}$. Therefore, in the basis given by the isotypic decomposition of $\mathbb{C}^3$, $\Delta(\lambda)$ block diagonalizes so that we have

$$\begin{pmatrix} \alpha + 2\beta & 0 & 0 \\ 0 & \alpha - \beta & 0 \\ 0 & 0 & \alpha - \beta \end{pmatrix} \begin{pmatrix} \tilde{w}_1 \\ \tilde{w}_2 \\ \tilde{w}_3 \end{pmatrix} = 0.$$

Hence, the eigenvalues are solutions to

$$\det \Delta(\lambda) = (\alpha + 2\beta)(\alpha - \beta)^2 = (\lambda - a_1 e^{-\lambda\tau_1} - 2a_2 e^{-\lambda\tau_2})(\lambda - a_1 e^{-\lambda\tau_1} + a_2 e^{-\lambda\tau_2})^2 = 0.$$

Each factor of the characteristic equation is of the same form as the characteristic equation for a scalar DDE. Therefore, by letting $\tilde{a}_1 = a_1$ and $\tilde{a}_2 = 2a_2$ in $(\lambda - a_1 e^{-\lambda\tau_1} - 2a_2 e^{-\lambda\tau_2})$, Theorem 2.1 applies directly. The same is true for the factor $(\lambda - a_1 e^{-\lambda\tau_1} + a_2 e^{-\lambda\tau_2})$, where we let $\tilde{a}_1 = a_1$ and $\tilde{a}_2 = -a_2$. Hence, for any choice of a set of complex numbers $\Lambda = \{i\omega_1, i\omega_2\}$ with $\omega_1, \omega_2 > 0$ and rationally independent, there exists a linear $\mathbf{D}_3$ symmetric coupled cell system including Λ in its spectrum.

In the context of bifurcation theory, the symmetry properties of the critical eigenspace depend on which factor contains the critical eigenvalue, and this leads to different bifurcation behavior. Two imaginary eigenvalues in the first factor correspond to a nonresonant Hopf/Hopf mode interaction (without symmetry), while the second case leads to a nonresonant $\mathbf{D}_3$ Hopf/Hopf mode interaction. Details of the unfolding of these bifurcations can be found, respectively, in Kuznetsov [22] and Golubitsky, Stewart, and Schaeffer [14].

Note that Theorem 2.1 is not sufficient to guarantee the existence of a linear $\mathbf{D}_3$-symmetric coupled cell system with $i\omega_1$ satisfying the first factor and $i\omega_2$ satisfying the second factor simultaneously. We characterize this situation as follows. Let $b_1^1 = b_1^2 = 1$, $b_2^1 = 2$, and $b_2^2 = -1$, and for fixed rationally independent $i\omega_1, i\omega_2$ (with $\omega_1, \omega_2 > 0$) we look for a_1, a_2 and τ_1, τ_2 such that

$$\begin{aligned} a_1 b_1^1 e^{-i\omega_1\tau_1} + a_2 b_2^1 e^{-i\omega_1\tau_2} &= i\omega_1, \\ a_1 b_1^2 e^{-i\omega_2\tau_1} + a_2 b_2^2 e^{-i\omega_2\tau_2} &= i\omega_2, \end{aligned} \tag{21}$$

and their complex conjugate equations are satisfied. This is the context of the next theorem, which is a generalization of Theorem 2.1. We state this result in a general form below and postpone the proof to section 4, as it follows similar steps as the proof of Theorem 2.1.

Note that in the proof of Theorem 2.1, the matrix $\mathcal{I}$ defined in (8) is nonsingular by construction, and this is a crucial step in the argument. For this more general result we shall present, the matrix which holds a similar role is denoted by $\mathcal{I}_B$ since it is a matrix consisting of $\pm$ the constants b_k^j which appear in (21). The form of this matrix is not relevant for the

moment, and the structure of the matrix is described in section 4. We are now ready to state the theorem.

Theorem 2.4. *Consider the factors*

$$\prod_{j=1}^{r}\left(\lambda-\sum_{k=1}^{n}a_k b_k^j e^{-\lambda\tau_k}\right) \tag{22}$$

of a characteristic polynomial, where the constants $b_k^j \in \mathbb{R}\setminus\{0\}$ *are fixed for all* $j=1,\dots,r$, $k=1,\dots,n$, *and suppose that* $\det \mathcal{I}_B \neq 0$. *Suppose that*

$$\omega_1^1,\dots,\omega_{\ell_1}^1,\omega_1^2,\dots,\omega_{\ell_2}^2,\dots,\omega_1^r,\dots,\omega_{\ell_r}^r$$

are positive and linearly independent over the rationals where $\ell_1+\cdots+\ell_r=n$. *Then there exist* $\tau_1>0$, $\tau_2>0$, $\dots$, $\tau_n>0$, $a_1\in\mathbb{R}$, $a_2\in\mathbb{R}$, $\dots$, $a_n\in\mathbb{R}$ *such that for all* $j=1,\dots,r$,

$$\left(\lambda-\sum_{k=1}^{n}a_k b_k^j e^{-\lambda\tau_k}\right)=0$$

has roots $i\omega_\ell^j$ *for* $\ell=1,\dots,\ell_j$.

This theorem is applied in the following section to the case of $\mathbf{D}_n$-symmetrically coupled one-dimensional cell systems. If n odd, it is easy to show that $b_k^j\neq 0$ holds, but for n even, some of the b_k^j's can be zero, and in those cases, Theorem 2.4 cannot be applied directly.

Example 2.5. Consider the case of a $\mathbf{D}_4$-symmetric ring of DDEs given by

$$\dot{x}_i = a_1 x_i(t-\tau_1)+a_2[x_{i+1}(t-\tau_2)+x_{i-1}(t-\tau_2)],$$

where $i=1,\dots,4$ and the indices are taken modulo 4. A calculation similar to the $\mathbf{D}_3$ case above yields the characteristic equation

$$\det\Delta(\lambda)=(\lambda-a_1e^{-\lambda\tau_1}-2a_2e^{-\lambda\tau_2})(\lambda-a_1e^{-\lambda\tau_1})(\lambda-a_1e^{-\lambda\tau_1}+2a_2e^{-\lambda\tau_2})^2=0.$$

Here Theorem 2.4 cannot be applied if we include the second factor of the characteristic equation since the b_k^j coefficient of a_2 is null. However, Theorem 2.4 can be applied if we are looking for critical eigenvalues distributed among the first and third factors.

3. Linear Γ-symmetric DDEs. We now look at the case of Γ-equivariant linear retarded functional differential equations (RFDEs) depending on ℓ discrete delays. For the results of this section, we find it convenient to introduce the well-known abstract setting (see, for instance, Hale and Verduyn-Lunel [20]), which is adapted to the symmetric case. Let $C_n = C([-\tau,0],\mathbb{C}^n)$ be the Banach space of continuous functions from the interval $[-\tau,0]$, into $\mathbb{C}^n$ ($\tau>0$) endowed with the norm of uniform convergence. Consider the linear homogeneous RFDE

$$\dot{z}(t)=\mathcal{L}_0(z_t), \tag{23}$$

where $\mathcal{L}_0$ is a bounded linear operator from C_n into $\mathbb{C}^n$. We write

$$\mathcal{L}_0(\varphi) = \int_{-\tau}^{0} d\eta(\theta)\varphi(\theta),$$

where η is an $n \times n$ matrix-valued function of bounded variation defined on $[-\tau, 0]$. The characteristic equation is

$$\det \Delta(\lambda) = 0, \qquad \text{where } \Delta(\lambda) = \lambda I_n - \int_{-\tau}^{0} d\eta(\theta) e^{\lambda\theta}, \tag{24}$$

where I_n is the $n \times n$ identity matrix. Note that $e^{\lambda\theta} = e^{\lambda\theta} I_n$.

Suppose that Γ is a compact group of transformations acting linearly on $\mathbb{C}^n$. We say that (23) is Γ-*equivariant* if

$$\gamma \cdot \eta(\theta) = \eta(\theta) \cdot \gamma \quad \forall\, \gamma \in \Gamma,\ \theta \in [-\tau, 0]. \tag{25}$$

The group action of Γ on $\mathbb{C}^n$ induces an isotypic decomposition of $\mathbb{C}^n$:

$$\mathbb{C}^n = V_1 \oplus V_2 \oplus \cdots V_k,$$

where $V_i = U_i \oplus \cdots U_i$ for irreducible representations U_i of Γ and $U_i \not\cong U_j$ for $i \neq j$. Since $\eta(\theta)$ commutes with the action of Γ, then

$$\eta(\theta) V_i \subset V_i$$

for all $i = 1, \ldots, k$.

Therefore, $\Delta(\lambda)$ also commutes with the representation of Γ. Indeed, for all $\gamma \in \Gamma$,

$$\begin{aligned}
\Delta(\lambda)\gamma &= \lambda I \gamma - \left[\textstyle\int_{-\tau}^{0} d\eta(\theta) e^{\lambda\theta}\right] \gamma \\
&= \gamma\lambda I - \left[\textstyle\int_{-\tau}^{0} d\eta(\theta)\gamma e^{\lambda\theta}\right] \\
&= \gamma\lambda I - \left[\textstyle\int_{-\tau}^{0} \gamma d\eta(\theta) e^{\lambda\theta}\right] \\
&= \gamma\left(\lambda I - \textstyle\int_{-\tau}^{0} d\eta(\theta) e^{\lambda\theta}\right) = \gamma\Delta(\lambda).
\end{aligned}$$

Thus,

$$\Delta(\lambda) V_i \subset V_i$$

for all $i = 1, \ldots, k$, and in the orthogonal basis given by the isotypic decomposition, the matrix $\Delta(\lambda)$ block diagonalizes and we write

$$\Delta(\lambda) = \operatorname{diag}(\Delta_1(\lambda), \ldots, \Delta_k(\lambda)).$$

The characteristic equation then becomes

$$\det \Delta(\lambda) = \prod_{i=1}^{k} \det \Delta_i(\lambda).$$

Therefore, we are led to the following result.

Proposition 3.1. *Suppose that $V_i = U_i$ and U_i is a one-dimensional irreducible representation of Γ. Then*

$$\det \Delta_i(\lambda) = \lambda - \sum_{j=1}^{\ell} a_j e^{-\lambda \tau_j}.$$

Corollary 3.2. *Theorem 2.1 applies to factors of the characteristic equation which correspond to the context of Proposition 3.1.*

3.1. Delay coupled cell systems with $\mathbf{D}_n$-symmetry, n odd. Our goal is to apply Theorem 2.4 to delay coupled cell systems with $\mathbf{D}_n$-symmetry. We focus on the case of n odd because the assumption $b_k^j \neq 0$ is satisfied for all j, k. As Example 2.5 shows, when n is even one cannot apply Theorem 2.4 in all cases because some coefficients b_k^j may be zero. Therefore, we perform the following analysis on the case in which n is odd only. Note that these computations are valid in the case in which n is even with minor modifications.

Multiple authors [6, 10, 15, 16, 17, 24, 25, 29, 30] have studied Hopf bifurcation in $\mathbf{D}_n$-symmetric rings of cells with delayed coupling where each cell is one-dimensional. The differential equation systems in those papers have the following general form. For $i = 1, \ldots, n$, the dynamics of cell i is given, respectively, for n odd:

$$\dot{x}_i(t) = f(X_i) + g(x_{i+1}, \ldots, x_{i+(n-1)/2}, x_{i-(n-1)/2}, \ldots, x_{i-1}), \tag{26}$$

where $X_i = (x_i(t - s_1), \ldots, x_i(t - s_m))$, $x_j = x_j(t - \tau_j)$ for $j \neq i$, $f : \mathbb{R}^m \to \mathbb{R}$, $g : \mathbb{R}^{n-1} \to \mathbb{R}$ are smooth functions, $\tau_j, s_\ell \in [0, \tau]$ for all $j \neq i$, and $s = 1, \ldots, m$. Here, f is called the internal dynamics and g is the coupling function.

3.1.1. Characterization of delayed $\mathbf{D}_n$ networks. We introduce a more general notation for delayed symmetrically coupled cell systems inspired by recent work on (not necessarily symmetric) coupled cell systems of ODEs; see Golubitsky and Stewart [13] for a survey of the theory. Note that all the results of this section are straightforward generalizations of the nondelayed case.

Suppose that each cell in the system has phase space $\mathbb{R}^k$. We generalize system (26) to

$$\dot{X}_i(t) = f(\widetilde{X}_i, \widetilde{X}_{i+1}, \ldots, \widetilde{X}_{i+(n-1)/2}, \widetilde{X}_{i-(n-1)/2}, \ldots, \widetilde{X}_{i-1}), \qquad i = 1, \ldots, n, \tag{27}$$

where

$$\widetilde{X}_j = (X_j(t - \tau_1), \ldots, X_j(t - \tau_m)),$$

$f : (\mathbb{R}^{km})^n \to \mathbb{R}^k$ is a smooth function, and the position of $\widetilde{X}_k$ corresponds to the coupling from cell k to cell i and $\tau_j \geq 0$ for $j = 1, \ldots, m$. The following proposition is a straightforward consequence of the coupled cell system theory [13]. The proof is given for completeness.

Proposition 3.3. *Any delay coupled network of n-odd identical cells depending on m delays can be written as (27).*

Proof. Since the cells are identical, each cell has the same dimension k and the dynamics of all cells is given by the same function f. The function f has arguments coming from every other cell in the network corresponding to possible connections from these other cells, and those depend possibly on the m delays by the definition of $\widetilde{X}_j$. ∎

We say that cells j and k have *identical coupling* to cell i if

$$f(X_i,\dots,u,\dots,v,\dots)=f(X_i,\dots,v,\dots,u,\dots), \tag{28}$$

where u and v are permuted from positions j and k. We rewrite system (27) as

$$\dot{X}=F(\widetilde{X}), \tag{29}$$

where $X=(X_1,\dots,X_n)^t$,

$$\widetilde{X} \;=\; \widetilde{X}_i,\widetilde{X}_{i+1},\dots,\widetilde{X}_{i+(n-1)/2},\widetilde{X}_{i-(n-1)/2},\dots,\widetilde{X}_{i-1},$$

and $F:(\mathbb{R}^{km})^n\to\mathbb{R}^{nk}$ has the ith component given by the formulas above for $\dot{X}_i(t)$.

Consider the group $\mathbf{D}_n$, with generators ρ and κ acting on $\mathbb{R}^{kn}$ as follows:

$$\begin{aligned}\rho.(X_1,\dots,X_n)&=(X_n,X_1,X_2,\dots,X_{n-1}),\\ \kappa.(X_1,\dots,X_n)&=(X_1,X_n,\dots,X_{n+2-j},\dots,X_{(n+3)/2},X_{(n+1)/2},\dots,X_j,\dots,X_2).\end{aligned} \tag{30}$$

Without loss of generality we assume that our networks are transitive. That is, all cells in the network can be reached from any other cell via the coupling.

We now characterize the connections in the network so that the delay-differential system is $\mathbf{D}_n$-symmetric. We think of each cell in the network as having $[(n-1)/2]$ neighbors on each side and an opposite cell if n is even. Graphically, it is clear that an n-cell network is $\mathbf{D}_n$-symmetric if for all cells in the network, all connections to and from the jth neighbor on each side (or the opposite cell if n is even) are all the same; that is, the coupling term and its delay must be the same for all those connections. This idea is formalized in the next result.

Proposition 3.4. *A transitive network of n coupled identical cells with delays is $\mathbf{D}_n$-equivariant if and only if it satisfies the conditions below.*

(i) *Suppose that cell 1 receives an input from cell j with delay $\epsilon\in[0,\tau]$; then every cell i in the network ($i=2,\dots,n$) receives an input from cell $(i+j-1)$ mod n with delay ϵ identical to the one received by cell 1.*

(ii) *For every connection in part* (i)*, there is an identical connection from cell i to cell $(i+j-1)$ mod n with delay ϵ.*

Proof. The result is true for n odd and even but we give the proof only for n odd. We begin by looking at ρ-equivariance. Denote by $[w]_i$ the ith row of vector w. Then,

$$[\rho F(\widetilde{X})]_i=f(\widetilde{X}_{i-1},\widetilde{X}_i,\dots,\widetilde{X}_{i-2}),$$

and since $\rho X=(X_n,X_1,\dots,X_{n-1})$ we have

$$[F(\rho\widetilde{X})]_i=f(\widetilde{X}_{i-1},\widetilde{X}_i,\dots,\widetilde{X}_{i-2}).$$

Thus, ρ-equivariance holds automatically by the structure of the equations.

Suppose that cell 1 receives an input from cell j. We look at the system of equations (27) and focus on the possible coupling from cell $(i+j-1)$ mod n to cell i. Moreover, consider the possible connection from cell i to cell $(i-j+1)$ mod n. Note that the connections from $(i+j-1)$ mod n to i and from i to $(i-j+1)$ mod n are obtained by taking the index and

subtracting $j-1$. Finally, consider the possible connection from cell $(i-j+1)$ mod n to cell i. We now show that $F(\widetilde{X})$ is κ-equivariant (and so $\mathbf{D}_n$-equivariant) if and only if the connections defined above are identical. An easy computation shows that for all $i=1,\dots,n$ we have

$$[\kappa F(\widetilde{X})]_{n+2-i} = f(\widetilde{X}_i,\dots,\widetilde{X}_{i+j-1},\dots,\widetilde{X}_{i-j+1},\dots)$$

and

$$[F(\kappa\widetilde{X})]_{n+2-i} = f(\widetilde{X}_i,\dots,\widetilde{X}_{i-j+1},\dots,\widetilde{X}_{i+j-1},\dots),$$

where $n+2-i$ is taken modulo n for $i=1$.

We now show that parts (i) and (ii) imply κ-equivariance. If part (i) holds, the coupling from cell $(i+j-1)$ mod n to cell i and the coupling from cell i to cell $(i-j+1)$ mod n are identical. Then, by part (ii), the coupling from cell $(i-j+1)$ mod n to cell i is identical to the coupling from cell i to cell $(i-j+1)$ mod n. Therefore, the coupling from cells $(i+j-1)$ mod n and $(i-j+1)$ mod n to i are identical. By definition of identical coupling given by (28) we have that

$$f(\widetilde{X}_i,\dots,\widetilde{X}_{i-j+1},\dots,\widetilde{X}_{i+j-1},\dots) = f(\widetilde{X}_i,\dots,\widetilde{X}_{i+j-1},\dots,\widetilde{X}_{i-j+1},\dots).$$

Since the dynamics of all cells is given by the same function f, this is true for all $i=1,\dots,n$. Thus, F is κ-equivariant.

Suppose now that F is κ-equivariant. Equality of both sides of the equivariance condition implies that for all $i=1,\dots,n$, the couplings from cells $(i+j-1)$ mod n and $(i-j+1)$ mod n to i are identical. Since the dynamics of all cells is given by the same function f, the coupling from cell $(i+j-1)$ mod n to cell i guarantees an identical coupling from cell i to cell $(i-j+1)$ mod n, and this proves (i). But, the coupling from cell $(i-j+1)$ mod n to i is therefore identical to the coupling from cell i to cell $(i-j+1)$ mod n. Hence there is an identical two-way coupling between cells i and $(i-j+1)$ mod n, which proves (ii). ∎

3.1.2. General form of the characteristic equation. We now focus our attention on delay coupled cell systems where each cell is one-dimensional, that is, $k=1$. The results of this section are also easy generalizations of the nondelayed case. We split the linear and nonlinear parts of system (27) and write the result in abstract form:

$$\dot{X} = LX_t + H(X_t),$$

where $X_t \in C([-\tau,0],\mathbb{R}^n)$, $L : C([-\tau,0],\mathbb{R}^n) \to \mathbb{R}^n$ is a bounded linear map, and H is a nonlinear mapping. Thus, L is $\mathbf{D}_n$-equivariant, $\eta(\theta)$ is an $n\times n$ $\mathbf{D}_n$-equivariant matrix of bounded variation, and

$$L\phi = \int_{-\tau}^{0} d\eta(\theta)\phi.$$

Proposition 3.5. *The matrix $\eta(\theta)$ is symmetric ($\eta(\theta)=\eta(\theta)^T$) with the following properties:*
(1) *for all $j=1,\dots,n$, $\eta_{jj}(\theta)=p(\theta)$ for some function p, and*
(2) *for all i,k with $i\neq k$, $\eta_{ki}(\theta)=\eta_{(2+n-k)i}(\theta)=\eta_{k1}(\theta)$.*

Proof. We use Proposition 3.4 to obtain information on η. By part (ii), the matrix $\eta(\theta)$ is symmetric. From the structure of (27), we deduce that for all $j = 1, \dots, n$, $\eta_{jj}(\theta) = p(\theta)$ for some function $p(\theta)$. We denote by $\eta_{ji}(\theta)$ the element of η corresponding to the coupling from cell j to i. Consider $\eta_{j1}(\theta)$; then there is an identical connection from cell 1 to cell $2 + n - j$ by part (i) and so $\eta_{j1}(\theta) = \eta_{1(2+n-j)}(\theta)$. By part (ii), the connection from cell $2 + n - j$ to cell 1 is identical to its reciprocal and so $\eta_{j1}(\theta) = \eta_{(2+n-j)1}(\theta)$. By part (i), we then have $\eta_{ki}(\theta) = \eta_{(2+n-k)i}(\theta) = \eta_{k1}(\theta)$ since the connections to cell i are identical to the connections to cell 1. ∎

Remark 3.6. This result can be obtained for higher-dimensional cells with a proof essentially similar to this one, but with more cumbersome notation. We decided to restrict ourselves to the one-dimensional case as this is the one which we study in detail in what follows.

The diagonalization of the linear equation is obtained using the results at the beginning of section 3 and are analogous to calculations for the $\mathbf{D}_n$-symmetric ODEs found in Golubitsky, Stewart, and Schaeffer [14, Chapter XVIII]. The details are left to the reader. One obtains for $j = 0, \dots, n-1$

$$A_j(\theta) := p(\theta) + \sum_{k=2}^{(n+1)/2} 2\cos(2\pi(k-1)j/n)\eta_{k1}(\theta). \tag{31}$$

Note that $A_j(\theta) = A_{n-j}(\theta)$ for $j = 1, \dots, [n/2]$. The block diagonalization of η is given by the terms $A_j(\theta)$ for $j = 0, \dots, n-1$. Hence, in the basis given by the isotypic decomposition, we have

$$\Delta(\lambda) = \lambda I_n - \int_{-\tau}^{0} d\eta(\theta) e^{\lambda\theta} = \lambda I_n - \int_{-\tau}^{0} \operatorname{diag}(dA_0(\theta)e^{\lambda\theta}, \dots, dA_{n-1}(\theta)e^{\lambda\theta}).$$

Let $\Delta_j(\lambda) = \lambda - \int_{-\tau}^{0} dA_j(\theta)e^{\lambda\theta}$; then

$$\Delta(\lambda) = \operatorname{diag}(\Delta_0(\lambda), \dots, \Delta_{n-1}(\lambda)).$$

Therefore, the characteristic equation has the decomposition

$$\det \Delta(\lambda) = \det \Delta_0(\lambda) \prod_{j=1}^{(n-1)/2} [\det \Delta_j(\lambda)]^2 = 0. \tag{32}$$

For completeness, the reader can verify that the corresponding formula for n even is

$$\det \Delta(\lambda) = \det \Delta_0(\lambda) \det \Delta_{n/2}(\lambda) \prod_{j=1}^{n/2-1} [\det \Delta_j(\lambda)]^2 = 0, \tag{33}$$

where the A_j formula is slightly different from the one above.

3.1.3. Application of main theorems to the $\mathbf{D}_n$ case. It is straightforward that Theorem 2.1 can be applied to any of the factors of the characteristic equations (32) or (33). Recall from Example 2.5 that Theorem 2.4 can possibly be applied in the case in which n is even if the chosen factors of the characteristic equation satisfy $b_k^j \neq 0$ for all j, k. We do not pursue this case here.

To apply Theorem 2.4 in the $\mathbf{D}_n$ case with n odd, we need to verify that the coefficients b_k^j in the factors of the characteristic equation are nonzero and that the nondegeneracy condition $\det \mathcal{I}_B \neq 0$ is satisfied. In fact, as shown in section 4, $\det \mathcal{I}_B \neq 0$ if and only if the matrix

$$\mathcal{B} := \begin{bmatrix} b_1^1 & b_{1+\mu_1}^1 & \cdots & b_{1+\mu_{r-1}}^1 \\ b_1^2 & b_{1+\mu_1}^2 & \cdots & b_{1+\mu_{r-1}}^2 \\ \vdots & \vdots & \ddots & \vdots \\ b_1^r & b_{1+\mu_1}^r & \cdots & b_{1+\mu_{r-1}}^r \end{bmatrix}$$

is nonsingular, where ℓ_j is the number of imaginary eigenvalues satisfying the jth term of the product of the characteristic equation (22) and $\mu_j = \sum_{i=1}^{j} \ell_i$, where $\mu_r = n$ and $\mu_0 := 0$. Note that row j of $\mathcal{B}$ contains coefficients belonging to the jth factor of the characteristic equation (22). We apply Theorem 2.4 to $\mathbf{D}_n$-symmetric coupled cell systems depending on an arbitrary number of finite delays.

The characteristic equation is

$$\det \Delta(\lambda) = \det \Delta_0(\lambda) \prod_{j=1}^{(n-1)/2} [\det \Delta_j(\lambda)]^2 = 0. \tag{34}$$

We can write

$$\Delta_j(\lambda) = \lambda - F(\lambda) - G_j(\lambda),$$

where

$$F(\lambda) = \sum_{i=1}^{p} a_i e^{-\lambda \tau_i}$$

are the terms coming from the internal dynamics of each cell and

$$G_j(\lambda) = \sum_{k=2}^{(n+1)/2} \left[2 \cos\left(\frac{2\pi(k-1)j}{n} \right) \right] \sum_{t=1}^{m_k} \alpha_t^k e^{-\lambda s_t^k}$$

are the contributions from the coupling where m_k is the number of delayed terms in the connection from cell k to 1 and α_t^k are the respective coupling coefficients.

Example 3.7. As an example, consider a delay coupled $\mathbf{D}_5$-symmetric cell. Let $u_s(\theta) = 0$ if $\theta = [-\tau, -s]$ and $u_s(\theta) = 1$ for $\theta \in (-s, 0]$, where $\tau \geq s$ for all delays s, and suppose

$$\eta(\theta) = \begin{bmatrix} p(\theta) & \eta_{21}(\theta) & \eta_{31}(\theta) & \eta_{41}(\theta) & \eta_{51}(\theta) \\ \eta_{21}(\theta) & p(\theta) & \eta_{51}(\theta) & \eta_{31}(\theta) & \eta_{41}(\theta) \\ \eta_{41}(\theta) & \eta_{21}(\theta) & p(\theta) & \eta_{51}(\theta) & \eta_{31}(\theta) \\ \eta_{31}(\theta) & \eta_{41}(\theta) & \eta_{21}(\theta) & p(\theta) & \eta_{51}(\theta) \\ \eta_{51}(\theta) & \eta_{41}(\theta) & \eta_{31}(\theta) & \eta_{21}(\theta) & p(\theta) \end{bmatrix},$$

where

$$p(\theta)=\sum_{i=1}^{2} a_i u_{\tau_i}(\theta), \quad \eta_{21}(\theta)=\sum_{\ell=1}^{3} \alpha_\ell^2 u_{s_\ell^2}(\theta), \quad \text{and} \quad \eta_{31}(\theta)=\sum_{\ell=1}^{2} \alpha_\ell^3 u_{s_\ell^3}(\theta)$$

with the conditions $\eta_{41}(\theta)=\eta_{31}(\theta)$ and $\eta_{51}(\theta)=\eta_{21}(\theta)$ given by Proposition 3.5, part (2). Then,

$$F(\lambda)=\sum_{i=1}^{2} a_i e^{-\lambda\tau_i}$$

and

$$G_j(\lambda)=\sum_{k=2}^{3}\left[2\cos\left(\frac{2\pi(k-1)j}{n}\right)\right]\sum_{t=1}^{m_k}\alpha_t^k e^{-\lambda s_t^k},$$

where $m_2=3$ and $m_3=2$.

Thus, all coefficients b_j^k of $\mathcal{I}_B$ are nonzero and it is convenient to set b_1^j to be the coefficient of a_1; that is, $b_1^j=1$ for $j=0,1,2,\dots,(n+1)/2$, and we keep this convention for the remainder of the paper.

We suppose that the characteristic equation $\det\Delta(\lambda)=0$ has purely imaginary roots coming from all factors; then for $r=(n-1)/2$ we have

$$\mathcal{B}:=\begin{bmatrix} b_1^1 & b_{1+\mu_1}^1 & \cdots & b_{1+\mu_{r-1}}^1 \\ b_1^2 & b_{1+\mu_1}^2 & \cdots & b_{1+\mu_{r-1}}^2 \\ \vdots & \vdots & \ddots & \vdots \\ b_1^r & b_{1+\mu_1}^r & \cdots & b_{1+\mu_{r-1}}^r \end{bmatrix},$$

and we assign the coefficients $b_{1+\mu_j}^i$ as follows. We suppose that the first row corresponds to the factor for the trivial representation, which means that

$$b_{1+\mu_j}^1=2, \qquad j=1,2,\dots,(n-1)/2.$$

Then, we set the remaining coefficients of each row to be equal to $\left[2\cos\left(\frac{2\pi(k-1)j}{n}\right)\right]$ for $k=2,3,\dots,(n+1)/2$, where row $j+1$ has the coefficients of Δ_j for $j=1,2,\dots,(n-1)/2$. This leads to the matrix

(35)
$$\mathcal{B}=\begin{bmatrix} 1 & 2 & 2 & \cdots & 2 & 2 \\ 1 & 2\cos\left(\frac{2\pi}{n}\right) & 2\cos\left(\frac{4\pi}{n}\right) & \cdots & 2\cos\left(\frac{(n-3)\pi}{n}\right) & 2\cos\left(\frac{(n-1)\pi}{n}\right) \\ 1 & 2\cos\left(\frac{4\pi}{n}\right) & 2\cos\left(\frac{8\pi}{n}\right) & \cdots & 2\cos\left(\frac{2(n-3)\pi}{n}\right) & 2\cos\left(\frac{2(n-1)\pi}{n}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2\cos\left(\frac{(n-1)\pi}{n}\right) & 2\cos\left(\frac{2(n-1)\pi}{n}\right) & \cdots & 2\cos\left(\frac{(n-3)(n-1)\pi}{2n}\right) & 2\cos\left(\frac{(n-1)^2\pi}{2n}\right) \end{bmatrix}.$$

Let $i_1<i_2<\cdots<i_s$ be a set of indices chosen from $\{0,\dots,(n-1)/2\}$ defining a combination of factors from the characteristic equation (32). We now construct the $s\times s$

matrix $\mathcal{B}$ by removing rows and columns of (35) not in the set $\{i_1, i_2, \dots, i_s\}$. Suppose that $i_1, \dots, i_s$ are chosen from $\{1, \dots, (n-1)/2\}$; then the matrix $\mathcal{B}$ is symmetric ($\mathcal{B}^T = \mathcal{B}$) and has the form

(36)
$$\mathcal{B} = \begin{bmatrix} 2\cos\left(\frac{2\pi i_1^2}{n}\right) & 2\cos\left(\frac{2\pi i_2 i_1}{n}\right) & \cdots & 2\cos\left(\frac{2\pi i_{s-1} i_1}{n}\right) & 2\cos\left(\frac{2\pi i_s i_1}{n}\right) \\ 2\cos\left(\frac{2\pi i_1 i_2}{n}\right) & 2\cos\left(\frac{2\pi i_2^2}{n}\right) & \cdots & 2\cos\left(\frac{2\pi i_{s-1} i_2}{n}\right) & 2\cos\left(\frac{2\pi i_s i_2}{n}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 2\cos\left(\frac{2\pi i_1 i_{s-1}}{n}\right) & 2\cos\left(\frac{2\pi i_2 i_{s-1}}{n}\right) & \cdots & 2\cos\left(\frac{2\pi i_{s-1}^2}{n}\right) & 2\cos\left(\frac{2\pi i_s i_{s-1}}{n}\right) \\ 2\cos\left(\frac{2\pi i_1 i_s}{n}\right) & 2\cos\left(\frac{2\pi i_2 i_s}{n}\right) & \cdots & 2\cos\left(\frac{2\pi i_{s-1} i_s}{n}\right) & 2\cos\left(\frac{2\pi i_s^2}{n}\right) \end{bmatrix}.$$

In the other case, $i_1 = 0$ and the matrix is of the form

(37)
$$\mathcal{B} = \begin{bmatrix} 1 & 2 & \cdots & 2 & 2 \\ 1 & 2\cos\left(\frac{2\pi i_2^2}{n}\right) & \cdots & 2\cos\left(\frac{2\pi i_{s-1} i_2 - 1}{n}\right) & 2\cos\left(\frac{2\pi i_s i_2}{n}\right) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2\cos\left(\frac{2\pi i_2 i_{s-1}}{n}\right) & \cdots & 2\cos\left(\frac{2\pi i_{s-1} i_{s-1}}{n}\right) & 2\cos\left(\frac{2\pi i_s i_{s-1}}{n}\right) \\ 1 & 2\cos\left(\frac{2\pi i_2 i_s}{n}\right) & \cdots & 2\cos\left(\frac{2\pi i_{s-1}^2}{n}\right) & 2\cos\left(\frac{2\pi i_s^2}{n}\right) \end{bmatrix}.$$

We can now state our result.

Theorem 3.8. *Consider a linear $\mathbf{D}_n$-symmetric coupled cell system with n odd depending on k delays $\tau_1, \dots, \tau_k$, and let $i_1 < i_2 < \cdots < i_s$ be indices chosen from $\{0, \dots, (n-1)/2\}$ defining a combination of factors from the characteristic equation (32). We assume that the matrix $\mathcal{B}$ given by (36) or (37) is nonsingular. Suppose*

$$\omega_1^1, \dots, \omega_{\ell_{i_1}}^1, \omega_1^2, \dots, \omega_{\ell_{i_2}}^2, \dots, \omega_1^s, \dots, \omega_{\ell_{i_s}}^s$$

are positive and linearly independent over the rationals, where $\ell_{i_1} + \cdots + \ell_{i_s} = k$. Then there exist $\tau_1 > 0, \dots, \tau_k > 0$ and real coefficients a_i such that for all $m = 1, \dots, s$

$$\det \Delta_{i_m}(\lambda) = 0$$

has solutions $i\omega_\ell^m$ for $\ell = 1, \dots, \ell_{i_m}$.

Proof. Since n is odd, the coefficients

$$b_k^j = 2\cos\left(\frac{2\pi(k-1)j}{n}\right)$$

are nonzero for all $k = 2, \ldots, (n+1)/2$ and $j = 0, \ldots, (n-1)/2$. Because $\mathcal{B}$ is assumed to be nonsingular, Theorem 2.4 applies and the result is obtained. ■

The condition that $\mathcal{B}$ is nonsingular does not always hold, as we show in the case $s = 2$. Consider the matrix (37) with $n = 9$ so that $i_2 \in \{1, 2, 3, 4\}$. Choosing $i_2 = 3$ we have the singular matrix

$$\mathcal{B} = \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}.$$

We now show that $\mathcal{B}$ is nonsingular in the case $s = 2$ if the matrix is given by (36); that is,

$$\mathcal{B} = \begin{pmatrix} 2\cos\left(\frac{2\pi i_1^2}{n}\right) & 2\cos\left(\frac{2\pi i_1 i_2}{n}\right) \\ 2\cos\left(\frac{2\pi i_2 i_1}{n}\right) & 2\cos\left(\frac{2\pi i_2^2}{n}\right) \end{pmatrix}.$$

We compute

$$\begin{aligned} \det \mathcal{B} &= 4\left[\cos\left(\frac{2\pi i_1^2}{n}\right)\cos\left(\frac{2\pi i_2^2}{n}\right) - \cos\left(\frac{2\pi i_1 i_2}{n}\right)^2\right] \\ &= 2\left[\cos\left(\frac{2\pi(i_1^2+i_2^2)}{n}\right) + \cos\left(\frac{2\pi(i_1^2-i_2^2)}{n}\right) - \cos\left(\frac{4\pi i_1 i_2}{n}\right) - 1\right]. \end{aligned}$$

We show a few cases explicitly. First, the case $n = 3$ is not relevant since $i_1 < i_2$, $(n-1)/2 = 1$, and $i_1 \neq 0$. We show the case $n = 5$, where we must have $i_1 = 1$ and $i_2 = 2$. This implies that $i_1^2 + i_2^2 = 5$, and so

$$\det \mathcal{B} = 2\left[\cos\left(\frac{4\pi}{5}\right) - \cos\left(\frac{8\pi}{5}\right)\right] \neq 0.$$

We now turn to the general case and show that the determinant cannot vanish. Because the three cosines are projections of nth roots of unity on the real axis for n odd, then

$$\cos\left(\frac{2\pi(i_1^2+i_2^2)}{n}\right) + \cos\left(\frac{2\pi(i_1^2-i_2^2)}{n}\right) - \cos\left(\frac{4\pi i_1 i_2}{n}\right) \neq 1.$$

So if the determinant is to vanish, one of the cosines must be equal to 1. Since $i_1 < i_2$, there is only one option and we must have $i_1^2 + i_2^2 = n$. Thus, $i_1^2 = n - i_2^2$ and

$$\cos\left(\frac{2\pi(i_1^2-i_2^2)}{n}\right) = \cos\left(\frac{2\pi(n-2i_2^2)}{n}\right) = \cos\left(\frac{4\pi i_2^2}{n}\right).$$

If

$$\cos\left(\frac{4\pi i_2^2}{n}\right) - \cos\left(\frac{4\pi i_1 i_2}{n}\right) = 0,$$

this would imply $i_1 = i_2$, but we know that $i_1 < i_2$ and so $\det \mathcal{B}$ cannot vanish. We summarize this result in the next theorem.

Theorem 3.9. *Consider a linear $\mathbf{D}_n$-symmetric coupled cell system with n odd depending on k delays $\tau_1, \ldots, \tau_k$, and let $i_1 < i_2$ be indices chosen from $\{1, \ldots, (n-1)/2\}$ defining a combination of factors from the characteristic equation (32). Suppose*

$$\omega_1^1, \ldots, \omega_{\ell_{i_1}}^1, \omega_1^2, \ldots, \omega_{\ell_{i_2}}^2$$

are positive and linearly independent over the rationals, where $\ell_{i_1} + \ell_{i_2} = k$. Then there exist $\tau_1 > 0$, ..., $\tau_k > 0$ and real coefficients $a_1 \in \mathbb{R}$, ..., $a_p \in \mathbb{R}$ such that for $m = 1$ and $m = 2$

$$\det \Delta_{i_m}(\lambda) = 0$$

has solutions $i\omega_\ell^m$ for $\ell = 1, \ldots, \ell_{i_m}$.

4. Proof of Theorem 2.4. Before we present the proof of Theorem 2.4, we describe in the next lemma the form of the matrix $\mathcal{I}_B$ which appears in the proof and compute its determinant.

Lemma 4.1. *Let $\ell_1, \ldots, \ell_r$ be positive integers and define $\mu_j = \sum_{i=1}^{j} \ell_i$, where $\mu_r = n$ and $\mu_0 := 0$. Consider the $n \times n$ matrix*

$$\mathcal{I}_B := [A_1 \cdots A_j \cdots A_r]^T,$$

where

$$A_j = \begin{bmatrix} b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & b_{2+\mu_{j-1}}^j & \cdots & b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \\ b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & b_{2+\mu_{j-1}}^j & \cdots & -b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & -b_{2+\mu_{j-1}}^j & \cdots & -b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \end{bmatrix}$$

is an $\ell_j \times n$ matrix and all elements are nonzero. Then,

$$\det \mathcal{I}_B = \pm \prod_{j=1}^{r} \left[(-2)^{\ell_j - 1} \prod_{s=2}^{\ell_j} b_{s+\mu_{j-1}}^j \right] \det \mathcal{B},$$

where

$$\mathcal{B} := \begin{bmatrix} b_1^1 & b_{1+\mu_1}^1 & \cdots & b_{1+\mu_{r-1}}^1 \\ b_1^2 & b_{1+\mu_1}^2 & \cdots & b_{1+\mu_{r-1}}^2 \\ \vdots & \vdots & \ddots & \vdots \\ b_1^r & b_{1+\mu_1}^r & \cdots & b_{1+\mu_{r-1}}^r \end{bmatrix}.$$

Proof. Substitute row k, denoted by R_k, of

$$A_j = \begin{bmatrix} b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & b_{2+\mu_{j-1}}^j & \cdots & b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \\ b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & b_{2+\mu_{j-1}}^j & \cdots & -b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & -b_{2+\mu_{j-1}}^j & \cdots & -b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \end{bmatrix}$$

for $k = 2, \ldots, \ell_j$ by $R_k - R_1$. The matrix A_j becomes

$$\tilde{A}_j := \begin{bmatrix} b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & b_{2+\mu_{j-1}}^j & b_{3+\mu_{j-1}}^j & \cdots & b_{-1+\mu_j}^j & b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & -2b_{\mu_j}^j & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & -2b_{-1+\mu_j}^j & -2b_{\mu_j}^j & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & -2b_{3+\mu_{j-1}}^j & \cdots & -2b_{-1+\mu_j}^j & -2b_{\mu_j}^j & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & -2b_{2+\mu_{j-1}}^j & -2b_{3+\mu_{j-1}}^j & \cdots & -2b_{-1+\mu_j}^j & -2b_{\mu_j}^j & 0 & \cdots & 0 \end{bmatrix}.$$

We compute the determinant of $\mathcal{I}_B$ by cofactor expansion starting with row 2 of $\tilde{A}_j$, which contains a unique nonzero element $-2b_{\mu_j}^j$. Denote by C_{ij} the (i,j)-cofactor matrix. The row $2+\mu_{j-1}$ of $C_{(2+\ell_{j-1},\mu_j)}$ has a unique nonzero element $-2b_{-1+\mu_j}$, and we perform a cofactor expansion along this row. The row $2+\mu_{j-1}$ of this new cofactor matrix also has a unique nonzero element $-2b_{-2+\mu_j}^j$, and we proceed with the same process removing successively columns $3+\mu_{j-1}$ to μ_j (and the appropriate rows) until the cofactor matrix has only two rows corresponding to the original $\tilde{A}_j$ matrix and the second row has the unique nonzero element $-2b_{2+\mu_{j-1}}$ which is used to perform a cofactor expansion. Performing this process successively on each matrix $\tilde{A}_j$ for $j = 1, \ldots, r$, leaves as a cofactor matrix the $r \times r$ matrix $\mathcal{B}$ defined in the statement. The formula in the lemma is written using $\mu_j = \mu_{j-1} + \ell_j$, and so the lemma is proved. ∎

We are now ready to prove our Theorem 2.4.

Proof of Theorem 2.4. The proof of this theorem is similar to the proof of Theorem 2.1. However, more notation is needed and the details of some calculations are more elaborate because we are now dealing with r factors of the characteristic equation and the vector ω is separated in r subvectors of possibly unequal length.

A necessary and sufficient condition for the conclusion of the theorem to hold is that the following algebraic system of $2n$ equations has a solution in the $2n$ unknowns $(\tau_1, \tau_2, \ldots, \tau_n,$

$a_1, a_2, \ldots, a_n$):

$$\text{(38)}\qquad \begin{cases} \displaystyle\sum_{k=1}^{n} a_k \left(b_k^1 e^{-i\omega_\ell^1 \tau_k}\right) = i\omega_\ell^1, & \ell = 1, \ldots, \ell_1, \\ \displaystyle\sum_{k=1}^{n} a_k \left(b_k^2 e^{-i\omega_\ell^1 \tau_k}\right) = i\omega_\ell^2, & \ell = 1, \ldots, \ell_2, \\ \qquad\vdots \\ \displaystyle\sum_{k=1}^{n} a_k \left(b_k^r e^{-i\omega_\ell^1 \tau_k}\right) = i\omega_\ell^r, & \ell = 1, \ldots, \ell_r, \end{cases}$$

$$\text{(39)}\qquad \begin{cases} \displaystyle\sum_{k=1}^{n} a_k \left(b_k^1 e^{i\omega_\ell^1 \tau_k}\right) = -i\omega_\ell^1, & k = 1, \ldots, \ell_1, \\ \displaystyle\sum_{k=1}^{n} a_k \left(b_k^2 e^{i\omega_\ell^1 \tau_k}\right) = -i\omega_\ell^2, & \ell = 1, \ldots, \ell_2, \\ \qquad\vdots \\ \displaystyle\sum_{k=1}^{n} a_k \left(b_k^r e^{i\omega_\ell^1 \tau_k}\right) = -i\omega_\ell^r, & \ell = 1, \ldots, \ell_r. \end{cases}$$

We introduce the following notation to describe the above system of equations in matrix form. Let

$$\omega = (\omega_1^1, \ldots, \omega_{\ell_1}^1, \ldots, \omega_1^r, \ldots, \omega_{\ell_r}^r),$$

and define (39) as

$$\text{(40)}\qquad \begin{pmatrix} P(\tau;\omega) \\ P(-\tau;\omega) \end{pmatrix} A^T = \begin{pmatrix} i\omega^T \\ -i\omega^T \end{pmatrix},$$

where $A = (a_1, \ldots, a_n)$, superscript T denotes transpose, and $P(\tau;\omega) = P(\tau_1, \ldots, \tau_n; \omega)$ is the $n \times n$ matrix of the form

$$P(\tau;\omega) = \begin{bmatrix} P_1(\tau;\omega) \\ P_2(\tau;\omega) \\ \vdots \\ P_r(\tau;\omega) \end{bmatrix},$$

whose entry at block j, row ℓ, and column k is

$$\left[P_j(\tau;\omega)\right]_{\ell k} = b_k^j e^{-i\omega_\ell^j \tau_k}.$$

Note that $\overline{P(\tau;\omega)} = P(-\tau;\omega)$.

As in the proof of Theorem 2.1, we use the fact that since the ω_ℓ^k are rationally independent, then $\omega\tau_k$ taken modulo 2π generates a dense orbit, denoted $\mathcal{O}_k$, on a torus $\mathbf{T}^n$.

Just as in Theorem 2.1, we embed the problem into a mapping F associated to (40). Let

$$V = \{\Phi = (\Phi^1, \ldots, \Phi^r) \mid \Phi^j = (\Phi_1^j, \ldots, \Phi_j^n),\ \Phi_j^k = (\varphi_{1k}^j, \ldots, \varphi_{\ell_j k}^j),\ j = 1, \ldots, r\},$$

and define

$$F : V \times \mathbb{R}^n \longmapsto \mathbb{R}^{2n}$$

as

$$F(\Phi^1, \dots, \Phi^r, A; \omega) = \begin{pmatrix} \widetilde{P}(\Phi^1, \dots, \Phi^r) \\ \widetilde{P}(-\Phi^1, \dots, -\Phi^r) \end{pmatrix} A^T - i \begin{pmatrix} \omega^T \\ -\omega^T \end{pmatrix}, \tag{41}$$

where A and ω are as previously defined and

$$\widetilde{P}(\Phi^1, \dots, \Phi^r) = \begin{bmatrix} \widetilde{P}_1(\Phi^1) \\ \widetilde{P}_2(\Phi^2) \\ \vdots \\ \widetilde{P}_r(\Phi^r) \end{bmatrix}$$

with

$$\left[\widetilde{P}_j(\Phi^1, \dots, \Phi^n) \right]_{\ell k} = b_k^j e^{-i\varphi_{\ell k}^j}$$

for $j = 1, \dots, r$, $\ell = 1, \dots, \ell_j$, and $k = 1, \dots, n$. The definition of $\widetilde{P}$ uses the following coordinates of V.

We single out some coordinates as follows to facilitate the use of the implicit function theorem. Let $\Psi_j = \Phi_n^j$ and $\Psi = (\Psi_1, \dots, \Psi_r)$, and we now write

$$\Phi = (\Phi_o^1, \dots, \Phi_o^{n-1}),$$

where

$$\Phi_o^j = (\Phi_1^j, \dots, \Phi_{n-1}^j).$$

Thus, the mapping (41) can be written as $F(\Phi, \Psi, A; \omega)$.

We now find an explicit solution of $F = 0$ using the vectors v_j defined in the proof of Theorem 2.1. Denote by $\mathcal{I}^\ell$ the (invertible) $\ell \times \ell$ matrix whose jth column is the vector v_j^T.

We define $\mu_0 := 0$, $\mu_j := \sum_{i=1}^j \ell_i$, and

$$\Theta_{\ell_j} := (\Phi_{1+\mu_{j-1}}^j, \dots, \Phi_{\mu_j}^j).$$

We use the following base point in V. For $j = 1, \dots, r$ define $\hat{\Phi}_j$ be the point given by

$$\hat{\Theta}_{\ell_j} = \frac{-\pi}{2}((v_1, \dots, v_{\ell_j - 1}), v_{\ell_j})$$

and $\hat{\Phi}_i^j = -\frac{\pi}{2} v_1$ for $i \notin \{\mu_{1+\mu_{j-1}}, \dots, \mu_j\}$. In particular, $\hat{\Phi}_{\mu_j}^j = -\frac{\pi}{2} v_{\ell_j}$.

We now evaluate $\widetilde{P}(\widehat{\Phi}, \widehat{\Psi})$ by computing $\widetilde{P}_j(\widehat{\Phi}, \widehat{\Psi})$ for $j = 1, \dots, r$:

$$\widetilde{P}_j(\widehat{\Phi}, \widehat{\Psi}) = i \begin{bmatrix} b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & b_{2+\mu_{j-1}}^j & \cdots & b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \\ b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & b_{2+\mu_{j-1}}^j & \cdots & -b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ b_1^j & \cdots & b_{\mu_{j-1}}^j & b_{1+\mu_{j-1}}^j & -b_{2+\mu_{j-1}}^j & \cdots & -b_{\mu_j}^j & b_{\mu_j+1}^j & \cdots & b_n^j \end{bmatrix}.$$

Thus,

$$\widetilde{P}(\widehat{\Phi},\widehat{\Psi}) = \begin{bmatrix} \widetilde{P}_1(\widehat{\Phi},\widehat{\Psi}) \\ \vdots \\ \widetilde{P}_j(\widehat{\Phi},\widehat{\Psi}) \\ \vdots \\ \widetilde{P}_r(\widehat{\Phi},\widehat{\Psi}) \end{bmatrix} := i\mathcal{I}_B,$$

where $\mathcal{I}_B$ is invertible by assumption. In particular, $\widetilde{P}(-\widehat{\Phi},-\widehat{\Psi}) = -i\mathcal{I}_B$. We define

$$\widehat{A}^T \equiv (\hat{a}_1,\ldots,\hat{a}_n)^T = \mathcal{I}_B^{-1}\omega^T,$$

which leads to the solution:

$$F(\widehat{\Phi},\widehat{\Psi},\widehat{A};\omega) = 0.$$

Because the ω_j are rationally independent, it follows that the components $\hat{a}_k$ of $\widehat{A}$ are all nonzero.

We now show that we can use the implicit function theorem at the point found above. We define the $\ell \times \ell$ invertible matrix $\mathcal{U}_j$ to be the diagonal matrix whose kth diagonal element is the kth component of the vector v_j (in particular, $\mathcal{U}_1$ is the identity matrix). Note also that $\mathcal{U}_j^{-1} = \mathcal{U}_j$, $j = 1,\ldots,r$. We compute

$$J \equiv D_{(\Psi,A)}F(\widehat{\Phi},\widehat{\Psi},\widehat{A};\omega) = \begin{pmatrix} \hat{U} & i\mathcal{I}_B \\ \hat{U} & -i\mathcal{I}_B \end{pmatrix},$$

where

$$\hat{U} = \operatorname{diag}(\hat{a}_n b_n^1 \mathcal{U}_1^1,\ldots,\hat{a}_n b_n^{r-1}\mathcal{U}_1^{r-1},\hat{a}_n b_n^r \mathcal{U}_{\ell_r}^r)$$

is an $n \times n$ matrix with diagonal blocks of dimensions $\ell_1 \times \ell_1$ to $\ell_r \times \ell_r$. By the implicit function theorem, there exist a neighborhood N of $\widehat{\Phi}$ in V_Φ and a unique smooth function

$$\begin{aligned} G : N &\longmapsto V_\Psi \times \mathbb{R}^n, \\ G : \Phi &\longmapsto G(\Phi) = (G_\Psi(\Phi), G_A(\Phi)) \end{aligned}$$

such that

$$G(\widehat{\Phi}) = (\widehat{\Psi},\widehat{A})$$

and

$$F(\Phi, G(\Phi);\omega) \equiv 0 \quad \forall\, \Phi \in N. \tag{42}$$

Now that we have identified a set of solutions for $F = 0$, we wish to identify within this set solutions which lie on the dense orbits $\mathcal{O}_k$.

We show now that G_Ψ is regular at $\widehat{\Phi}$. A computation shows that

$$K \equiv D_\Phi F(\widehat{\Phi},\widehat{\Psi},\widehat{A};\omega) = \begin{pmatrix} \hat{K} \\ \hat{K} \end{pmatrix},$$

where

$$\hat{K} = \operatorname{diag}(\hat{K}_1, \ldots, \hat{K}_r)$$

is an $n \times (n-1)n$ matrix where the block $\hat{K}_j$ has dimensions $\ell_j \times (n-1)\ell_j$ and is of the form

$$\hat{K}_j = \Big(\begin{array}{cccccccc} \hat{a}_1\, b_1^j\, \mathcal{U}_1^j & \cdots & \hat{a}_{1+\mu_{j-1}}\, b_{1+\mu_{j-1}}^j\, \mathcal{U}_1^j & \hat{a}_{2+\mu_{j-1}}\, b_{2+\mu_{j-1}}^j\, \mathcal{U}_2^j & \cdots & \hat{a}_{\mu_j}\, b_{\mu_j}^j\, \mathcal{U}_{\ell_j}^j \\ \hat{a}_{1+\mu_j} b_{1+\mu_j}^j \mathcal{U}_1^j & \cdots & \hat{a}_{n-1} b_{n-1}^j \mathcal{U}_1^j & & & & & \end{array} \Big).$$

The matrix J is invertible and its inverse is

$$J^{-1} = \begin{pmatrix} \frac{1}{2}\hat{U}^{-1} & \frac{1}{2}\hat{U}^{-1} \\ \frac{-i}{2}\mathcal{I}_B^{-1} & \frac{i}{2}\mathcal{I}_B^{-1} \end{pmatrix}.$$

Implicit differentiation of (42) yields

$$\begin{aligned} DG(\widehat{\Phi}) &= \begin{pmatrix} DG_\Psi(\widehat{\Phi}) \\ DG_A(\widehat{\Phi}) \end{pmatrix} = -J^{-1}K \\ &= \begin{pmatrix} \operatorname{diag}(\mathcal{M}_1, \ldots, \mathcal{M}_r) \\ * \end{pmatrix}, \end{aligned} \tag{43}$$

where $*$ is not important for our purposes and the first component is an $n \times (n-1)n$ matrix composed of $n-1$ block matrices

$$\mathcal{M}_j = \Big(\begin{array}{cccccccc} -\frac{\hat{a}_1 b_1^j}{\hat{a}_{\mu_j} b_{\mu_j}^j} (\mathcal{U}_1^j)^2 & \cdots & -\frac{\hat{a}_{1+\mu_{j-1}} b_{1+\mu_{j-1}}^j}{\hat{a}_{\mu_j} b_{\mu_j}^j} (\mathcal{U}_1^j)^2 & -\frac{\hat{a}_{2+\mu_{j-1}} b_{2+\mu_{j-1}}^j}{\hat{a}_{\mu_j} b_{\mu_j}^j} \mathcal{U}_1^j \mathcal{U}_2^j & \cdots \\ -\frac{\hat{a}_{\mu_j} b_{\mu_j}^j}{\hat{a}_{\mu_j} b_{\mu_j}^j} \mathcal{U}_1^j \mathcal{U}_{\ell_j}^j & -\frac{\hat{a}_{1+\mu_j} b_{1+\mu_j}^j}{\hat{a}_{\mu_j} b_{\mu_j}^j} (\mathcal{U}_1^j)^2 & \cdots & -\frac{\hat{a}_{n-1} b_{n-1}^j}{\hat{a}_{\mu_j} b_{\mu_j}^j} (\mathcal{U}_1^j)^2 & & & & \end{array} \Big)$$

of dimension $\ell_j \times (n-1)\ell_j$, where $j = 1, \ldots, r-1$. Recall that $\mu_r = n$, so that we have

$$\mathcal{M}_r = \Big(\begin{array}{cccccccc} -\frac{\hat{a}_1 b_1^r}{\hat{a}_{\mu_r} b_{\mu_r}^r} \mathcal{U}_{\ell_r}^r \mathcal{U}_1^r & \cdots & -\frac{\hat{a}_{\mu_{r-1}} b_{\mu_{r-1}}^r}{\hat{a}_{\mu_r} b_{\mu_r}^r} \mathcal{U}_{\ell_r}^r \mathcal{U}_1^r & -\frac{\hat{a}_{1+\mu_{r-1}} b_{1+\mu_{r-1}}^r}{\hat{a}_{\mu_r} b_{\mu_r}^r} \mathcal{U}_{\ell_r}^r \mathcal{U}_1^r \\ -\frac{\hat{a}_{2+\mu_{r-1}} b_{2+\mu_{r-1}}^r}{\hat{a}_{\mu_r} b_{\mu_r}^r} \mathcal{U}_{\ell_r}^r \mathcal{U}_2^r & \cdots & -\frac{\hat{a}_{n-1} b_{n-1}^r}{\hat{a}_{\mu_r} b_{\mu_r}^r} \mathcal{U}_{\ell_r}^r \mathcal{U}_{\ell_r-1}^r & & & & & \end{array} \Big).$$

Consequently,

$$DG_\Psi(\widehat{\Phi}) = \operatorname{diag}(\mathcal{M}_1, \ldots, \mathcal{M}_r) \tag{44}$$

is nonsingular and it follows that the mapping

$$G_\Psi : N \longrightarrow V_\Psi$$

is regular at $\widehat{\Phi}$.

From the density of $\mathcal{O}_k$ in $\mathbf{T}^n$ for $k = 1, \ldots, n-1$, we know that for every $\epsilon > 0$ there exists $(\tau_{1,\epsilon}, \ldots, \tau_{n-1,\epsilon})$ such that

$$\Phi_\epsilon^* = (\tau_{1,\epsilon}\omega, \ldots, \tau_{n-1,\epsilon}\omega) \bmod 2\pi$$

is in an ϵ-neighborhood of $\widehat{\Phi}$ in N, and we define a small $(n-1)$-dimensional surface in V_Φ based at Φ^*_ϵ by

$$S^h_{\Phi^*_\epsilon} = \{(\tau_1\omega, \dots, \tau_{n-1}\omega) \bmod 2\pi \mid \tau_j \in (\tau_{j,\epsilon} - h, \tau_{j,\epsilon} + h)\}$$

with $\epsilon, h > 0$ small enough so that $S^h_{\Phi^*_\epsilon} \subset N$. We now show that the image of $S^h_{\Phi^*_\epsilon}$ by G_Ψ has a nontrivial transversal intersection with $\mathcal{O}_\Psi$.

Consider the following $n-1$ vectors in $(\mathbb{R}^{\ell_1})^{n-1} \times (\mathbb{R}^{\ell_2})^{n-1} \times \cdots \times (\mathbb{R}^{\ell_r})^{n-1} \simeq (\mathbb{R}^n)^{n-1}$:

$$\begin{aligned}
W_1 &= (\omega^1, 0, \dots, 0; \omega^2, 0, \dots, 0; \dots; \omega^r, 0, \dots, 0),\\
W_2 &= (0, \omega^1, \dots, 0; 0, \omega^2, \dots, 0; \dots, 0, \omega^r, \dots, 0),\\
&\vdots\\
W_{n-2} &= (0, \dots, \omega^1, 0; 0, \dots, \omega^2, 0; \dots; 0, \dots, \omega^r, 0),\\
W_{n-1} &= (0, \dots, 0, \omega^1; 0, \dots, 0, \omega^2; \dots, 0, \dots, 0, \omega^r),
\end{aligned}$$

where 0 represents the 0 vector in the respective space $\mathbb{R}^{\ell_j}$, and we recall that

$$(\omega^1, \dots, \omega^r) = (\omega^1_1, \dots, \omega^1_{\ell_1}, \dots, \omega^r_1, \dots, \omega^r_{\ell_r}).$$

The set $\{W_1, \dots, W_{n-1}\}$ is linearly independent. We consider the function

$$\mathcal{T} : N \longrightarrow \mathbb{R}$$

defined by

$$\mathcal{T}(\Phi) = \det\begin{pmatrix} DG_\Psi(\Phi)\cdot W_1^T & DG_\Psi(\Phi)\cdot W_2^T & \cdots & DG_\Psi(\Phi)\cdot W_{n-1}^T & \omega^T \end{pmatrix}, \tag{45}$$

and recalling that $(\mathcal{U}^i_j)^2 = I$ for all j, i we compute

$$\begin{aligned}
\mathcal{T}(\hat{\Phi}) &= \det\begin{pmatrix} DG_\Psi(\hat{\Phi})\cdot W_1^T & DG_\Psi(\hat{\Phi})\cdot W_2^T & \cdots & DG_\Psi(\hat{\Phi})\cdot W_{n-1}^T & \omega^T \end{pmatrix}\\
&= \det(\alpha_{jk}),
\end{aligned} \tag{46}$$

where $j = 1, \dots, r$, $k = 1, \dots, n$. The elements of the matrix (α_{jk}) are

$$\alpha_{1k} = \begin{cases} -\frac{\hat{a}_k b^1_k}{\hat{a}_{\mu_1} b^1_{\mu_1}} \mathcal{U}^1_1 \mathcal{U}^1_k (\omega^1)^T, & k = 1, \dots, \ell_1,\\ -\frac{\hat{a}_k b^1_k}{\hat{a}_{\mu_1} b^1_{\mu_1}} (\mathcal{U}^1_1)^2 (\omega^1)^T, & k = 1 + \ell_1, \dots, n-1,\\ \frac{\hat{a}_{\mu_1} b^1_{\mu_1}}{\hat{a}_{\mu_1} b^1_{\mu_1}} (\mathcal{U}^1_1)^2 (\omega^1)^T, & k = n, \end{cases}$$

and for $j = 2, \dots, r-1$

$$\alpha_{jk} = \begin{cases} -\frac{\hat{a}_k b^j_k}{\hat{a}_{\mu_j} b^j_{\mu_j}} (\mathcal{U}^j_1)^2 (\omega^j)^T, & k = 1, \dots, \mu_{j-1}, \quad \text{and} \quad k = \mu_j + 1, \dots, n-1,\\ -\frac{\hat{a}_k b^j_k}{\hat{a}_{\mu_j} b^j_{\mu_j}} \mathcal{U}^j_1 \mathcal{U}^j_{k-\mu_{j-1}} (\omega^j)^T, & k = 1 + \mu_{j-1}, \dots, \mu_j,\\ \frac{\hat{a}_{\mu_j} b^j_{\mu_j}}{\hat{a}_{\mu_j} b^j_{\mu_j}} (\mathcal{U}^j_1)^2 (\omega^j)^T, & k = n, \end{cases}$$

and finally

$$\alpha_{rk} = \begin{cases} -\frac{\hat{a}_k b^r_k}{\hat{a}_{\mu_r} b^r_{\mu_r}} \mathcal{U}^r_{\ell_r} \mathcal{U}^r_1 (\omega^r)^T, & k = 1, \dots, \mu_{r-1}, \\ -\frac{\hat{a}_k b^r_k}{\hat{a}_{\mu_r} b^r_{\mu_r}} \mathcal{U}^r_{\ell_r} \mathcal{U}^r_{k-\mu_{r-1}} (\omega^r)^T, & k = 1 + \mu_{r-1}, \dots, \mu_r - 1, \\ \frac{\hat{a}_{\mu_r} b^r_{\mu_r}}{\hat{a}_{\mu_r} b^r_{\mu_r}} (\mathcal{U}^r_{\ell_r - 1})^2 (\omega^r)^T, & k = n, \end{cases}$$

where we recall that $\mu_r = n$. Note that the elements of the last column are rewritten so as to lead to the significant simplification of the determinant to the following form:

(47)
$$\mathcal{T}(\hat{\Phi}) = \frac{(-1)^{n-1} \omega^1_1 \cdots \omega^1_{\ell_1} \cdots \omega^r_1 \cdots \omega^r_{\ell_r} \hat{a}_1 \cdots \hat{a}_{n-1}}{(\hat{a}_{\mu_1} b^1_{\mu_1})^{\ell_1} \cdots (\hat{a}_{\mu_r} b^r_{\mu_r})^{\ell_r}} \det(\operatorname{diag}(\mathcal{U}^1_1, \mathcal{U}^2_1, \dots, \mathcal{U}^{r-1}_1, \mathcal{U}^r_{\ell_r})) \det \mathcal{I}'_B,$$

where

$$\mathcal{I}'_B = \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_r \end{pmatrix},$$

and for $j = 1, \dots, r$

$$Q_j = \begin{bmatrix} b^j_1 & \cdots & b^j_{\mu_{j-1}} & b^j_{1+\mu_{j-1}} & b^j_{2+\mu_{j-1}} & \cdots & b^j_{\mu_j} & b^j_{\mu_j+1} & \cdots & \hat{a}_{\mu_j} b^j_{\mu_j} \\ b^j_1 & \cdots & b^j_{\mu_{j-1}} & b^j_{1+\mu_{j-1}} & b^j_{2+\mu_{j-1}} & \cdots & -b^j_{\mu_j} & b^j_{\mu_j+1} & \cdots & \hat{a}_{\mu_j} b^j_{\mu_j} \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ b^j_1 & \cdots & b^j_{\mu_{j-1}} & b^j_{1+\mu_{j-1}} & -b^j_{2+\mu_{j-1}} & \cdots & -b^j_{\mu_j} & b^j_{\mu_j+1} & \cdots & \hat{a}_{\mu_j} b^j_{\mu_j} \end{bmatrix}$$

is an $\ell_j \times n$ matrix. Moreover, $\det \mathcal{I}'_B \neq 0$ since $\det \mathcal{I}'_B = \hat{a}_n \det \mathcal{I}_B$ and $\hat{a}_n \neq 0$. Thus $\mathcal{T}(\hat{\Phi}) \neq 0$, and the conclusion is exactly the same as in Theorem 2.1; hence the theorem is proved. ∎

5. Conclusion. We have shown in this paper that n nonresonant eigenvalues on the imaginary axis can be realized by a scalar DDE with n delays. Moreover, the same is true for any collection of n imaginary eigenvalues in a neighborhood of an n-tuple of nonresonant imaginary eigenvalues. We have also shown how these results can be applied to nonscalar DDEs in the context of symmetric DDEs, where the characteristic equation decomposes according to the isotypic decomposition. We apply our result to delay coupled $\mathbf{D}_n$-symmetric cell systems with n odd.

There are several ways of extending the main result of our paper. One question we did not address in this paper is if n nonresonant nonzero imaginary eigenvalues constitute an upper bound for the realizability by a scalar equation with n delay. The case $n = 1$ is one such example since an easy calculation shows that we can have at most one imaginary eigenvalue on the imaginary axis. It is likely, but unknown, that this is also true for general n.

One may want to study whether k zero eigenvalues in a single Jordan block and ℓ nonresonant nonzero imaginary eigenvalues can be realized in a scalar DDE with $k + \ell$ delays. This problem may be feasible by modifying the proof of Theorem 2.1 since the nonresonance of the ℓ eigenvalues is again present. However, we expect the argument used in this paper to

break down for n nonzero imaginary eigenvalues with resonance. We can also study the same problem as in this paper but for higher-dimensional delay equations. One problem would be to find out if n nonresonant nonzero imaginary eigenvalues can be realized by an m-dimensional system with k delays. For instance, it is known that a pair of nonzero imaginary eigenvalues can be realized by a two-dimensional equation with one delay [8]. In this case $n = 2$, $m = 2$, $k = 1$, and so $n = mk$; is it possible to realize three nonresonant nonzero eigenvalues, or does the relationship $n = mk$ provide a bound to realizability in general?

Another problem which can be studied is whether a restriction in the class of delay equation can change the realizability. For instance, the characteristic equation for a general two-dimensional system with one-delay τ is

$$\lambda^2 + a\lambda + b\lambda e^{-\lambda\tau} + c + de^{-\lambda\tau} = 0,$$

while for a second-order equation with one delay in the feedback term we must set $b = 0$. We know in this case that two nonresonant nonzero imaginary eigenvalues can be realized by a second-order equation with a unique delay in the feedback term [4]. In fact, two imaginary eigenvalues with $1 : 2$ resonance have been found in such an equation [5]. The obvious question is to see if n nonresonant (and resonant) nonzero imaginary eigenvalues can be realized within the class of nth-order scalar equations with one delay in the feedback term.

An extension of our main result in a direction relevant for studying bifurcations is to find out whether the n nonresonant nonzero imaginary eigenvalues can be realized by a scalar delay equation such that the remaining eigenvalues have negative real parts; that is, the multiple Hopf point lies at the boundary of the stability region for the equilibrium solution. This is typically verified in stability analysis of specific equations. For scalar problems see Bélair and Campbell [1] for a thorough analysis of a two-delay case and a review of several other cases. Yuan and Campbell [31] study the stability regions for $\mathbf{D}_n$-symmetric rings of scalar cells with nearest neighbor delay coupling and simultaneously obtain the location of multiple Hopf bifurcation points at the boundary of the stability region.

Finally, let us mention the case of linear T-periodic equations with $N + 1$ delays:

$$\dot{x} = \sum_{j=0}^{N} a_j(t)x(t - \tau_j), \tag{48}$$

where each $a_j(t)$ is a T-periodic $n \times n$ matrix. Hale [19] states the following open problem:

> *Is it possible to give a precise upper bound in terms of N on the number of Floquet multipliers of (48) that can have moduli 1?*

If we restrict (48) to scalar equations, we can pose a possibly simpler problem which is related to the main result of our paper:

> *Is it possible to realize $N + 1$ complex numbers $e^{\pm i\omega_1}, \ldots, e^{\pm i\omega_{N+1}}$ with $\omega_1, \ldots, \omega_{N+1}$ positive and rationally independent as Floquet multipliers of the scalar equation (48)?*

This problem is automatically solved by Theorem 2.1 if a Floquet theorem can be applied to (48); that is, the Floquet exponents of the Floquet multipliers of (48) are eigenvalues of a scalar equation (1) with $N + 1$ delays. Such a theorem has not been proved in general; however, it may hold true given that some conditions are imposed on (48).

Acknowledgment. We would like to thank the referees for their suggestions concerning the contents and presentation of the paper.

REFERENCES

[1] J. Bélair and S. A. Campbell, *Stability and bifurcations of equilibria in a multiple-delayed differential equation*, SIAM J. Appl. Math., 54 (1994), pp. 1402–1424.
[2] A. Beuter, J. Bélair, and C. Labrie, *Feedback and delays in neurological diseases: A modeling study using dynamical systems*, Bull. Math. Biol., 55 (1993), pp. 525–541.
[3] A. Beuter, L. Glass, M. Mackey, and M. Titcombe, eds., *Nonlinear Dynamics in Physiology and Medicine*, Interdiscip. Appl. Math. 25, Springer-Verlag, New York, 2003.
[4] S. A. Campbell, J. Bélair, T. Ohira, and J. Milton, *Limit cycles, tori, and complex dynamics in a second-order differential equation with delayed negative feedback*, J. Dynam. Differential Equations, 7 (1995), pp. 213–236.
[5] S. A. Campbell and V. G. LeBlanc, *Resonant Hopf-Hopf interactions in delay differential equations*, J. Dynam. Differential Equations, 10 (1998), pp. 327–346.
[6] S. A. Campbell, Y. Yuan, and S. Bungay, *Equivariant Hopf bifurcation in a ring of identical cells with delayed coupling*, Nonlinearity, 18 (2005), pp. 2827–2846.
[7] Y. Choi and V.G. LeBlanc, *Toroidal normal forms for bifurcations in retarded functional differential equations. I. Multiple Hopf and transcritical/multiple Hopf interaction*, J. Differential Equations, 227 (2006), pp. 166–203.
[8] K. Cooke and Z. Grossman, *Discrete delay, distributed delay and stability switches*, J. Math. Anal. Appl., 86 (1982), pp. 592–627.
[9] O. Diekmann, S. A. van Gils, S. M. Verduyn-Lunel, and H. O. Walther, *Delay-Equations, Functional, Complex, and Nonlinear Analysis*, Appl. Math. Sci. 110, Springer-Verlag, New York, 1995.
[10] R. Dodla, A. Sen, and G. L. Johnston, *Phase-locked patterns and amplitude death in a ring of delay-coupled limit cycle oscillators*, Phys. Rev. E (3), 69 (2004), 056217.
[11] T. Faria and L. T. Magalhães, *Realisation of ordinary differential equations by retarded functional-differential equations in neighbourhoods of equilibrium points*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 759–776.
[12] L. Glass and M. C. Mackey, *Oscillations and chaos in physiological control systems*, Science, 197 (1977), pp. 287–289.
[13] M. Golubitsky and I. Stewart, *Nonlinear dynamics of networks: The groupoid formalism*, Bull. Amer. Math. Soc. (N.S.), 43 (2006), pp. 305–364.
[14] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory, Vol.* II, Appl. Math. Sci. 69, Springer-Verlag, New York, 1988.
[15] S. Guo, *Spatio-temporal patterns of nonlinear oscillations in an excitatory ring network with delay*, Nonlinearity, 18 (2005), pp. 2391–2407.
[16] S. Guo and L. Huang, *Hopf bifurcating periodic orbits in a ring of neurons with delays*, Phys. D, 183 (2003), pp. 19–44.
[17] S. Guo and L. Huang, *Stability of nonlinear waves in a ring of neurons with delays*, J. Differential Equations, 236 (2007), pp. 343–374.
[18] M. S. Gurney, S. P. Blythe, and R. M. Nisbee, *Nicholson's blowflies revisited*, Nature, 287 (1980), pp. 17–21.
[19] J. K. Hale, *Some problems in FDE*, in Topics in Functional Differential and Difference Equations, Fields Inst. Commun. 29, AMS, Providence, RI, 2001, pp. 195–222.
[20] J. K. Hale and S. M. Verduyn-Lunel, *Introduction to Functional Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
[21] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Math. Sci. Engrg. 191, Academic Press, Boston, 1993.
[22] Y. Kuznetsov, *Elements of Applied Bifurcation Theory*, 3rd ed., Appl. Math. Sci. 112, Springer-Verlag, New York, 2004.

[23] R. Lang and K. Kobayashi, *External optical feedback effects on semiconductor injection laser properties*, IEEE J. Quantum Electronics, 16 (1980), pp. 347–355.
[24] M. Peng, *Bifurcation and stability analysis of nonlinear waves in* $\mathbf{D}_n$*-symmetric delay-differential systems*, J. Differential Equations, 232 (2007), pp. 521–543.
[25] M. Peng and Y. Yuan, *Complex dynamics in discrete delayed models with* $\mathbf{D}_4$*-symmetry*, Chaos Solitons Fractals, 37 (2008), pp. 393–408.
[26] B. F. Redmond, V. G. LeBlanc, and A. Longtin, *Bifurcation analysis of a class of first-order nonlinear delay-differential equations with reflectional symmetry*, Phys. D, 166 (2002), pp. 131–146.
[27] E. Stone and A. Askari, *Nonlinear models of chatter in drilling processes*, Dyn. Syst., 17 (2002), pp. 65–85.
[28] M. J. Suarez and P. L. Schopf, *A delayed action oscillator for ENSO*, J. Atmospheric Sci., 45 (1988), pp. 3283–3287.
[29] J. Wu, *Symmetric functional-differential equations and neural networks with memory*, Trans. Amer. Math. Soc., 350 (1998), pp. 4799–4838.
[30] J. Wu, T. Faria, and Y. S. Huang, *Synchronization and stable phase-locking in a network of neurons with memory*, Math. Comput. Modelling, 30 (1999), pp. 117–138.
[31] Y. Yuan and S. A. Campbell, *Stability and synchronization of a ring of identical cells with delayed coupling*, J. Dynam. Differential Equations, 16 (2004), pp. 709–744.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1355–1377
© 2008 Society for Industrial and Applied Mathematics

Singular Hopf Bifurcation in Systems with Two Slow Variables*

John Guckenheimer†

Abstract. Hopf bifurcations have been studied intensively in two dimensional vector fields with one slow and one fast variable [É. Benoît et al., *Collect. Math.*, 31 (1981), pp. 37–119; F. Dumortier and R. Roussarie, *Mem. Amer. Math. Soc.*, 121 (577) (1996); W. Eckhaus, in *Asymptotic Analysis* II, Lecture Notes in Math. 985, Springer-Verlag, Berlin, 1983, pp. 449–494; M. Krupa and P. Szmolyan, *SIAM J. Math. Anal.*, 33 (2001), pp. 286–314; J. Guckenheimer, in *Normal Forms, Bifurcations and Finiteness Problems in Differential Equations*, NATO Sci. Ser. II Math. Phys. Chem. 137, Kluwer, Dordrecht, The Netherlands, 2004, pp. 295–316]. *Canard explosions* are associated with these *singular Hopf* bifurcations [S. M. Baer and T. Erneux, *SIAM J. Appl. Math.*, 46 (1986), pp. 721–739; S. M. Baer and T. Erneux, *SIAM J. Appl. Math.*, 52 (1992), pp. 1651–1664; B. Braaksma, *J. Nonlinear Sci.*, 8 (1998), pp. 457–490; Y. Lijun and Z. Xianwu, *J. Differential Equations*, 206 (2004), pp. 30–54], manifested by a very rapid growth in the amplitude of periodic orbits. There has been less analysis of Hopf bifurcations in slow-fast systems with two slow variables where singular Hopf bifurcation occurs simultaneously with *type* II *folded saddle-nodes* [A. Milik and P. Szmolyan, in *Multiple-Time-Scale Dynamical Systems*, IMA Vol. Math. Appl. 122, Springer-Verlag, New York, 2001, pp. 117–140; M. Wechselberger, *SIAM J. Appl. Dyn. Syst.*, 4 (2005), pp. 101–139]. This work contributes to our understanding of these Hopf bifurcations in five ways: (1) it computes the first Lyapunov coefficient of the bifurcation in terms of a normal form, (2) it describes global features of the flow that constrain the types of trajectories found in the system near the bifurcation, (3) it identifies codimension two bifurcations that occur as coefficients in the normal form vary, (4) it exhibits complex solutions that occur in the vicinity of the bifurcation for some values of the normal form coefficients, and (5) it identifies singular Hopf bifurcation as a mechanism for the creation of mixed-mode oscillations. A subtle aspect of the normal form is that terms of higher order contribute to the first Lyapunov coefficient of the bifurcation in an essential way.

Key words. Hopf bifurcation, mixed mode oscillation, singular perturbation

AMS subject classifications. 37M20, 34E13, 37G10

DOI. 10.1137/080718528

1. Introduction.

Slow-fast vector fields have the form

$$\begin{aligned} \varepsilon \dot{x} &= f(x, y, \varepsilon), \\ \dot{y} &= g(x, y, \varepsilon), \end{aligned} \tag{1.1}$$

with $x \in R^m$ as the fast variable, $y \in R^n$ as the slow variable, and ε as a small parameter that represents the ratio of time scales. The set of points satisfying $f = 0$ is the *critical manifold* of the system: slow motion of trajectories can occur only near the critical manifold. Fenichel theory [14] establishes that there are invariant slow manifolds of the system near portions of the critical manifold, where $D_x f$ is hyperbolic. Moreover, the trajectories on the

*Received by the editors March 14, 2008; accepted for publication (in revised form) by T. Kaper July 10, 2008; published electronically October 31, 2008.
http://www.siam.org/journals/siads/7-4/71852.html
†Mathematics Department, Cornell University, Ithaca, NY 14853 (jmg16@cornell.edu).

slow manifold approach trajectories of the *slow flow* $\dot{y} = g(h(y), y, 0)$ on the critical manifold where h is defined implicitly by $f(h(y), y, 0) = 0$. Points of the critical manifold where $D_x f$ is singular are called *fold* points. In much of the literature on slow-fast models of neural systems, the fold points are called "knees" [29].

Hopf bifurcation occurs in the following slow-fast system with one slow and one fast variable:

$$\begin{aligned} \varepsilon \dot{x} &= y - x^2/2 - x^3/3, \\ \dot{y} &= \mu - x. \end{aligned} \tag{1.2}$$

This system has an equilibrium point at $(\mu, \mu^2/2 + \mu^3/3)$ that undergoes a supercritical Hopf bifurcation as μ decreases through zero. The Hopf equilibrium is at the fold of the system. We also note that the slow flow along the critical manifold has a stable equilibrium on a stable branch of the critical manifold for $\mu > 0$ but an unstable equilibrium on the unstable branch of the critical manifold for $-1 < \mu < 0$. The periodic orbits that emerge from the Hopf bifurcation grow explosively from an amplitude that is $O(\varepsilon^{1/2})$ to an amplitude that is $O(1)$ over a range of values of μ that has length $O(\exp(-c/\varepsilon))$ for a constant $c > 0$ independent of ε. This *canard explosion* was discovered by Benoît et al. [7] and subsequently analyzed by Eckhaus [13], Dumortier and Roussarie [12], and others [23, 15].

Consider now another system with one slow and one fast variable:

$$\begin{aligned} \varepsilon \dot{x} &= y - x^2, \\ \dot{y} &= \mu - x + ay. \end{aligned} \tag{1.3}$$

This system also has a Hopf bifurcation, but it occurs when $x = \varepsilon a/2$, $y = \varepsilon^2 a^2/4$, and $\mu = \varepsilon a/2 - \varepsilon^2 a^3/4$. Its point of Hopf bifurcation is on the critical manifold but displaced from the fold by a distance that is $O(\varepsilon)$. When $a \neq 0$, the periodic orbits in the canard explosion of this system grow monotonically with variations of μ like those of system (1.2), but they become unbounded as μ varies over a finite interval. Whether the bifurcation is subcritical or supercritical is determined by the sign of a. When $a = 0$, the singular Hopf bifurcation at $\mu = 0$ is totally degenerate: system (1.3) has a family of periodic orbits that are level curves of the function $H(x, y) = \exp(-2y/\varepsilon)(y - x^2 + \varepsilon/2)$. In this case, the parabola $H = 0$ contains the stable and unstable slow manifolds of the system, and it bounds the family of periodic orbits.

The system (1.3) can be rescaled by $x = \varepsilon^{1/2} X$, $y = \varepsilon Y$, and $t = \varepsilon^{1/2} T$ to give

$$\begin{aligned} X' &= Y - X^2, \\ Y' &= \varepsilon^{-1/2}\mu - X + \varepsilon^{1/2} aY. \end{aligned} \tag{1.4}$$

This system can be transformed to the Hopf normal form

$$\begin{aligned} r' &= \frac{\varepsilon^{1/2} a}{16\sqrt{1 - \varepsilon a^2}} r^3 + o(r^3), \\ \theta' &= \sqrt{(1 - \varepsilon a^2)} + o(r) \end{aligned} \tag{1.5}$$

at the equilibrium.

System (1.3) is representative of generic singular Hopf bifurcations [2, 3, 8, 25] with one slow and one fast variable. There are three aspects of the bifurcation that are directly influenced by multiple time scales:

- The bifurcation occurs at a distance that is $O(\varepsilon)$ from a fold point.
- The periodic orbits emanating from the Hopf bifurcation undergo a canard explosion.
- The slow stable and unstable manifolds of the system cross each other as a varies.

Tangential intersections of the slow stable and unstable manifolds are not bifurcations in traditional terms, but rather a degeneracy in the slow-fast decomposition of the system [15] comparable to a homoclinic/heteroclinic bifurcation. Generically, such tangencies occur at different parameter values from those where the equilibrium point is on a fold curve or at the Hopf bifurcation parameter value. The crossing marks the transition from parameters at which the slow stable manifold converges to the equilibrium or periodic orbit and parameters for which it jumps along the fast direction after approaching the vicinity of the equilibrium. This transition is one of the most significant changes in dynamical behavior associated with the singular Hopf bifurcation.

Singular Hopf bifurcations with two slow variables and one fast variable are analogous to system (1.3) with a single slow variable. There are counterparts to each of the three properties listed above. Equilibrium points of a system with two slow variables lie on its two dimensional critical manifold. The folds of the critical manifold form a curve. A stable equilibrium point of the system (1.1) may approach and cross the fold curve in a generic manner when a single parameter is varied. If it does so, Hopf bifurcation occurs at a distance $O(\varepsilon)$ from the fold curve. Canard explosions also occur, but the dimension of the state space is now large enough to allow period-doubling and torus bifurcations as the periodic orbits grow. Section 3 gives examples of each of these bifurcations. In systems with two slow variables and one fast variable, the slow stable and unstable manifolds are each two dimensional and therefore can intersect transversally along a trajectory in the three dimensional state space. These intersections occur for open sets of parameters and are a common feature of systems near singular Hopf bifurcations. The location of the slow stable and unstable manifold intersections helps determine whether there are bounded attractors near the singular Hopf bifurcation.

This paper explores the dynamics of singular Hopf bifurcation via analysis of normal forms. Coordinate changes and scaling suggest a normal form analogous to (1.3), but the normal form has four coefficients that cannot be scaled to fixed values. If these coefficients are regarded as parameters (or *moduli*), then degenerate Hopf bifurcations occur in the system for some values of these coefficients. Regarding one of the coefficients as a second parameter, the theory of codimension two bifurcations [18] can be used to investigate the dynamics. In some cases, higher order terms in ε must be retained in a rescaled normal form to obtain nondegeneracy of the codimension two bifurcations.

Singular Hopf bifurcation with two slow variables has been studied previously in other papers [3, 8, 25, 27]. The normal form used by Braaksma [8] differs from the one used here: one difference is that Braaksma's normal form has "global returns" of trajectories that leave the vicinity of the equilibrium point in the flow. System (1.2) also has global returns but just one slow variable. This paper emphasizes the role of singular Hopf bifurcation in the creation of certain types of mixed-mode oscillations (MMOs). MMOs are oscillations in which there are small and large amplitude cycles in each period of the oscillation. Singular Hopf bifurca-

tions produce small oscillations near the equilibrium point that can be combined with global returns to create MMOs. MMOs appear to have been studied first in chemically reacting systems [4, 5, 22, 27] and then in lasers [21]. More recently, MMOs have been studied in neural oscillations [9], where they are sometimes associated with folded nodes [31, 33, 17] as well as singular Hopf bifurcations. Some of the subsidiary bifurcations analyzed here have been observed in the models of chemical oscillators, but their relationship to singular Hopf bifurcation does not seem to have been noticed previously. Section 4 discusses the "autocatalator" model analyzed by Petrov, Scott, and Showalter [28] and Milik and Szmolyan [27].

2. Coordinate changes and normal forms. The goal of this section is to derive a normal form for a generic system with two slow variables and one fast variable with an equilibrium point that crosses a simple fold transversally. We denote x as the fast variable and (y, z) as the slow variables. The fast equation for such a system near a simple fold can be reduced to $\varepsilon\dot{x} = y - x^2$ [1], perhaps using a rescaling of time. This is our starting point for deriving a normal form for singular Hopf bifurcation. We approximate the system by truncating nonlinear terms in the Taylor series expressions for $\dot{y}$ and $\dot{z}$. The truncated system is further reduced by noting that if $\dot{y} = \alpha + \beta x + \gamma y + \delta z$, then replacing z by $w = \alpha + \gamma y + \delta z$ makes $\dot{y} = \beta x + w$ while $\dot{w}$ is still an affine function of (x, y, z). We relabel w as z. Hopf bifurcation occurs when $\beta < 0$. Rescaling (x, y, z, t) by $(|\beta|^{1/2}, |\beta|, |\beta|^{3/2}, |\beta|^{-1/2})$ makes a further reduction to the case that $\beta = -1$. These coordinate transformations yield a truncated system of the form

$$\begin{aligned} \varepsilon\dot{x} &= y - x^2, \\ \dot{y} &= z - x, \\ \dot{z} &= -\mu - ax - by - cz. \end{aligned} \tag{2.1}$$

Note that a detailed study of higher degree normal forms for singular Hopf bifurcation with two slow variables does not appear in the literature. The term singular point is used in Arnold et al. [1] to refer to folded singularities [15] (pseudosingularities in Benoît [6]) that are regular points of the vector field (1.1) when $\varepsilon > 0$. A final rescaling $(x, y, z, t) = (\varepsilon^{1/2}X, \varepsilon Y, \varepsilon^{1/2}Z, \varepsilon^{1/2}T)$ and $(A, B, C) = (\varepsilon^{1/2}a, \varepsilon b, \varepsilon^{1/2}c)$ eliminates ε from the system:

$$\begin{aligned} X' &= Y - X^2, \\ Y' &= Z - X, \\ Z' &= -\mu - AX - BY - CZ. \end{aligned} \tag{2.2}$$

Our numerical studies and bifurcation analysis will be conducted largely with system (2.2). Note that A, B, and C tend to zero as $\varepsilon \to 0$ and that B tends to zero faster than A and C. The extent to which nonlinear terms in the equations for $\dot{y}$ and $\dot{z}$ that have order $\varepsilon^{1/2}$ following rescaling influence the dynamics described in this paper has not been investigated.

The "desingularized" slow flow of system (2.1) is

$$\begin{aligned} \dot{z} &= -2x(\mu + ax + bx^2 + cz), \\ \dot{x} &= z - x. \end{aligned} \tag{2.3}$$

This equation is obtained from (2.1) by setting $\varepsilon = 0$, differentiating the resulting equation $y - x^2 = 0$ to obtain $\dot{y} = 2x\dot{x}$, then eliminating $\dot{y}$ from the second equation and finally rescaling the equation by $2x$.

3. Normal form dynamics and flow maps. This section investigates the dynamics of the systems (2.1) and (2.2). As μ varies near zero in system (2.1), the equilibrium point crosses the fold curve of the critical manifold at the origin. This crossing has been called a *folded saddle-node type* II by Milik and Szmolyan [27] because the slow flow (2.3) has a degenerate equilibrium point at this parameter value. The bifurcation in the slow flow is a transcritical bifurcation. The origin is always a folded singularity (or pseudosingularity) that is a saddle when $\mu < 0$, a node when $0 < \mu < 1/8$, and a focus when $1/8 < \mu$. While the folded saddle-node appears as the main change in the dynamics of the slow flow, Hopf bifurcations of the systems (2.1) and (2.2) typically occur at nonzero values of μ. The equilibrium point of system (2.1) undergoes Hopf bifurcation at a value of μ that is $O(\varepsilon)$. Much of the analysis in this section is devoted to exploring this Hopf bifurcation and the family of periodic orbits emerging from it.

3.1. Hopf bifurcation. Equilibria of system (2.2) occur at points (X_e, X_e^2, X_e) with $\mu = -AX_e - BX_e^2 - CX_e$. The Jacobian at this equilibrium is the matrix

$$\begin{pmatrix} -2X_e & 1 & 0 \\ -1 & 0 & 1 \\ -A & -B & -C \end{pmatrix},$$

whose characteristic polynomial is

$$P(s) = s^3 + (C + 2X_e)s^2 + (B + 2X_eC + 1)s + (A + 2X_eB + C).$$

Thus Hopf bifurcation of the system occurs where $(B + 2X_eC + 1) > 0$ and

$$(C + 2X_e)(B + 2X_eC + 1) = (A + 2X_eB + C).$$

Note that $(B+2X_eC+1) > 0$ is satisfied when B and C are small. Thus, the Hopf bifurcation locus of system (2.2) is parametrized by the equations

$$\begin{aligned} A &= BC + 2X_eC^2 + 4X_e^2C + 2X_e, \\ \mu &= -AX_e - BX_e^2 - CX_e \end{aligned} \tag{3.1}$$

in terms of the equilibrium position X_e and the parameters B, C. If $B = 0$, then $A = 2X_eC^2 + 4X_e^2C + 2X_e$ and $\mu = -(2X_eC^2 + 4X_e^2C + 2X_e + C)X_e$. Since B is $O(\varepsilon)$ while A and C are $O(\varepsilon^{1/2})$, zero eigenvalues of the equilibrium occur near the origin only if $a + c$ is small in system (2.1).

The program Maple [26] has been used to compute the Hopf normal form of system (2.2). Consider first the case $B = 0$. System (2.2) can then be transformed to its Hopf normal form by rational coordinate changes if A, C, and μ are parametrized by X_e and ω, the magnitude of its imaginary eigenvalue. Whether the bifurcation is subcritical or supercritical is determined

by the sign of the first Lyapunov coefficient.[1] Maple yields the following expression for the first Lyapunov coefficient:

$$4\,\frac{{X_e}^3\left(12\,\omega^2{X_e}^2-4\,{X_e}^2-2\,\omega^2+\omega^4+1\right)}{\left(1-2\,\omega^2+\omega^4+12\,\omega^2{X_e}^2-8\,{X_e}^2+16\,{X_e}^4\right)\left(16\,{X_e}^4-8\,{X_e}^2+24\,\omega^2{X_e}^2-2\,\omega^2+\omega^4+1\right)}.$$

Substituting $\omega = \sqrt{1+2X_eC}$ and $X_e = 1/8\,\frac{-2\,C^2-2+2\,\sqrt{C^4+2\,C^2+1+4\,CA}}{C}$ gives the first Lyapunov coefficient in terms of A and C. The first Lyapunov coefficient is divisible by A and the leading order term of its Taylor series is $A/4$. The leading order term in the expansion of μ at the Hopf bifurcation is $-A(A+C)/2$. Following the rescaling of (2.1), the first Lyapunov coefficient is $O(\varepsilon^{1/2})$ and the value of μ is $O(\varepsilon)$.

Interestingly, there is a second term in the first Lyapunov coefficient of system (2.2) that is $O(\varepsilon^{1/2})$ in the case that $B \neq 0$ even though B is $O(\varepsilon)$. The coordinate transformations to Hopf normal form are rational if the system is parametrized by ω, X_e, and r, the magnitude of the real eigenvalue. Maple computes the first Lyapunov coefficient as a rational function P/Q of r, X_e, and ω. The leading terms in the Taylor series expansion of P and Q as functions of A, B, and C are $16C^5(2B + A^2 + AC + 2B^2)$ and $64C^5(A + C)$, respectively. If A and C are $O(\varepsilon^{1/2})$ and B is $O(\varepsilon)$, the first Lyapunov coefficient is $\frac{A}{4} + \frac{B}{2(A+C)} + o(\varepsilon^{1/2})$. Thus, even though B has higher order than A and C in terms of ε, it plays a significant role in the dynamics associated with the Hopf bifurcations of system (2.2).

Two different ways in which the nondegeneracy conditions for the Hopf bifurcation can fail are that the real eigenvalue r vanishes and that the first Lyapunov coefficient vanishes. Both of these degeneracies occur for small values of the parameters (A, B, C). When $r = 0$, a codimension two saddle-node Hopf bifurcation occurs if appropriate nondegeneracy conditions are met. The first Lyapunov coefficient vanishes along a surface in the parameter space that is asymptotic to $B = -A(A + C)/2$ as $\varepsilon \to 0$. This produces a *generalized* Hopf bifurcation, first analyzed by Bautin. The saddle-node Hopf and generalized Hopf bifurcations of system (2.2) are discussed in the next two subsections.

3.2. Saddle-node Hopf bifurcations. Saddle-node Hopf bifurcations (also called fold-Hopf bifurcations) occur at equilibria with both a zero eigenvalue and a pair of pure imaginary eigenvalues. The parameter values for which system (2.2) has such an equilibrium are given by $A = C(B - 1)$ and $\mu = BC^2/4$. Alternatively, these can be expressed by the equations $B = (A + C)/C$ and $\mu = C(A + C)/4$. Since B is of higher order than A and C in ε, these bifurcations are located where $A \approx -C$.

The truncated normal form of the saddle-node Hopf bifurcation [18] can be written in polar coordinates as

$$\begin{aligned}\dot r &= a_1 r z,\\ \dot z &= b_1 r^2 + b_2 z^2,\\ \dot\theta &= \omega.\end{aligned}$$

[1]The magnitude of the first Lyapunov coefficient depends upon the coordinates used in the eigenspace of the pure imaginary eigenvalues. Guckenheimer and Holmes [18] and Kuznetsov [24] use different coordinate systems that yield expressions which differ by a factor of 2. The expression here follows the conventions of Guckenheimer and Holmes [18].

The three coefficients a_1, b_1, b_2 determine the main features of the dynamics of its unfolding, for example, whether invariant tori occur close to the bifurcation. Maple calculations yield the expressions $\omega = 1 + B - C^2$ and

$$a_1 = \frac{C^2 - 1}{\omega^2},$$
$$b_1 = \frac{-B}{2\omega^2},$$
$$b_2 = \frac{-2B}{\omega^2}.$$

If $B > 0$, then all of these coefficients are negative, while if $B < 0$, then a_1 is negative and b_1, b_2 are both positive. Since $|b_i| < |a_1|$, these correspond to the cases IVb and III of Guckenheimer and Holmes [18, section 7.4]. Small invariant tori and chaotic solutions occur in generic unfoldings of case III.

3.3. Degenerate Hopf bifurcations. Hopf bifurcations are degenerate when their first Lyapunov coefficient vanishes [24]. Takens [32] described the unfolding of codimension two degenerate Hopf bifurcations, assuming that the second Lyapunov coefficient does not vanish. In the unfolding, the Hopf bifurcation makes a transition between subcritical and supercritical and there is a region with two periodic orbits that annihilate each other in a saddle-node of limit cycles bifurcation [18].

To find parameters where degenerate Hopf bifurcation might occur, we express the first Lyapunov coefficient as a rational function of X_e, A, B, C. Denote its numerator by P. We then compute the resultant of P and the Hopf polynomial $BC + 2X_e + 4CX_e^2 + 2X_eC^2 - A$ as functions of X_e to obtain a polynomial $R_H(A, B, C)$ that vanishes at degenerate Hopf points:

$$\begin{aligned}
R_H(A,B,C) = 256\,C^2(&-12\,A^2B - 15\,B^2A^2 + 20\,C^5A + 16\,C^2A^2 + 76\,C^4A^2 + 92\,C^3A^3\\
&+ 36\,C^2A^4 + 8\,C^7A + 16\,C^6A^2 + 8\,C^5A^3 + 8\,CA^3 + 16\,C^2B + 56\,BC^6\\
&+ 54\,C^4B - 120\,C^4B^2 - 65\,C^2B^2 + 36\,C^4B^5 + 91\,C^2B^4 + 8\,C^4B^3\\
&- 26\,C^6B^3 - 10\,C^4B^4 - 85\,C^6B^2 - 10\,C^8B^2 + 46\,C^2B^3 + 4\,B^2C^{10}\\
&- 32\,B^3C^8 + 55\,C^6B^4 + 26\,C^8B + 4\,C^{10}B + 8\,C^3A - 24\,B^2 - 24\,B^3\\
&+ 100\,C^3BA^3 - 144\,C^5B^3A - 124\,CB^3A - 156\,CB^2A + 22\,C^2BA^2\\
&- 32\,C^2B^3A^2 - 64\,CA^3B^2 + 60\,C^3B^4A + 76\,C^7AB - 52\,BCA^3\\
&+ 50\,BC^6A^2 + 4\,C^9AB - 95\,C^4B^2A^2 + 172\,C^3B^3A - 170\,C^3B^2A\\
&- 244\,C^5B^2A\\
&+ 128\,C^3BA + 148\,C^5BA + 4\,BCA + 192\,C^4BA^2\\
&- 210\,C^2B^2A^2 + 14\,C^7B^2A).
\end{aligned}$$

The leading order terms of $R_H(A, B, C)$ are

$$-1024\,C^2B\left(6\,B + 3\,A^2 - CA + 6\,B^2 - 4\,C^2\right),$$

implying that $B \approx (A+C)(4C-3A)/6$ at degenerate Hopf points with (A,B,C) small. For example, two approximate solutions of $R_H(A,B,C) = 0$ are $(-0.01, 0.00013335, 0.02)$ and $(-0.01, 0.00005, 0.02)$.

3.4. Periodic orbits. When $A = B = 0$ but $C \neq 0$, the family of periodic orbits at $\mu = 0$ is normally hyperbolic. Normal hyperbolicity implies that this surface of periodic orbits in (X, Y, Z, μ) space deforms but does not disappear when A and/or B are perturbed from zero. For most values of the parameters, the periodic orbits are isolated in the (X, Y, Z) state space. Continuation methods implemented in AUTO [11] and MATCONT [10] track the periodic orbits and locate saddle-node, period-doubling, and torus bifurcations as a single parameter is varied. Continuation methods further track curves of these bifurcations as two parameters are varied. This subsection presents some results obtained with MATCONT. Numerical integration has been used to check these continuation calculations and visualize complex trajectories from the family (2.2).

The first Lyapunov coefficient of the Hopf bifurcation in system (2.2) is $\frac{A}{4} + \frac{B}{2(A+C)} + o(\varepsilon^{1/2})$. When the first Lyapunov coefficient is negative, the bifurcation is supercritical, and stable periodic orbits emerge from the equilibrium. The periodic orbits can bifurcate as they grow in amplitude. Figure 1 shows four periodic orbits at the beginning of a period-doubling cascade computed with $(A,B,C) = (-0.05, -0.01, 0.1)$ and μ taking the values 0.0082, 0.0084, 0.0086, and 0.008618. As μ increases, three period-doubling bifurcations give successive transitions from the blue orbit to the green, then the red, and finally the thin blue orbit. The Hopf value of μ for these values of (A,B,C) is approximately 0.0008. Figure 2 shows a cross-section (green) to a quasi-periodic trajectory (blue) with parameter values $(A,B,C,\mu) = (-0.08, 0, 0.1, 0.001)$.

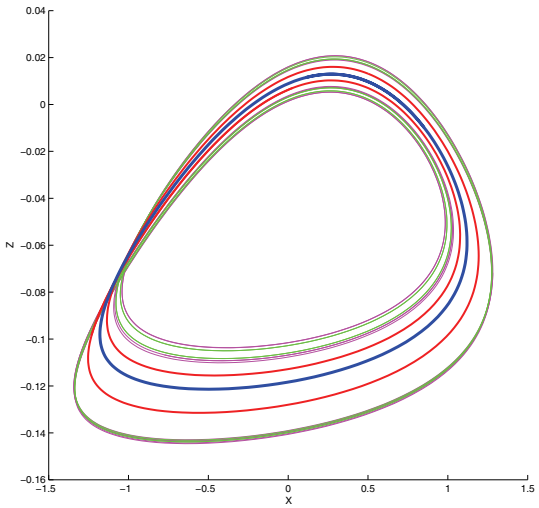


Figure 1. *Four periodic orbits of system (2.2) at the beginning of a period-doubling cascade, projected onto the (X,Z) plane. The heavy blue periodic orbit undergoes a period-doubling bifurcation to give rise to the red orbit. Two further period-doubling bifurcations yield the thin green and magenta. Parameter values are $(A,B,C) = (-0.05, -0.01, 0.1)$ and $\mu = 0.0082, 0.0084, 0.0086, 0.008618$ for the successive orbits.*

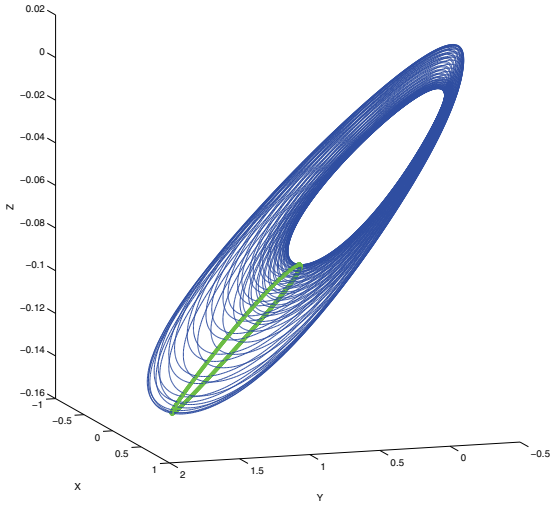
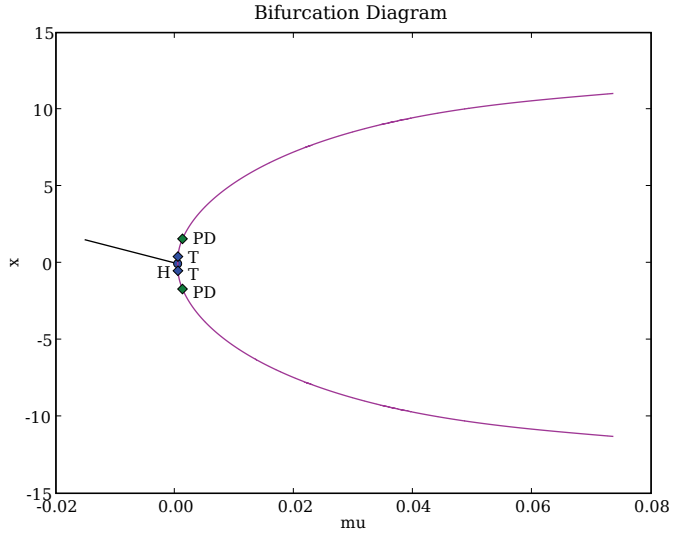


Figure 2. *A cross-section to a quasiperiodic trajectory of system (2.2). Parameter values are* $(A, B, C, \mu) = (-0.08, 0, 0.1, 0.001)$. *A portion of the trajectory is drawn as a blue curve. Intersections of the full computed trajectory with the plane* $X = 0$ *with* X *increasing are plotted in green.*

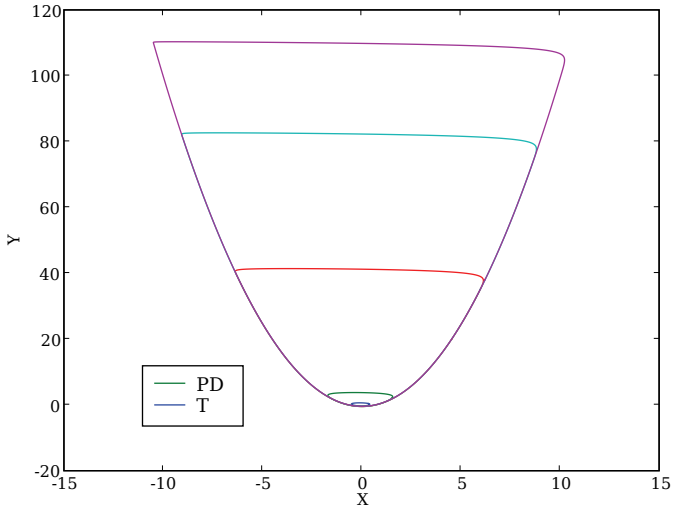
Figure 3 shows a bifurcation diagram for periodic orbits that emerge from a Hopf bifurcation as μ varies with $(A, B, C) = (-0.09, 0, 0.1)$. In this figure, the maximum and minimum values of x of the periodic orbits are plotted as a magenta curve. Points of period-doubling and torus bifurcations along this branch are marked and labeled "PD" and "T." Since $B = 0$, the system has a single equilibrium point, and the amplitude of the periodic orbits continues to grow as μ increases. The calculations are inconclusive as to whether this family of periodic orbits extends to ∞. When $B \neq 0$, system (2.2) has a second equilibrium on its critical manifold. If $A + C \neq 0$, then the second equilibrium is at finite distance from the origin. It appears that homoclinic orbits to this equilibrium can terminate families of periodic orbits. If $A + C \neq 0$, then the second equilibrium of system (2.1) is at finite distance from the origin and does not play a role in the local behavior of the singular Hopf bifurcation.

Branches of period-doubling and torus bifurcations with varying μ and B were computed with MATCONT 2.3.3 [10] and are shown in blue and green in Figure 4. Low order resonances of the torus bifurcations are marked by red dots. The three that occur for values of $\mu < 0.03$ are labeled with the order of the resonance; the point labeled R2 is the intersection of the two curves. The curve of torus bifurcations has a sharp bend near $\mu = 0.1$, where MATCONT detects several resonances of different orders as well as a fold of torus bifurcations very close to each other along the branch. The location of these resonances is indicated by the red marker at the right side of the figure.

4. Invariant manifolds. Invariant slow manifolds lie within an $O(\varepsilon)$ neighborhood of normally hyperbolic critical manifolds of slow-fast systems [14]. In typical settings, invariant manifolds are not unique, but their distance from each other is $O(\exp(-c/\varepsilon))$ for a suitable $c > 0$. The critical manifold $y = x^2$ of system (2.1) is normally hyperbolic away from the fold



(a)



(b)

Figure 3. (a) *A bifurcation diagram showing the growth of periodic orbits emerging from a Hopf bifurcation (labeled "H") of system* (2.2) *as* μ *is varied. The maxima and minima of* X *along the periodic orbits are drawn as a magenta curve. Torus and period-doubling bifurcations along the family of periodic orbits are labeled "T" and "PD." The parameters* $(A, B, C) = (-0.09, 0, 0.1)$. (b) *Five periodic orbits within the family, including those at the torus and period-doubling bifurcations.*

curve $x = y = 0$, with a stable sheet in the half space $x > 0$ and an unstable sheet in the half space $x < 0$. The slow stable and unstable manifolds associated to these sheets of the critical

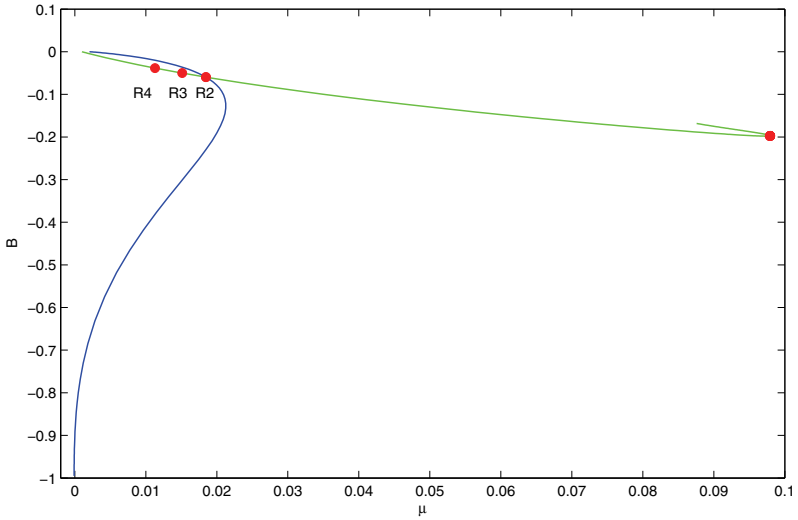


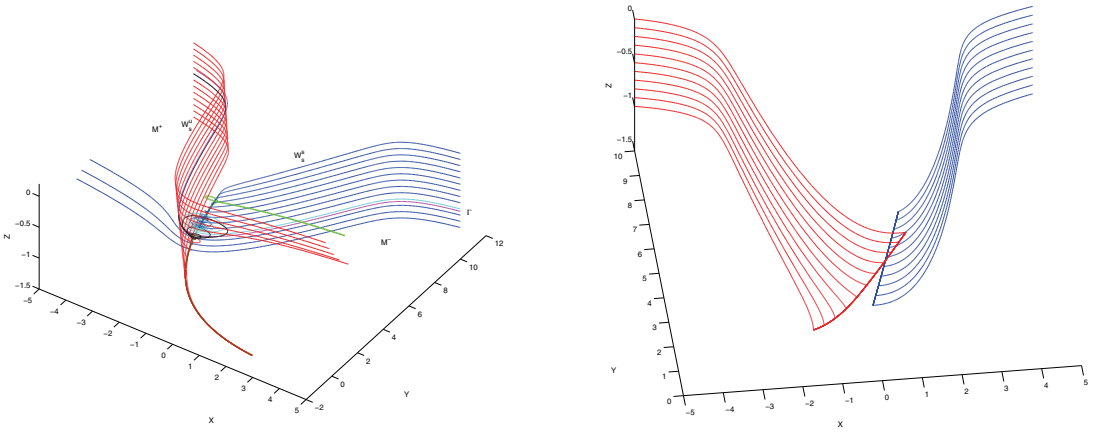
Figure 4. *Curves of torus and period-doubling bifurcations in a two dimensional slice of the parameter space with varying* (μ, B)*. The values of* A *and* C *are* 0.1 *and* -0.08*. The curve of torus bifurcations is drawn in green and points of second* (R2)*, third* (R3)*, and fourth* (R4) *order resonance are marked as red dots along the curve. Additional resonances occur near the red dot at the right-hand bend in this curve. The curve of period-doubling bifurcations is drawn in blue. The two bifurcation curves intersect at the point of second order resonance.*

manifold are important objects in the phase portrait of the system. Away from the critical manifold, the vector field is almost parallel to the x axis. The critical manifold and the slow stable and unstable manifolds separate trajectories on which x decreases rapidly from those on which x increases rapidly. The region of trajectories flowing from $x = +\infty$ to the stable slow manifold W_s^s is denoted M^-, and the region of trajectories flowing toward $x = -\infty$ from the unstable slow manifold W_s^u is denoted M^+. On the fast time scale, trajectories are drawn toward W_s^s and away from W_s^u. In many cases, parts of W_s^s and W_s^u lie on the boundary of the domain of attraction for bounded attractors. This section visualizes these manifolds, examining their intersections with each other and with the stable and unstable manifolds of the equilibrium.

4.1. Intersections of stable and unstable slow manifolds. Numerical investigations are more convenient with the rescaled system (2.2) than with system (2.1). The stable and unstable slow manifolds of system (2.2) lie close to the parabolic cylinder $Y = X^2$ for large values of $|X|$, though the theory does not specify how large. The stable slow manifold W_s^s is computed by forward numerical integration starting with initial conditions on a curve parallel to the Z axis with X suitably larger than $\sqrt{Y}$, while the unstable slow manifold W_s^u is computed by backward numerical integration starting with initial conditions on a curve parallel to the Z axis with X suitably smaller than $-\sqrt{Y}$. In the examples below, the initial conditions are chosen with $X = \pm 5$ and $Y = 10$. These trajectories approach W_s^s and W_s^u exponentially fast, so beyond a transient they give good approximations to the manifolds. Estimates for

how close the trajectories are to W_s^s and W_s^u can be obtained by comparing their distance from trajectories with initial conditions on the critical manifold since the critical manifold lies on the opposite side of the slow manifolds from the curves of initial conditions.

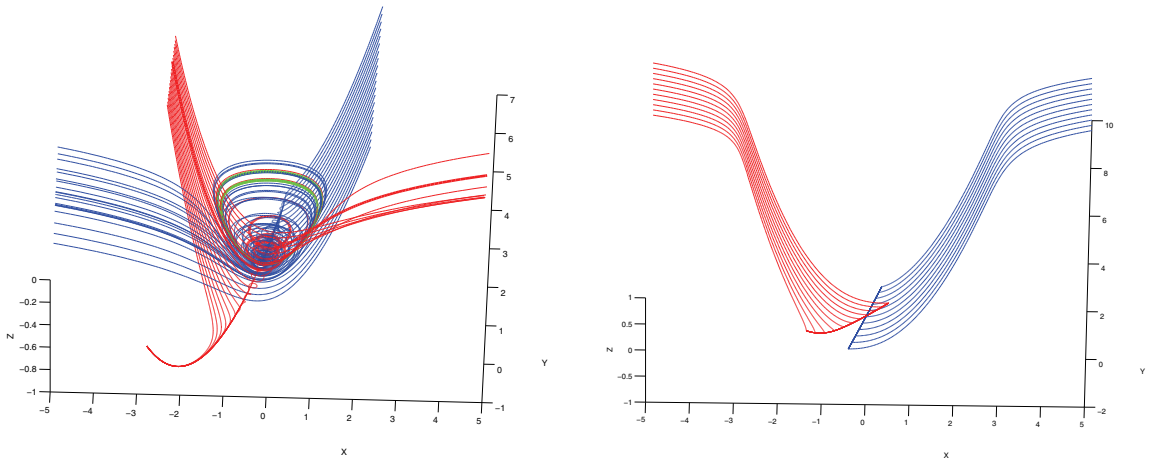
Figure 5(a) visualizes portions of the slow stable manifold W_s^s (blue) and the slow unstable manifold W_s^u (red) of system (2.2) as bundles of trajectories that begin on the lines $X = \pm 5$, $Y = 10$ until they reach $X = \mp 5$, $Y = 11$, or $T = 500$. Note that these stopping criteria extend the stable and unstable slow manifolds beyond the region where they lie close to the critical manifold. The parameter values used in Figure 5 are $(\mu, A, B, C) = (0, -0.05, -0.01, 0.1)$. The equilibrium is at the origin for these parameter values. It is on the fold curve and is stable with eigenvalues approximately $-0.0506, -0.0247 \pm 0.9934i$. The strong stable manifold of the equilibrium tangent to the eigenvector of its real eigenvalue is drawn in green. One branch of the strong stable manifold with X and Z negative approaches the slow unstable manifold while the other branch tends to ∞ in the X direction after its projection onto the (X, Y) plane makes a loop. The manifolds W_s^s and W_s^u appear to intersect transversally along a single trajectory Γ whose intersection with the plane $Z = X$ is depicted in Figure 5(b). In Figure 5(b), $Y' = Z - X = 0$ is the stopping criterion for the trajectories and piecewise



(a) (b)

Figure 5. (a) *Trajectories approaching and flowing along the slow stable manifold W_s^s are drawn in blue; trajectories approaching and flowing along the slow unstable manifold W_s^u are drawn in red. The initial conditions for these trajectories lie on the lines defined by $X = \pm 5$, $Y = 10$. The magenta curve approximates the intersection Γ of these two manifolds. The strong stable manifold of the equilibrium point is drawn in green. The equilibrium point is the forward limit set of the cyan trajectory and the dark blue trajectories above it. The blue trajectories below the magenta trajectory are unbounded, tending to $x = -\infty$ in finite time. Trajectories in the unstable manifold above the magenta trajectory tend to $x = +\infty$ as time decreases. The trajectories in the unstable manifold below the magenta trajectory tend to a branch of the strong stable manifold of the equilibrium which itself approaches the slow unstable manifold as time decreases. The slow unstable manifold W_s^u bounds the basin of attraction of the equilibrium point.* (b) *The same trajectories approaching W_s^s and W_s^u are drawn up to their intersection with the plane $Y' = Z - X = 0$. It is apparent that the intersection of these manifolds is transverse. The parameter values are $(\mu, A, B, C) = (0, -0.05, -0.01, 0.1)$.*

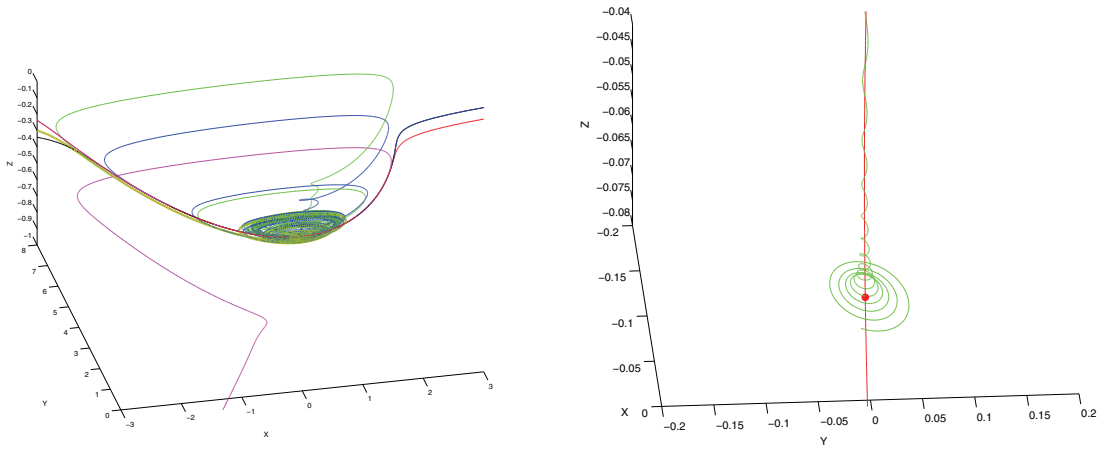
linear interpolations of the endpoints of these trajectories are drawn. An approximation to Γ is plotted as a magenta curve in Figure 5(a). Trajectories in W_s^s above Γ approach the equilibrium point, while trajectories below Γ tend to $-\infty$ along the X direction. The lowest trajectory in W_s^s above Γ is drawn in cyan to distinguish it from the others; the highest trajectory in W_s^u below Γ is drawn in black. In W_s^u, the trajectories above Γ tend to ∞ along the X direction while the trajectories below Γ spiral around the strong stable manifold of the equilibrium point. These observations motivate the conjecture that W_s^u is the boundary of the basin of attraction of the equilibrium point.



(a) (b)

Figure 6. (a) *Trajectories approaching and flowing along the slow stable manifold W_s^s are drawn in blue; trajectories approaching and flowing along the slow unstable manifold W_s^u are drawn in red. The initial conditions for these trajectories lie on the lines defined by $X = \pm 5$, $Y = 10$. A stable periodic orbit is drawn in green. All of the computed trajectories in W_s^s reach the plane $X = -5$ on their way to $X = -\infty$. The thick red trajectory shows that some of the trajectories in W_s^u that tend to $X = \infty$ oscillate before doing so.* (b) *The slow stable and unstable manifolds W_s^s and W_s^u intersect transversally. The parameter values are* $(\mu, A, B, C) = (0.0084, -0.05, -0.01, 0.1)$.

4.2. Intersections of the unstable slow manifold with the unstable manifold of the equilibrium point. As μ increases, the phase portraits of system (2.2) become more complicated. The equilibrium point has a supercritical Hopf bifurcation near $\mu = 0.0008$ and the periodic orbits born in this Hopf bifurcation enter a cascade of period-doubling bifurcations near $\mu = 0.008$, as illustrated in Figure 1. While μ increases, the asymptotic properties of the slow stable and unstable manifolds also change. Figure 6(a) visualizes portions of the slow stable (blue) and unstable (red) manifolds of system (2.2) as bundles of trajectories that begin on the lines $X = \pm 5$, $Y = 10$ and end on the plane defined by $Z' = 0$. The parameter values are $(\mu, A, B, C) = (0.0084, -0.05, -0.01, 0.1)$ used for the green period-doubled orbit displayed in Figure 1. This periodic orbit is also drawn in green here. As shown in Figure 6(b), the manifolds W_s^s and W_s^u intersect transversally, as they do when $\mu = 0$. However, in contrast to



(a) (b)

Figure 7. (a) *Three pairs of trajectories for system* (2.2) *with parameter values* $(\mu, A, B, C) = (0.003686, -0.05, -0.01, 0.1)$. *The black and blue trajectories are forward trajectories that approach the slow stable manifold* W_s^s*; the magenta and red trajectories are backward trajectories that approach the slow unstable manifold* W_s^u*; and the olive and green trajectories are forward trajectories starting close to the unstable manifold of the equilibrium point. Note that the olive and green trajectories approach* W_s^u *but then diverge from it in opposite directions.* (b) *A detailed view showing how the green trajectory spirals around the red stable manifold of the equilibrium point before approaching the periodic orbit.*

the situation with $\mu = 0$, trajectories above the intersection in W_s^s escape the bounded region containing the periodic orbits, and some trajectories above the intersection in W_s^u oscillate before they tend to $X = \infty$. The dynamical events that produce these qualitative changes in W_s^s and W_s^u as μ increases from 0 to 0.0084 are hardly clear.

For values of μ slightly larger than the Hopf bifurcation value, the equilibrium is a saddle with a two dimensional unstable manifold W_p^u bounded by the periodic orbit. As μ increases, W_p^u begins to spiral around the periodic orbit as the eigenvalues of its return map become complex. Near $\mu = 0.003686$, it appears that W_p^u begins to intersect W_s^u, the unstable slow manifold. Figure 7 presents evidence for this intersection. Figure 7(a) plots three pairs of trajectories, each of which is separated by the slow manifolds. The black and blue trajectories are forward trajectories with initial conditions $(5, 10, -0.704948)$ and $(5, 10, -0.704947)$. The black trajectory lies below the intersection of W_s^s and W_s^u and flows to $X = -\infty$, while the blue trajectory turns back toward positive values of X and then appears to spiral around the stable manifold of the equilibrium point before approaching the periodic orbit. The magenta and red trajectories have initial conditions $(-5, 10, -0.291832)$ and $(-5, 10, -0.291831)$ and are followed backward. The magenta orbit lies below the intersection of W_s^s and W_s^u and flows backward to $X = -\infty$ while the red trajectory flows backward to $X = \infty$. The olive and green trajectories have approximate initial conditions $(-0.073363697, 0.005235108, -0.072670595)$ and $(-0.073363704, 0.005235153, -0.072670597)$, points that lie close to the unstable manifold of the equilibrium. These trajectories approach the unstable slow manifold W_s^u and follow it to near its intersection with $Y = 6$ before separating. The olive trajectory then tends to

$X = -\infty$ while the green trajectory follows a similar path as the blue trajectory, spiraling around the stable manifold of the equilibrium and then approaching the periodic orbit. One might conjecture that the equilibrium point has a homoclinic orbit for a value of μ close to 0.003686. Figure 7(b) shows the green trajectory spiraling around the red stable manifold of the equilibrium point in more detail. However, the close approach of the trajectory to the equilibrium point does not imply that the parameters are close to those with a homoclinic orbit. As explained by Guckenheimer and Willms [19], the stable manifold of the equilibrium may be transversally stable as one moves away from the equilibrium, and large volumes of the state space may flow close to the stable manifold. This example demonstrates that qualitative changes in the intersections of invariant manifolds for system (2.2) typically occur at different parameter values than those where there are local bifurcations of the equilibrium point or periodic orbits of the system.

4.3. Intersections of the stable slow manifold with the stable manifold of the equilibrium point. When the equilibrium point has a one dimensional stable manifold, intersections of that manifold with the slow stable manifold might be expected to occur as codimension one bifurcations. This section presents evidence for this bifurcation by examining parameters with $A = B = \mu = 0$ and $C > 0$. For these parameters, the equilibrium is at the origin, the plane $Z = 0$ is invariant under the flow and time reversible, and there is a family of periodic orbits surrounding the origin and bounded by the parabola $Y - X^2 = -1/2$ in the plane $Z = 0$. The orbits below this parabola are unbounded, tending to $X = -\infty$ in finite time as t increases and to $X = \infty$ as t decreases. The family of periodic orbits is normally hyperbolic: each orbit has a strong stable manifold consisting of trajectories that tend toward it as $t \to \infty$. Figure 8(a) shows a branch of the stable manifold of the origin for six values of C, namely $(0.01, 0.1002, 0.1004, 0.1006, 0.1008, 0.101)$. Figure 8(b) shows the intersection of the stable manifolds with the plane $X = 3$ for 51 equally spaced values in the C interval $[0.01, 0.0101]$. Figure 8(c) shows the intersections from a much finer mesh of 5001 parameter values in this interval. It is evident that the stable manifold W^s_p of the origin oscillates in the (X, Y) plane as Z increases. These oscillations cease and X tends to ∞ for values of Z that depend upon C. Since trajectories tend to ∞ in finite time, the values of (Y, Z) typically approach finite limits along W^s_p. However, there are values of C where these limits appear to jump. These are produced by small ranges of C in which W^s_p crosses W^s_s. Figure 8(d) visualizes one crossing. The intersection of W^s_p with the plane $Y = 3$ is plotted in red as C varies through a regular mesh of 11 points in the interval $[0.01000965, 0.01000975]$. As C varies in this interval, the intersections of W^s_s with the plane $Y = 3$, drawn as a set of 11 blue curves in Figure 8(d), hardly move. The three dimensional manifold in (X, Y, Z, C) space swept out by W^s_s and the surface swept out by W^s_p clearly intersect transversally. When A, B, and μ are perturbed so that the equilibrium point becomes a saddle, the transverse intersection persists.

4.4. Contrasts between systems with one and two slow variables. The figures of this section hardly begin a systematic analysis of the global bifurcations of the invariant manifolds of system (2.2). Since the system likely has chaotic attractors for some parameter values, a complete analysis does not seem feasible. The behavior displayed here contrasts with the simpler two dimensional flows of singular Hopf bifurcations in systems with one slow variable and one fast variable. There, the slow stable and unstable manifolds are each a single trajectory

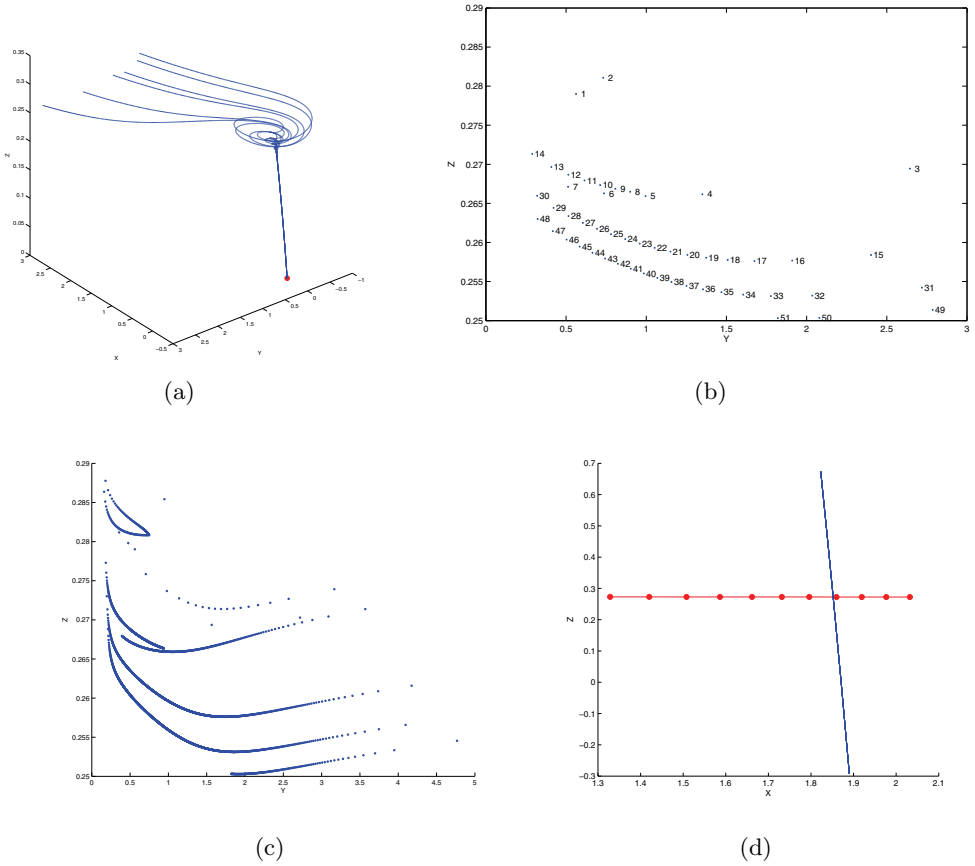


Figure 8. (a) *Trajectories lying in the stable manifold of the equilibrium point of system* (2.2) *for six different values of the parameter* C*:* $0.01, 0.1002, 0.1004, 0.1006, 0.1008, 0.101$*. The remaining parameters are zero.* (b) *Numbered intersections of the stable manifold of the equilibrium point of system* (2.2) *with the plane* $X = 3$ *for varying values of the parameter* C *in the interval* $[0.01, 0.0101]$*. Parameters* μ, A, B *are all zero.* (c) *Intersections from a mesh of* 5001 *parameter values in the interval* $[0.01, 0.0101]$*.* (d) *Intersections with the plane* $Y = 3$ *of the stable manifold of the equilibrium (red) for a mesh of* 11 *values of* C *in the interval* $[0.01000965, 0.01000975]$ *and intersections of the slow stable manifold (blue) with the plane* $Y = 3$*. There are* 11 *blue curves that are indistinguishable at this resolution.*

and the global bifurcation happens when the manifolds coincide. In the three dimensional setting investigated here, the slow stable and unstable manifolds are two dimensional and appear to intersect transversally in the vicinity of singular Hopf bifurcations. These intersections separate portions of the slow manifolds that turn in different directions. In some cases, parts of the manifolds become entangled with periodic orbits and the stable and unstable manifolds of the equilibrium. Sometimes the trajectories of these tangles remain bounded and sometimes they reemerge from the region of entanglement and proceed to $X = \pm\infty$. Further analysis of the intersections of these invariant sets is not pursued in this paper.

5. Mixed-mode oscillations in an example.

Mixed-mode oscillations (MMOs) have been observed and studied in chemical systems, for example, the Belousov–Zhabotinsky reac-

tion [20], and in the oxidation of carbon monoxide on platinum catalysts [22]. Several models have been proposed for these systems, but previous analysis has not identified that many of the properties seen in both the experimental data and models can be produced by singular Hopf bifurcations. Barkley [4] suggested that the minimum dimension of a system that fit the characteristics of MMOs in the Belousov–Zhabotinsky reaction was four. The more recent literature on MMOs in neural systems has focused upon MMOs produced by folded nodes [9, 30, 16], but some MMOs associated with singular Hopf bifurcations have characteristics that differ from those seen in the folded-node MMOs. Note that systems with singular Hopf bifurcations also have folded nodes, so singular Hopf bifurcations may produce MMOs that pass through folded nodes as well as ones that do not.

This section revisits one of the simplest models for MMOs—the autocatalator studied by Petrov, Scott, and Showalter [28] and Milik and Szmolyan [27]. The equations for this model are

$$\begin{aligned} \dot{a} &= \mu(\kappa + c) - ab^2 - a, \\ \varepsilon\dot{b} &= ab^2 + a - b, \\ \dot{c} &= b - c. \end{aligned} \tag{5.1}$$

In the studies of this system cited above, $\kappa = 2.5$ was held fixed and the parameters μ and/or ε were varied. The critical manifold of this system is given by $a = b/(1+b^2)$ and its fold curve is defined by $a = 1/2$, $b = 1$. At equilibrium points, $b = c$, so the equilibrium is on the fold curve when $b = c = 1$, $a = 1/2$, implying that $\mu = 1/(1+\kappa)$. In general, if we parametrize the equilibria of the system by c, then the curve of equilibrium points is given by $a = c/(1+c^2)$, $b = c$, $\mu = c/(c+\kappa)$. Computing the Jacobian of the system (5.1), we find that the criterion for Hopf bifurcation of the system is a polynomial expression that is affine in κ and quadratic in ε, so we can readily parametrize the Hopf bifurcation as a function of the variables c and ε. In addition to the equilibrium equations,

$$\begin{aligned} &\kappa_{hopf} \\ &= -\frac{c\left(9\,c^2\varepsilon + 2 - c^2 + 5\,\varepsilon + 6\,c^4\varepsilon + 5\,c^6\varepsilon^2 + 9\,c^4\varepsilon^2 + 7\,c^2\varepsilon^2 + 3\,c^6\varepsilon + 2\,\varepsilon^2 + c^8\varepsilon^2 + c^8\varepsilon - c^6\right)}{2 + 3\,c^4\varepsilon + 5\,c^6\varepsilon^2 + c^8\varepsilon^2 + 6\,c^2\varepsilon + 9\,c^4\varepsilon^2 + 7\,c^2\varepsilon^2 + 2\,c^6\varepsilon + c^8\varepsilon - c^2 + 4\,\varepsilon + 2\,\varepsilon^2 - c^6}. \end{aligned}$$

The function κ_{hopf} is singular at $c = 1, \varepsilon = 0$. For fixed κ, the Hopf criterion defines c as a smooth function of ε that vanishes at $c = 1$ and has slope $(3 + 2\kappa)/(1 + \kappa)$. Thus, there is indeed a singular Hopf bifurcation in this system. This does not appear in the analysis of Milik and Szmolyan [27] because they transform the parameters to set $\mu = \varepsilon\bar{\mu} + 1/(1+\kappa)$ and then use $\bar{\mu}$ and ε as the parameters they vary. In this representation, the Hopf bifurcations have parameter values that are close to $\bar{\mu} = 0.4375$ and are apparent as $\varepsilon \to 0$ only if $\bar{\mu}$ is also varied in the region near this Hopf value. Indeed, they do not analyze properties of the equilibrium point at all except at the values at which the Hopf bifurcation lies on the fold curve, a point termed a folded saddle-node in their work.

Petrov, Scott, and Showalter [28] studied the periodic orbits of system (5.1) using AUTO [11]. They work with two values of ε, namely, $\varepsilon = 0.01$ and $\varepsilon = 0.013$. For $\varepsilon = 0.013$, they observe that there is a supercritical Hopf bifurcation at $\mu \approx 0.29202$ and a second supercritical Hopf bifurcation at $\mu \approx 0.77372$. The first of these is the singular Hopf bifurcation: the

value of (a, b) is approximately $(0.49977, 1.031080)$. Petrov, Scott, and Showalter [28] observe that there is a narrow band of values of $\mu \in [0.297, 0.303]$ where the system has complex dynamics. The periodic orbits born at the singular Hopf bifurcation undergo a period-doubling cascade to a small amplitude chaotic attractor. Near $\mu = 0.29795$, the chaotic attractor disappears, and trajectories starting near the previous attractor approach a periodic MMO. Milik and Szmolyan use geometric and singular perturbation methods to study this system, producing return maps for some of the attractors. Figures 9 and 10 extend this analysis using the insights into the singular Hopf bifurcation described in this paper.

MMOs are formed from trajectories which concatenate small and large amplitude oscillations. In system (5.1), the large amplitude oscillations come from trajectories that pass "outside" the unstable slow manifold; i.e., they have larger values of a. To test whether trajectories with small amplitude oscillations flow to the outside of the unstable slow manifold, trajectories in the unstable manifold of the equilibrium were computed, similar to the calculations of the singular Hopf normal form illustrated in Figures 5, 6, and 7. Figures 9(a) and 9(b) display trajectories on the unstable manifold of the equilibrium point in blue and trajectories on the unstable slow manifold in red for parameter values $(\varepsilon, \kappa, \mu) = (0.013, 2.5, 0.2963)$ and $(\varepsilon, \kappa, \mu) = (0.013, 2.5, 0.2964)$, respectively. It appears that as μ increases from 0.2963 to 0.2964, the unstable manifold of the equilibrium point begins to intersect the unstable slow manifold. The intersection of these invariant manifolds seems to be intimately related to the formation of MMOs. Nonetheless, it is difficult to make definitive statements about these dynamics because the periodic orbits of the system have followed a period-doubling route to chaotic attractors for smaller values of μ, similar to the behavior displayed by the singular Hopf normal form (2.2) for parameters $(A, B, C) = (-0.05, -0.01, 0.1)$ and increasing μ (cf. Figure 1). Here, Petrov, Scott, and Showalter [28] showed that there are several families of MMOs as well as the small amplitude chaotic attractors in parameter ranges close to those displayed here.

Figure 9 suggests that intersections of the unstable slow manifold with the basins of small amplitude attractors are critical to the formation of MMOs. The value of ε used in this figure makes the system only moderately stiff. Figure 10 displays similar calculations for the smaller value $\varepsilon = 0.001$. Subfigure (a) shows trajectories in the unstable manifold of the equilibrium (blue) and the unstable slow manifold (blue) for $(\varepsilon, \kappa, \mu) = (0.001, 2.5, 0.2864)$. There is a stable periodic orbit, and this orbit forms the boundary of the unstable manifold of the equilibrium point. Subfigure (b) shows analogous information for $(\varepsilon, \kappa, \mu) = (0.001, 2.5, 0.2865)$. The stable periodic orbit persists, but some trajectories near the equilibrium point flow to the outside of the unstable slow manifold and generate MMOs. Figure 11 shows a portion of one of these MMOs as it passes close to the equilibrium point. The large amplitude excursions of the trajectory approach the stable manifold of the equilibrium point closely, and these are followed by slowly growing small amplitude oscillations similar to those that can appear in the aftermath of a subcritical Hopf bifurcation [19]. For these parameter values, the birth of MMOs is clearly the direct result of the intersections of the unstable slow manifold with the unstable manifold of the equilibrium.

6. Discussion. As a slow-fast system, the equations for singular Hopf bifurcation are reduced to a three dimensional vector field that can be rescaled so that the Hopf frequency

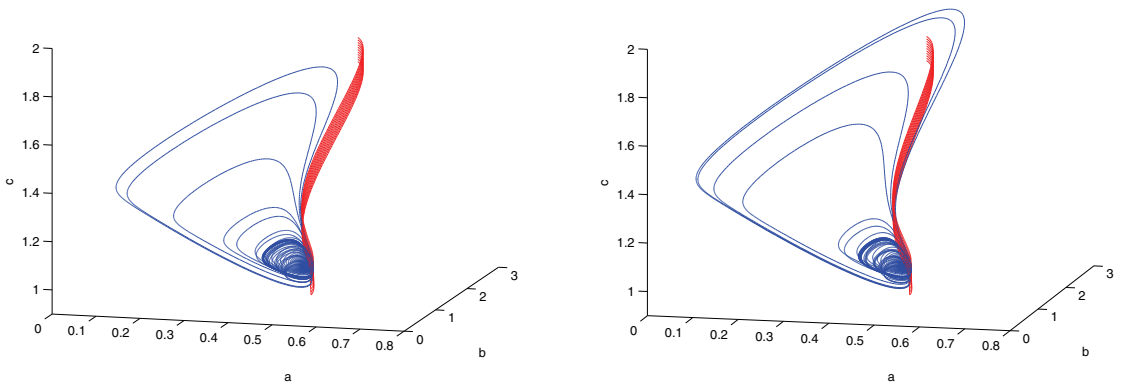


Figure 9. (a) *Trajectories on the unstable manifold of the equilibrium point (blue) and unstable slow manifold (red) of system* (5.1) *are drawn for* $(\varepsilon, \kappa, \mu) = (0.0013, 2.5, 0.2963)$. *The unstable manifold of the equilibrium point remains to the left side of the slow unstable manifold.* (b) *Analogous trajectories are drawn for* $(\varepsilon, \kappa, \mu) = (0.0013, 2.5, 0.2963)$. *Here the unstable manifold of the equilibrium point intersects the unstable slow manifold. Some trajectories with initial conditions near the equilibrium point make large excursions before approaching the small amplitude attractor.*

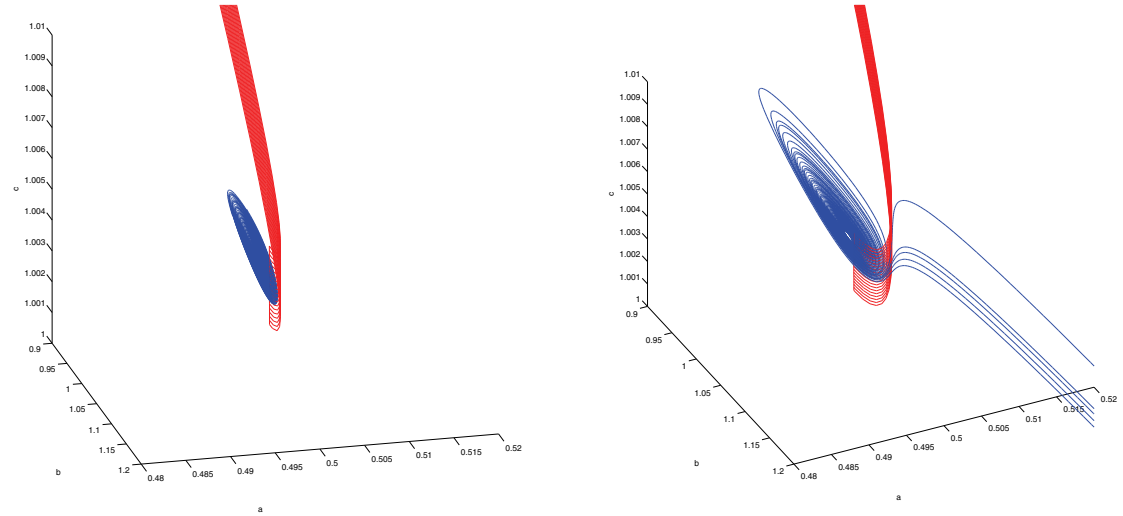
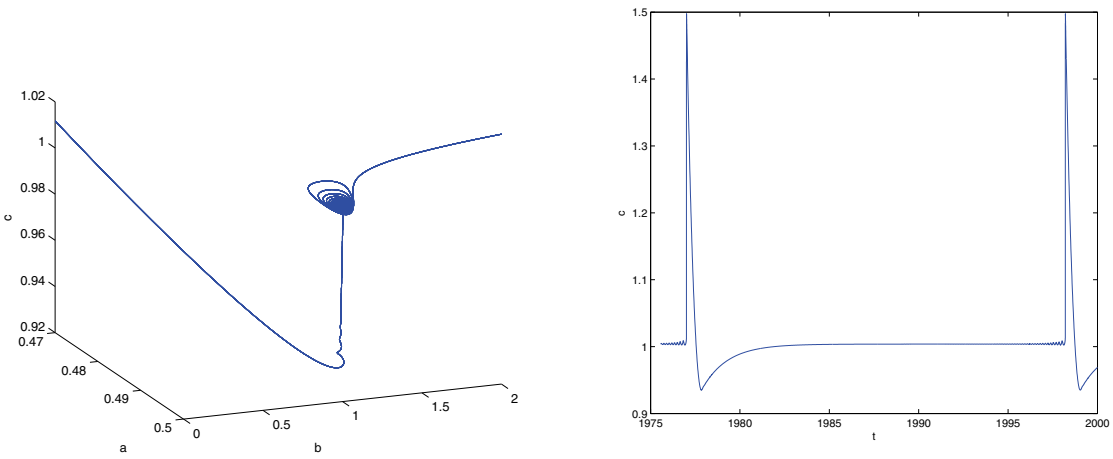


Figure 10. (a) *Trajectories on the unstable manifold of the equilibrium point (blue) and unstable slow manifold (red) of system* (5.1) *are drawn for* $(\varepsilon, \kappa, \mu) = (0.001, 2.5, 0.2864)$. *The unstable manifold of the equilibrium point remains to the left side of the slow unstable manifold and lies in the basin of attraction of a stable periodic orbit.* (b) *Analogous trajectories are drawn for* $(\varepsilon, \kappa, \mu) = (0.001, 2.5, 0.2865)$. *Here the unstable manifold of the equilibrium point intersects the unstable slow manifold. Some trajectories with initial conditions near the equilibrium point make large excursions and approach MMOs.*



(a) (b)

Figure 11. (a) *A portion of an MMO trajectory of system* (5.1) *is drawn for* $(\varepsilon, \kappa, \mu) = (0.001, 2.5, 0.2865)$ *and initial condition* $(0.5, 1, 1)$. *The trajectory was computed to time* 2000, *and its intersections with a region around the equilibrium point were plotted for the time interval* $[1700, 2000]$. *The trajectory approaches the stable manifold of the equilibrium and flows out along its unstable manifold with slowly growing small amplitude oscillations before making another large excursion. The trajectory is approximately periodic, but the calculations do not conclusively rule out the possibility that there is a more complicated attractor that is very "thin."* (b) *The final cycle of the time series displaying the c coordinate of the trajectory displayed in subfigure* (a).

remains close to one as the singular perturbation parameter ε tends to zero. This scaling emphasizes the fast time scale whose singular limit is a vector field with an equilibrium point with pure imaginary and zero eigenvalues and a one parameter family of periodic orbits emanating from the equilibrium. Two coefficients of the Taylor expansion of the rescaled vector field, the real eigenvalue of its equilibrium and the first Lyapunov coefficient of its Hopf bifurcation, are $O(\varepsilon^{1/2})$. An interesting aspect of the normal form analysis is that an $O(\varepsilon)$ term in the Taylor expansion of the rescaled system still contributes to the first Lyapunov coefficient of the Hopf bifurcation at $O(\varepsilon^{1/2})$. The truncated normal form used in this paper includes this $O(\varepsilon)$ term. Thus the normal form has five parameters: the singular perturbation parameter, a primary parameter that drives the equilibrium point across the fold curve of the critical manifold, and three secondary parameters that can be regarded as moduli.

This paper highlights the complexity of Hopf bifurcation in multiple time scale systems with two slow variables and one fast variable. Numerical simulations and continuation calculations with the normal form demonstrate that periodic orbits near a singular Hopf bifurcation can have secondary bifurcations that produce quasiperiodic or chaotic trajectories of these systems in an $O(\varepsilon)$ neighborhood of the equilibrium undergoing Hopf bifurcation. The dependence of the secondary bifurcations on the moduli in the normal form is clearly very complicated. There are additional global bifurcations that separate parameter regimes with only small amplitude attractors from parameter regimes in which trajectories starting near the equilibrium can make large excursions. These transitions have been studied here by test-

ing for intersections of the two dimensional unstable manifold of the equilibrium point with the unstable slow manifold. From a pragmatic point of view, the boundary between trajectories that remain in the vicinity of the equilibrium point of the system and those that leave a neighborhood of the equilibrium point is an important aspect of the dynamics of singular Hopf bifurcation.

This paper is partly motivated by attempts to understand the mechanisms that create MMOs in slow-fast systems. MMOs have been observed in diverse physical systems and dynamical models. These MMOs appear to fall into different dynamical classes that have yet to be clearly delineated or analyzed. One class that has been identified and studied are MMOs associated with flow through a folded node [9, 16]. This paper identifies singular Hopf bifurcation as another mechanism for generating MMOs. As illustrated with the model chemical system (5.1), the intersections of the unstable slow manifold with the unstable manifold of a saddle-focus equilibrium point can produce MMOs. These intersections are a byproduct of singular Hopf bifurcation. The small oscillations of MMOs associated with singular Hopf bifurcation often begin with very small amplitude as they approach a saddle-focus equilibrium along its stable manifold and depart with growing oscillations along its unstable manifold. In contrast, the trajectories that pass through a folded node have oscillations that first decrease and then increase in amplitude. Figure 12 displays a trajectory of the system

$$\begin{aligned}\dot{x} &= y - x^2,\\ \dot{y} &= z - x,\\ \dot{z} &= -0.002\end{aligned}$$

with initial conditions $(50, 395, 0.16)$. The oscillations of this trajectory typify the small oscillations that one finds for MMOs produced by folded nodes. Compare this figure with Figure 11, showing an MMO associated with singular Hopf bifurcation in the autocatalator. In the normal form for the folded node, the numbers of oscillations with decreasing and increasing amplitude are equal. Further work to analyze the dynamical origins of MMOs from experimental observations of chemically reacting systems might be of interest [20]. It

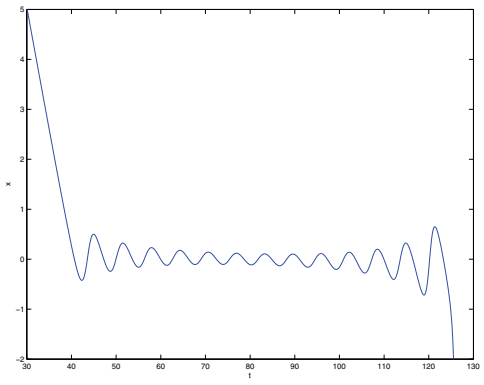


Figure 12. *Oscillations of a trajectory passing through a folded node.*

seems likely that MMOs associated with folded nodes and those associated with singular Hopf bifurcations both occur as well as MMOs that are far from these bifurcations.

REFERENCES

[1] V. I. Arnold, V. S. Afrajmovich, Yu. S. Il'yashenko, and L. P. Shil'nikov, *Dynamical Systems* V, Encyclopaedia Math. Sci., Springer-Verlag, Berlin, 1994.

[2] S. M. Baer and T. Erneux, *Singular Hopf bifurcation to relaxation oscillations*, SIAM J. Appl. Math., 46 (1986), pp. 721–739.

[3] S. M. Baer and T. Erneux, *Singular Hopf bifurcation to relaxation oscillations* II, SIAM J. Appl. Math., 52 (1992), pp. 1651–1664.

[4] D. Barkley, *Slow manifolds and mixed mode oscillations in the Belousov-Zhabotinskii reaction*, J. Chem. Phys., 89 (1988), pp. 3812–3820.

[5] D. Barkley, J. Ringland, and J. Turner, *Observations of a torus in a model of the Belousov-Zhabotinskii reaction*, J. Chem. Phys., 87 (1987), pp. 5547–5559.

[6] É. Benoît, *Canards et enlacements*, Inst. Hautes Études Sci. Publ. Math., 72 (1990), pp. 63–91.

[7] É. Benoît, J. L. Callot, F. Diener, and M. Diener, *Chasse au canards*, Collect. Math., 31 (1981), pp. 37–119.

[8] B. Braaksma, *Singular Hopf bifurcation in systems with fast and slow variables*, J. Nonlinear Sci., 8 (1998), pp. 457–490.

[9] M. Brøns, M. Krupa, and M. Wechselberger, *Mixed mode oscillations due to the generalized canard phenomenon*, in Bifurcation Theory and Spatio-Temporal Pattern Formation, Fields Inst. Commun. 49, AMS, Providence, RI, 2006, pp. 39–63.

[10] A. Dhooge, W. Govaerts, and Yu. A. Kuznetsov, *MATCONT: A MATLAB package for numerical bifurcation analysis of ODEs*, ACM Trans. Math. Software, 29 (2003), pp. 141–164. Also available online from http:// www.matcont.ugent.be/.

[11] E. Doedel, *AUTO: Software for Continuation and Bifuracation Problems in Ordinary Differential Equations*, http://indy.cs.concordia.ca/auto/.

[12] F. Dumortier and R. Roussarie, *Canard cycles and center manifolds*, Mem. Amer. Math. Soc., 121 (577) (1996).

[13] W. Eckhaus, *Relaxation oscillations, including a standard chase on French ducks*, in Asymptotic Analysis II, Lecture Notes in Math. 985, Springer-Verlag, Berlin, 1983, pp. 449–494.

[14] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–225.

[15] J. Guckenheimer, *Bifurcations of relaxation oscillations*, in Normal Forms, Bifurcations and Finiteness Problems in Differential Equations, NATO Sci. Ser. II Math. Phys. Chem. 137, Kluwer, Dordrecht, The Netherlands, 2004, pp. 295–316.

[16] J. Guckenheimer, *Return maps of folded nodes and folded saddle-nodes*, Chaos, 18 (2008), 015108.

[17] J. Guckenheimer and R. Haiduc, *Canards at folded nodes*, Mosc. Math. J., 5 (2005), pp. 91–103.

[18] J. Guckenheimer and P. J. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[19] J. Guckenheimer and A. Willms, *Asymptotic analysis of subcritical Hopf-homoclinic bifurcation*, Phys. D, 139 (2000), pp. 195–216.

[20] J. Hudson, M. Hart, and D. Marinko, *An experimental study of multiple peak periodic and nonperiodic oscillations in the Belousov-Zhabotinskii reaction*, J. Chem. Phys., 71 (1979), pp. 1601–1606.

[21] G. Kozyreff and T. Erneux, *Singular Hopf bifurcation to strongly pulsating oscillations in lasers containing a saturable absorber*, European J. Appl. Math., 14 (2003), pp. 407–420.

[22] K. Krischer, M. Lübke, M. Eiswirth, W. Wolf, J. L. Hudson, and G. Ertl, *A hierarchy of transitions to mixed mode oscillations in an electrochemical system*, Phys. D, 62 (1993), pp. 123–133.

[23] M. Krupa and P. Szmolyan, *Extending geometric singular perturbation theory to nonhyperbolic points—fold and canard points in two dimensions*, SIAM J. Math. Anal., 33 (2001), pp. 286–314.

[24] Yu. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 3rd ed., Appl. Math. Sci. 112, Springer-Verlag, New York, 2004.

[25] Y. Lijun and Z. Xianwu, *Stability of singular Hopf bifurcations*, J. Differential Equations, 206 (2004), pp. 30–54.
[26] http://www.maplesoft.com/.
[27] A. Milik and P. Szmolyan, *Multiple time scales and canards in a chemical oscillator*, in Multiple-Time-Scale Dynamical Systems (Minneapolis, 1997), IMA Vol. Math. Appl. 122, Springer-Verlag, New York, 2001, pp. 117–140.
[28] V. Petrov, S. Scott, and K. Showalter, *Mixed-mode oscillations in chemical systems*, J. Chem. Phys., 97 (1992), pp. 6191–6198.
[29] J. Rinzel, *A formal classification of bursting mechanisms in excitable systems*, in Proceedings of the International Congress of Mathematicians, Vols. 1, 2 (Berkeley, 1986), AMS, Providence, RI, 1987, pp. 1578–1593.
[30] J. Rubin and M. Wechselberger, *Giant squid—hidden canard: The 3D geometry of the Hodgkin-Huxley model*, Biol. Cybernet., 97 (2007), pp. 5–32.
[31] P. Szmolyan and M. Wechselberger, *Canards in R^3*, J. Differential Equations, 177 (2001), pp. 419–453.
[32] F. Takens, *Unfoldings of certain singularities of vector fields: Generalized Hopf bifurcations*, J. Differential Equations, 14 (1973), pp. 476–493.
[33] M. Wechselberger, *Existence and bifurcation of canards in $\mathbb{R}^3$ in the case of a folded node*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 101–139.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1378–1396
© 2008 Society for Industrial and Applied Mathematics

Eulerian Equilibria of a Gyrostat in Newtonian Interaction with Two Rigid Bodies*

J. A. Vera†

Abstract. In this paper the noncanonical Hamiltonian dynamics of a gyrostat in the three body problem will be examined. By means of geometric-mechanics methods we will study the approximate dynamics that arises when we develop the potential in Legendre series and truncate the series to the second harmonics. Some relative equilibria, called Eulerian, of the dynamics of a gyrostat in Newtonian interaction with two rigid bodies will be studied. Taking advantage of the results obtained in previous papers, working on the reduced problem, we will study the bifurcations of these relative equilibria. The instability of Eulerian relative equilibria if the gyrostat is close to a sphere is proven. The rotational Poisson dynamics of the gyrostat placed in an Eulerian equilibrium and the study of the nonlinear stability of some equilibria are considered. The analysis is done in vectorial form avoiding the use of canonical variables and the tedious expressions associated with these variables. In this way, the classic results on equilibria of the three body problem, many of them obtained by other authors who had made used of more classic techniques, are generalized.

Key words. three body problem, gyrostat, Eulerian, stability, energy-Casimir

AMS subject classifications. 37N05, 70F15, 70E55

DOI. 10.1137/060671711

1. Introduction.

In the study of configurations of relative equilibria by differential geometry methods or by more classical methods we will mention here the papers of Wang, Krishnaprasad, and Maddocks [11], about the problem of a rigid body in a central Newtonian field, and Maciejewski [4], about the problem of two rigid bodies in mutual Newtonian attraction. These papers have been generalized to the case of two gyrostats by Mondéjar and Vigueras [5].

For the problem of three rigid bodies, we would like to mention that Vidyakin [9] and Dubochine [1] proved the existence of Euler and Lagrange configurations of equilibria when the bodies possess symmetries; Zhuravlev and Petrutskii [13] reviewed the results up to 1990. These works use canonical variables for the deduction of their results.

On the other hand, if a rigid solid model is used to represent celestial bodies instead of the point masses model, the possibility of internal or relative motions in the celestial bodies will not be considered. This hypothesis is not appropriate in many cases, as shown by Volterra [10] in his study on the variation of the latitude in the Earth's surface.

In relation to the previous point, let us remember what is known as a gyrostat, which is a mechanical system S, composed of a rigid body S', and other bodies S'' (deformable or rigid) connected to it, in such a way that their relative motion with respect to its rigid part

*Received by the editors October 7, 2006; accepted for publication (in revised form) by C. Wayne July 3, 2008; published electronically October 31, 2008. This research was partially supported by the Consejería de Educación y Cultura de la Comunidad Autónoma de la Región de Murcia (Spain) (Project 05783/PI/07).
http://www.siam.org/journals/siads/7-4/67171.html
†Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, 30203 Cartagena (Murcia), Spain (juanantonio.vera@upct.es).

does not change the distribution of mass of the total system S (see Leimanis [3] for details). Examples of such systems are bodies that represent mobile internal cavities (for example, fluids) or bodies with coupled symmetrical rotors that can be activated by remote control.

In Vera [7] and a recent paper of Vera and Vigueras [8] we study the noncanonical Hamiltonian dynamics of $n+1$ bodies in Newtonian attraction, where n of them are rigid bodies with spherical distribution of mass or material points and the other body is a triaxial gyrostat. Using the symmetries of the system, we carried out two reductions, giving in each step the Poisson structure of the reduced space. Then, we obtained the equations of motion, the Casimir function of the system, and the equations that determine the equilibria and global conditions for the existence of the equilibria.

This paper is a concrete application of the general methods of [8] to the study of certain types of equilibria of a gyrostat in Newtonian attraction with two rigid bodies (see Figure 1). We describe the approximate dynamics that arises in a natural way when we take the Legendre development of the potential function and truncate this to the second harmonics. If the involved bodies are at much more mutual distances than the individual dimensions of the involved bodies, this approximate dynamics is a good description of the full dynamics of the problem.

Additionally, we suppose that the attitude dynamics of the two rigid bodies is the same as a rigid body in torque free motion. The two rigid bodies have revolution symmetry about the third axis of inertia. On the other hand, the Newtonian interaction of the gyrostat with the rigid bodies is the same as that of a material particle in Newtonian interaction with two rigid bodies. The Newtonian interaction among the two rigid bodies is the same as that of two material particles.

Under these hypotheses, we give global conditions on the existence of relative equilibria, and analogous to classic results on the topic (see [12] for details), we study the existence of relative equilibria that we will denominate of *Euler type* in the case in which S_1, S_2 are spherical or symmetrical bodies and S_0 is a gyrostat. Necessary and sufficient conditions for their existence in this approximate dynamics are obtained, and we give explicit expressions of the Eulerian equilibria, useful for the later study of their stability. A complete study of the number of the Eulerian equilibria is made when S_1, S_2 are spherical rigid bodies. Concerning to the stability of these equilibria, the instability of the Eulerian equilibria is proven.

On the other hand, the rotational Poisson dynamics of the gyrostat placed in an Eulerian equilibrium is considered. The nonlinear stability of some rotational equilibria, called cylindrical equilibria, is studied by means of the energy-Casimir method. By means of the tangent flow of this dynamics in the cylindrical equilibrium we obtain necessary conditions for the nonlinear stability of the cylindrical equilibria.

This analysis was done in vectorial form, giving to this problem a very compact treatment which avoids the use of canonical variables (Eulerian or Andoyer–Deprit variables) and the tedious expressions associated with these variables. This is a typical characteristic of the classic literature on these systems that the paper overcomes with this vectorial approach. Contrary to the canonical variables, this analysis is free of singularities.

We should note that the studied system has potential interest both in astrodynamics (dealing with spacecrafts) as well as in the understanding of the evolution of planetary systems recently found (and more to appear), where some of the planets may be modeled like gyrostats

rather than rigid bodies. In fact, the equilibria reported might well be compared with those taken for the "parking areas" of the space missions (GENESIS, SOHO, DARWIN, etc.) around the Eulerian points of the Sun-Earth and the Earth-Moon systems (see [2] for details).

To finish this introduction, we describe the structure of the article. The paper is organized into seven sections, two appendices, and the bibliography. In these sections we study the equations of motion, the Casimir function and integrals of the system, and the relative equilibria and the existence of Eulerian equilibria, in particular the study of the bifurcations of Eulerian equilibria in the approximate dynamics. The rotational Poisson dynamics of the gyrostat placed in an Eulerian equilibrium and the study of the stability of some equilibria are considered.

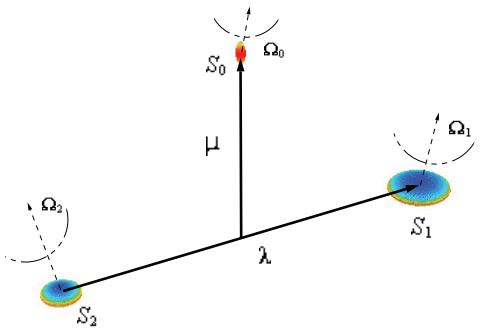


Figure 1. *Gyrostat in the three body problem.*

2. Equations of motion. Let S_0 be a gyrostat of mass m_0, and let S_1, S_2 be two symmetrical rigid bodies of masses m_1 and m_2, respectively; $\mathfrak{I} = \{\mathbf{O}, \mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is an inertial reference frame; $\mathfrak{J} = \{\mathbf{C}_0, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ is a body frame fixed at the center of mass $\mathbf{C}_0$ of S_0 (see Vera and Vigueras [8] for details).

The following notation is used:

$$M_2 = m_1 + m_2, \quad M_1 = m_1 + m_2 + m_0, \quad g_1 = \frac{m_1 m_2}{M_2}, \quad g_2 = \frac{m_0 M_2}{M_1}.$$

For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$, $\mathbf{u} \cdot \mathbf{v}$ is the dot product; $|\mathbf{u}|$ is the Euclidean norm of the vector $\mathbf{u}$; $\mathbf{u} \times \mathbf{v}$ is the cross product; $\mathbf{I}_{\mathbb{R}^3}$ is the identity matrix; and $\mathbf{0}$ is the zero matrix of order three. Consider $\mathbb{I}_0 = \operatorname{diag}(A_0, B_0, C_0)$ the diagonal tensor of inertia of the gyrostat and $\mathbb{I}_i = \operatorname{diag}(A_i, A_i, C_i)$ the diagonal tensors of inertia for the rigid bodies S_i, $i = 1, 2$. The generic expression $\mathbf{z} = (\mathbf{\Pi}_1, \mathbf{\Pi}_2, \mathbf{\Pi}_0, \boldsymbol{\lambda}, \mathbf{p}_{\boldsymbol{\lambda}}, \boldsymbol{\mu}, \mathbf{p}_\mu) \in \mathbb{R}^{21}$ is a vector of the twice reduced problem obtained by applying the symmetries of the system. The vector $\mathbf{z}$ is obtained by a Poisson reduction by means of the group $\mathbf{SO(3)}$ and a symplectic reduction by means of the translation group; see Vera and Vigueras [8] for details. The vector $\mathbf{\Pi}_0 = \mathbb{I}_0 \mathbf{\Omega}_0 + \mathbf{l}_r$ is the total rotational angular momentum vector of the gyrostat in the body frame, which is attached to its rigid part and whose axes have the direction of the principal axes of inertia of S_0; the vector $\mathbf{l}_r = (0, 0, l)$ is the constant gyrostatic momentum (this vector models the internal motions of the gyrostat) and $\mathbf{\Pi}_i = \mathbb{I}_i \mathbf{\Omega}_i$ $(i = 1, 2)$ are the total rotational angular

momentum vectors for the two rigid bodies. The elements $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\mathbf{p}_{\boldsymbol{\lambda}}$, and $\mathbf{p}_{\mu}$ are, respectively, the barycentric coordinates and the linear momenta expressed in the body frame $\mathfrak{J}$.

Following the results of Vera and Vigueras [8], according to the hypotheses formulated in the introduction of this paper, a good approximation to the potential of the system is expressed by

$$\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2,$$

where

$$\mathcal{V}_1 = -\left(\frac{Gm_1m_2}{|\boldsymbol{\lambda}|} + \frac{Gm_1m_0}{|\boldsymbol{\mu} - \frac{m_2}{M_2}\boldsymbol{\lambda}|} + \frac{Gm_2m_0}{|\boldsymbol{\mu} + \frac{m_1}{M_2}\boldsymbol{\lambda}|} \right),$$

$$\mathcal{V}_2 = -\frac{1}{2}\left(\frac{Gm_0\alpha_1}{|\boldsymbol{\mu} - \frac{m_2}{M_2}\boldsymbol{\lambda}|^3} + \frac{Gm_0\alpha_2}{|\boldsymbol{\mu} + \frac{m_1}{M_2}\boldsymbol{\lambda}|^3} - \frac{3Gm_0f_1}{|\boldsymbol{\mu} - \frac{m_2}{M_2}\boldsymbol{\lambda}|^5} - \frac{3Gm_0f_2}{|\boldsymbol{\mu} + \frac{m_1}{M_2}\boldsymbol{\lambda}|^5} \right),$$

and

$$\alpha_1 = 2A_1 + C_1, \quad \alpha_2 = 2A_2 + C_2,$$

$$f_1(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{\mu} \cdot \mathbb{I}_1\boldsymbol{\mu} - \frac{2m_2}{M_2}\boldsymbol{\lambda} \cdot \mathbb{I}_1\boldsymbol{\mu} + \left(\frac{m_2}{M_2}\right)^2 \boldsymbol{\lambda} \cdot \mathbb{I}_1\boldsymbol{\lambda},$$

$$f_2(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{\mu} \cdot \mathbb{I}_2\boldsymbol{\mu} + \frac{2m_1}{M_2}\boldsymbol{\lambda} \cdot \mathbb{I}_2\boldsymbol{\mu} + \left(\frac{m_1}{M_2}\right)^2 \boldsymbol{\lambda} \cdot \mathbb{I}_2\boldsymbol{\lambda}.$$

The Hamiltonian function of the system adopts the form

$$\mathcal{H}(\mathbf{z}) = \frac{|\mathbf{p}_{\boldsymbol{\lambda}}|^2}{2g_1} + \frac{|\mathbf{p}_{\boldsymbol{\mu}}|^2}{2g_2} + \frac{1}{2}\Pi_0^t\mathbb{I}_0^{-1}\Pi_0 - \mathbf{l}_r \cdot \mathbb{I}_0^{-1}\Pi + \frac{1}{2}\Pi_1^t\mathbb{I}_1^{-1}\Pi_1 + \frac{1}{2}\Pi_2^t\mathbb{I}_2^{-1}\Pi_2 + \mathcal{V}(\boldsymbol{\lambda}, \boldsymbol{\mu}).$$

Let $(\mathbf{M}, \{\ ,\ \}, \mathcal{H})$ with $\mathbf{M} = \mathbb{R}^{21}$ be the Poisson manifold, where $\{\ ,\ \}$ is the Poisson bracket defined by means of the Poisson tensor

$$\mathbf{B}(\mathbf{z}) = \begin{pmatrix} \widehat{\boldsymbol{\Pi}_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{\Pi}_2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{\boldsymbol{\Pi}_0} & \widehat{\boldsymbol{\lambda}} & \widehat{\mathbf{p}_{\boldsymbol{\lambda}}} & \widehat{\boldsymbol{\mu}} & \widehat{\mathbf{p}_{\mu}} \\ \mathbf{0} & \mathbf{0} & \widehat{\boldsymbol{\lambda}} & \mathbf{0} & \mathbf{I}_{\mathbb{R}^3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{\mathbf{p}_{\boldsymbol{\lambda}}} & -\mathbf{I}_{\mathbb{R}^3} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{\boldsymbol{\mu}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{\mathbb{R}^3} \\ \mathbf{0} & \mathbf{0} & \widehat{\mathbf{p}_{\mu}} & \mathbf{0} & \mathbf{0} & -\mathbf{I}_{\mathbb{R}^3} & \mathbf{0} \end{pmatrix}.$$

In $\mathbf{B}(\mathbf{z})$, $\widehat{\mathbf{v}}$ is considered to be the image of the vector $v \in \mathbb{R}^3$ by the standard isomorphism between the Lie algebras $\mathbb{R}^3$ and $\mathfrak{so}(3)$, i.e.,

$$\widehat{\mathbf{v}} = \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix}.$$

The equations of the motion are given by

$$\frac{d\mathbf{z}}{dt} = \{\mathbf{z}, \mathcal{H}(\mathbf{z})\} = \mathbf{B}(\mathbf{z})\boldsymbol{\nabla}_{\mathbf{z}}\mathcal{H}(\mathbf{z}),$$

with $\boldsymbol{\nabla}_{\mathbf{u}}\mathcal{V}$ being the gradient of $\mathcal{V}$ with respect to an arbitrary vector $\mathbf{u}$.

Calculating $\{\mathbf{z}, \mathcal{H}(\mathbf{z})\}$, the following group of vectorial equations of motion can be written as

$$\begin{aligned}
&\frac{d\boldsymbol{\Pi}_0}{dt} = \boldsymbol{\Pi}_0 \times \boldsymbol{\Omega}_0 + \boldsymbol{\lambda} \times \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V} + \boldsymbol{\mu} \times \boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V},\\
&\frac{d\boldsymbol{\lambda}}{dt} = \frac{\mathbf{p}_{\boldsymbol{\lambda}}}{g_1} + \boldsymbol{\lambda} \times \boldsymbol{\Omega}_0, \quad \frac{d\mathbf{p}_{\boldsymbol{\lambda}}}{dt} = \mathbf{p}_{\boldsymbol{\lambda}} \times \boldsymbol{\Omega}_0 - \boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V},\\
&\frac{d\boldsymbol{\mu}}{dt} = \frac{\mathbf{p}_{\boldsymbol{\mu}}}{g_2} + \boldsymbol{\mu} \times \boldsymbol{\Omega}_0, \quad \frac{d\mathbf{p}_{\boldsymbol{\mu}}}{dt} = \mathbf{p}_{\boldsymbol{\mu}} \times \boldsymbol{\Omega}_0 - \boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V},\\
&\frac{d\boldsymbol{\Pi}_1}{dt} = \boldsymbol{\Pi}_1 \times \boldsymbol{\Omega}_1, \quad \frac{d\boldsymbol{\Pi}_2}{dt} = \boldsymbol{\Pi}_2 \times \boldsymbol{\Omega}_2.
\end{aligned} \tag{1}$$

Important elements of $\mathbf{B}(\mathbf{z})$ are the associated Casimir functions. The vector

$$\mathbf{L}_0 = \boldsymbol{\Pi}_0 + \boldsymbol{\lambda} \times \mathbf{p}_{\boldsymbol{\lambda}} + \boldsymbol{\mu} \times \mathbf{p}_{\mu}$$

is a part of the total angular momentum $\mathbf{L}$ given by

$$\mathbf{L} = \boldsymbol{\Pi}_2 + \boldsymbol{\Pi}_1 + \mathbf{L}_0.$$

Then the following result can be concluded.

Proposition 1. *If φ_i $(i = 0, 1, 2)$ are real smooth functions, then $\varphi_0(\frac{|\mathbf{L}_0|^2}{2})$, $\varphi_i(\frac{|\boldsymbol{\Pi}_i|^2}{2})$ $(i = 1, 2)$ are Casimir functions of the Poisson tensor $\mathbf{B}(\mathbf{z})$. Furthermore, $\operatorname{Ker}\mathbf{B}(\mathbf{z}) = \langle \boldsymbol{\nabla}_{\mathbf{z}}\varphi_0, \boldsymbol{\nabla}_{\mathbf{z}}\varphi_1, \boldsymbol{\nabla}_{\mathbf{z}}\varphi_2 \rangle$. We also have $\frac{d\mathbf{L}}{dt} = \mathbf{0}$, which means the total angular momentum vector remains constant. If $\boldsymbol{\Pi}_0 = (\boldsymbol{\pi}_0^1, \boldsymbol{\pi}_0^2, \boldsymbol{\pi}_0^3)$, then $\boldsymbol{\pi}_0^3$ is an integral of the motion.*

3. Relative equilibria. If $\mathbf{z}_e = (\boldsymbol{\Pi}_2^e, \boldsymbol{\Pi}_1^e, \boldsymbol{\Pi}_0^e, \boldsymbol{\lambda}^e, \mathbf{p}_{\boldsymbol{\lambda}}^e, \boldsymbol{\mu}^e, \mathbf{p}_{\boldsymbol{\mu}}^e)$ is a generic relative equilibrium, the following vectorial equations are verified:

$$\begin{aligned}
&\boldsymbol{\Pi}_0^e \times \boldsymbol{\Omega}_0^e + \boldsymbol{\lambda}^e \times (\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V})_e + \boldsymbol{\mu} \times (\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V})_e = \mathbf{0},\\
&\frac{\mathbf{p}_{\boldsymbol{\lambda}}^e}{g_1} + \boldsymbol{\lambda}^e \times \boldsymbol{\Omega}_0^e = \mathbf{0}, \quad \mathbf{p}_{\boldsymbol{\lambda}}^e \times \boldsymbol{\Omega}_0^e = (\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V})_e,\\
&\frac{\mathbf{p}_{\boldsymbol{\mu}}^e}{g_2} + \boldsymbol{\mu}^e \times \boldsymbol{\Omega}_0^e = \mathbf{0}, \quad \mathbf{p}_{\boldsymbol{\mu}}^e \times \boldsymbol{\Omega}_0^e = (\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V})_e,\\
&\boldsymbol{\Pi}_1^e \times \boldsymbol{\Omega}_1^e = \mathbf{0}, \quad \boldsymbol{\Pi}_2^e \times \boldsymbol{\Omega}_2^e = \mathbf{0},
\end{aligned} \tag{2}$$

where $(\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V})_e$ and $(\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V})_e$ are the values of $\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V}$ and $\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V}$ at $\mathbf{z}_e$.

According to the relationships provided by Vera and Vigueras [8], the following results are obtained.

Lemma 2. *Whenever $\mathbf{z}_e = (\boldsymbol{\Pi}_2^e, \boldsymbol{\Pi}_1^e, \boldsymbol{\Pi}_0^e, \boldsymbol{\lambda}^e, \mathbf{p}_{\boldsymbol{\lambda}}^e, \boldsymbol{\mu}^e, \mathbf{p}_{\boldsymbol{\mu}}^e)$ is a relative equilibrium, the following relationships are satisfied:*

$$|\boldsymbol{\Omega}_0^e|^2|\boldsymbol{\lambda}^e|^2 - (\boldsymbol{\lambda}^e \cdot \boldsymbol{\Omega}_0^e)^2 = \frac{1}{g_1}(\boldsymbol{\lambda}^e \cdot (\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V})_e),$$
$$|\boldsymbol{\Omega}_0^e|^2|\boldsymbol{\mu}^e|^2 - (\boldsymbol{\mu}^e \cdot \boldsymbol{\Omega}_0^e)^2 = \frac{1}{g_2}(\boldsymbol{\mu}^e \cdot (\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V})_e).$$

The previous two identities will be used to obtain necessary conditions for the existence of relative equilibria.

Certain relative equilibria will be studied assuming that vectors $\boldsymbol{\Omega}_0^e$, $\boldsymbol{\lambda}^e$, and $\boldsymbol{\mu}^e$ satisfy special geometric properties.

Definition 1. $\mathbf{z}_e$ *is said to be an* Eulerian relative equilibrium *when* $\boldsymbol{\lambda}^e$ *and* $\boldsymbol{\mu}^e$ *are proportional and* $\boldsymbol{\Omega}_e$ *is perpendicular to the straight line that they generate.*

From the above definitions, the following property is deduced.

Proposition 3. *In an Eulerian relative equilibrium, the forces that derive the potential do not exercise moments on the gyrostat.*

Next, necessary and sufficient conditions for the existence of Eulerian relative equilibria will be obtained.

4. Eulerian relative equilibria.

According to the relative position of the gyrostat S_0 with respect to S_1 and S_2, there are three possible equilibrium configurations: (a) $S_0S_2S_1$, (b) $S_2S_0S_1$, and (c) $S_2S_1S_0$ (see Figure 2 for details on the configuration $S_2S_1S_0$).

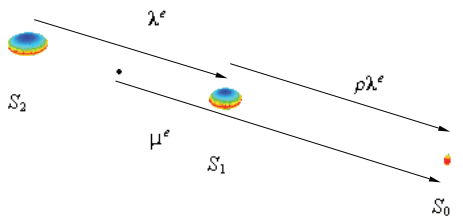


Figure 2. *Eulerian relative equilibria in configuration* $S_2S_1S_0$.

4.1. Necessary conditions of existence.

The following lemma is a direct consequence of the geometry of the problem.

Lemma 4. *If* $\mathbf{z}_e = (\boldsymbol{\Pi}_2^e, \boldsymbol{\Pi}_1^e, \boldsymbol{\Pi}_0^e, \boldsymbol{\lambda}^e, \mathbf{p}_{\boldsymbol{\lambda}}^e, \boldsymbol{\mu}^e, \mathbf{p}_{\mu}^e)$ *is a relative equilibrium of Euler type, then for the configuration* $S_0S_2S_1$

$$\left|\boldsymbol{\mu}^e - \frac{m_2}{M_2}\boldsymbol{\lambda}^e\right| = |\boldsymbol{\lambda}^e| + \left|\boldsymbol{\mu}^e + \frac{m_1}{M_2}\boldsymbol{\lambda}^e\right|.$$

In a similar way, for the configuration $S_2S_0S_1$

$$|\boldsymbol{\lambda}^e| = \left|\boldsymbol{\mu}^e - \frac{m_2}{M_2}\boldsymbol{\lambda}^e\right| + \left|\boldsymbol{\mu}^e + \frac{m_1}{M_2}\boldsymbol{\lambda}^e\right|.$$

Finally, for the configuration $S_2S_1S_0$

$$\left|\boldsymbol{\mu}^e + \frac{m_1}{M_2}\boldsymbol{\lambda}^e\right| = \left|\boldsymbol{\mu}^e - \frac{m_2}{M_2}\boldsymbol{\lambda}^e\right| + |\boldsymbol{\lambda}^e|.$$

If $\mathbf{z}_e$ is an Eulerian relative equilibrium, then

$$g_1|\mathbf{\Omega}_e^0|^2|\boldsymbol{\lambda}^e|^2 = \boldsymbol{\lambda}^e \cdot (\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V})_e,$$
$$g_2|\mathbf{\Omega}_e^0|^2|\boldsymbol{\mu}^e|^2 = \boldsymbol{\mu}^e \cdot (\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V})_e,$$

with

$$\boldsymbol{\mu}^e - \frac{m_2}{M_2}\boldsymbol{\lambda}^e = \rho\boldsymbol{\lambda}^e, \quad \boldsymbol{\mu}^e + \frac{m_1}{M_2}\boldsymbol{\lambda}^e = (1+\rho)\boldsymbol{\lambda}^e, \quad \boldsymbol{\mu}^e = \frac{((1+\rho)m_2 + \rho m_1)}{M_2}\boldsymbol{\lambda}^e,$$

where $\rho \in (-\infty, -1)$ in the configuration (a), $\rho \in (-1, 0)$ in the configuration (b), and $\rho \in (0, +\infty)$ in the configuration (c). Moreover, it is possible to obtain

$$(\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V})_e = h_1(\rho)\boldsymbol{\lambda}^e, \quad (\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V})_e = h_2(\rho)\boldsymbol{\lambda}^e$$

with

$$h_1(\rho) = \frac{Gm_1m_2}{|\boldsymbol{\lambda}^e|^3} + \frac{Gm_0m_1 sgn(1+\rho)}{M_2|\boldsymbol{\lambda}^e|^3}\left(\frac{m_2}{(1+\rho)^2} + \frac{\beta_2}{(1+\rho)^4|\boldsymbol{\lambda}^e|^2}\right) - \frac{Gm_0m_2 sgn(\rho)}{M_2|\boldsymbol{\lambda}^e|^3}\left(\frac{m_1}{\rho^2} + \frac{\beta_1}{\rho^4|\boldsymbol{\lambda}^e|^2}\right) \tag{3}$$

and

$$h_2(\rho) = \frac{Gm_0 sgn(1+\rho)}{|\boldsymbol{\lambda}^e|^3}\left(\frac{m_2}{(1+\rho)^2} + \frac{\beta_2}{|\boldsymbol{\lambda}^e|^2(1+\rho)^4}\right) + \frac{Gm_0 sgn(\rho)}{|\boldsymbol{\lambda}^e|^3}\left(\frac{m_1}{\rho^2} + \frac{\beta_1}{|\boldsymbol{\lambda}^e|^2\rho^4}\right), \tag{4}$$

where $\beta_1 = 3(C_1 - A_1)/2$, $\beta_2 = 3(C_2 - A_2)/2$, and

$$sgn(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x \leq 0. \end{cases}$$

Now, from the identities

$$\boldsymbol{\lambda}^e \cdot (\boldsymbol{\nabla}_{\boldsymbol{\lambda}}\mathcal{V})_e = |\boldsymbol{\lambda}^e|^2 h_1(\rho),$$
$$\boldsymbol{\mu}^e \cdot (\boldsymbol{\nabla}_{\boldsymbol{\mu}}\mathcal{V})_e = \frac{((1+\rho)m_2 + \rho m_1)}{M_2}|\boldsymbol{\lambda}^e|^2 h_2(\rho),$$

the following equations are deduced:

$$|\mathbf{\Omega}_0^e|^2 = \frac{(m_1+m_2)h_1(\rho)}{m_1m_2},$$
$$|\mathbf{\Omega}_0^e|^2 = \frac{(m_0+m_1+m_2)h_2(\rho)}{m_0\left((1+\rho)m_2 + \rho m_1\right)}.$$

Then for an Eulerian relative equilibrium, ρ must be a real root of the following equation:

$$m_0(m_1+m_2)\left((1+\rho)m_2 + \rho m_1\right)h_1(\rho) = m_1m_2(m_0+m_1+m_2)h_2(\rho). \tag{5}$$

The following proposition summarizes all these results.

Proposition 5. *If* $\mathbf{z}_e = (\mathbf{\Pi}_2^e, \mathbf{\Pi}_1^e, \mathbf{\Pi}_0^e, \boldsymbol{\lambda}^e, \mathbf{p}_{\boldsymbol{\lambda}}^e, \boldsymbol{\mu}^e, \mathbf{p}_{\mu}^e)$ *is an Eulerian relative equilibrium, the equation* (5) *has at least one real root in which the functions* $h_1(\rho)$ *and* $h_2(\rho)$ *are given by the expressions* (3), (4). *The modulus of the angular velocity of the gyrostat is*

$$|\mathbf{\Omega}_0^e|^2 = \frac{(m_1+m_2)h_1(\rho)}{m_1 m_2}. \tag{6}$$

Remark 1. *When* $|\boldsymbol{\lambda}_e|$ *has a fixed value, if an Eulerian relative equilibrium exists, then* (5) *has real solutions. The number of real roots of* (5) *will depend, obviously, on the parameters which exist in the system.*

4.2. Sufficient conditions of existence. The following proposition indicates how to calculate solutions of (2).

Proposition 6. *When* $|\boldsymbol{\lambda}^e|$ *has a fixed value, let* ρ *be a solution of* (5), *where the functions* $h_1(\rho)$ *and* $h_2(\rho)$ *are given by the expressions* (3), (4). *Then* $\mathbf{z}_e = (\mathbf{\Pi}_2^e, \mathbf{\Pi}_1^e, \mathbf{\Pi}_0^e, \boldsymbol{\lambda}^e, \mathbf{p}_{\boldsymbol{\lambda}}^e, \boldsymbol{\mu}^e, \mathbf{p}_{\mu}^e)$ *are given by*

$$\boldsymbol{\lambda}^e = (\lambda^e, 0, 0), \quad \boldsymbol{\mu}^e = (\mu^e, 0, 0), \quad \mathbf{\Omega}_0^e = (0, 0, \omega_0^e),$$
$$\mathbf{p}_{\boldsymbol{\lambda}}^e = (0, g_1\omega_0^e\lambda^e, 0), \quad \mathbf{p}_{\boldsymbol{\mu}}^e = (0, g_2\omega_0^e\mu^e, 0), \quad \mathbf{\Pi}_0^e = (0, 0, C_0\omega_0^e + l),$$

where

$$\boldsymbol{\mu}^e = \frac{((1+\rho)m_1 + \rho m_2)}{M_2}\boldsymbol{\lambda}^e, \quad (\omega_0^e)^2 = \frac{(m_1+m_2)h_1(\rho)}{m_1 m_2}$$

is an Eulerian relative equilibrium. The total angular momentum of the system is expressed by

$$\mathbf{L} = (0, 0, C_2\omega_2^e + C_1\omega_1^e + C_0\omega_0^e + l + g_1(\omega_0^e)^2\lambda^e + g_2(\omega_0^e)^2\mu^e),$$

with l *being the gyrostatic momentum. The vectors* $\mathbf{\Pi}_2^e$, $\mathbf{\Pi}_1^e$ *verify the vectorial equations*

$$\mathbf{\Pi}_1^e \times \mathbf{\Omega}_1^e = \mathbf{0}, \quad \mathbf{\Pi}_2^e \times \mathbf{\Omega}_2^e = \mathbf{0}.$$

4.3. Eulerian relative equilibria when S_2 and S_1 are spherical rigid bodies. Consider the existence and number of solutions for Eulerian relative equilibria when S_2 and S_1 are spherical rigid bodies. In this case $C_1 = A_1$, $C_2 = A_2$, and (5) is equivalent to the following polynomial equation:

$$\begin{aligned}(m_1+m_2)\rho^5 + (3m_2+2m_1)\rho^4 + (3m_2 + m_1 + m_0(sgn(1+\rho)\\ - sgn(\rho)))\rho^3 + (m_2 - m_2 sgn(1+\rho) - m_1 sgn(\rho) - 3m_0 sgn(\rho))\rho^2\\ - (3sgn(\rho)m_0 + 2sgn(\rho)m_1)\rho - (m_0 sgn(\rho) + m_1 sgn(\rho)) = 0.\end{aligned} \tag{7}$$

This equation, by Descartes's rule of signs, has a unique real solution in the intervals $(-\infty, -1)$, $(-1, 0)$, and $(0, +\infty)$. Therefore, only one Eulerian relative equilibrium exists.

On the other hand,

$$|\mathbf{\Omega}_0^e|^2 = \frac{G(m_1+m_2)}{|\boldsymbol{\lambda}^e|^3}\left(1 + \frac{m_0}{(m_1+m_2)}\left(\frac{sgn(1+\rho)}{(1+\rho)^2} - \frac{sgn(\rho)}{\rho^2}\right)\right),$$

where ρ is the only solution of (7).

Proposition 5 gathers the results about Eulerian relative equilibria when S_2 and S_1 are spherical rigid bodies in the configurations (a), (b), and (c).

Proposition 7. 1. *If ρ is the unique positive root of the equation*

$$(m_1+m_2)\rho^5+(3m_2+2m_1)\rho^4+(3m_2+m_1)\rho^3 \\ -(3m_0+m_1)\rho^2-(3m_0+2m_1)\rho-(m_0+m_1)=0$$

with

$$|\mathbf{\Omega}_0^e|^2=\frac{G(m_1+m_2)}{|\boldsymbol{\lambda}^e|^3}\left(1+\frac{m_0}{(m_1+m_2)}\left(\frac{1}{(1+\rho)^2}-\frac{1}{\rho^2}\right)\right),$$

then $\mathbf{z}_e=(\mathbf{\Pi}_2^e,\mathbf{\Pi}_1^e,\mathbf{\Pi}_0^e,\boldsymbol{\lambda}^e,\mathbf{p}_{\boldsymbol{\lambda}}^e,\boldsymbol{\mu}^e,\mathbf{p}_{\mu}^e)$, given by

$$\boldsymbol{\lambda}^e=(\lambda^e,0,0),\quad \boldsymbol{\mu}^e=(\mu^e,0,0),\quad \mathbf{\Omega}_0^e=(0,0,\omega_0^e), \\ \mathbf{p}_{\boldsymbol{\lambda}}^e=(0,g_1\omega_0^e\lambda^e,0),\quad \mathbf{p}_{\boldsymbol{\mu}}^e=(0,g_2\omega_0^e\mu^e,0),\quad \mathbf{\Pi}_0^e=(0,0,C_0\omega_0^e+l), \tag{8}$$

is the unique solution of relative equilibrium of Euler type in the configuration $S_2S_1S_0$.

2. *If $\rho\in(-1,0)$ is the unique root of the equation*

$$(m_1+m_2)\rho^5+(3m_2+2m_1)\rho^4+(3m_2+m_1+2m_0)\rho^3 \\ +(3m_0+m_1)\rho^2+(3m_0+2m_1)\rho+(m_0+m_1)=0$$

with

$$|\mathbf{\Omega}_0^e|^2=\frac{G(m_1+m_2)}{|\boldsymbol{\lambda}^e|^3}\left(1+\frac{m_0}{(m_1+m_2)}\left(\frac{1}{(1+\rho)^2}+\frac{1}{\rho^2}\right)\right),$$

then $\mathbf{z}_e=(\mathbf{\Pi}_2^e,\mathbf{\Pi}_1^e,\mathbf{\Pi}_0^e,\boldsymbol{\lambda}^e,\mathbf{p}_{\boldsymbol{\lambda}}^e,\boldsymbol{\mu}^e,\mathbf{p}_{\mu}^e)$ given by (8) is the unique solution of relative equilibrium of Euler type in the configuration $S_2S_0S_1$.

3. *If $\rho\in(-\infty,-1)$ is the unique root of the equation*

$$(m_1+m_2)\rho^5+(3m_2+2m_1)\rho^4+(3m_2+m_1)\rho^3 \\ +(3m_0+m_1+2m_2)\rho^2+(3m_0+2m_1)\rho+(m_0+m_1)=0,$$

where

$$|\mathbf{\Omega}_0^e|^2=\frac{G(m_1+m_2)}{|\boldsymbol{\lambda}^e|^3}\left(1+\frac{m_0}{(m_1+m_2)}\left(\frac{1}{\rho^2}-\frac{1}{(1+\rho)^2}\right)\right),$$

then $\mathbf{z}_e=(\mathbf{\Pi}_2^e,\mathbf{\Pi}_1^e,\mathbf{\Pi}_0^e,\boldsymbol{\lambda}^e,\mathbf{p}_{\boldsymbol{\lambda}}^e,\boldsymbol{\mu}^e,\mathbf{p}_{\mu}^e)$ given by (8) is the unique solution of relative equilibrium of Euler type in the configuration $S_0S_2S_1$.

These results agree with the classical Newtonian three body problem; see [12] for details.

4.4. Eulerian relative equilibria when S_2 and S_1 are not spherical rigid bodies. In the present case, after carrying out the appropriate calculations, we reduce (5) to the study of the positive real roots of the nine degree equation

$$\beta_2 q(\rho)-m_1m_2a^2\rho^2(\rho+1)^2p(\rho)=0, \tag{9}$$

where

$$p(\rho) = (m_1+m_2)\rho^5 + (3m_2+2m_1)\rho^4 + (3m_2+m_1+m_0(sgn(1+\rho) - sgn(\rho)))\rho^3 + (m_2 - m_2 sgn(1+\rho) - m_1 sgn(\rho) - 3m_0 sgn(\rho))\rho^2 - (3sgn(\rho)m_0 + 2sgn(\rho)m_1)\rho - (m_0 sgn(\rho) + m_1 sgn(\rho)) \tag{10}$$

and

$$q(\rho) = m_0(ksgn(\rho)m_2 - sgn(1+\rho)m_1)\rho^5 + m_2(sgn(1+\rho)m_1 + 5sgn(\rho)m_0 k + km_1 sgn(\rho))\rho^4 + 2m_2 ksgn(\rho)(2m_1+5m_0)\rho^3 + 2m_2 ksgn(\rho)(3m_1+5m_0)\rho^2 + ksgn(\rho)m_2(4m_1+5m_0)\rho + ksgn(\rho)m_2(m_0+m_1),$$

where

$$\beta_1 = 3(C_1 - A_1)/2, \quad \beta_2 = 3(m_1+m_2)(C_2 - A_2)/2$$

with $\beta_1 = k\beta_2$, $a = |\boldsymbol{\lambda}_e|$, and $k \in \mathbb{R}$.

In order to study the number of real roots of the polynomial (9), the rational function

$$\beta_2 = R(\rho) = \frac{m_1 m_2 a^2 \rho^2 (\rho+1)^2 p(\rho)}{q(\rho)}$$

will be studied.

In practical applications m_0 is very small; then up to first order in m_0

$$\beta_2 = R_1(\rho) = \frac{a^2\rho^2(\rho+1)^2 p_1(\rho)}{q_1(\rho)} + o(m_0),$$

where

$$p_1(\rho) = \rho^5 + (2+\mu)\rho^4 + (1+2\mu)\rho^3 + (\mu - \mu sgn(1+\rho) - (1-\mu)sgn(\rho))\rho^2 - 2sgn(\rho)(1-\mu)\rho - (1-\mu)sgn(\rho)$$

and

$$q_1(\rho) = (sgn(1+\rho) + sgn(\rho)k)\rho^4 + 4sgn(\rho)k\rho^3 + 6sgn(\rho)k\rho^2 + 4sgn(\rho)k\rho + sgn(\rho)k,$$

where $\mu = \frac{m_2}{m_2+m_1}$.

The polynomial q_1 has no roots in $(0,+\infty)$ and $(-1,0)$ if $k > 0$ and $k < 0$, respectively. On the other hand, q_1 has only one root, ρ_1, in $(0,+\infty)$ and $(-1,0)$ if $k < 0$ and $k > 0$, respectively. ρ_0 will be denoted as the only root of p_1. The implicit curve $Res(k,\mu) = 0$, where Res is the resultant of the polynomials p_1 and q_1, is used to study the graph of R_1. When μ_0 has a fixed value, the only k_0 which verifies $Res(k_0,\mu_0) = 0$ exists according to the implicit function theorem. Consider the only root, $\widetilde{\rho_1} = \rho_0(\mu_0)$, of q_1 for k_0 and μ_0. The expressions

$$(\rho_{\max}, \xi_1(k) = R_1(\rho_{\max})), \quad (\rho_{\min}, \xi_2(k) = R_1(\rho_{\min}))$$

are the local maximum and minimum of the function R_1.

In the configuration $S_2S_1S_0$, for any value of μ fixed

$$\lim_{k\to+\infty}\xi_2(k)=0,\quad \lim_{k\to0^+}\xi_2(k)=-\infty$$

if $k>0$. For $k<0$

$$\lim_{k\to+\infty}\xi_1(k)=0,\quad \lim_{k\to0^+}\xi_2(k)=+\infty$$

if $\rho_1>\widetilde{\rho_1}$. If $\rho_1\le\widetilde{\rho_1}$, then R_1 is strictly increasing.

For the configuration $S_2S_0S_1$, if $\widetilde{\rho_1}\le\rho_1$ and $k>0$, then the function R_1 has just one minimum, which verifies

$$\lim_{k\to0^+}\xi_2(k)=\xi_0.$$

If $\rho_1<\widetilde{\rho_1}$, then

$$\lim_{k\to+\infty}\xi_2(k)=0.$$

If $k<0$, then R_1 verifies that

$$\lim_{k\to0^-}\xi_2(k)=\xi_0,\quad \lim_{k\to0^-}\xi_1(k)=+\infty,$$
$$\lim_{k\to-\infty}\xi_2(k)=0,\quad \lim_{k\to-\infty}\xi_1(k)=0.$$

The results for the configuration $S_0S_2S_1$ will be deduced from the configuration $S_2S_1S_0$.

According to these statements, the following proposition can be stated.

Proposition 8. *In the configuration $S_2S_1S_0$ for $k>0$,*

1. *if $\beta_2<R_1(\rho_{\min})$, then Eulerian relative equilibria do not exist.*
2. *If $\beta_2=R_1(\rho_{\min})$, a unique 2-parametric family of Eulerian relative equilibria exists.*
3. *If $R_1(\rho_{\min})<\beta_2<0$, two 2-parametric families of Eulerian relative equilibria exist.*
4. *If $\beta_2>0$, a unique 2-parametric family of Eulerian relative equilibria exists. For $k_0<k<0$ and $\beta_2>0$,*
5. *if $\beta_2\in(\xi_1(k),\xi_2(k))$, Eulerian relative equilibria do not exist.*
6. *If $\beta_2=\xi_1(k)$ or $\beta_2=\xi_2(k)$, then a unique 2-parametric family of Eulerian relative equilibria exists.*
7. *If $\beta_2>\xi_2(k)$, two 2-parametric families of Eulerian relative equilibria exist.*
8. *If $0<\beta_2<\xi_1(k)$, two 2-parametric families of Eulerian relative equilibria exist. For $k_0<k<0$ and $\beta_2<0$,*
9. *a unique 2-parametric family of Eulerian relative equilibria exists. For $k<k_0$ and $\beta_2>0$,*
10. *two 2-parametric families of Eulerian relative equilibria exist. For $k<k_0$ and $\beta_2<0$,*
11. *a unique 2-parametric family of Eulerian relative equilibria exists.*

Similar results are obtained for the configuration $S_2S_0S_1$. Figures 3, 4, 5, and 6 illustrate the bifurcations of the equilibria.

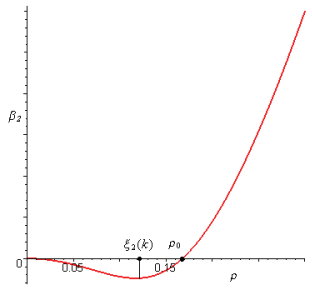


Figure 3. *Function $R_1(\rho)$ for the configuration $S_2S_1S_0$ and $k > 0$.*

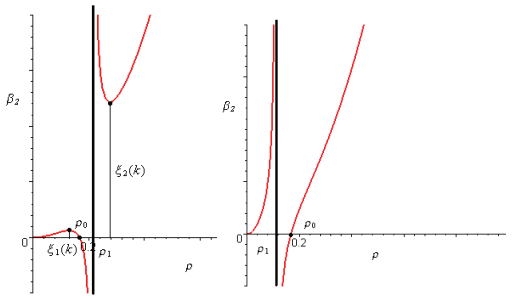


Figure 4. *Function $R_1(\rho)$ for the configuration $S_2S_1S_0$ for $k_0 < k < 0$ and $\beta_2 > 0$, $\beta_2 < 0$, respectively.*

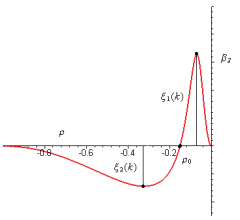


Figure 5. *Function $R_1(\rho)$ for the configuration $S_2S_0S_1$.*

5. Stability of Eulerian relative equilibria. The tangent flow of (1) in an Eulerian relative equilibrium $\mathbf{z}_e$ is expressed by

$$\frac{d\delta\mathbf{z}}{dt} = \mathfrak{U}(\mathbf{z}_e)\delta\mathbf{z},$$

with $\delta\mathbf{z} = \mathbf{z} - \mathbf{z}_e$ and $\mathfrak{U}(\mathbf{z}_e)$ being the Jacobian matrix of (1) in $\mathbf{z}_e$.

The characteristic polynomial $\mathfrak{U}(\mathbf{z}_e)$ is expressed as

$$P(X) = X^3(X^2+\Phi_0^2)(X^2+\Phi_1^2)(X^2+\Phi_2^2)(X^4+mX^2+n)(X^8+pX^6+qX^4+rX^2+s), \tag{11}$$

with $\Phi_i^2 = \frac{(C_i-A_i)\omega_i^e+l}{A_i}$. The coefficients present in the above polynomial are functions of the parameters of the problem and ρ, where ρ is taken as the root of (5).

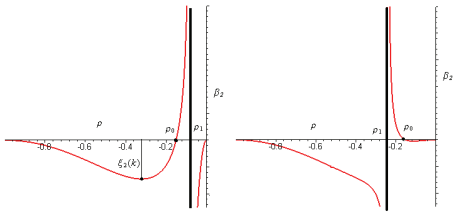


Figure 6. *Function $R_1(\rho)$ for the configuration $S_2S_0S_1$.*

5.1. S_2 and S_1 are spherical rigid bodies. The characteristic polynomial (11) of $\mathfrak{U}(\mathbf{z}_e)$ simplifies to

$$P(X) = X^5(X^2+\Phi_0^2)(X^2+\Phi_1^2)(X^2+\Phi_2^2)(X^2+(\omega_0^e)^2)^2(X^2+p)(X^4+qX^2+r) \tag{12}$$

with coefficients shown in Appendix B.

If $p \geq 0$, $q \geq 0$, $r \geq 0$, $q^2-4r \geq 0$, then $\mathbf{z}_e$ is spectrally stable. These conditions are not verified since $r < 0$.

Proposition 9. *If $\mathbf{z}_e$ is the only relative equilibrium in the configuration $S_0S_2S_1$ of the zero order approximate dynamics, then it is unstable.*

5.2. S_2 and S_1 are close to a sphere. The case in which S_i $(i=1,2)$ are close to a sphere will now be analyzed. In this case $C_i - A_i \approx 0$, and this is the reason why, by applying the implicit function theorem, $\mathbf{z}_e$ is unstable.

If $C_i - A_i$ is not close to zero, the coefficients of the polynomial (11) have very complicated expressions. Numeric calculations prove that linear stable Eulerian relative equilibria exist for certain values of the parameters $C_i - A_i$ $(i=1,2)$ (see Vera and Vigueras [8] for details). *These results are also applicable to configurations $S_2S_0S_1$ and $S_2S_1S_0$.*

6. Rotational Poisson dynamics in an Eulerian equilibrium. To describe the rotational Poisson dynamics in an Eulerian equilibrium, we consider the gyrostat fixed frame $\mathfrak{J} = \mathfrak{R}_{orb} = \{\mathbf{e}, \mathbf{b}, \mathbf{n}\}$ with origin in the mass center of the gyrostat. The versors $\mathbf{e}$ and $\mathbf{n}$ denote the radial and the orbital angular velocity directions, respectively, and $\mathbf{b} = \mathbf{n} \times \mathbf{e}$. We consider $\mathbf{\Omega}_{orb} = \mathbf{\Omega}_0 - \omega_0^e \mathbf{n}$ with ω_0^e given by (6). The rotational equations of the motion of a gyrostat placed in an Eulerian equilibrium are

$$\frac{d\mathbf{\Pi}_0}{dt} = \mathbf{\Pi}_0 \times \mathbf{\Omega}_0 + \mathfrak{M}, \quad \frac{d\mathbf{e}}{dt} = \mathbf{e} \times \mathbf{\Omega}_{\mathbf{orb}}, \quad \frac{d\mathbf{b}}{dt} = \mathbf{b} \times \mathbf{\Omega}_{\mathbf{orb}}, \quad \frac{d\mathbf{n}}{dt} = \mathbf{n} \times \mathbf{\Omega}_{\mathbf{orb}},$$

$\mathfrak{M}$ being the gravitational torque acting on the gyrostat. The following formula is verified for the gravitational torque:

$$\mathfrak{M} = \nabla_{\mathbf{e}} \mathcal{U},$$

with

$$\mathcal{U}(\mathbf{e}) = \frac{3k}{2} \mathbf{e} \cdot \mathbb{I}_0 \mathbf{e}.$$

The parameter k is given by

$$k = \frac{G}{|\boldsymbol{\lambda}^e|^3} \left(\frac{m_1}{|1+\rho|^3} + \frac{m_2}{|\rho|^3} \right),$$

and ρ is determined by (5).

Keeping in mind that $\mathbf{\Omega}_0 = \mathbb{I}_0^{-1}\mathbf{\Pi} - \mathbf{l}_r$ with $\mathbf{l}_r = \mathbb{I}^{-1}\widetilde{\mathbf{l}_r}$, the equations of motion are

$$\begin{aligned}
\frac{d\mathbf{\Pi}_0}{dt} &= \mathbf{\Pi}_0 \times \mathbb{I}_0^{-1}\mathbf{\Pi}_0 - \mathbf{\Pi}_0 \times \mathbf{l_r} + 3k\,\mathbf{e} \times \mathbb{I}\mathbf{e},\\
\frac{d\mathbf{e}}{dt} &= \mathbf{e} \times \left(\mathbb{I}_0^{-1}\mathbf{\Pi}_0 - \mathbf{l_r} - \omega_0^e\mathbf{n}\right),\\
\frac{d\mathbf{b}}{dt} &= \mathbf{b} \times \left(\mathbb{I}_0^{-1}\mathbf{\Pi}_0 - \mathbf{l_r} - \omega_0^e\mathbf{n}\right),\\
\frac{d\mathbf{n}}{dt} &= \mathbf{n} \times \left(\mathbb{I}_0^{-1}\mathbf{\Pi}_0 - \mathbf{l}_r\right).
\end{aligned}$$

As can be easily deduced, we obtain the closed system of nine equations of motion

$$\begin{aligned}
\frac{d\mathbf{\Pi}_0}{dt} &= \mathbf{\Pi}_0 \times \mathbb{I}^{-1}\mathbf{\Pi}_0 - \mathbf{\Pi}_0 \times \mathbf{l_r} + 3k\,\mathbf{e} \times \mathbb{I}\mathbf{e},\\
\frac{d\mathbf{e}}{dt} &= \mathbf{e} \times \left(\mathbb{I}^{-1}\mathbf{\Pi}_0 - \mathbf{l_r} - \omega_0^e\mathbf{n}\right),\\
\frac{d\mathbf{n}}{dt} &= \mathbf{n} \times \left(\mathbb{I}^{-1}\mathbf{\Pi}_0 - \mathbf{l}_r\right).
\end{aligned} \tag{13}$$

The system (13) is a Poisson system in $\mathbb{R}^9$. For it, given two arbitrary functions $f, g \in C^\infty(\mathbb{R}^9)$ and $\mathbf{u} \in \mathbb{R}^9$, we define the Poisson bracket $\{\cdot,\cdot\}$ by $\{f,g\} = (\nabla_{\mathbf{u}} f)^t \mathbf{\Gamma}(\mathbf{u}) \nabla_{\mathbf{u}} g$, where $\mathbf{u} = (\mathbf{\Pi}, \mathbf{e}, \mathbf{n})^t$ and $\mathbf{\Gamma}$ is

$$\mathbf{\Gamma}(\mathbf{u}) = \begin{pmatrix} \widehat{\mathbf{\Pi}} & \widehat{\mathbf{e}} & \widehat{\mathbf{n}} \\ \widehat{\mathbf{e}} & \mathbf{0} & \mathbf{0} \\ \widehat{\mathbf{n}} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Then, system (13) can be written as

$$\frac{d\mathbf{u}}{dt} = \mathbf{\Gamma}(\mathbf{u}) \boldsymbol{\nabla}_{\mathbf{u}} \mathcal{H}$$

with

$$\mathcal{H} = \frac{1}{2}\mathbf{\Pi}_0^t \mathbb{I}_0^{-1}\mathbf{\Pi}_0 - \mathbf{l_r} \cdot \mathbf{\Pi}_0 - \omega_0^e\,\mathbf{\Pi}_0 \cdot \mathbf{n} + \mathcal{U} \tag{14}$$

and $\mathcal{U}(\mathbf{e}) = \frac{3k}{2}\mathbf{e} \cdot \mathbb{I}\mathbf{e}$.

We can therefore conclude that (13) is a Hamiltonian system in the Poisson manifold $(\mathbb{R}^9, \{\cdot,\cdot\})$ with Hamiltonian function (14). The Poisson structure is noncanonical; that is, there exist nonconstant Casimir functions which are the following geometric integrals:

$$\Phi_1(\mathbf{u}) = \mathbf{e} \cdot \mathbf{e}, \quad \Phi_2(\mathbf{u}) = \mathbf{n} \cdot \mathbf{n}, \quad \Phi_3(\mathbf{u}) = \mathbf{e} \cdot \mathbf{n}.$$

6.1. Rotational equilibrium conditions. We know that a necessary and sufficient condition in order for $\mathbf{u}_e = (\mathbf{\Pi}_e, \mathbf{e}_e, \mathbf{n}_e)^t$ to be a equilibrium of the equations of motion (13) is that there must exist $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}$ such that $\mathbf{d}(\widetilde{\mathcal{H}}) = \mathbf{d}(\mathcal{H} - \lambda_1\Phi_1 - \lambda_2\Phi_2 - \lambda_3\Phi_3) = \mathbf{0}$. From this, we deduce the following vectorial equations:

$$\begin{aligned} \mathbb{I}^{-1}\mathbf{\Pi}_0^e - \mathbf{l}_r - \omega_0^e\mathbf{n}_e &= \mathbf{0}, \\ 3k\mathbb{I}\mathbf{e}_e - 2\lambda_1\mathbf{e}_e - \lambda_3\mathbf{n}_e &= \mathbf{0}, \\ -\omega_0^e\mathbf{\Pi}_e - 2\lambda_2\mathbf{n}_e - \lambda_3\mathbf{e}_e &= \mathbf{0}. \end{aligned}$$

From these equations we obtain the relations

$$\mathbf{\Omega}_0^e = \omega_0^e\mathbf{n}_e, \quad \lambda_1 = \tfrac{3k}{2}\mathbf{e}_e \cdot \mathbb{I}_0^{-1}\mathbf{e}_e, \quad \lambda_2 = -\tfrac{\omega_0^e}{2}\mathbf{\Pi}_0^e \cdot \mathbf{n}_e, \quad \lambda_3 = 0.$$

We call *cylindrical equilibrium* to the following expression:

$$\begin{aligned} &\mathbf{\Pi}_0^e = (0, 0, C_0\omega_0^e + l), \quad \mathbf{e}_e = (1, 0, 0), && \mathbf{n}_e = (0, 0, 1), \\ &\lambda_1 = \frac{3k}{2A}, && \lambda_2 = -\frac{\omega_e(C_0\omega_0^e + l)}{2}, \quad \lambda_3 = 0. \end{aligned}$$

In this paper we obtain necessary and sufficient conditions for the nonlinear stability of the cylindrical equilibrium. Other equilibria and stability conditions are studied in an upcoming paper.

6.2. Necessary and sufficient conditions of stability of the cylindrical equilibrium. By means of the tangent flow of this dynamics in the cylindrical equilibrium, we obtain necessary conditions for the nonlinear stability. After some standard calculations we obtain the following expressions:

$$\begin{aligned} &C_0 - A_0 \geq 0, \\ &p = C_0[\omega_e^2(B_0 - C_0) + \omega_e l] + A[((\omega_0^e)^2 - 3k)(B_0 - A_0) + \omega_0^e l] \\ &+ [\omega_0^e(B_0 - C_0 - A_0) + l]^2 \geq 0, \\ &q = [(\omega_0^e)^2(B_0 - C_0) + \omega_e l][((\omega_0^e)^2 + 3k)(B_0 - A_0) + \omega_0^e l] \geq 0, \\ &p^2 - 4qC_0A_0 \geq 0. \end{aligned}$$

By means of the energy-Casimir method (see [6] for details), we obtain conditions for nonlinear stability of the cylindrical equilibria. Following the process, we will need to evaluate

$$\mathbf{d}^2(\widetilde{\mathcal{H}})(\mathbf{u}_e) = \begin{pmatrix} \mathbb{I}^{-1} & 0 & -\omega_0^e\mathbf{I}_{\mathbb{R}^3} \\ 0 & 3k\mathbb{I} - 2\lambda_1\mathbf{I}_{\mathbb{R}^3} & 0 \\ -\omega_0^e\mathbf{I}_{\mathbb{R}^3} & 0 & -2\lambda_2\mathbf{I}_{\mathbb{R}^3} \end{pmatrix}$$

with $\lambda_1 = 3k/2A$ and $\lambda_2 = -\omega_e(C\omega_e + l)/2$. This matrix restricted to

$$\mathbf{W} = \mathrm{Ker}(\mathbf{d}\Phi_1(\mathbf{u}_e)) \cap \mathrm{Ker}(\mathbf{d}\Phi_2(\mathbf{u}_e)) \cap \mathrm{Ker}(\mathbf{d}\Phi_3(\mathbf{u}_e))$$

is

$$\mathbf{d}^2(\widetilde{\mathcal{H}})_{|\mathbf{W}}(\mathbf{u}_e) = \begin{pmatrix} \frac{1}{A_0} & 0 & 0 & 0 & -\omega_0^e & 0 \\ 0 & \frac{1}{B_0} & 0 & 0 & 0 & -\omega_0^e \\ 0 & 0 & \frac{1}{C_0} & 0 & 0 & 0 \\ 0 & 0 & 0 & 3k(B_0 - A_0) & 0 & 0 \\ -\omega_0^e & 0 & 0 & 0 & 3k(C_0 - A_0) + \omega_0^e(C_0\omega_0^e + l) & 0 \\ 0 & -\omega_0^e & 0 & 0 & 0 & \omega_0^e(C_0\omega_0^e + l) \end{pmatrix}.$$

Using Sylvester's theorem, we obtain the following conditions for nonlinear stability of the cylindrical equilibrium:

$$\begin{aligned} &C_0 - A_0 > 0, \\ &(\omega_0^e)^2(B_0 - C_0) + \omega_0^e l > 0, \\ &((\omega_0^e)^2 + 3k)(B_0 - A_0) + \omega_0^e l > 0. \end{aligned}$$

7. Conclusions and future work. In this paper we have investigated some important periodic solutions of the dynamics of a gyrostat in Newtonian interaction with two symmetric rigid bodies. With the hypotheses formulated in the introduction, working in the double reduced space of configuration of the problem, both the equations of motion and those which determine the relative equilibria have been derived. The Eulerian relative equilibria have been completely determined by a polynomial equation of degree nine. The obtained results generalize those of [12, 1, 9]. The bifurcations of these equilibria have been carried out when m_0 is very small. The instability of Eulerian relative equilibria has been proven if the gyrostat S_0 is close to a sphere. The rotational Poisson dynamics of the gyrostat placed in an Eulerian equilibrium is considered. The nonlinear stability of some rotational equilibria, called cylindrical equilibria, is studied by means of the energy-Casimir method. By means of the tangent flow of this dynamics in the cylindrical equilibrium, we obtain necessary conditions for the nonlinear stability of the cylindrical equilibria. Diverse results, which had been obtained by means of classic methods in previous works, have been obtained and generalized in a different way. The methods employed in this work are suitable for use in similar problems. Numerous problems are open; among them, it is necessary to consider the study of the "inclined" relative equilibria, in which $\mathbf{\Omega}_0^e$ forms an angle $\alpha \neq 0$ and $\pi/2$ with the vector $\boldsymbol{\lambda}^e$.

Appendix A. Some numerical results. In order to obtain the values of $C_i - A_i$ $(i = 1, 2)$, the following relationships will be utilized:

$$\begin{aligned} C_1 - A_1 &= (1 - \mu)\left(\frac{e_{S_1}}{Z}\right)^2 J_2^{S_1}, \\ C_2 - A_2 &= \mu\left(\frac{e_{S_2}}{Z}\right)^2 J_2^{S_2}, \end{aligned}$$

where e_{S_i} and p_{S_i} represent the equatorial and polar radii of S_i $(i = 1, 2)$, $J_2^{S_i} = \frac{2}{5}\varepsilon_i$, respectively, and $\varepsilon_i = \frac{e_{S_i} - p_{S_i}}{e_{S_i}}$. S_1, S_2 are considered to be homogeneous ellipsoids. The distances are measured in kilometers.

$S_2S_1S_0$ $(m_0 \to 0)$	Without oblat.	Oblat. of S_2	Oblat. of S_2 and S_1
Earth-Moon-S_0	448879.206	448879.221	448879.251
Mars-Phobos-S_0	9414.945	9414.958	9414.958
$S_0S_2S_1$ $(m_0 \to 0)$	Without oblat.	Oblat. of S_2	Oblat. of S_2 and S_1
S_0-Earth-Moon	381679.691	381679.763	381679.763
S_0-Mars-Phobos	9310.642	9310.666	9310.668
$S_2S_0S_1$ $(m_0 \to 0)$	Without oblat.	Oblat. of S_2	Oblat. of S_2 and S_1
Earth-S_0-Moon	326409.744	326409.780	326409.751
Mars-S_0-Phobos	9339.156	9339.196	9339.196

Appendix B. Coefficients of the characteristic polynomial in Eulerian relative equilibria $S_0S_2S_1$.

The coefficients of the characteristic polynomial (12) are

$$\omega_e^2 = \frac{G((m_2+m_1)\rho^4 + (2m_1+2m_2)\rho^3 + (m_2+m_1)\rho^2 - 2m_0\rho - m_0)}{\lambda_e^3(1+\rho)^2\rho^2},$$

$$p = \frac{G((m_2+4m_0+m_1)\rho^3 + (3m_2+6m_0)\rho^2 + (4m_0+3m_2)\rho + m_0 + m_2)}{(1+\rho)^3\rho^3\lambda_e^3},$$

$$\begin{aligned} q = {} & \frac{G((-2m_1\rho^4 m_2 + (-2m_0m_1 + m_1^2 + m_2^2 - 2m_1m_2 - 2m_0m_2)\rho^3}{((1+\rho)^3\rho^3\lambda_e^3)} \\ & + \frac{(3m_2^2 + m_1m_2 - 6m_0m_1)\rho^2 + (-m_1m_2 + 3m_2^2 + 2m_0m_2 - 4m_0m_1)\rho}{((1+\rho)^3\rho^3\lambda_e^3)} \\ & + \frac{m_2^2 - m_0m_1 + m_0m_2 - m_1m_2))}{((1+\rho)^3\rho^3\lambda_e^3)}, \end{aligned}$$

$$r = \frac{G^2(a_1\rho^4 + a_2\rho^4 + a_3\rho^2 + a_4\rho + a_5)}{((1+\rho)^8\rho^8\lambda_e^9)}.$$

B.1. Coefficients a_i $(i = 1, \ldots, 5)$.

$$\begin{aligned} a_1 = {} & -42m_2^7m_1 - 48m_2^7m_0 - 147m_2^6m_1^2 - 336m_2^6m_1m_0 - 129m_2^6m_0^2 \\ & - 207m_2^5m_1^3 - 782m_2^5m_1^2m_0 - 673m_2^5m_1m_0^2 - 81m_2^5m_0^3 - 150m_2^4m_1^4 \\ & - 869m_2^4m_1^3m_0 - 1325m_2^4m_1^2m_0^2 - 378m_2^4m_1m_0^4 - 64m_2^3m_1^5 \\ & - 513m_2^3m_1^4m_0 - 1270m_2^3m_1^3m_0^2 - 702m_2^3m_1^2m_0^3 - 14m_2^2m_1^6 \\ & - 165m_2^2m_1^5m_0 - 610m_2^2m_1^4m_0^2 - 648m_2^2m_1^3m_0^3 - 24m_2m_1^6m_0 \\ & - 119m_2m_1^5m_0^2 - 297m_2m_1^4m_0^3 + 2m_1^6m_0^2 - 54m_1^5m_0^3, \end{aligned}$$

$$a_2 = -60m_2^7m_1 - 54m_2^7m_0 - 243m_2^6m_1^2 - 474m_2^6m_1m_0 - 173m_2^6m_0^2 - 399m_2^5m_1^3 - 1345m_2^5m_1^2m_0 - 999m_2^5m_1m_0^2 - 135m_2^5m_0^3 - 329m_2^4m_1^4 - 1846m_2^4m_1^3m_0 - 2223m_2^4m_1^2m_0^2 - 648m_2^4m_1m_0^3 - 138m_2^3m_1^5 - 1364m_2^3m_1^4m_0 - 2506m_2^3m_1^3m_0^2 - 1242m_2^3m_1^2m_0^3 - 24m_2^2m_1^6 - 536m_2^2m_1^5m_0 - 1530m_2^2m_1^4m_0^2 - 1188m_2^2m_1^3m_0^3 - 90m_2m_1^6m_0 - 477m_2m_1^5m_0^2 - 567m_2m_1^4m_0^3 - 56m_1^6m_0^2 - 108m_1^5m_0^3,$$

$$a_3 = -42m_2^7m_1 - 36m_2^7m_0 - 183m_2^6m_1^2 - 342m_2^6m_1m_0 - 93m_2^6m_0^2 - 349m_2^5m_1^3 - 1097m_2^5m_1^2m_0 - 630m_2^5m_1m_0^2 - 81m_2^5m_0^3 - 358m_2^4m_1^4 - 1776m_2^4m_1^3m_0 - 166m_2^4m_1^2m_0^2 - 405m_2^4m_1m_0^3 - 189m_2^3m_1^5 - 1614m_2^3m_1^4m_0 - 2256m_2^3m_1^3m_0^2 - 810m_2^3m_1^2m_0^3 - 31m_2^2m_1^6 - 827m_2^2m_1^5m_0 - 1683m_2^2m_1^4m_0^2 - 810m_2^2m_1^3m_0^3 - 6m_2m_1^7 - 228m_2m_1^6m_0 - 666m_2m_1^5m_0^2 - 405m_2m_1^4m_0^3 - 30m_1^7m_0 - 81m_1^5m_0^3 - 111m_1^6m_0^2,$$

$$a_4 = -12m_2^7m_1 - 12m_2^7m_0 - 56m_2^6m_1^2 - 114m_2^6m_1m_0 - 24m_2^6m_0^2 - 130m_2^5m_1^3 - 387m_2^5m_1^2m_0 - 162m_2^5m_1m_0^2 - 179m_2^4m_1^4 - 687m_2^4m_1^3m_0 - 432m_2^4m_1^2m_0^2 - 140m_2^3m_1^5 - 693m_2^3m_1^4m_0 - 588m_2^3m_1^3m_0^2 - 52m_2^2m_1^6 - 387m_2^2m_1^5m_0 - 432m_2^2m_1^4m_0^2 - 6m_2m_1^7 - 108m_2m_1^6m_0 - 162m_2m_1^5m_0^2 - 12m_1^7m_0 - 24m_1^6m_0^2,$$

$$a_5 = -(m_0+m_2)(18m_0m_2^6 + 12m_1m_2^6 + 94m_2^5m_0m_1 + 36m_1^2m_2^5 + 81m_2^4m_0^2m_1 + 168m_2^4m_0m_1^2 + 42m_2^4m_1^3 + 128m_2^3m_0m_1^3 + 27m_2^3m_1^4 + 15m_2^2m_1^5 + 31m_2^2m_0m_1^4 + 126m_2^2m_0^2m_1^3 + 18m_0^2m_2^5 + 54m_2m_0^2m_1^4 + 12m_2m_0m_1^5 + 5m_2m_1^6 + 7m_1^6m_0 + 9m_0^2m_1^5 + 144m_2^3m_0^2m_1^2).$$

Acknowledgment. The author is grateful to the anonymous referees for their useful suggestions and comments, which improved the quality of the paper.

REFERENCES

[1] G. N. Dubochine, *The problem of three rigid bodies*, Celestial Mech. Dynam. Astronom., 33 (1981), pp. 31–47.

[2] D. W. Dunham and R. W. Farquhar, *Libration point missions:* 1978–2002, in Proceedings of the Conference on Libration Point Orbits and Applications, World Scientific, River Edge, NJ, 2002, pp. 45–74.

[3] E. Leimanis, *The General Problem of the Motion of Coupled Rigid Bodies about a Fixed Point*, Springer-Verlag, Berlin, 1965.

[4] A. Maciejewski, *Reduction, relative equilibria and potential in the two rigid bodies problem*, Celestial Mech. Dynam. Astronom., 63 (1995), pp. 1–28.

[5] F. Mondéjar and A. Vigueras, *The Hamiltonian dynamics of the two gyrostats problem*, Celestial Mech. Dynam. Astronom., 73 (1999), pp. 303–312.
[6] J. P. Ortega and T. S. Ratiu, *Stability of Hamiltonian relative equilibria*, Nonlinearity, 12 (1999), pp. 693–720.
[7] J. A. Vera, *Reducciones, equilibrios y estabilidad en dinámica de sólidos rígidos y giróstatos*, Ph.D. dissertation, Universidad Politécnica de Cartagena, Cartagena, Spain, 2004.
[8] J. A. Vera and A. Vigueras, *Hamiltonian dynamics of a gyrostat in the n-body problem: Relative equilibria*, Celestial Mech. Dynam. Astronom., 94 (2006), pp. 289–315.
[9] V. V. Vidyakin, *Euler solutions in the problem of translational-rotational motion of three-rigid bodies*, Celestial Mech., 16 (1977), pp. 509–526.
[10] V. Volterra, *Sur la théorie des variations des latitudes*, Acta Math., 22 (1899), pp. 201–357.
[11] L. S. Wang, P. S. Krishnaprasad, and J. H. Maddocks, *Hamiltonian dynamics of a gyrostat in a central gravitational field*, Celestial Mech. Dynam. Astronom., 50 (1991), pp. 349–386.
[12] A. Wintner, *The Analytical Foundations of Celestial Mechanics*, Princeton University Press, Princeton, NJ, 1941.
[13] S. G. Zhuravlev and A. A. Petrutskii, *Current state of the problem of translational-rotational motion of three-rigid bodies*, Soviet Astronom., 34 (1990), pp. 299–304.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1397–1420
© 2008 Society for Industrial and Applied Mathematics

Snaking of Multiple Homoclinic Orbits in Reversible Systems*

J. Knobloch† and T. Wagenknecht‡

Abstract. We study N-homoclinic orbits near a heteroclinic cycle in a reversible system. The cycle is assumed to connect two equilibria of saddle-focus type. Using Lin's method, we establish the existence of infinitely many N-homoclinic orbits for each N near the cycle. In particular, these orbits exist along snaking curves, thus mirroring the behavior of 1-homoclinic orbits. The general analysis is illustrated by numerical studies for a Swift–Hohenberg system.

Key words. bifurcation, multiple homoclinic orbits, heteroclinic cycle, homoclinic snaking, Lin's method

AMS subject classifications. 34C37, 37C29, 37G20

DOI. 10.1137/070695800

1. Introduction.

Spatially localized structures, such as solitary pulses, appear in many systems described by higher-order nonlinear partial differential equations (PDEs). Particular examples have been found in structural mechanics [11], nonlinear optics [20], and water wave problems [4]. A common feature of these cases is the onset of the localized patterns in a sequence of fold bifurcations, which are connected by a snaking curve.

In one spatial dimension this phenomenon can be explained by a sequence of bifurcations in the associated ordinary differential equation (ODE) for traveling waves. Localized patterns correspond to homoclinic solutions of this ODE, and it has been found that infinitely many of such orbits can exist near a heteroclinic cycle in the ODE. These homoclinic orbits all lie on a snaking curve, along which they undergo infinitely many fold bifurcations, while thereby getting wider and developing new oscillations about their center at each fold. An important requirement for this scenario to happen is the time-reversibility of the traveling-wave ODE.

In addition to time-reversibility, the type of the heteroclinic cycle plays an important role for the dynamics. Homoclinic snaking has been observed in a neighborhood of a heteroclinic cycle between an equilibrium p_1 and a saddle-focus equilibrium p_2 (EE cycle) and near cycles connecting an equilibrium p_1 and a periodic orbit P (EP cycle). The homoclinic orbits occurring along a snaking curve are asymptotic to p_1.

In the case of an EE cycle the snaking occurs locally around some critical value of a family parameter at which a codimension one heteroclinic cycle exists. This feature allows one to study the scenario using a local bifurcation analysis. In an earlier paper [14], it was shown rigorously by the authors that heteroclinic cycles between symmetric equilibria of saddle-focus type generate a snaking behavior. In contrast to that behavior, the snaking related to an EP

*Received by the editors June 28, 2007; accepted for publication (in revised form) by B. Sandstede June 18, 2008; published electronically November 7, 2008.
http://www.siam.org/journals/siads/7-4/69580.html
†Department of Mathematics, TU-Ilmenau, D-98684 Ilmenau, Germany (juergen.knobloch@tu-ilmenau.de).
‡School of Mathematics, University of Manchester, Sackville Street, Manchester, M60 1QD, UK. Current address: Department of Applied Mathematics, University of Leeds, Leeds, LS2 9JT, UK (thomas@maths.leeds.ac.uk).

cycle is generically a global phenomenon in parameter space; that is, it can be observed in a region in parameter space. (This difference is also apparent in the shape of corresponding snaking curves (see Figures 1 and 7).) For a geometric explanation for why homoclinic snaking occurs in this case, we refer the reader to [23, 7]; see also [1] for a recent analytical description.

If p_1 is also of saddle-focus type, then general results by Häerterich [8] (see also [3, 18]) show that homoclinic orbits to p_1 will be accompanied by a plethora of N-homoclinic orbits, i.e., homoclinic orbits to p_1 that pass p_2 N times before closing the loop. For each N there exist infinitely many N-homoclinic orbits, which are distinguished by the times they spend near p_1. In the situation studied here, one may now expect these orbits to snake under variation of the parameter, too. Our goal is to describe this snaking.

More precisely, we understand a snaking curve as a graph $\{(\omega, \lambda(\omega)), \omega \in I\}$ that intersects a line $\{(\omega, \lambda^*), \omega \in I\}$ infinitely many times. Moreover, in the case of an EE cycle, $\lambda(\omega)$ tends to λ^* as ω tends to infinity. So, a snaking curve looks qualitatively similar to the one in Figure 1. Here ω is some intrinsic parameter characterizing the N-homoclinic orbit, taken from some infinite interval I, and $\lambda(\omega)$ is the family parameter of the ODE at which the N-homoclinic orbit exists. Roughly speaking, ω is the length of stay near p_2 during a certain passage of the N-homoclinic orbit near p_2. In the analysis in section 4, we will particularly focus on the case where ω is the first passage past p_2, and we will discuss other possibilities only briefly.

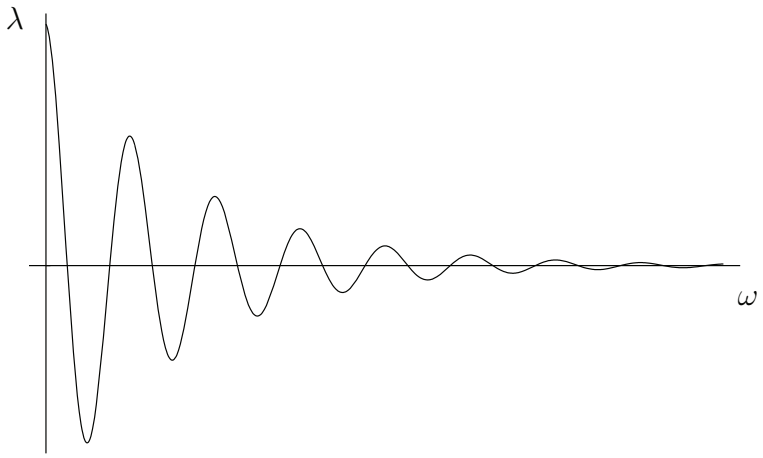


Figure 1. *Snaking curve for symmetric N-homoclinic orbits.*

This paper can be seen as a follow-up to [14], where we discussed 1-homoclinic orbits near an EE cycle. There it was also shown that p_2 has to be of saddle-focus type in order to find snaking behavior of 1-homoclinic orbits. Furthermore, it was shown there that p_1 has to be of saddle-focus type, too, if N-homoclinic orbits to this equilibrium are to exist.

Similar to the procedure in [14] we will use Lin's method [12, 17] to derive bifurcation equations for N-homoclinic orbits near the cycle. The general setup will be introduced in section 3. Under certain genericity assumptions it will be shown in section 4 that the cycle is accompanied by a multitude of N-homoclinic orbits, which exist on snaking curves. In section 5 we briefly discuss snaking near EP cycles. Numerical results for this case disclose

a remarkable difference in the behavior of 2-homoclinic orbits in comparison with the results for EE cycles, derived in the present paper.

Before the general bifurcation analysis, however, we will illustrate the problem we are interested in by numerical results for a generalized Swift–Hohenberg equation in the next section.

2. N-pulses in a generalized Swift–Hohenberg equation. We consider the generalized Swift–Hohenberg equation studied in [10]

$$\frac{\partial u}{\partial t} = ru - \left(\partial_x^2 + q_c^2\right)^2 u + vu^2 - gu^3. \tag{2.1}$$

Stationary localized solutions of this equation, that is, homoclinic solutions to 0 of the fourth-order equation

$$u'''' + 2q_c^2 u'' + (q_c^4 - r)u - vu^2 + gu^3 = 0, \tag{2.2}$$

have been discussed in detail in [2]. In particular, several snaking scenarios have been found to occur in (2.2).

We stress the fact that (2.2) is reversible. This means that there is a linear involution R such that the corresponding first-order system is invariant under the transformation $((u, u', u'', u'''), x) \mapsto (R(u, u', u'', u'''), -x)$. Here the involution R is defined by $(u, u', u'', u''') \mapsto (u, -u', u'', -u''')$. The reversibility plays an important role for the dynamics of (2.2). In particular, it leads to the robust existence of those homoclinic solutions, for which u is an even function. As is common, we will call such solutions *symmetric*. We refer the reader to section 3 for more general comments about reversible systems.

Note that, in addition to being reversible, (2.2) is also conservative and preserves a first integral. However, both in our computations for (2.2) and in the general analysis afterward, we will focus on symmetric homoclinics, and for these the reversibility of (2.2) is the most important property.

Following computations in [2], we set $q_c = 0.5$, $v = 0.75$, and $g = 1$ and consider (2.2) as an equation depending only on r. Then the 0 equilibrium is a saddle focus for all $r < 0$; i.e., the linear part of the vector field has a quadruple of complex eigenvalues. Furthermore, there are two additional equilibria $u_\pm = (3 \pm \sqrt{5 + 64r})/8$ if $r > -5/64$.

We compute symmetric homoclinic orbits to 0 by shooting for orbits in the unstable manifold of this equilibrium, which intersect the symmetry section $\mathrm{Fix}(R) = \{(u, u', u'', u''') : u' = u''' = 0\}$. The behavior and bifurcations of these orbits under variation of r are studied using the software package AUTO/HomCont [24].

Figure 2 shows a bifurcation diagram for a particular 1-homoclinic solution H_1. In this diagram we plot the L^2-norm of the solution vector (u, u', u'', u''') for H_1 against the parameter r. We see that H_1 emerges at $r = 0$ in a local bifurcation of the 0-equilibrium and then undergoes a sequence of fold bifurcations along a snaking curve which accumulate at $r^* = -1/16$.

Inspection of the solutions, contained in the accompanying boxes, shows that along the snaking curve the middle part of the homoclinic orbit approaches the equilibrium u_+, and

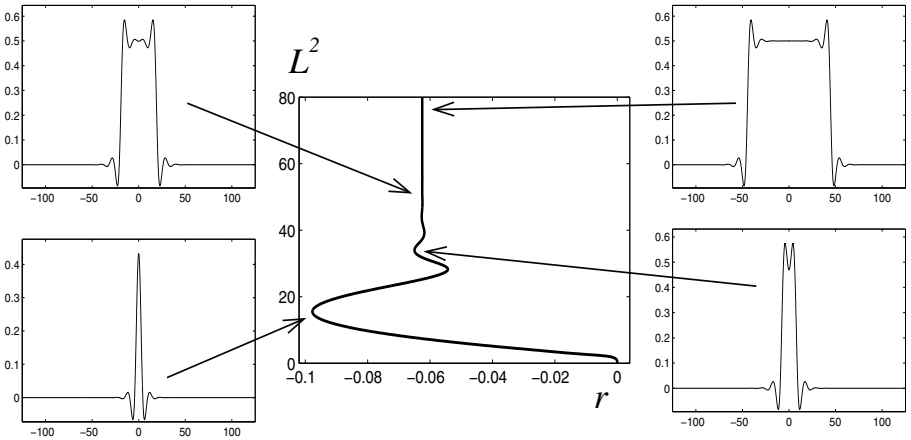


Figure 2. *Snaking curve for* 1*-homoclinic orbits of* (2.2).

indeed a heteroclinic cycle Γ between 0 and u_+ is found to exist at r^*; see also [2]. Note that Γ is invariant under R; it is the limit of symmetric homoclinic orbits.

Remark 2.1. *In all of the numerically obtained diagrams in this paper we plot the L^2-norm of the solution u against the parameter because this solution measure is numerically convenient. We note, however, that this norm is also directly related to the transition times, which will be used in the general analysis (compare with Figure 1). Indeed, the part of the solution that is close to u_+ is the one that predominantly contributes to its L^2-norm.*

Note that for $r = r^*$ there exist infinitely many homoclinic orbits to 0. Since 0 is of saddle-focus type for that parameter value, the results by Härterich [8] suggest the existence of N-homoclinic orbits to 0. These N-homoclinic orbits are computed using a homoclinic branch-switching method developed in [16]. We find a multitude of symmetric N-homoclinic orbits and consider only a few of them with $N = 2, 3$ here.

Figure 3 contains a bifurcation diagram for two symmetric 2-homoclinic orbits existing near the cycle Γ. As before, we plot the L^2-norm against the parameter r and find a snaking curve for each orbit, which accumulates at $r = r^*$. Note that the snaking curves in the diagram stop at some finite norm. This is caused by numerical difficulties in continuing the solutions.

Some solutions along the green curve are shown to the right of the diagram. As for H_1, we find that along the snaking curve the pulses widen, and their middle parts approach u_+. It is interesting to observe that the central part of the solutions between the pulses remains unchanged.

We note that the diagram in Figure 3 shows only half of the L^2-norm of the 2-homoclinic solutions, in order to allow for comparison with the snaking curve for the 1-homoclinic orbit. (The snaking curve for H_1 is shown in grey.) The green snaking curve is not close to the 1-homoclinic snaking curve, but we find a much better approximation in the red curve. This red curve corresponds to a 2-homoclinic solution for which the two pulses are further separated; compare also with the solution plots to the left of the diagram. This suggests that the snaking curves move closer together if the pulses become further separated. However, we encounter numerical difficulties in finding and continuing such solutions.

Next we discuss numerical results for 3-homoclinic orbits near Γ. In Figure 4 we present

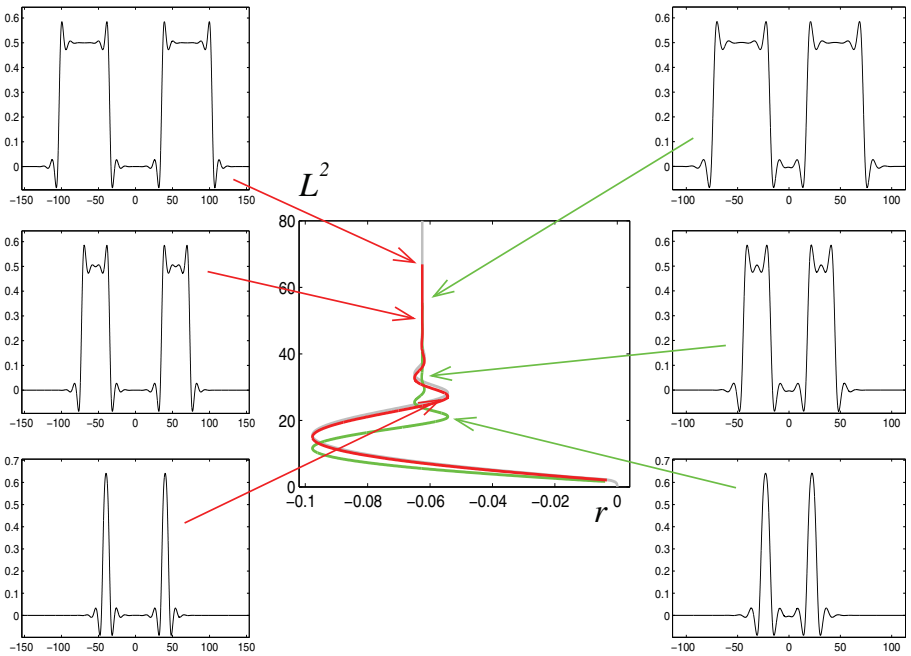


Figure 3. *Snaking curves for* 2*-homoclinic orbits of* (2.2).

two bifurcation diagrams for 3-homoclinic orbits near H_1. The orbits in this figure are created in saddle-node bifurcations and follow only parts of the snaking curve of H_1. Nevertheless, as the L^2-norm of the solutions increases, the curves approach the parameter value r^*.

But now the solution plots show that there are two different types of behavior along the snaking curves. Along the red snaking curves the outer pulses develop additional oscillations, whereas the middle pulse hardly changes. On the other hand, for orbits on the green curve it is the middle pulse that spreads out, and the outer pulses remain unchanged. In fact, solutions along the green curve approach a 2-*heteroclinic cycle* between 0 and u_+. The existence of such cycles will be investigated in section 4.4. We note that for both types of snaking the times separating the pulses are again virtually constant along all snaking curves.

As before, we rescale the L^2-norm of the 3-homoclinic orbits in Figure 4 to make the curves comparable with the one for H_1, whose snaking curve (in parts) is again shown in grey. Note that we again plot half of the L^2-norm of the solution vector, in order to accommodate the behavior along the red curves. This means that we plot an approximation to the L^2-norm of one of the outer pulses. Hence, there is a better match of the red and grey curves.

In summary, we have found 2- and 3-homoclinic orbits near H_1 which mirror the behavior of this 1-homoclinic solution in that they exist on snaking curves, along which certain pulses become wider and approach a steady state u_+. Furthermore, we expect to find different types of behavior along snaking curves for N-homoclinic orbits with $N > 2$.

3. Notation and setup. We aim to understand the numerical results above in a more general context and are thus interested in bifurcations from a heteroclinic cycle between two symmetric equilibria of saddle-focus type in the class of time-reversible systems. Let us

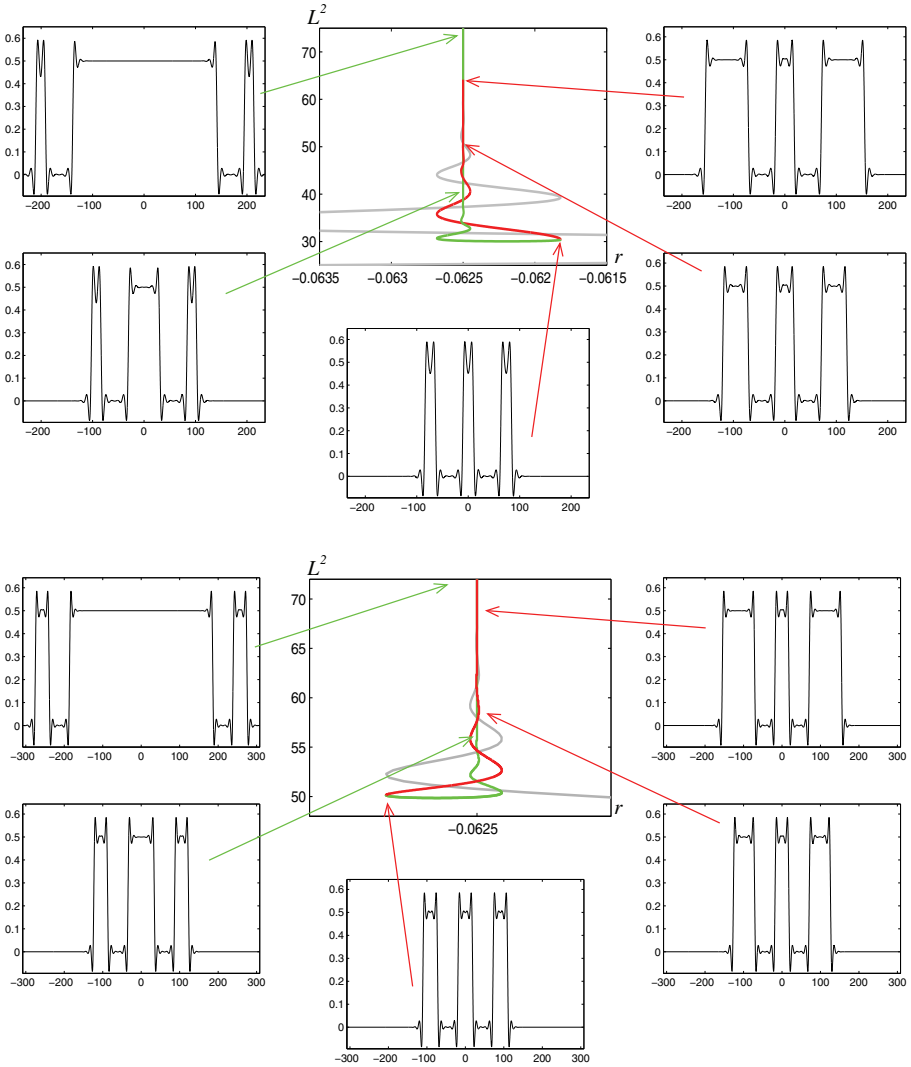


Figure 4. *Two bifurcation diagrams for 3-homoclinic orbits near the heteroclinic cycle. Two different types of snaking behavior are encountered along the red and green curves, respectively.*

describe this configuration in detail.

We consider a system of ODEs

$$\dot{x} = f(x, \lambda), \quad x \in \mathbb{R}^{2n}, \ \lambda \in \mathbb{R}, \tag{3.1}$$

with a smooth vector field f, which is assumed to be (time-)reversible; that is, the vector fields anticommute with some linear involution R:

$$Rf(x, \lambda) = -f(Rx, \lambda).$$

In a reversible system, the image RX of an orbit X is also an orbit. Orbits for which $RX = X$ are called *symmetric*. It is well known that orbits of a reversible system are symmetric if and

only if they intersect the fixed space $\mathrm{Fix}(R) := \{x \in \mathbb{R}^{2n} : Rx = x\}$ of the involution R. We refer the reader to [15, 21] for a collection of fundamental results about reversible systems.

We will be concerned with bifurcations from a heteroclinic cycle Γ in (3.1). More precisely, Γ is a collection of two equilibria p_1, p_2 and heteroclinic orbits connecting p_1 to p_2 and p_2 to p_1 in this order, respectively. Let us first discuss the equilibria.

We assume that (3.1) possesses two symmetric equilibria of saddle-focus type. More precisely, we assume that there are $p_1, p_2 \in \mathrm{Fix}(R)$ such that

$$f(p_k, 0) = 0, \quad k = 1, 2,$$

and the stable spectrum of $D_x f(p_k, 0)$ has the structure

$$\sigma^s(D_x f(p_k, 0)) = \{-\mu_k(0) \pm \varphi_k(0)i\} \cup \sigma_k^{ss}, \quad \text{with } \mu_k(0), \varphi_k(0) > 0. \tag{3.2}$$

In (3.2), σ_k^{ss} denotes the strong stable spectrum such that $\Re\mu < -\mu_k$ for all $\mu \in \sigma_k^{ss}$. Moreover, the principal eigenvalues $-\mu_k(0) \pm \varphi_k(0)i$ are assumed to be simple. Due to the symmetry of the equilibria p_k, we find for the unstable spectrum that

$$\sigma^u(D_x f(p_k, 0)) = -\sigma^s(D_x f(p_k, 0)). \tag{3.3}$$

By hyperbolicity the equilibria will persist as symmetric equilibria for small λ, and thus, with no loss of generality, we may assume that $f(p_k, \lambda) = 0$ for all λ. Furthermore, the leading eigenvalues of the linearized vector field vary smoothly with λ; this means that both $\mu_k(\cdot)$ and $\varphi_k(\cdot)$ are smooth functions of λ.

Let $W^{s(u)}(p_k, \lambda)$ denote the (un)stable manifold of p_k with respect to $f(p_k, \lambda)$. Again, due to the symmetry and hyperbolicity of the equilibria, these manifolds are n-dimensional; see also (3.3). Moreover, reversibility implies $W^s(p_k, \lambda) = RW^u(p_k, \lambda)$.

Finally, we assume the existence of a heteroclinic orbit $\Gamma_1 = \{\gamma_1(t) : t \in \mathbb{R}\}$ connecting p_1 to p_2 for $\lambda = 0$. By reversibility, this orbit is part of a heteroclinic cycle Γ, together with $\Gamma_2 = R\Gamma_1$ and the equilibria p_1 and p_2. Our analysis will require certain nondegeneracy conditions to be fulfilled. These will be imposed on Γ_1, and reversibility ensures that they are fulfilled along Γ_2, too.

First, we assume Γ_1 to be nondegenerate; that is, we assume

$$\dim\left(T_{\gamma_1(0)}W^u(p_1, 0) \cap T_{\gamma_1(0)}W^s(p_2, 0)\right) = 1, \tag{3.4}$$

where T_qM denotes the tangent space of a manifold M at the point q. As a consequence of (3.4), the equation $\dot{v} = -D_x f(\gamma_1(t), 0)^* v$ has a unique bounded solution ψ_1. We assume that both γ_1 and ψ_1 converge along the leading directions to the equilibria and zero, respectively; that is, we assume

$$\lim_{t\to-\infty} e^{\mu_1 t}\|\gamma_1(t) - p_1\| \neq 0, \qquad \lim_{t\to\infty} e^{\mu_2 t}\|\gamma_1(t) - p_2\| \neq 0, \tag{3.5}$$

$$\lim_{t\to-\infty} e^{\mu_2 t}\|\psi_1(t)\| \neq 0, \qquad \lim_{t\to\infty} e^{\mu_1 t}\|\psi_1(t)\| \neq 0. \tag{3.6}$$

Conditions (3.5) and (3.6) are known as nonorbit flip and noninclination flip conditions, respectively [18]; see also [13] for an equivalent geometric statement.

Finally, we also assume a generic unfolding of the heteroclinic connection Γ_1, which should break up under variation of the parameter λ. This can be ensured by assuming that a Melnikov integral M does not vanish [18]:

$$M := \int_{\infty}^{\infty} \langle \psi_1(t), D_\lambda f(\gamma_1(t), 0) \rangle \, dt \neq 0. \tag{3.7}$$

3.1. The main result. We are interested in N-homoclinic orbits to p_1 that lie in a sufficiently small neighborhood of the cycle Γ. Our main result reads as follows.

Theorem 3.1. *Consider* (3.1) *near the heteroclinic cycle* Γ *under the conditions set up in section* 3. *Let* $N \geq 1$ *be fixed.*

At $\lambda = 0$ *there exist countably many symmetric* N*-homoclinic orbits* $\mathcal{H}$ *to* p_1 *which can be continued on a snaking curve* $\lambda_{\mathcal{H}}(\cdot)$ *defined on* $(\Omega_{\mathcal{H}}, \infty)$. *The functions* $\lambda_{\mathcal{H}}$ *have countably infinitely many zeros, and* $|\lambda_{\mathcal{H}}(\omega)|$ *tends exponentially quickly to zero as* ω *tends to infinity.*

The graphs of $\lambda_{\mathcal{H}}$ are depicted in Figure 1. As already mentioned, we will prove Theorem 3.1 for the case where ω is the length of stay of the *first* passage of $\mathcal{H}$ near p_2. For a precise definition of ω, we refer the reader to section 4. However, the numerical results for 3-homoclinic orbits of (2.2) demonstrate that there are other possibilities for the choice of ω. We will discuss these issues briefly in section 4.4.

Note that for $N = 1$ the theorem has already been proved in [14]. In this case all symmetric 1-homoclinic orbits (to p_1) lie on the same snaking curve.

For $N > 1$ it turns out that the lengths of all the other passages near p_1 and p_2 remain bounded as ω tends to infinity. This phenomenon can also be observed in the numerical bifurcation diagrams; see, for example, Figure 4. Along the red branch in that figure, the outer pulses (representing the first and last passages near p_2) become and wider as the L^2-norm increases, while both the inner pulse and the time intervals, by which the pulses are separated, remain nearly unchanged.

Remark 3.2. *In addition, for fixed* N *there are infinitely many different snaking curves. For example, two* N*-homoclinic orbits* $\mathcal{H}$ *and* $\mathcal{H}'$ *(which might exist for the same* λ*) are not on the same snaking curve if their numbers of rotations around* p_1 *and* p_2 *are "too different."*

4. The analysis. We will analyze the existence of N-homoclinic orbits to p_1 using Lin's method [17, 12]. In a first step we determine *N-homoclinic Lin orbits* near the cycle Γ. An N-homoclinic Lin orbit to p_1 is a piecewise continuous orbit that starts in the unstable manifold of p_1, follows the cycle Γ, and finishes after N loops in the stable manifold of p_1. Thereby, the discontinuities are allowed to lie only in certain places and have well-defined jump directions. Figure 5 shows an impression of a 2-homoclinic Lin orbit near Γ.

4.1. Setup of the bifurcation equation. In the following we work exclusively in a neighborhood $\mathcal{U}$ of Γ. To define N-homoclinic Lin orbits precisely, let Σ_1 be a hyperplane intersecting Γ_1 transversally at $\gamma_1(0)$, and let $\Sigma_2 := R\Sigma_1$. We are concerned with four different types of partial orbits. First, let $X^- = \{x^-(t) : t \in (-\infty, 0]\}$, such that $x^-(0) \in \Sigma_1 \cap W^u(p_1, \lambda)$ and $x^-(t) \notin \Sigma_k$ for all $t < 0$. Similarly, let $X^+ = \{x^+(t) : t \in [0, \infty)\}$, such that $x^+(0) \in \Sigma_2 \cap W^s(p_1, \lambda)$ and $x^+(t) \notin \Sigma_k$ for all $t > 0$. Finally, for positive numbers ω_2^i, ω_1^j, we consider orbits $X_2^i = \{x_2^i(t) : t \in [0, 2\omega_2^i]\}$ and $X_1^j = \{x_1^j(t) : t \in [0, 2\omega_1^j]\}$, such that

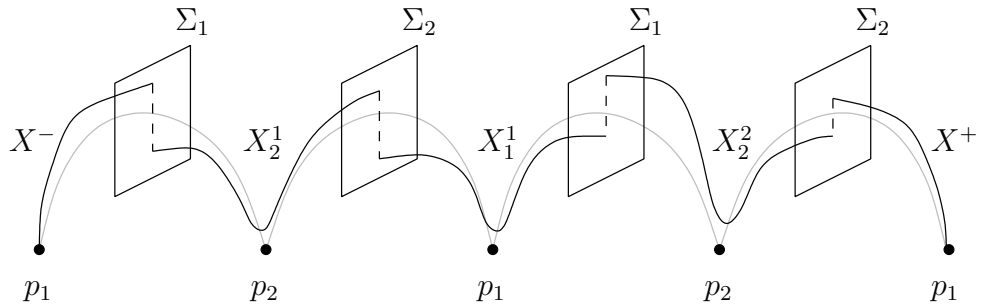


Figure 5. *A 2-homoclinic Lin orbit near the cycle* Γ. *The original heteroclinic orbits are shown in grey. Note that for the purpose of illustration* Γ *is shown as a heteroclinic chain.*

$$x_2^i(0) \in \Sigma_1,\ x_2^i\left(2\omega_2^i\right) \in \Sigma_2,\ \text{and}\ x_2^i(t) \notin \Sigma_2\ \text{for}\ t \in \left(0, 2\omega_2^i\right),$$
$$x_1^j(0) \in \Sigma_2,\ x_1^j\left(2\omega_1^j\right) \in \Sigma_1,\ \text{and}\ x_1^j(t) \notin \Sigma_1\ \text{for}\ t \in \left(0, 2\omega_1^j\right).$$

Now, we introduce a space $Z_1 \subset T_{\gamma_1(0)}\Sigma_1$, complementary to $T_{\gamma_1(0)}W^u(p_1,0)+T_{\gamma_1(0)}W^s(p_2,0)$. Note that because of (3.4) we have $\dim Z_1 = 1$. Furthermore, let $Z_2 = RZ_1 \subset \Sigma_2$. Then a collection of partial orbits

$$\mathcal{L} = \left(X^-, X_2^1, X_1^1, X_2^2, \ldots, X_1^{N-1}, X_2^N, X^+\right)$$

is called an N-homoclinic Lin orbit to p_1 if the jump between two consecutive partial orbits is parallel to Z_1 or Z_2, respectively.

Note that the lower index k of X_k^i indicates that this partial orbit passes p_k while the upper index i counts the number of passages past the equilibrium p_k. The indices in the corresponding quantities x_k^i and ω_k^i have the same meaning.

Lin orbits can be characterized by the times ω_2^i and ω_1^j, and by the parameter λ. More precisely, we have the following result.

Lemma 4.1 (see [17, 12]). *There are positive numbers* $\hat{\lambda}$ *and* $\hat{\omega}$ *such that for each* $|\lambda| < \hat{\lambda}$ *and each set* $\boldsymbol{\omega}_1 = \left\{\omega_1^1, \ldots, \omega_1^{N-1}\right\}$ *and* $\boldsymbol{\omega}_2 = \left\{\omega_2^1, \ldots, \omega_2^N\right\}$ *with*

$$\min\left\{\omega_1^j, \omega_2^i : j = 1, \ldots, N-1,\ i = 1, \ldots, N\right\} > \hat{\omega}$$

there exists a unique N*-homoclinic Lin orbit* $\mathcal{L}\left(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda\right)$ *as introduced above.*

The detection of N-homoclinic orbits near Γ now amounts to finding those Lin orbits without discontinuities (jumps), which are given by

$$\begin{aligned}
\Xi_1^1 &= x^-(0) - x_2^1(0),\\
\Xi_1^i &= x_1^{i-1}\left(2\omega_1^{i-1}\right) - x_2^i(0), \quad i = 2, \ldots, N,\\
\Xi_2^i &= x_2^i\left(2\omega_2^i\right) - x_1^i(0), \quad i = 1, \ldots, N-1,\\
\Xi_2^N &= x_2^N\left(2\omega_2^N\right) - x^+(0).
\end{aligned}$$

The lower index k of Ξ_k^i indicates in which cross-section the jump takes place, and the upper index i counts the jumps in these cross-sections.

Setting $\boldsymbol{\Xi}_1 = (\Xi_1^1, \dots, \Xi_1^N)$ and $\boldsymbol{\Xi}_2 = (\Xi_2^1, \dots, \Xi_2^N)$, we find by Lemma 4.1 that $\boldsymbol{\Xi}_i = \boldsymbol{\Xi}_i(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda)$. In order to detect actual N-homoclinic orbits to p_1 we have to solve the bifurcation equations

$$\boldsymbol{\Xi}_1(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) = 0, \quad \boldsymbol{\Xi}_2(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) = 0.$$

Within the general framework of Lin's method we can derive expressions for the terms $\boldsymbol{\Xi}_k$, $k = 1, 2$. We will do this for $\boldsymbol{\Xi}_1$ only, since it turns out that only this is needed when we focus on symmetric orbits.

For the jumps $\Xi_1^i = \Xi_1^i(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda)$ it has been shown in [17, 12] that

$$\Xi_1^i(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) = \xi^\infty(\lambda) + \xi_1^i(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda),$$

where $\xi^\infty(\lambda)$ measures the splitting of $W^u(p_1, \lambda)$ and $W^s(p_2, \lambda)$ in the direction Z_1. Obviously, $\xi^\infty(0) = 0$, and according to assumption (3.7) we have $D\xi^\infty(0) \neq 0$. Hence, with no loss of generality we may assume that

$$\xi^\infty(\lambda) = \lambda.$$

While ξ^∞ depends only on λ, the terms ξ_1^i measure the influence of the finite transition times between the Σ_i, too. The leading terms in the ξ_1^i depend on the asymptotic behavior near the equilibria p_1 and p_2. Applying general results from [17, 12], we find the following lemma.

Lemma 4.2. *Assuming the nondegeneracy conditions (3.2), (3.5), and (3.6) the jumps ξ_1^i have the following representation:*

$$\begin{aligned}
\xi_1^1(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) &= L_2(\omega_2^1, \lambda) + \mathcal{R}^1, \\
\xi_1^i(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) &= L_1(\omega_1^{i-1}, \lambda) + L_2(\omega_2^i, \lambda) + \mathcal{R}^i, \quad i = 2, \dots, N,
\end{aligned}$$

where

$$\begin{aligned}
&L_k(\omega, \lambda) := c_k(\lambda) e^{-2\mu_k(\lambda)\omega} \sin\left(2\varphi_k(\lambda)\omega + \vartheta_k(\lambda)\right), \quad k = 1, 2, \\
&\mathcal{R}^1 = \mathcal{O}\left(e^{-2\alpha\mu_2(\lambda)\omega_2^1}\right), \quad \mathcal{R}^i = \mathcal{O}\left(e^{-2\alpha\mu_1(\lambda)\omega_1^{i-1}}\right) + \mathcal{O}\left(e^{-2\alpha\mu_2(\lambda)\omega_2^i}\right),
\end{aligned}$$

$i = 2, \dots, N$. Here α is some real number greater than one. The quantities c_1, c_2 and ϑ_1, ϑ_2 depend smoothly on λ, and we have $c_1(0) \neq 0$, $c_2(0) \neq 0$.

Also, from [17, 12] we know that the jumps ξ_k^i are differentiable, and moreover we have the following estimates of the derivatives.

Lemma 4.3. *Under the assumptions of Lemma 4.2, the mappings $\xi_1^i : \mathbb{R}^{N-1} \times \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ are smooth and the partial derivatives $D_j\xi_1^i$, $j \in \{1, 2, 3\}$, can be estimated as follows:*

$$\begin{aligned}
D_j\xi_1^1(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) &= D_j L_2(\omega_2^1, \lambda) + o\left(e^{-2\mu_2(\lambda)\omega_2^1}\right), \\
D_j\xi_1^i(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) &= D_j\left(L_1(\omega_1^{i-1}, \lambda) + L_2(\omega_2^i, \lambda)\right) \\
&\quad + \mathcal{O}\left(e^{-2\alpha\mu_1(\lambda)\omega_1^{i-1}}\right) + \mathcal{O}\left(e^{-2\alpha\mu_2(\lambda)\omega_2^i}\right),
\end{aligned}$$

$i = 2, \dots, N$. Again α is some real number greater than one.

In the following we will focus on symmetric N-homoclinic orbits to p_1. Those orbits correspond to symmetric N-homoclinic Lin orbits to p_1, which are characterized by

$$\begin{aligned} &X^- = RX^+, \quad X_1^i = RX_1^{N-i}, \quad i = 1, \dots, \lfloor N/2 \rfloor, \quad \text{and} \\ &X_2^i = RX_2^{N+1-i}, \quad i = 1, \dots, \lfloor (N+1)/2 \rfloor, \end{aligned}$$

where $\lfloor r \rfloor$ denotes the integer part of the real number r. If N is even, then the partial orbit $X_1^{N/2}$ is symmetric and the N-homoclinic Lin orbit intersects $\operatorname{Fix} R$ near p_1, while for N odd $X_2^{(N+1)/2}$ is symmetric and the N-homoclinic Lin orbit intersects $\operatorname{Fix} R$ near p_2.

In particular, this implies that

$$\omega_1^i = \omega_1^{N-i}, \ i = 1, \dots \lfloor \tfrac{N}{2} \rfloor, \quad \text{and} \quad \omega_2^i = \omega_2^{N+1-i}, \ i = 1, \dots, \lfloor \tfrac{N+1}{2} \rfloor. \tag{4.1}$$

Taking these particulars of the transition times into consideration, we will henceforth write

$$\omega := \omega_2^1, \quad \boldsymbol{\omega} := (\omega_1^1, \dots, \omega_1^{\lfloor \frac{N}{2} \rfloor}, \omega_2^2, \dots, \omega_2^{\lfloor \frac{N+1}{2} \rfloor}).$$

Furthermore, the symmetry of an N-homoclinic Lin orbit implies

$$\Xi_1^i = R\,\Xi_2^{N+1-i}, \quad i = 1, \dots, N. \tag{4.2}$$

Hence, $\boldsymbol{\Xi}_1 = 0$ if and only if $\boldsymbol{\Xi}_2 = 0$, and the bifurcation equation for symmetric N-homoclinic orbits to p_1 reads

$$\boldsymbol{\Xi}(\omega, \boldsymbol{\omega}, \lambda) := \boldsymbol{\Xi}_1(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \lambda) = 0.$$

In what follows we just write Ξ^i instead of Ξ_1^i. With this notation we have

$$\begin{aligned} \Xi^1 &= \lambda + L_2(\omega_2^1, \lambda) + \mathcal{R}^1, \\ \Xi^i &= \lambda + L_1(\omega_1^{i-1}, \lambda) + L_2(\omega_2^i, \lambda) + \mathcal{R}^i, \quad i = 2, \dots, \lfloor \tfrac{N+1}{2} \rfloor, \\ \Xi^i &= \lambda + L_1(\omega_1^{N+1-i}, \lambda) + L_2(\omega_2^{N+1-i}, \lambda) + \mathcal{R}^i, \quad i = \lfloor \tfrac{N+1}{2} \rfloor + 1, \dots, N. \end{aligned} \tag{4.3}$$

Note that all Ξ^i and $\mathcal{R}^i$ depend on $(\omega, \boldsymbol{\omega}, \lambda)$.

4.2. Reformulation of the bifurcation equation.

We define

$$r_k^i := e^{-2\mu_k(0)\omega_k^i}, \quad r := r_2^1, \quad \mathbf{r} := (r_1^1, \dots, r_1^{\lfloor \frac{N}{2} \rfloor}, r_2^2, \dots, r_2^{\lfloor \frac{N+1}{2} \rfloor}). \tag{4.4}$$

We want to emphasize that by definition all r as well as all components of $\mathbf{r}$ are greater than zero.

With that we write the jumps as quantities depending on $(r, \mathbf{r}, \lambda)$:

$$\hat{\boldsymbol{\Xi}}(r, \mathbf{r}, \lambda) = (\hat{\Xi}^1(r, \mathbf{r}, \lambda), \dots, \hat{\Xi}^N(r, \mathbf{r}, \lambda)) := \boldsymbol{\Xi}(\omega(r), \boldsymbol{\omega}(\mathbf{r}), \lambda).$$

Corollary 4.4. *The $(r,\mathbf{r})$-dependent jumps read as follows:*

$$\begin{aligned}
\hat{\Xi}^1 &= \lambda + \hat{\mathcal{R}}^1,\\
\hat{\Xi}^i &= \lambda + \hat{L}_1(r_1^{i-1},\lambda) + \hat{L}_2(r_2^i,\lambda) + \hat{\mathcal{R}}^i, \quad i = 2,\dots,\lfloor\tfrac{N+1}{2}\rfloor,\\
\hat{\Xi}^i &= \lambda + \hat{L}_1(r_1^{N+1-i},\lambda) + \hat{L}_2(r_2^{N+1-i},\lambda) + \hat{\mathcal{R}}^i, \quad i = \lfloor\tfrac{N+1}{2}\rfloor + 1,\dots,N-1,\\
\hat{\Xi}^N &= \lambda + \hat{L}_1(r_1^1,\lambda) + \hat{\mathcal{R}}^N,
\end{aligned}$$

where $\hat{\mathcal{R}}^i = \hat{\mathcal{R}}^i(r,\mathbf{r},\lambda)$ and

$$\hat{L}_k(s,\lambda) := c_k(\lambda) s^{\frac{\mu_k(\lambda)}{\mu_k(0)}} \sin\left(-\frac{\varphi_k(\lambda)}{\mu_k(0)}\ln s + \vartheta_k(\lambda)\right), \quad k = 1,2,$$

and, with some $\alpha > 1$,

$$\begin{aligned}
\hat{\mathcal{R}}^1 &= \hat{L}_2(r,\lambda) + \mathcal{O}\big(r^\alpha\big),\\
\hat{\mathcal{R}}^i &= \mathcal{O}\left((r_1^{i-1})^\alpha\right) + \mathcal{O}\big((r_2^i)^\alpha\big), \quad i = 2,\dots,N-1,\\
\hat{\mathcal{R}}^N &= \hat{L}_2(r,\lambda) + \mathcal{O}\left((r_1^{N-1})^\alpha\right) + \mathcal{O}\big((r_2^N)^\alpha\big).
\end{aligned}$$

And similarly, we get estimates for the derivatives of the residual terms $\hat{\mathcal{R}}^i = \hat{\mathcal{R}}^i(r,\mathbf{r},\lambda)$ from Lemma 4.3.

Corollary 4.5. *The statement of Lemma 4.3 for the $(r,\mathbf{r})$-dependent jumps reads as follows:*

$$\begin{aligned}
&D_1\hat{\mathcal{R}}^1 = D_1\hat{L}_2(r,\lambda) + \mathcal{O}\big(r^{\alpha-1}\big), \quad D_2\hat{\mathcal{R}}^1 = \mathcal{O}\big(r^\alpha\big),\\
&D_3\hat{\mathcal{R}}^1 = D_3\hat{L}_2(r,\lambda) + \mathcal{O}\big(r^\alpha\big),\\
&D_2\hat{\mathcal{R}}^i = \mathcal{O}\left((r_1^{i-1})^{\alpha-1}\right) + \mathcal{O}\big((r_2^i)^{\alpha-1}\big), \quad i = 2,\dots,N,\\
&D_3\hat{\mathcal{R}}^i = \mathcal{O}\left((r_1^{i-1})^\alpha\right) + \mathcal{O}\big((r_2^i)^\alpha\big), \quad i = 2,\dots,N-1,\\
&D_3\hat{\mathcal{R}}^N = D_3\hat{L}_2(r,\lambda) + \mathcal{O}\left((r_1^{N-1})^\alpha\right) + \mathcal{O}\big((r_2^N)^\alpha\big).
\end{aligned}$$

Our goal is to rewrite the bifurcation equation $\hat{\boldsymbol{\Xi}} = 0$ as a fixed point equation. For this we introduce

$$\begin{aligned}
\hat{\mathbf{L}}(\mathbf{r},\lambda) &:= \big(\hat{L}_1(r_1^1,\lambda),\dots,\hat{L}_1(r_1^{\lfloor\frac{N}{2}\rfloor},\lambda),\hat{L}_2(r_2^2,\lambda),\dots,\hat{L}_2(r_2^{\lfloor\frac{N+1}{2}\rfloor},\lambda)\big)^T,\\
\hat{\boldsymbol{\mathcal{R}}} &:= \big(\hat{\mathcal{R}}^1,\dots,\hat{\mathcal{R}}^N\big)^T.
\end{aligned}$$

There is an invertible constant $(N\times N)$-matrix $\mathcal{M}$ such that

$$\hat{\boldsymbol{\Xi}}(r,\mathbf{r},\lambda) = \mathcal{M}\begin{pmatrix}\hat{\mathbf{L}}(\mathbf{r},\lambda)\\ \lambda\end{pmatrix} + \hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda).$$

We comment on properties of $\mathcal{M}$ in the remark at the end of this section.

That $\hat{\boldsymbol{\Xi}}(r,\mathbf{r},\lambda) = 0$ is equivalent to

$$\begin{pmatrix}\hat{\mathbf{L}}(\mathbf{r},\lambda)\\ \lambda\end{pmatrix} = -\mathcal{M}^{-1}\hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda). \tag{4.5}$$

In the next step we rewrite (4.5) into a fixed point equation. For that we choose $\hat{\mathbf{r}}$ such that $\hat{\mathbf{L}}(\hat{\mathbf{r}},0)=0$. Note that there are infinitely many candidates for such $\hat{\mathbf{r}}$ which accumulate at zero. However, for any such $\hat{\mathbf{r}}$ the partial derivative with respect to the first variable $D_1\hat{\mathbf{L}}(\hat{\mathbf{r}},0)$ is an invertible diagonal matrix $\mathcal{D}$, and the absolute values of the entries in the diagonal are either $|\frac{c_1(0)\varphi_1(0)}{\mu_1(0)}|$ or $|\frac{c_2(0)\varphi_2(0)}{\mu_2(0)}|$. We want to emphasize that these quantities do not depend on the particular choice of $\hat{\mathbf{r}}$ (as long as $\hat{\mathbf{L}}(\hat{\mathbf{r}},0)=0$).

With that, the Taylor expansion of $\hat{\mathbf{L}}$ at $(\mathbf{r},\lambda)=(\hat{\mathbf{r}},0)$ with second order residual term $\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}$ reads

$$\hat{\mathbf{L}}(\mathbf{r},\lambda)=\mathcal{D}(\mathbf{r}-\hat{\mathbf{r}})+D_2\hat{\mathbf{L}}(\hat{\mathbf{r}},0)\lambda+\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda). \tag{4.6}$$

Combining (4.5) and (4.6), we find the following fixed point equation for $\mathbf{r}$ which is equivalent to $\hat{\boldsymbol{\Xi}}(r,\mathbf{r},\lambda)=0$:

$$\begin{aligned}\begin{pmatrix}\mathbf{r}\\ \lambda\end{pmatrix}&=\begin{pmatrix}\hat{\mathbf{r}}-\mathcal{D}^{-1}(D_2\hat{\mathbf{L}}(\hat{\mathbf{r}},0)\lambda+\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda))\\ 0\end{pmatrix}-(\mathcal{M}\hat{\mathcal{D}})^{-1}\hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda)\\ &=:\mathcal{T}_{\hat{\mathbf{r}}}(r,\mathbf{r},\lambda),\end{aligned} \tag{4.7}$$

where $\hat{\mathcal{D}}=\left(\begin{smallmatrix}\mathcal{D}&0\\0&1\end{smallmatrix}\right)$. The right-hand side $\mathcal{T}_{\hat{\mathbf{r}}}$ of this equation can be read as a mapping

$$\mathcal{T}_{\hat{\mathbf{r}}}:\mathbb{R}_+\times\mathbb{R}_+^{N-1}\times\mathbb{R}\to\mathbb{R}_+^{N-1}\times\mathbb{R}.$$

The lower index "+" denotes the restriction to positive numbers.

Remark 4.6. *Let $\mathcal{M}=(m_{i,j})$ be the above-defined $(N\times N)$-matrix. The entries $m_{i,j}$ are either one or zero. From the representation of $\hat{\boldsymbol{\Xi}}$ given in Corollary 4.4 we find that exactly the following entries are equal to one:*

$$\begin{aligned}&m_{1,N},\\ &m_{i,i-1},\quad m_{i,\lfloor\frac{N}{2}\rfloor+i-1},\quad m_{i,N},\quad i=2,\dots,\lfloor\tfrac{N+1}{2}\rfloor,\\ &m_{i,N+1-i},\quad m_{i,\lfloor\frac{N}{2}\rfloor+N-i},\quad m_{i,N},\quad i=\lfloor\tfrac{N+1}{2}\rfloor+1,\dots,N-1,\\ &m_{N,1},\quad m_{N,N}.\end{aligned}$$

To show that $\mathcal{M}$ is indeed nonsingular, we simply compute its determinant by using Laplace's formula. We expand alternately along the first or the last line of the corresponding minor arising in the course of this procedure. This finally yields that the absolute value of the determinant of $\mathcal{M}$ is equal to one; in more detail,

$$|\det\mathcal{M}|=\begin{cases}m_{1,N}\cdot m_{N,1}\cdot m_{2,\lfloor\frac{N}{2}\rfloor+1}\cdot m_{N-1,2}\cdot\dots\cdot m_{\lfloor\frac{N+1}{2}\rfloor,\lfloor\frac{N+1}{2}\rfloor-1}, & N\text{ even},\\ m_{1,N}\cdot m_{N,1}\cdot m_{2,\lfloor\frac{N}{2}\rfloor+1}\cdot m_{N-1,2}\cdot\dots\cdot m_{\lfloor\frac{N+1}{2}\rfloor,N-1}, & N\text{ odd}.\end{cases}$$

4.3. Proof of Theorem 3.1. Our goal is to solve the reformulated bifurcation equation (4.7) for $(\mathbf{r}_{\hat{\mathbf{r}}},\lambda_{\hat{\mathbf{r}}})(r)$ near $(\mathbf{r},\lambda)=(\hat{\mathbf{r}},0)$, $r\in(0,\epsilon)$. We will do this by applying the Banach fixed point theorem. Our strategy is as follows. First we construct for the "principal part"

$$\begin{pmatrix}\mathbf{r}\\ \lambda\end{pmatrix}=\begin{pmatrix}\hat{\mathbf{r}}-\mathcal{D}^{-1}(D_2\hat{\mathbf{L}}(\hat{\mathbf{r}},0)\lambda+\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda))\\ 0\end{pmatrix}=:\mathcal{T}_{red,\hat{\mathbf{r}}}(r,\mathbf{r},\lambda)$$

of that equation a domain $\mathcal{C}_{\hat{\mathbf{r}}} \times \mathcal{B}_{\hat{\mathbf{r}}}$, which will be mapped contractively into itself by $\mathcal{T}_{red,\hat{\mathbf{r}}}$. Then we make clear that for small $\hat{\mathbf{r}}$ this can be carried forward to $\mathcal{T}_{\hat{\mathbf{r}}}$.

In the next step we consider $\hat{L}_k$, which represents the components of $\hat{\mathbf{L}}$. To simplify matters we omit the lower index k in further considerations:

$$\hat{L}(s,\lambda) := c(\lambda) s^{\frac{\mu(\lambda)}{\mu(0)}} \sin\left(-\frac{\varphi(\lambda)}{\mu(0)} \ln s + \vartheta(\lambda)\right).$$

The zeros s_n of $\hat{L}(\cdot,0)$ are explicitly given by $s_n = e^{-\frac{\mu(0)}{\varphi(0)}(n\pi - \vartheta(0))}$, $n \in \mathbb{Z}$. For our purpose we are interested only in those zeros which are close to $s = 0$, so we may assume $n \in \mathbb{N}$ sufficiently large.

Let $\delta := D_1\hat{L}(s_n, 0)$. Note that δ represents an element in the diagonal of $\mathcal{D}$ and that $|\delta|$ does not depend on n.

By $\hat{L}_{res,n}$ we denote the second-order residual term of the Taylor expansion of $\hat{L}$ at $(s_n, 0)$. Let $B[s,\rho]$ be the closed ball centered at s with radius ρ.

Lemma 4.7. *There are $\overline{\beta}, \overline{K} > 0$ such that for all sufficiently large $n \in \mathbb{N}$ the following estimates hold true: For all $s \in B[s_n, \rho_n]$, $\rho_n := s_n\overline{\beta}$, and all $\lambda \in B[0, l_{s_n}]$, $l_{s_n} = \overline{K} s_n$,*

$$|D_1\hat{L}_{res,n}(s_n\beta,\lambda)| < \frac{\delta}{3}, \quad |D_2\hat{L}_{res,n}(s_n\beta,\lambda)| < \frac{\delta}{3}, \tag{4.8}$$

and

$$\left|\delta^{-1}\left(D_2\hat{L}(s_n,0)\lambda + \hat{L}_{res,n}(s,\lambda)\right)\right| < \frac{2}{3}\rho_n. \tag{4.9}$$

Proof. First we prove the estimates regarding the derivatives of $L_{res,n}$. With $A(s,\lambda) := -\frac{\varphi(\lambda)}{\mu(0)} \ln s + \vartheta(\lambda)$ we get

$$\hat{L}_{res,n}(s,\lambda) = c(\lambda) s^{\frac{\mu(\lambda)}{\mu(0)}} \sin(A(s,\lambda)) - \delta(s - s_n) - D_2\hat{L}(s_n,0)\lambda,$$

and therefore

$$D_1\hat{L}_{res,n}(s,\lambda) = c(\lambda) s^{\frac{\mu(\lambda)}{\mu(0)}-1}\left[\tfrac{\mu(\lambda)}{\mu(0)} \sin(A(s,\lambda)) - \tfrac{\varphi(\lambda)}{\mu(0)} \cos(A(s,\lambda))\right] - \delta.$$

Obviously, $D_1\hat{L}_{res,n}(s_n,0) = 0$. Writing $s = s_n\beta$ one finds that there exist $\overline{\beta}$ and $\tilde{l}_{s_n}$ such that for all $\beta \in [1-\overline{\beta}, 1+\overline{\beta}]$ and all $\lambda \in B[0,\tilde{l}_{s_n}]$ the first estimate in (4.8) holds true. Note that for those β the corresponding $s = s_n\beta$ belong to $B[s_n, \rho_n]$.

For the derivative of $\hat{L}_{res,n}$ with respect to λ we find

$$\begin{aligned} D_2\hat{L}_{res,n}(s,\lambda) = {} & \left[c'(\lambda) s^{\frac{\mu(\lambda)}{\mu(0)}} + c(\lambda) s^{\frac{\mu(\lambda)}{\mu(0)}-1} \tfrac{\mu'(\lambda)}{\mu(0)}\right] \sin(A(s,\lambda)) \\ & + c(\lambda) s^{\frac{\mu(\lambda)}{\mu(0)}} \cos(A(s,\lambda)) D_2A(s,\lambda) - D_2\hat{L}(s_n,0). \end{aligned}$$

Because $\lim_{s\to+0} s^\alpha \ln s = 0$ for $\alpha > 0$, we have that $c(\lambda) s^{\frac{\mu(\lambda)}{\mu(0)}} \cos(A(s,\lambda)) D_2A(s,\lambda)$ tends to zero as $s \to 0$, uniformly in λ, and also that $\lim_{n\to\infty} D_2\hat{L}(s_n,0) = 0$. So, again we find that

there are a $\overline{\beta}$ and an $\tilde{l}_{s_n}$ such that for all $\beta \in [1-\overline{\beta}, 1+\overline{\beta}]$ and all $\lambda \in B[0, \tilde{l}_{s_n}]$ the estimate for $D_2\hat{L}_{res,n}$ holds true.

Note that indeed in both estimates $\tilde{l}_{s_n}$ may depend on n. The decisive term in this respect is $s^{\frac{\mu(\lambda)}{\mu(0)}-1}$. Using differentiability of μ at $\lambda = 0$, this term can be written as $s_n^{\mathcal{O}(\lambda)}\beta^{\mathcal{O}(\lambda)}$. If $|\lambda| < \overline{K}s_n$ for some $\overline{K}$, then $s_n^{\mathcal{O}(\lambda)}$ tends to 1 as n tends to infinity. This finally proves the estimates given in (4.8).

Our above consideration shows that $D_2\hat{L}(s_n, 0)$ remains bounded (as n tends to infinity) and that $D_2\hat{L}_{res,n}(s,\lambda)$ remains bounded if s and λ are close to zero. Applying the mean value theorem yields

$$\begin{aligned}|\hat{L}_{res,n}(s,\lambda)| &\le \sup_{\substack{s\in B[s_n,\rho_n]\\ \lambda\in B[0,l_{s_n}]}} |D_1\hat{L}_{res,n}(s,\lambda)|\,|s-s_n| + \sup_{\substack{s\in B[s_n,\rho_n]\\ \lambda\in B[0,l_{s_n}]}} |D_2\hat{L}_{res,n}(s,\lambda)|\,|\lambda| \\ &\le \tfrac{\delta}{3}(|s-s_n| + |\lambda|).\end{aligned}$$

Hence,

$$\left|\delta^{-1}\big(D_2\hat{L}(s_n,0)\lambda + \hat{L}_{res,n}(s,\lambda)\big)\right| \le |\delta|^{-1}\left(|D_2\hat{L}(s_n,0)|\,|\lambda| + \frac{\delta}{3}(|s-s_n|+|\lambda|)\right).$$

Recall that $|D_2\hat{L}(s_n,0)|$ tends to zero as $n \to \infty$, and therefore there is a K such that $|D_2\hat{L}(s_n,0)| < K$ for all sufficiently large n. Then with the particular

$$\overline{K} = \frac{\delta\overline{\beta}}{3}\left(K + \frac{\delta}{3}\right)^{-1},$$

(4.9) is also proved. ■

Now we return to the original fixed point equation (4.7). First we introduce some simplifying notation.

Let $\mathbf{r} := (r_1^1, \dots, r_1^{\lfloor\frac{N}{2}\rfloor}, r_2^2, \dots, r_2^{\lfloor\frac{N+1}{2}\rfloor}) \in \mathbb{R}_+^{N-1}$. We define projections π_k^i as follows:

$$\pi_k^i : \mathbb{R}_+^{N-1} \to \mathbb{R}_+, \quad \pi_k^i\big((r_1^1, \dots, r_1^{\lfloor\frac{N}{2}\rfloor}, r_2^2, \dots, r_2^{\lfloor\frac{N+1}{2}\rfloor})\big) = r_k^i.$$

Now let $\hat{\mathbf{L}}(\hat{\mathbf{r}}, 0) = 0$. By Lemma 4.7, each $\hat{r}_k^i$ has a $\rho_k^i = \hat{r}_k^i\overline{\beta}_k$ and a corresponding $B_k^i := B[\hat{r}_k^i, \rho_k^i]$ assigned to it. With these we define the cylinder

$$\mathfrak{C}_{\hat{\mathbf{r}}} := B_1^1 \times \cdots \times B_1^{\lfloor\frac{N}{2}\rfloor} \times B_2^2 \times \cdots \times B_2^{\lfloor\frac{N+1}{2}\rfloor}.$$

Further, we define

$$\overline{\beta} := \max\{\overline{\beta}_1, \overline{\beta}_2\}, \quad \underline{\beta} := \min\{\overline{\beta}_1, \overline{\beta}_2\}$$

and similarly

$$\overline{\rho}(\hat{\mathbf{r}}) := \max\{\rho_k^i\}, \quad \underline{\rho}(\hat{\mathbf{r}}) := \min\{\rho_k^i\}.$$

For $\mathbf{r} = (r^1, \dots, r^{N-1}) \in \mathbb{R}_+^{N-1} \subset \mathbb{R}^{N-1}$ we introduce

$$\overline{\mathbf{r}} := \max\{r^i,\, i = 1, \dots, N-1\} = \|\mathbf{r}\|, \quad \underline{\mathbf{r}} := \min\{r^i,\, i = 1, \dots, N-1\},$$

and with that we define

$$\kappa(\mathbf{r}) := \underline{\mathrm{r}}/\overline{\mathrm{r}} = \underline{\mathrm{r}}/\|\mathbf{r}\|.$$

From these definitions it follows that

$$\hat{\underline{\mathrm{r}}}\,\underline{\beta} \le \underline{\rho}(\hat{\mathbf{r}}) \le \overline{\rho}(\hat{\mathbf{r}}) \le \|\hat{\mathbf{r}}\|\overline{\beta}. \tag{4.10}$$

Therefore, $\sup\{\|\mathbf{r}\| : \mathbf{r} \in \mathfrak{C}_{\hat{\mathbf{r}}}\} \le \|\hat{\mathbf{r}}\|(1+\overline{\beta})$.

In accordance with Lemma 4.7 there is a constant $\hat{K}$ such that $l_{\hat{\underline{\mathrm{r}}}} = \hat{K}\underline{\rho}(\hat{\mathbf{r}})$. Finally, we define $\mathfrak{B}_{\hat{\mathbf{r}}} := B[0, l_{\hat{\underline{\mathrm{r}}}}]$.

Lemma 4.8. *Let $\kappa^* \in (0,1]$. There is an $\eta = \eta(\kappa^*)$ such that for all $\hat{\mathbf{r}}$ with $\kappa(\hat{\mathbf{r}}) \ge \kappa^*$, $\|\hat{\mathbf{r}}\| < \eta$, there is an ϵ_r such that for all $r \in (0, \epsilon_r)$ the fixed point equation (4.7) has a unique fixed point $(\mathbf{r}_{\hat{\mathbf{r}}}, \lambda_{\hat{\mathbf{r}}})(r) \in \mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$. Moreover, $\mathbf{r}_{\hat{\mathbf{r}}}$ and $\lambda_{\hat{\mathbf{r}}}$ depend smoothly on r.*

Proof. First we show that there is an appropriate $\tilde{\eta}$ such that $\mathcal{T}_{\hat{\mathbf{r}}}(r,\cdot,\cdot)$ is a contraction on $\mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$. We introduce for $(\mathbf{r},\lambda) \in \mathbb{R}^{N-1} \times \mathbb{R}$ the norm $\|(\mathbf{r},\lambda)\| := \|\mathbf{r}\| + |\lambda|$. Then, from the definition of $\mathcal{T}_{\hat{\mathbf{r}}}$ (see (4.7)) we find

$$\begin{aligned}\|D_{(2,3)}\mathcal{T}_{\hat{\mathbf{r}}}(r,\mathbf{r},\lambda)\| \le{}& \|\mathcal{D}^{-1}D_2\hat{\mathbf{L}}(\hat{\mathbf{r}},0)\| \\ &+ \max\{\|\mathcal{D}^{-1}\,D_1\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda)\|,\ \|\mathcal{D}^{-1}\,D_2\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda)\|\} \\ &+ \|(\mathcal{M}\hat{\mathcal{D}})^{-1}\|\,\|D_{(2,3)}\hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda)\|.\end{aligned}$$

Since $\lim_{n\to\infty} D_2\hat{L}(s_n,0) = 0$, there is an η_1 such that $\|\hat{\mathbf{r}}\| < \eta_1/1+\overline{\beta}$ implies

$$\|\mathcal{D}^{-1}D_2\hat{\mathbf{L}}(\hat{\mathbf{r}},0)\| < \frac{1}{3}.$$

Moreover, due to Lemma 4.7, the constant η_1 can be chosen such that for $\|\hat{\mathbf{r}}\| < \eta_1/1+\overline{\beta}$ and $(\mathbf{r},\lambda) \in \mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$ also

$$\|\mathcal{D}^{-1}\,D_1\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda)\| < \frac{1}{3}, \quad \|\mathcal{D}^{-1}\,D_2\hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda)\| < \frac{1}{3}.$$

According to Corollary 4.5, there are an $\tilde{\epsilon}_r$ and an $\eta_2 < \eta_1$ such that for all $r < \tilde{\epsilon}_r$, all $\mathbf{r}$ with $\|\mathbf{r}\| < \eta_2$, and all $\lambda \in \mathcal{B}_{\hat{\mathbf{r}}}$, we have

$$\|(\mathcal{M}\mathcal{D})^{-1}\|\,\|D_{(2,3)}\hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda)\| < \frac{1}{6}. \tag{4.11}$$

Therefore, if $\|\hat{\mathbf{r}}\|(1+\overline{\beta}) < \eta_2$, then (4.11) holds true for all $\mathbf{r} \in \mathfrak{C}_{\hat{\mathbf{r}}}$.

Thus, we have shown that if $\|\hat{\mathbf{r}}\| < \eta_2/1+\overline{\beta}$, then $\|D_{(2,3)}\mathcal{T}_{\hat{\mathbf{r}}}(r,\mathbf{r},\lambda)\| < 1$ for all $(\mathbf{r},\lambda) \in \mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$, $r < \tilde{\epsilon}_r$. In other words, the mapping $\mathcal{T}_{\hat{\mathbf{r}}}(r,\cdot,\cdot)$ is a contraction on $\mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$.

Next we verify that $\mathcal{T}_{\hat{\mathbf{r}}}(r,\cdot,\cdot)$ maps $\mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$ into itself. Lemma 4.7 provides that for all $(\mathbf{r},\lambda) \in \mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$

$$\left|\pi_k^i\big(\hat{\mathbf{r}} - \mathcal{D}^{-1}(D_2\hat{\mathbf{L}}(\hat{\mathbf{r}},0)\lambda + \hat{\mathbf{L}}_{res,\hat{\mathbf{r}}}(\mathbf{r},\lambda))\big)\right| < \tfrac{2}{3}\rho_k^i. \tag{4.12}$$

It remains (see (4.7)) to consider the term $(\mathcal{M}\hat{\mathcal{D}})^{-1}\hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda)$. Let $(\mathbf{r},\lambda) \in \mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$. In accordance with Corollary 4.4, there are constants C and $\hat{C}$ such that

$$\|\hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda)\| \leq Cr + \hat{C}\|\hat{\mathbf{r}}\|^{\alpha}.$$

For any given $\hat{\mathbf{r}}$ we can choose $\epsilon_r = \epsilon_r(\hat{\mathbf{r}}) < \tilde{\epsilon}_r$ small such that

$$\|(\mathcal{M}\hat{\mathcal{D}})^{-1}\|\, Cr < \min\{(1/6)\underline{\rho}(\hat{\mathbf{r}}), (1/2)l_{\underline{\hat{\mathbf{r}}}}\}.$$

Further, due to (4.10) we find that $\underline{\rho}(\hat{\mathbf{r}}) \geq \underline{\beta}\kappa^*\|\hat{\mathbf{r}}\|$, and because of $l_{\underline{\hat{\mathbf{r}}}} = \hat{K}\underline{\rho}(\hat{\mathbf{r}})$ we have $l_{\underline{\hat{\mathbf{r}}}} \geq \hat{K}\underline{\beta}\kappa^*\|\hat{\mathbf{r}}\|$. Hence there is an $\eta_3 \leq \eta_2$ such that for $\hat{\mathbf{r}}$, $\|\hat{\mathbf{r}}\|(1+\overline{\beta}) < \eta_3$,

$$\|(\mathcal{M}\hat{\mathcal{D}})^{-1}\|\, \hat{C}\|\hat{\mathbf{r}}\|^{\alpha} < \min\{(1/6)\underline{\rho}(\hat{\mathbf{r}}), (1/2)l_{\underline{\hat{\mathbf{r}}}}\}.$$

This finally shows for those $\hat{\mathbf{r}}$ and $r < \epsilon_r$ that

$$\|(\mathcal{M}\hat{\mathcal{D}})^{-1}\|\, \|\hat{\boldsymbol{\mathcal{R}}}(r,\mathbf{r},\lambda)\| \leq \min\{(1/3)\underline{\rho}(\hat{\mathbf{r}}), l_{\underline{\hat{\mathbf{r}}}}\}.$$

Together with (4.12) this implies that $\mathcal{T}_{\hat{\mathbf{r}}}(r,\cdot,\cdot)$ maps $\mathfrak{C}_{\hat{\mathbf{r}}} \times \mathfrak{B}_{\hat{\mathbf{r}}}$ into itself.

Now we can apply the Banach fixed point theorem to prove the existence of $(\mathbf{r}_{\hat{\mathbf{r}}}, \lambda_{\hat{\mathbf{r}}})(r)$.

Finally, let $(\mathbf{r},\lambda) = \mathcal{T}_{\hat{\mathbf{r}}}(r,\mathbf{r},\lambda)$. Applying the implicit function theorem provides the smooth dependence of $(\mathbf{r}_{\hat{\mathbf{r}}}, \lambda_{\hat{\mathbf{r}}})$ on r. ∎

Remark 4.9. *The mapping $\hat{\boldsymbol{\Xi}}(\cdot,\mathbf{r},\lambda)$ and hence also $\mathcal{T}_{\hat{\mathbf{r}}}(\cdot,\mathbf{r},\lambda)$ can be continuously extended for $r = 0$. Lemma 4.8 remains true in this case. As a consequence of the uniform contraction principle (see [5]), we find that $(\mathbf{r},\lambda)(\cdot)$ is continuous in $r = 0$.*

To complete the proof of Theorem 3.1, we consider the function $\lambda_{\hat{\mathbf{r}}}$. Our above considerations show that, in particular,

$$\hat{\Xi}^1(r, \mathbf{r}_{\hat{\mathbf{r}}}(r), \lambda_{\hat{\mathbf{r}}}(r)) = \lambda_{\hat{\mathbf{r}}}(r) + \hat{L}_2(r, \lambda_{\hat{\mathbf{r}}}(r)) + \mathcal{O}(r^{\alpha}) \equiv 0.$$

Due to the structure of $\hat{L}_2$ the function $\lambda_{\hat{\mathbf{r}}}$ has infinitely many zeros. Moreover, it follows that $\lim_{r\to 0}\lambda_{\hat{\mathbf{r}}}(r) = 0$. Let r_0 be such that $\lambda_{\hat{\mathbf{r}}}(r_0) = 0$. Then $(r_0, \mathbf{r}_{\hat{\mathbf{r}}}(r_0))$ corresponds to an N-homoclinic orbit $\mathcal{H}_{\hat{\mathbf{r}}}(r_0)$ (to p_1) of the vector field $f(\cdot,0)$.

Let $\lambda_{\mathcal{H}_{\hat{\mathbf{r}}}(r_0)}(\omega) := \lambda_{\hat{\mathbf{r}}}(r(\omega))$, defined on some interval (a,∞). Then $\lambda_{\mathcal{H}_{\hat{\mathbf{r}}}(r_0)}$ has infinitely many zeros, and $|\lambda_{\mathcal{H}_{\hat{\mathbf{r}}}(r_0)}(\omega)|$ tends exponentially quickly to zero as ω tends to infinity.

To conclude, we return to the remark given at the end of section 3. The analysis in the proof of Lemma 4.8 was performed for fixed $\hat{\mathbf{r}}$. However, there is a sequence $(\hat{\mathbf{r}}_i)_{i\in\mathbb{N}}$ such that all $\hat{\mathbf{r}}_i$ are in accordance with the assumptions of Lemma 4.8, $\hat{\mathbf{r}}_i \to 0$, and $\mathfrak{C}_{\hat{\mathbf{r}}_i} \cap \mathfrak{C}_{\hat{\mathbf{r}}_j} = \emptyset$, $i \neq j$. With this, the sets $\mathcal{H}_{\hat{\mathbf{r}}_i}(r)$ and $\mathcal{H}_{\hat{\mathbf{r}}_j}(r)$ are disjoint for $i \neq j$, meaning that $\mathcal{H}_{\hat{\mathbf{r}}_i}(r)$ and $\mathcal{H}_{\hat{\mathbf{r}}_j}(r)$ are not on the same snaking curve.

4.4. Further homoclinic orbits. So far we have proved the existence of infinitely many N-homoclinic solutions to the equilibrium p_1 which lie on snaking curves and whose "outer" pulses become wider along the snaking curves. In what follows we denote the corresponding homoclinic orbits by $\mathcal{H}^{out}$. The numerical experiments for the generalized Swift–Hohenberg equation, however, show that there are also snaking curves different from the ones we described

so far analytically. In Figure 4 there are 3-homoclinic orbits depicted where the snaking is due to the middle part; in other words, the snaking curve is parametrized by ω_2^2. We denote these homoclinic orbits by $\mathcal{H}^{mid}$. In what follows we shall attempt to explain these observations in light of Theorem 3.1.

In the sketch of our further analysis, we restrict ourselves to the consideration of symmetric 3-homoclinic orbits and show that two different types of behavior along snaking curves can occur. However, the arguments are easily generalized to arbitrary N-homoclinic orbits, for which we expect to find $\lfloor (N+1)/2 \rfloor$ different types of behavior along snaking curves.

To simplify notation we write

$$(r_2^1, r_1^1, r_2^2) =: (r_1, r_2, r_3).$$

In accordance with Corollary 4.4 the bifurcation equation for 3-homoclinic orbits reads

$$\begin{aligned} 0 &= \hat{\Xi}^1 = \lambda + \hat{L}_2(r_1, \lambda) + \hat{\mathcal{R}}^1, \\ 0 &= \hat{\Xi}^2 = \lambda + \hat{L}_1(r_2, \lambda) + \hat{L}_2(r_3, \lambda) + \hat{\mathcal{R}}^2, \\ 0 &= \hat{\Xi}^3 = \lambda + \hat{L}_1(r_2, \lambda) + \hat{L}_2(r_1, \lambda) + \hat{\mathcal{R}}^3, \end{aligned} \tag{4.13}$$

where $\hat{\mathcal{R}}^1 = \mathcal{O}(r_1^\alpha)$, $\hat{\mathcal{R}}^2 = \mathcal{O}(r_2^\alpha) + \mathcal{O}(r_3^\alpha)$, and $\hat{\mathcal{R}}^3 = \mathcal{O}(r_2^\alpha) + \mathcal{O}(r_1^\alpha)$ for some $\alpha > 1$. In section 4.3 we have solved this equation for $(r_2, r_3, \lambda) = (r_2, r_3, \lambda)^{out}(r_1)$. The corresponding snaking curve is parametrized by $\omega_2^1 =: \omega_1$ (resp., r_1), which we call the *snaking parameter*. In the limit $r_1 \to 0$ (4.13) becomes

$$\begin{aligned} 0 &= \hat{\Xi}^1 = \lambda, \\ 0 &= \hat{\Xi}^2 = \lambda + \hat{L}_1(r_2, \lambda) + \hat{L}_2(r_3, \lambda) + \hat{\mathcal{R}}^2_{r_1=0}, \\ 0 &= \hat{\Xi}^3 = \lambda + \hat{L}_1(r_2, \lambda) + \hat{\mathcal{R}}^3_{r_1=0}. \end{aligned} \tag{4.14}$$

The lower index $r_1 = 0$ should indicate that these $\hat{\mathcal{R}}^{2/3}_{r_1=0}$ do not depend on r_1. We will call the orbits corresponding to solutions of (4.14) the *snaking limit*.

Roughly speaking, for $\omega_1 \to \infty$ ($r_1 \to 0$, respectively) the system (4.13) decouples: $\Xi^1 = \lambda = 0$ models the break-up of the original cycle Γ, and the solutions of $\Xi^2 = 0$, $\Xi^3 = 0$ correspond to symmetric 2-homoclinic orbits to p_2. Note that, in addition to the homoclinic orbits to p_1, there also exist a multitude of homoclinic orbits to p_2.

Let $\mathcal{H}^{out}_{\hat{\mathbf{r}}}(r_1)$ be the 3-homoclinic orbits belonging to the snaking curve $\lambda^{out}_{\hat{\mathbf{r}}}(r_1)$. Hence the snaking limit of $\mathcal{H}^{out}_{\hat{\mathbf{r}}}(r_1)$ for $r_1 \to 0$ (or, equivalently, for $\omega_1 \to \infty$) is the union of Γ and a particular (symmetric) 2-homoclinic orbit $\mathcal{H}_{2,2}(\hat{\mathbf{r}})$ to p_2. We write

$$\lim_{r \to 0} \mathcal{H}^{out}_{\hat{\mathbf{r}}}(r) = \Gamma \cup \mathcal{H}_{2,2}(\hat{\mathbf{r}}).$$

Indeed, this limit exists in the sense of the Hausdorff metric.

In the same way (4.13) can be solved for $(r_2, r_1, \lambda) = (r_2, r_1, \lambda)^{mid}(r_3)$. Then the corresponding snaking curve is parametrized by $\omega_2^2 =: \omega_3$ (resp., r_3). The right-hand side of (4.13) tends to the right-hand side of (4.15) as r_3 tends to zero:

$$
\begin{aligned}
0 = \hat{\Xi}^1 &= \lambda + \hat{L}_2(r_1, \lambda) + \hat{\mathcal{R}}^1_{r_3=0},\\
0 = \hat{\Xi}^2 &= \lambda + \hat{L}_1(r_2, \lambda) + \hat{\mathcal{R}}^2_{r_3=0},\\
0 = \hat{\Xi}^3 &= \lambda + \hat{L}_1(r_2, \lambda) + \hat{L}_2(r_1, \lambda) + \hat{\mathcal{R}}^3_{r_3=0}.
\end{aligned} \tag{4.15}
$$

Here, $\hat{\mathcal{R}}^1_{r_3=0} = \mathcal{O}\big((r_1)^\alpha\big)$, $\hat{\mathcal{R}}^2_{r_3=0} = \mathcal{O}\big((r_2)^\alpha\big)$, and $\hat{\mathcal{R}}^3_{r_3=0} = \mathcal{O}\big((r_2)^\alpha\big) + \mathcal{O}\big((r_1)^\alpha\big)$; now the $\hat{\mathcal{R}}$ do not depend on r_3.

Note that (4.15) is the bifurcation equation for 2-heteroclinic cycles connecting p_1 and p_2. First we make clear that near Γ there are infinitely many of these cycles. In the same way as outlined in section 4.2, system (4.15) can be written as a fixed point equation:

$$
(r_2, r_1, \lambda) = T_{\check{\mathbf{r}}}(r_2, r_1, \lambda). \tag{4.16}
$$

Let $\mathbf{r} := (r_2, r_1)$, and with that let $\check{\mathbf{r}}$ have the same meaning as $\hat{\mathbf{r}}$ in section 4.2. Using the arguments of the proof of Lemma 4.8, we find that (4.16) has a unique fixed point $(\mathbf{r}_{\check{\mathbf{r}}}, \lambda_{\check{\mathbf{r}}})$ in $\mathfrak{C}_{\check{\mathbf{r}}} \times [-\epsilon_{\check{\mathbf{r}}}, \epsilon_{\check{\mathbf{r}}}]$. (So, depending on the choice of $\check{\mathbf{r}}$ (4.16) has infinitely many solutions.)

Remark 4.10. *Note that if the vector field f is conservative, as in the example of the Swift–Hohenberg equation (2.2), then $\lambda_{\check{\mathbf{r}}} = 0$ for all $\check{\mathbf{r}}$. Indeed, for $\lambda \neq 0$ the equilibrium points p_1 and p_2 are in different level sets, and therefore they cannot be connected by a heteroclinic orbit.*

So, the snaking limit of $\mathcal{H}^{mid}_{\check{\mathbf{r}}}(r_3)$ is a 2-heteroclinic cycle. Note that this cycle depends on the particular choice of $\check{\mathbf{r}}$ which again is strongly related to the transition times near p_1 and p_2 (of both the cycle and the "nonsnaking part" of the considered 3-homoclinic orbit).

We can also discuss this behavior from a different perspective. Each of the symmetric 2-heteroclinic cycles can be seen as a primary cycle in its own right. Then, due to [14], there are 1-homoclinic orbits (to p_1) having this cycle as a snaking limit. But, related to Γ (in the present context), these orbits are the observed 3-homoclinic orbits.

Finally, we note that $\omega^1_1 =: \omega_2$ (resp., r_2) cannot serve as a snaking parameter; that is, the times separating the pulses stay almost constant along snaking curves. This follows from the fact that (4.13) cannot be solved for $(r_1, r_3, \lambda)(r_2)$, $r_2 \in (0, \epsilon)$. To see this let us assume the opposite. Then the corresponding snaking limit consists of two nonsymmetric 1-homoclinic orbits to p_1, which are R-images of each other, and a symmetric 1-homoclinic orbit to p_1. Indeed, $\hat{\boldsymbol{\Xi}} = (\hat{\Xi}^1, \hat{\Xi}^2, \hat{\Xi}^3)$ given in (4.13) can be extended continuously for $r_2 = 0$ and the corresponding limit provides the equations for those orbits. Further, the limit $\lim_{r_2 \to 0} (r_1, r_3, \lambda)(r_2)$ also exists (see Remark 4.9) and coincides with the solutions of (4.13) with $r_2 = 0$. But there are no nonsymmetric 1-homoclinic orbits near Γ: First note that nonsymmetric orbits come in pairs. In the language of section 4.1 both (nonsymmetric) 1-homoclinic orbits correspond to (different) 1-homoclinic Lin orbits $\{X^-, X^1_2, X^+\}$ with the same transition time ω_1. But, due to Lemma 4.1, there is only one (unique) Lin orbit for given transition time.

5. Conclusions.

In this paper we have studied the emergence of N-homoclinic orbits near heteroclinic cycles between equilibria of saddle-focus type (EE cycles) in reversible systems. Let us make some further comments and address topics for future research.

The complete bifurcation diagram for the SH equation. Our analysis has concerned the behavior for large transition times $\omega^i_{1,2}$, or, equivalently, small $r^i_{1,2}$, and has shown that for

each N there are infinitely many N-homoclinic orbits, which exist on distinct snaking curves near the cycle.

The numerical results for homoclinic orbits in the Swift–Hohenberg equation (2.2) show that two different snaking curves can be connected globally in that they can emerge in fold bifurcations close to a local bifurcation of the equilibrium at $r = 0$. This is illustrated in the results for 2-homoclinic orbits in Figure 3. Note that we do not show the other branches involved in the fold bifurcations in this figure.

A second possibility is revealed in the diagrams for 3-homoclinic orbits in Figure 4. Here, the red and green snaking curves merge in fold bifurcations close to the primary snaking curve. Such behavior seems to be in contrast to the results presented in section 4.3 above. In fact, due to our analysis in that section, both $r_2^{out}(r_1)$ and $r_3^{out}(r_1)$ are bounded away from zero as $r_1 \to 0$, as they stay in a neighborhood of $\hat{r}_2$ and $\hat{r}_3$. (We use the same notation as in section 4.4 above.) Therefore, for $r_1 \to 0$, we have that, in particular, $r_1 \ll r_3^{out}(r_1)$. Similarly we find that $r_3 \ll r_1^{mid}(r_3)$, as $r_3 \to 0$. Consequently the tails of the curves $\mathcal{H}_{\hat{\mathbf{r}}}^{out}(r_1)$ and $\mathcal{H}_{\check{\mathbf{r}}}^{mid}(r_3)$ do not intersect (independently on the concrete choice of $\hat{\mathbf{r}}$ and $\check{\mathbf{r}}$).

At the moment it is not clear whether fold bifurcations as in Figure 4, where the red and green branches meet, can be covered by our analysis. But we note that their analysis deserves further attention. For example, for 3-homoclinic orbits we have found that bifurcation curves $\mathcal{H}_{\check{\mathbf{r}}}^{mid}$ and $\mathcal{H}_{\hat{\mathbf{r}}}^{out}$ are globally connected. For N-homoclinic orbits with $N > 4$ there are more different types of snaking behavior and thus more possibilities for such global connections of snaking curves. It is of interest to investigate these bifurcation scenarios more closely.

Besides being reversible, (2.2) is also conservative, and thus nonsymmetric N-homoclinic solutions will also exist robustly near the primary orbit; see, for example, [3] for a related investigation. For the related case of snaking near an EP cycle (see also the last paragraph below), nonsymmetric solutions have recently been found to bifurcate in pitchfork bifurcations from the symmetric branches, creating a "snakes and ladder" structure; see [2]. Thus, our future work on N-homoclinic orbits will also include nonsymmetric solutions in order to see if similar structures can be found near EE cycles.

Stability of multipulses. For the related PDE (2.1) the N-homoclinic orbits discussed in this paper describe N-pulse solutions bifurcating from the fronts that connect the two equilibrium states. In this context the stability of the multipulses is of importance, since only stable patterns will be observed in numerical simulations or experiments. It is well known that the primary one-pulse solution alternates between being stable and unstable along the snaking curve and that this change of stability occurs when the fold points are crossed; see, for example, Figure 19 in [2]. In particular, note that at the bifurcation values, infinitely many stable 1-pulse homoclinic orbits exist.

The stability of N-pulse solutions near fronts or pulses is a more delicate issue. We only note here that stable N-pulse solution can exist near stable primary pulses. Of course, this cannot be investigated within the finite-dimensional framework of this paper, but methods are available to decide about stability rigorously; see [19].

For the Swift–Hohenberg equation simulations reveal striking similarities between the behavior of 1-pulse and 2-pulse solutions. Some results are presented in Figure 6, where we show the plots for 1- and 2-pulse orbits of (2.1). The simulations are based on a pseudospectral code with the linear terms integrated exactly using the ETD method [6] due to stiffness within

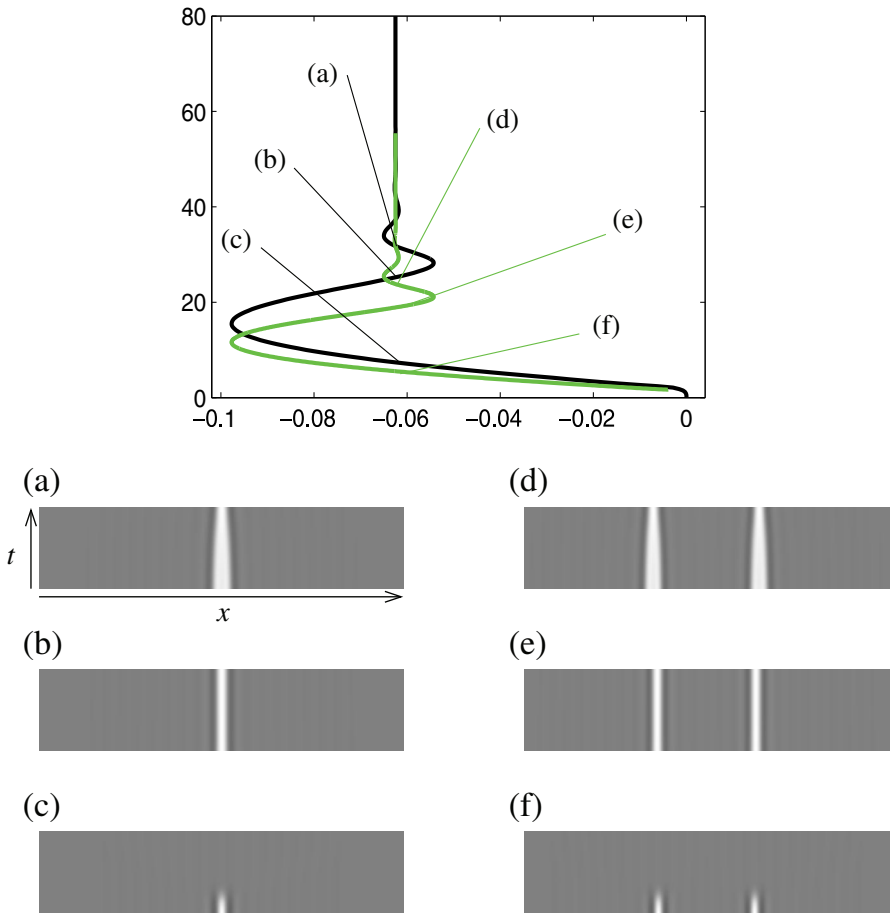


Figure 6. *Simulations of* 1*-pulse and* 2*-pulse solutions in the Swift–Hohenberg equation* (2.1).

the system. They have been performed for the critical parameter value $r = -0.0625$ over a time interval $t \in [0, 250]$. As initial profiles we have used perturbations of the 1-homoclinic orbit (panels (a)–(c)), and of the 2-homoclinic solutions along the green branch in Figure 3, panels (d)–(f).

The panels below the diagram in Figure 6 show the time evolution of the pulses viewed from the top with lighter regions corresponding to larger amplitudes. It can be seen that the pulses in panels (a), (d) and (c), (f), respectively, are unstable, with the amplitude of the pulses decaying to zero. Note that the decay in panels (a) and (d) is very slow. However, the pulses in panels (b) and (e) show stable behavior under the evolution in time. Because of the similarities in the evolution of 1-pulse and 2-pulse solutions, we can expect the emergence of infinitely many stable multipulse solutions in snaking scenarios in the Swift–Hohenberg equation.

N-homoclinic orbits near EP cycles. Finally, it is interesting to note that the behavior of N-homoclinic orbits reveals a striking difference in the dynamics near EE cycles and EP cycles. Note first that EP cycles can exist robustly in reversible systems, since the corresponding stable

and unstable manifolds can intersect transversally. Therefore, the fold points along snaking curves of homoclinic orbits in their neighborhood do not converge to a single parameter value but rather oscillate between two values which converge to the boundaries of the parameter interval in which the cycle exists; see also Figure 7.

This global character of bifurcations near an EP cycle makes a rigorous analytical treatment difficult. Recently, an analysis of 1-homoclinic orbits was presented in [1]. In addition, numerical results for N-homoclinic orbits near such a cycle have been described in [9, 22].

Below we present a corresponding result for the generalized Swift–Hohenberg equation (2.2). Similar to section 2 we view the equation as depending on r, this time setting $q_c = 0.5$, $v = 0.41$, and $g = 1$. And, as in that section, 1-homoclinic orbits to 0 are computed using shooting, and two-homoclinic orbits by the branch-switching method.

In Figure 7 we present a bifurcation diagram for symmetric 2-homoclinic orbits to 0 near an EP cycle, which exists for $r \in [-0.0146, -0.0125]$. This EP cycle generates two snaking curves of 1-homoclinic orbits, one which is shown in grey in the figure. In addition, the rescaled continuation curve for a 2-homoclinic orbit is shown in red in the figure. And although this curve follows (parts of) the snaking curve, we see a clear difference. In contrast to the EE cycle case the 2-homoclinic orbit does not exist on a snaking curve, but rather on an *isola* in the diagram.

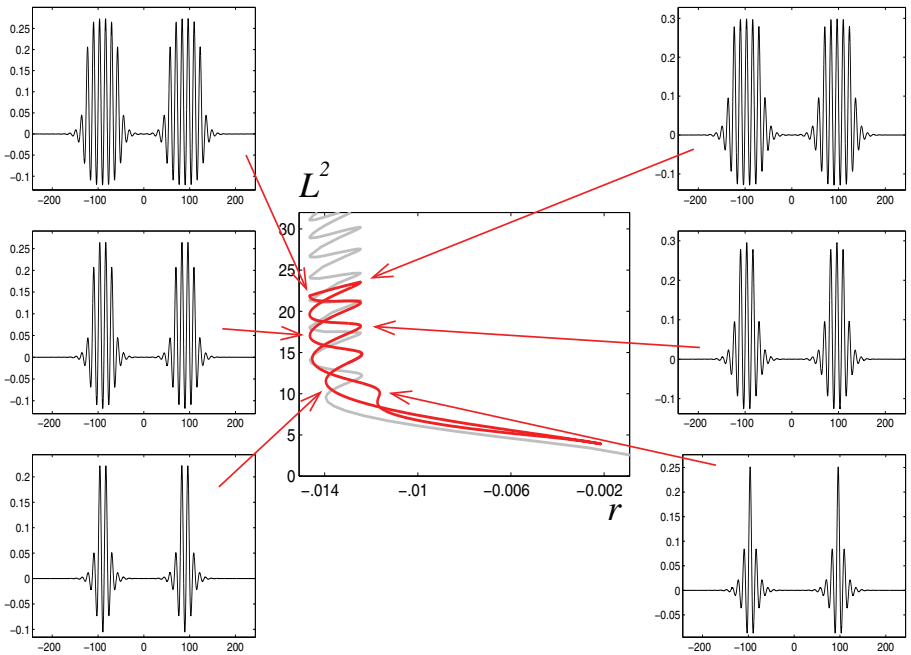


Figure 7. 2-*homoclinic orbits near an EP cycle exist along an isola in* (2.2).

A reason for this becomes apparent when we consider the plots for the 2-homoclinic solutions along the snaking curve. Moving up on the bifurcation curve, we see that the pulses of the solutions become wider, but in contrast to the EE cycle case, the pulses grow symmetrically about their center such that they do not stay separated but approach each other. (Note that instead of approaching an equilibrium solution the pulses now come close

to a periodic orbit such that they grow additional oscillations.) Hence, in the notation of section 4, the time ω_1^1 decreases along the snaking curve. And this process cannot be repeated ad infinitum, since the two pulses would have to meet, and therefore the 2-homoclinic orbit cannot follow the full snaking curve. Similar behavior has been found for different examples in [9, 22]. Indeed, so far, 2-homoclinic orbits near EP cycles have been found to lie only on isolas and not on snaking curves.

The precise character of these isolas is the subject of current research. General bifurcation results will be presented in a forthcoming paper. Furthermore, we also aim to understand the geometrical reason for the different behavior of N-homoclinic orbits near EE cycles and EP cycles, respectively.

Acknowledgments. We thank Steve Houghton for letting us use his numerical code for the PDE simulations. We are also grateful to the referees whose comments helped to improve the presentation of our results.

REFERENCES

[1] M. Beck, J. Knobloch, D. J. B. Lloyd, B. Sandstede, and T. Wagenknecht, *Snakes, Ladders, and Isolas of Localised Patterns*, preprint, University of Surrey, Guildford, UK, 2008; available online from http://www.dam.brown.edu/people/sandsted/publications/snaking-1d.pdf.

[2] J. Burke and E. Knobloch, *Localized states in the generalized Swift-Hohenberg equation*, Phys. Rev. E (3), 73 (2006), 056211.

[3] A. R. Champneys, *Subsidiary homoclinic orbits to a saddle-focus for reversible systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 1447–1482.

[4] M. Chen, *Solitary-wave and multi-pulsed travelling-wave solutions of Boussinesq systems*, Appl. Anal., 75 (2000), pp. 213–240.

[5] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1996.

[6] S. M. Cox and P. C. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.

[7] P. Coullet, C. Riera, and C. Tresser, *Stable static localized structures in one dimension*, Phys. Rev. Lett., 84 (2000), pp. 3069–3072.

[8] J. Härterich, *Cascades of homoclinic orbits in reversible dynamical systems*, Phys. D, 112 (1998), pp. 187–200.

[9] G. H. M. van der Heijden, A. R. Champneys, and J. M. T. Thompson, *Spatially complex localisation in twisted elastic rods constrained to a cylinder*, Internat. J. Solids Structures, 39 (2002), pp. 1863–1883.

[10] M. F. Hilali, S. Métens, P. Borckmans, and G. Dewel, *Pattern selection in the generalized Swift-Hohenberg model*, Phys. Rev. E, 51 (1995), pp. 2046–2052.

[11] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. Ahmer Wadee, C. J. Budd, and G. J. Lord, *Cellular buckling in long structures*, Nonlinear Dynam., 21 (2000), pp. 3–29.

[12] J. Knobloch, *Lin's Method for Discrete and Continuous Dynamical Systems and Applications*, TU Ilmenau, Ilmenau, Germany, 2004, http://imath.mathematik.tu-ilmenau.de/~knobi.

[13] J. Knobloch, J. S. W. Lamb, and K. Webster, *Shift Dynamics near T-point Heteroclinic Cycles*, preprint, TU Ilmenau, Ilmenau, Germany, 2008; available online from http://www.tu-ilmenau.de/fakmn/fileadmin/template/ifm/user/Knobloch/publikationen/t-point200707.pdf.

[14] J. Knobloch and T. Wagenknecht, *Homoclinic snaking near a heteroclinic cycle in reversible systems*, Phys. D, 206 (2005), pp. 82–93.

[15] J. S. W. Lamb and J. A. G Roberts, *Time-reversal symmetry in dynamical systems: A survey*, Phys. D, 112 (1998), pp. 1–39.

[16] B. E. Oldeman, A. R. Champneys, and B. Krauskopf, *Homoclinic branch switching: A numerical implementation of Lin's method*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 13 (2003), pp. 2977–3000.

[17] B. Sandstede, *Verzweigungstheorie homokliner Verdopplungen*, WIAS Report 7, WIAS, Berlin, Germany, 1993; see also http://www.dam.brown.edu/people/sandsted/publications/dissertation.pdf.

[18] B. Sandstede, *Instability of localized buckling modes in a one-dimensional strut model*, Philos. Trans. Roy. Soc. London Ser. A, 355 (1997), pp. 2083–2097.

[19] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 983–1055.

[20] M. Tlidi, P. Mandel, and R. Lefever, *Localized structures and localized patterns in optical bistability*, Phys. Rev. Lett., 73 (1994), pp. 640–643.

[21] A. Vanderbauwhede and B. Fiedler, *Homoclinic period blow-up in reversible and conservative systems*, Z. Angew. Math. Phys., 43 (1992), pp. 292–318.

[22] M. K. Wadee, C. D. Coman, and A. P. Bassom, *Solitary wave interaction phenomena in the strut buckling model incorporating restabilisation*, Phys. D, 163 (2002), pp. 26–48.

[23] P. D. Woods and A. R. Champneys, *Heteroclinic tangles and homoclinic snaking in the unfolding of a degenerate reversible Hamiltonian–Hopf bifurcation*, Phys. D, 129 (1999), pp. 147–170.

[24] E. J. Doedel, A. R. Champneys, T. R. Fairgrieve, Yu. A. Kuznetsov, B. Sandstede, and X. J. Wang, *AUTO97: Continuation and Bifurcation Software for Ordinary Differential Equations*, 1997. Available by anonymous ftp from ftp.cs.concordia.ca, directory pub/doedel/auto.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1421–1444
© 2008 Society for Industrial and Applied Mathematics

Hysteresis in a Rotating Differentially Heated Spherical Shell of Boussinesq Fluid*

Gregory M. Lewis† and William F. Langford‡

Abstract. A mathematical model of convection of a Boussinesq fluid in a rotating spherical shell is analyzed using numerical computations guided by bifurcation theory. The fluid is differentially heated on its inner spherical surface, with the temperature increasing from both poles to a maximum at the equator. The model is assumed to be both rotationally symmetric about the polar axis and reflectionally symmetric across the equator. This work is an extension to spherical geometry of previous work on the differentially heated rotating annulus. The spherical geometry is motivated by applications to planetary atmospheres. As the temperature gradient increases from zero, large Hadley cells extending from equator to poles form immediately. For larger temperature differences, two or three convection cells appear in each hemisphere. An organizing center is shown to exist, at which two saddle-node bifurcations come together in a codimension-2 hysteresis bifurcation (or cusp) point, providing a mechanism for hysteretic transitions between different cell patterns as the temperature gradient is varied.

Key words. cusp point, hysteresis bifurcation, flow transitions, Boussinesq fluid, flow in a rotating spherical shell, numerical computation, large-scale geophysical flow

AMS subject classifications. 37N10, 76U05, 37M99

DOI. 10.1137/070697306

1. Introduction.

The atmosphere of a planet may be idealized as a spherical shell of fluid surrounding the spherical surface of the planet. Many factors affect the circulation of the atmosphere; chief among these are the rotation of the planet, the differential heating of the surface and atmosphere by the planet's sun, and the thickness and composition of the atmosphere itself. In this paper we construct an idealized mathematical model of such a planet, ignoring all local variations of surface features such as oceans, continents, mountains, and glaciers. The symmetries of this model are exploited to make the computations more tractable.

On the inner boundary surface a temperature profile is prescribed to reflect the differential heating of the atmosphere by the sun, as follows. The average annual flux of solar radiation on a planet whose axis of rotation is tilted approximately 20° with respect to the plane that is perpendicular to the solar rays is nearly proportional to $-\cos(2\theta)$ (see [19]), where θ is the polar angle, which differs from the latitude only in its range (in particular, the latitude is given by $\pi/2 - \theta$). This flux is independent of the azimuthal variable φ and is a maximum at the

*Received by the editors July 15, 2007; accepted for publication (in revised form) by D. Barkley July 13, 2008; published electronically November 7, 2008. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada and by the Fields Institute for Research in Mathematical Sciences.
http://www.siam.org/journals/siads/7-4/69730.html
†Faculty of Science, University of Ontario Institute of Technology, 2000 Simcoe Street North, Oshawa, ON, L1H 7K4 Canada (greg.lewis@uoit.ca).
‡Department of Mathematics and Statistics, University of Guelph, 50 Stone Road East, Guelph, ON, N1G 2W1 Canada (wlangfor@uoguelph.ca).

equator ($\theta = \pi/2$) and a minimum at the poles ($\theta = 0, \pi$). This profile is similar to the annual average of solar flux for many of the planets in our solar system, including Earth. Therefore, we choose the temperature on the inner boundary surface to be fixed at $T = T_r - \Delta T \cos(2\theta)$, where T_r is a reference temperature and the difference in the temperature from equator to pole is $2\Delta T$. On a real planet, a flux of radiation of this form would not result exactly in this prescribed temperature distribution on the surface, but we expect this model to capture the gross effects of the differential heating. In the model, the spherical shell is filled with a Boussinesq fluid, which means that its density varies linearly with temperature. Thus, the gravitational force acting radially inward drives convective motions of the fluid when $\Delta T > 0$.

We assume that both the inner and outer spheres are rigid, and that the fluid satisfies no-slip conditions on both boundaries. For the temperature at the outer sphere, we assume insulating boundary conditions. Inside the spherical shell, the fluid satisfies the Navier–Stokes equations in the Boussinesq approximation. Complete details of the model are given in section 2. The nonlinear equations for a steady state are solved using Newton iteration with Keller continuation from the trivial solution at $\Delta T = 0$. The linear stability problem is solved using the implicitly restarted Arnoldi method following a generalized Cayley transformation. The full methodology for analysis of the model is described in sections 3 and 4, and the results are presented in sections 5 and 6.

It is not possible in an Earth-bound laboratory to perform an experimental study of convection in a spherical shell as described here, because the gravitational force cannot be directed radially towards the common center of the spheres. (However, experiments to study spherical convection with a central force field under weightless conditions in a space lab are an interesting possibility; see, for example, [11, 3].) Many laboratory experiments have been performed on a differentially heated rotating cylindrical annulus, with Earth's gravity acting downward parallel to the cylinder axis and with centrifugal force acting radially. These experiments typically used water as the working fluid. Much has been learned about the dynamics of large-scale geophysical fluids from these laboratory experiments, even though the Reynolds number is significantly smaller in the experimental fluids than in actual geophysical flows [12, 24]. The Boussinesq approximation has been used as a basis for a mathematical model of the differentially heated rotating annulus, using laboratory-scale parameters and with water as the fluid [13, 20, 21]. The model has provided good agreement with the corresponding laboratory experiments.

Therefore, in this prototype mathematical model of a differentially heated rotating spherical shell of fluid, we adopt a philosophy of choosing the parameters of the fluid to be those of water and using the laboratory scale rather than the geophysical. In this way, the computations are tractable and the results obtained can be compared to both experiments and theoretical calculations for the better understood case of a differentially heated rotating annulus. This may provide insight into how the geometry of the system leads to the observed flow transitions, and may serve as a step towards understanding the dynamical structure of planetary systems. Although we have made quantitative choices for the fluid and the boundary conditions, we expect that these choices will have little effect on qualitative features of the results. This is supported by the numerous studies that have been performed on the differentially heated rotating annulus. Many experiments have been performed with different forms of heating, different geometries of the apparatus, and with different fluid properties,

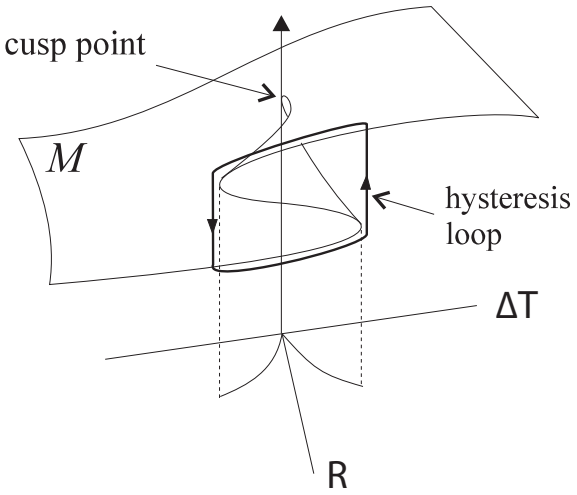


Figure 1. *The codimension-2 hysteresis bifurcation, showing the hysteresis loop as ΔT varies back and forth. The cusp shown in the $(R, \Delta T)$ parameter plane is the projection of the two curves of fold bifurcations onto this plane.*

and it is found that these have little qualitative effect on the types of transitions and form of the transition curves that are observed.

The goal of this paper is to determine the flow patterns, their stabilities and their transitions (bifurcations) for a differentially heated rotating spherical shell, consistent with our modeling assumptions. There are three bifurcation parameters of interest: the temperature difference ΔT, the shell gap width R, and the rotation rate Ω. The results presented in section 5 show that the qualitative features of the flow patterns do not change for moderate changes of Ω, although the stability with respect to non–rotationally symmetric perturbations is affected. However, for even very small $\Delta T > 0$, the fluid immediately begins to flow in a large stable convection cell with easterly flow toward the equator near the inner surface. This flow is similar to the Hadley cell, which exists in the atmosphere between approximately the equator and 30° latitude (in each hemisphere), except that it extends from equator to pole. In this paper, we refer to this convection cell as the Hadley cell. For larger values of ΔT, two or even three convection cells may appear between the equator and pole. The Hadley cell shrinks, always keeping one edge at the equator, while the additional cells appear between it and the pole. When a second cell exists next to the Hadley cell, it is characterized by a weaker counter-rotating flow, with a westerly component. The third cell, when it exists, has the same sense of rotation as the Hadley cell but lies near the pole. In every case, there is a region of high velocity azimuthal flow at high altitudes and midlatitudes, resembling Earth's *jet stream*.

Furthermore, we show that there exists a critical value of the pair $(R, \Delta T)$ that is a *hysteresis bifurcation point* (also called a *cusp point*); see Figure 1. This is a codimension-2 steady-state bifurcation that is well understood in mathematical bifurcation theory. The existence of this hysteresis point is demonstrated by an explicit calculation of its defining conditions in section 6, where the study of this hysteresis bifurcation is presented in detail. Only

its main features are outlined here. Figure 1 represents a generic hysteresis bifurcation. Here M is a manifold of steady states, existing for given parameters $(R, \Delta T)$, and the vertical axis represents the amplitude of a steady state (determined by the stream function in section 6). In a neighborhood of a hysteresis bifurcation point, the number of steady states (on M) can vary from one to three as the parameters $(R, \Delta T)$ vary. When three steady states coexist in Figure 1, the middle one is unstable, but all other steady states not on the fold lines in Figure 1 remain stable (attracting). Furthermore, there are abrupt upward and downward transitions at two *fold bifurcations* (also called limit points or saddle-node bifurcations), as shown in Figure 1, that occur at both ends of an interval of the bifurcation parameter ΔT, for appropriate fixed R. If ΔT is varied back and forth through this interval, the system traces a *hysteresis loop*, as shown in Figure 1. Between these two abrupt transitions there is an interval of ΔT that exhibits *bistability*; that is, two solutions are stable simultaneously.

1.1. Relationship to previous work. The fundamental problem of convection in spherical shells was formulated by Chandrasekhar [4, Chap. 6], who also solved the stability problem in the Boussinesq approximation for the spherically symmetric case, in terms of spherical harmonics. There is a large literature on spherically symmetric convection that is motivated by Earth's core and mantle, where the boundary conditions have full spherical symmetry (in particular, a constant temperature on the inner sphere); see, for example, [3, 5, 22, 31]. The present work differs from all of these because of the latitudinal temperature gradient.

Motivation for the present work is provided also by the classical Taylor–Couette experiment, in which the flow of a fluid (usually water) between two differentially rotating long coaxial cylinders is studied; see, for example, [2, 6, 8, 17]. In that experiment many interesting flow patterns may form, including *Taylor vortices*, which are invariant tori stacked coaxially between the cylinders in pairs with alternating clockwise and counterclockwise helical flows. There is no differential heating in the Taylor–Couette experiment; even so, the Taylor vortices resemble Hadley cells. Marcus and Tuckerman [25, 26] performed numerical simulations of the flow between two differentially rotating spheres (without heating) and found bifurcations to different numbers of cells that they called Taylor vortices, thus showing that Hadley-like cells may form even without differential heating, if the spheres are rotated differentially. See also [14].

2. Model equations. We use the Navier–Stokes equations in the Boussinesq approximation to model the fluid flow within the spherical shell. In the Boussinesq approximation, the variations of all fluid properties except the density are considered to be negligible, and the equation of state of the fluid is assumed to be

$$\rho = \rho_0 \left(1 - \alpha \left(T - T_r\right)\right), \tag{2.1}$$

where ρ is the density of the fluid, T is the temperature, α is the (constant) coefficient of thermal expansion, and ρ_0 is the density at a reference temperature T_r. The dimensionless quantity $\alpha\left(T - T_r\right)$ is assumed to be small. In the Boussinesq approximation the fluid is considered to be incompressible, which is a significant simplification.

The fluid is contained within a spherical shell with inner sphere of radius r_a and outer sphere of radius r_b. We assume gravity acts everywhere in the radial direction. The spherical shell rotates at rate Ω about the polar axis, and the inner and outer spheres rotate at the

same rate. The equations are written in spherical polar coordinates in a frame of reference corotating at rate Ω with the shell. The radial, polar, and azimuthal coordinates are denoted r, θ, and φ, respectively, with unit vectors $\mathbf{e}_r$, $\mathbf{e}_\theta$, and $\mathbf{e}_\varphi$.

The Navier–Stokes Boussinesq equations describing the evolution of the vector fluid velocity, $\mathbf{u} = \mathbf{u}(r,\theta,\varphi,t) = w\mathbf{e}_r + v\mathbf{e}_\theta + u\mathbf{e}_\varphi$, and the temperature of the fluid, $T = T(r,\theta,\varphi,t)$, are

$$\frac{\partial \mathbf{u}}{\partial t} = \nu\nabla^2\mathbf{u} - 2\boldsymbol{\Omega}\times\mathbf{u} + \left[g\mathbf{e}_r + \boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{r})\right]\alpha\,(T - T_r) - \frac{1}{\rho_0}\nabla p - (\mathbf{u}\cdot\nabla)\,\mathbf{u}, \tag{2.2}$$

$$\frac{\partial T}{\partial t} = \kappa\nabla^2 T - (\mathbf{u}\cdot\nabla)\,T, \tag{2.3}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{2.4}$$

where $\boldsymbol{\Omega} = \Omega\,(\cos\theta\mathbf{e}_r - \sin\theta\mathbf{e}_\theta)$ is the rotation vector, $\Omega = |\boldsymbol{\Omega}|$ is the rate of rotation about the polar axis, p is the pressure deviation from $p_0 = \rho_0 g(R - r) + \rho_0\Omega^2 r^2\sin^2\theta/2$, $\mathbf{r} = r\mathbf{e}_r + \theta\mathbf{e}_\theta + \varphi\mathbf{e}_\varphi$, ν is the kinematic viscosity, κ is the coefficient of thermal diffusivity, g is the gravitational acceleration, ∇ is the usual gradient operator in spherical coordinates, u is the azimuthal fluid velocity, often referred to as the zonal velocity, v is the polar fluid velocity, and w is the radial fluid velocity. The spatial domain is defined by $r_a < r < r_b$, $0 \le \varphi < 2\pi$, and $0 < \theta < \pi$. Thus, $\theta = 0, \pi$ correspond to the north and south poles of the shell, respectively, while $\theta = \pi/2$ corresponds to the equator. The equations can be rewritten in planetary coordinates by performing the change of variable $\theta \to \pi/2 - \theta$. The values of ν and κ are chosen to be those of the fluid at the reference temperature T_r, and it is assumed that the difference between the temperature of the fluid and T_r is everywhere small enough so that ν and κ can be considered as constants. We have included the effects of centrifugal buoyancy in the equations via the term $\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{r})$. All dimensional quantities are measured in CGS units.

As described in the introduction, the boundary conditions are

$$\begin{aligned} &\mathbf{u} = 0, \qquad T = T_r - \Delta T\cos(2\theta) \qquad \text{on} \quad r = r_a, \\ &\mathbf{u} = 0, \qquad \frac{\partial T}{\partial r} = 0 \qquad\quad \text{on} \quad r = r_b, \end{aligned} \tag{2.5}$$

with 2π-periodicity in the azimuthal variable φ.

In this paper we investigate flows that preserve the symmetries of the model, that is, flows that are invariant under rotation about the polar axis (i.e., axisymmetric flows) and that are invariant under reflection across the equator (i.e., across the line defined by $\theta = \pi/2$). Therefore, we study solutions of (2.2)–(2.5) in the form

$$\begin{aligned} &u = u(r,\theta,t) = u(r,\pi-\theta,t), \qquad v = v(r,\theta,t) = v(r,\pi-\theta,t), \\ &w = w(r,\theta,t) = w(r,\pi-\theta,t), \quad T = T(r,\theta,t) = T(r,\pi-\theta,t). \end{aligned} \tag{2.6}$$

The assumed symmetries significantly simplify the analysis. We may use the analysis of the symmetric system as a starting point for an analysis of the full system. Although it is not written explicitly, the solutions also depend on the parameters.

If we scale the radial coordinate as

$$r \to Rr', \tag{2.7}$$

where $R = r_b - r_a$ is the gap width, write the temperature as

$$T \to T' + T_r - \Delta T \cos(2\theta), \tag{2.8}$$

substitute these into (2.2)–(2.4), and drop the primes, we obtain the following axisymmetric equations describing the evolution of the fluid velocity $\mathbf{u} = w(r,\theta,t)\mathbf{e}_r + v(r,\theta,t)\mathbf{e}_\theta + u(r,\theta,t)\mathbf{e}_\varphi$, pressure deviation $p = p(r,\theta,t)$, and temperature deviation $T = T(r,\theta,t)$:

$$\begin{aligned}\frac{\partial u}{\partial t} &= \nu_s \nabla_0^2 u - \nu_s \frac{1}{r^2 \sin^2\theta} u - 2\Omega\left(\sin\theta w + \cos\theta v\right) \\ &\quad - \frac{1}{R}\left[\left(\mathbf{u}\cdot\nabla_0\right) u + \frac{\cos\theta}{r\sin\theta} uv + \frac{uw}{r}\right],\end{aligned} \tag{2.9}$$

$$\begin{aligned}\frac{\partial v}{\partial t} &= \nu_s \nabla_0^2 v - \nu_s\left(\frac{1}{r^2\sin^2\theta} v - \frac{2}{r^2}\frac{\partial w}{\partial \theta}\right) + 2\Omega\cos\theta u - \frac{1}{\rho_0 R r}\frac{\partial p}{\partial \theta} \\ &\quad - \left(\alpha\Omega^2 R r \sin\theta\cos\theta\right)\left(T - \Delta T\cos 2\theta\right) - \frac{1}{R}\left[\left(\mathbf{u}\cdot\nabla_0\right) v - \frac{\cos\theta}{r\sin\theta} u^2 + \frac{vw}{r}\right],\end{aligned} \tag{2.10}$$

$$\begin{aligned}\frac{\partial w}{\partial t} &= \nu_s \nabla_0^2 w - \nu_s\left(\frac{2}{r^2}\frac{\cos\theta}{\sin\theta} v + \frac{2}{r^2}\frac{\partial v}{\partial\theta} + \frac{2}{r^2} w\right) + 2\Omega\sin\theta u - \frac{1}{\rho_0 R r}\frac{\partial p}{\partial r} \\ &\quad - \alpha\left(\Omega^2 R r\sin^2\theta + g\right)\left(T - \Delta T\cos 2\theta\right) - \frac{1}{R}\left[\left(\mathbf{u}\cdot\nabla_0\right) w - \frac{1}{r}\left(u^2 + v^2\right)\right],\end{aligned} \tag{2.11}$$

$$\frac{\partial T}{\partial t} = \kappa_s \nabla_0^2 T + \frac{4\Delta T \kappa_s}{r^2}\left(\cos 2\theta + \cos^2\theta\right) + \frac{2\Delta T}{Rr}\sin 2\theta v - \frac{1}{R}\left(\mathbf{u}\cdot\nabla_0\right) T, \tag{2.12}$$

$$\nabla_0 \cdot \mathbf{u} = \frac{\partial w}{\partial r} + \frac{2}{r} w + \frac{1}{r}\frac{\partial v}{\partial\theta} + \frac{\cos\theta}{r\sin\theta} v \;=\; 0, \tag{2.13}$$

where $\nu_s = \nu/R^2$, $\kappa_s = \kappa/R^2$,

$$\nabla_0^2 = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{r^2\sin\theta}\frac{\partial}{\partial\theta},$$

$$\nabla_0 = \mathbf{e}_r\frac{\partial}{\partial r} + \mathbf{e}_\theta\frac{1}{r}\frac{\partial}{\partial\theta},$$

and

$$\left(\mathbf{u}\cdot\nabla_0\right) f = w\frac{\partial f}{\partial r} + \frac{v}{r}\frac{\partial f}{\partial\theta} \tag{2.14}$$

for any scalar function $f = f(r,\theta,t)$. The domain is now expressed as $r_a/R < r < r_b/R$, $0 \le \theta < \pi/2$, and the boundary conditions become

$$\begin{aligned}&\mathbf{u} = 0, \qquad T = 0 \qquad \text{on} \quad r = \frac{r_a}{R}, \\ &\mathbf{u} = 0, \qquad \frac{\partial T}{\partial r} = 0 \qquad \text{on} \quad r = \frac{r_b}{R}.\end{aligned} \tag{2.15}$$

The symmetry assumptions not only reduce the domain size (i.e., we now have $0 \le \theta \le \pi/2$), but also effectively introduce new boundary conditions at the equator and the pole. In order to satisfy the symmetries, there can be no flow of fluid or heat across the equator or the pole. In addition, the condition $u = 0$ at the pole is necessary to ensure that no discontinuity occurs in the fluid velocity at the pole. Thus, we have the additional boundary conditions

$$\begin{aligned} & u, v = 0, \qquad \frac{\partial w}{\partial \theta} = 0, \qquad \frac{\partial T}{\partial \theta} = 0 \qquad \text{on} \quad \theta = 0, \\ & v = 0, \qquad \frac{\partial u}{\partial \theta} = \frac{\partial w}{\partial \theta} = 0, \qquad \frac{\partial T}{\partial \theta} = 0 \qquad \text{on} \quad \theta = \frac{\pi}{2}. \end{aligned} \tag{2.16}$$

It is possible to write the equations completely in terms of dimensionless variables. However, this would not simplify the analysis, so we choose to work with the equations in the form (2.9)–(2.13). This follows previous numerical work on similar problems, e.g., [13, 30, 20].

3. Analysis.

3.1. Nonlinear equations for steady solution. The analysis begins with the computation of steady axisymmetric solutions that are invariant with respect to reflection across the equator; that is, we seek solutions of (2.9)–(2.13) that satisfy the boundary conditions (2.15)–(2.16) and are independent of time.

The method of stream functions is used to solve the steady equations. If v and w are written in terms of a (Stokes) stream function ξ, defined by

$$v = -\frac{1}{r \sin\theta} \frac{\partial \xi}{\partial r}, \qquad w = \frac{1}{r^2 \sin\theta} \frac{\partial \xi}{\partial \theta}, \tag{3.1}$$

then the incompressibility condition (2.13) is automatically satisfied [1]. After using (3.1) to replace v and w in the equations, the pressure terms can be eliminated. Subsequently, the steady solution can be found from the resulting three equations in the three unknown functions u, ξ, and T. These equations are found using the Maple symbolic computation software package and are sufficiently complicated that no insight is gained by explicitly writing them here. The boundary conditions for u and T are given by (2.15)–(2.16), as before, while the conditions on v and w will be satisfied if ξ satisfies the boundary conditions

$$\begin{aligned} & \xi = 0, \qquad \frac{\partial \xi}{\partial \theta} = 0, \qquad \frac{\partial^3 \xi}{\partial \theta^3} = 0 \quad \text{on} \quad \theta = 0, \\ & \xi = 0, \qquad \frac{\partial^2 \xi}{\partial \theta^2} = 0 \qquad \text{on} \quad \theta = \frac{\pi}{2}, \\ & \xi = 0, \qquad \frac{\partial \xi}{\partial r} = 0 \qquad \text{on} \quad r = \frac{r_a}{R}, \frac{r_b}{R}. \end{aligned} \tag{3.2}$$

The numerical algorithm to solve this system of nonlinear equations is described in section 4, and the steady axisymmetric solutions obtained are presented in section 5.

3.2. Linear stability analysis. The linear stability of a steady solution is defined in terms of the eigenvalues of the linearization of the dynamical equations about that solution. If the real parts of *all* the eigenvalues are negative, then all perturbations from the steady solution

will decay in the linearized equations. In this case, the solution is said to be linearly stable. If any of the eigenvalues has positive real part, then some small perturbations will grow, and the solution is linearly unstable. If there exist only eigenvalues with zero real part and negative real part, the solution is called neutrally stable. If the real part of an eigenvalue crosses the imaginary axis as a parameter is varied, then a qualitative change in the solution occurs; i.e., a bifurcation takes place. Flows corresponding to linearly stable steady solutions can be observed physically if the noise that is naturally present is sufficiently small. If a steady solution is linearly unstable, then the corresponding flow cannot be observed because some small perturbations due to the noise will tend to grow. Thus, it is expected that bifurcations correspond to transitions in the observed flow.

We compute the steady axisymmetric solution and its linear stability as the parameters of the system change, and we seek locations in the space of parameters where an eigenvalue crosses the imaginary axis. We are primarily interested in solutions that do not break the assumed symmetry, and therefore we initially require that this eigenvalue be associated with an eigenfunction that respects the symmetry.

If we write

$$u = u' + u_0, \qquad \xi = \xi' + \xi_0, \qquad T = T' + T_0, \tag{3.3}$$

where u_0, ξ_0, T_0 is a steady solution, and substitute into the three equations for u, ξ, and T, we obtain the perturbation equations in u', ξ', and T'. The trivial solution satisfies the perturbation equations, and it corresponds to u_0, ξ_0, T_0. If the perturbation equations are linearized and we assume that the unknown functions may be written as

$$u'(r,\theta,t) = e^{\lambda t}\psi_u(r,\theta), \quad \xi'(r,\theta,t) = e^{\lambda t}\psi_\xi(r,\theta), \quad T'(r,\theta,t) = e^{\lambda t}\psi_T(r,\theta), \tag{3.4}$$

then a linear eigenvalue problem is obtained. Consequently, the eigenvalues λ can be found from the generalized eigenvalue problem of the form

$$\lambda \mathbf{A}_0 \Psi = \mathbf{L}_0 \Psi, \tag{3.5}$$

where

$$\Psi = \begin{pmatrix} \psi_u \\ \psi_\xi \\ \psi_T \end{pmatrix}$$

is the eigenfunction and $\mathbf{A}_0$ and $\mathbf{L}_0$ are 3×3 matrices of linear differential operators.

The perturbations u', ξ', and T' correspond to axisymmetric perturbations. If all eigenvalues corresponding to these perturbations have negative real part, then the steady solution u_0, ξ_0, T_0 is a linearly stable solution of the axisymmetric equations (2.9)–(2.13). For this steady solution to be a corresponding linearly stable solution of the full three-dimensional model (2.2)–(2.4), we must also compute the eigenvalues corresponding to the nonaxisymmetric perturbations. To do this, we linearize the perturbation equations corresponding to the three-dimensional model, and we assume that the eigenfunctions have the form

$$\Phi(r,\theta,\varphi) = \hat{\Phi}_m(r,\theta)e^{im\varphi},$$

where $m = 1, 2, \ldots$ is the azimuthal wave number. Unlike in the axisymmetric case, it is not possible to solve this problem using a stream function approach. However, due to the form of the eigenfunctions, it is possible to use the incompressibility condition and the equation for the azimuthal velocity to eliminate the pressure and the azimuthal velocity, resulting in a generalized eigenvalue problem for each wave number m. The eigenvalue problem has the form (3.5) with $\hat{\Psi}_m$ replacing Ψ, and where

$$\hat{\Psi}_m = \begin{pmatrix} \hat{v}_m \\ \hat{w}_m \\ \hat{T}_m \end{pmatrix},$$

and the functions $\hat{v}_m, \hat{w}_m, \hat{T}_m$ depend only on r and θ.

To ensure the continuity of solutions, we require that, for $m \neq 1$, $\hat{v}_m, \hat{w}_m, \hat{T}_m$ vanish at the pole. Furthermore, to ensure that the solutions have continuous first derivatives at the pole, we also require that, for $m \neq 1$,

$$\frac{\partial \hat{v}_m}{\partial \theta} = \frac{\partial \hat{w}_m}{\partial \theta} = \frac{\partial \hat{T}_m}{\partial \theta} = 0 \qquad \text{on} \quad \theta = 0.$$

For $m = 1$, the condition of continuity also requires that $\hat{w}_1$ and $\hat{T}_1$ vanish at the pole. However, no such restriction applies to the meridional velocity, because the condition that $v(r, \theta, \varphi) = -v(r, \theta, \varphi + \pi)$ in the limit as $\theta \to 0$ allows for continuity in this case. However, to avoid the difficulty of computing nonzero solutions at $\theta = 0$, we look for stability only with respect to perturbations that satisfy $\hat{v}_1(r, \theta = 0) = 0$; this corresponds to perturbations that do not exhibit flow across the pole. We consider this additional boundary condition to be an additional simplifying assumption of the model. Furthermore, to reduce computational requirements we also compute the stability only with respect to perturbations that do not break the reflectional symmetry about the equator.

4. Numerical methods.

4.1. Discretization.

Because it is not possible to find analytic solutions for either the steady solution or the eigenvalue problem, the solutions are approximated numerically. Second order centered finite differencing is used to discretize the spatial derivatives. We approximate the value of the unknown functions at the locations of $N \times N$ uniformly spaced grid points in the interior of the domain. The values of T on the outer boundary, on the equator, and at the pole are not determined by the boundary conditions and must also be considered as unknowns. This leads to discretized solution vectors of size $3N^2 + 3N$. Discretization of the steady equations for u, ξ, and T leads to a system of nonlinear algebraic equations that can be solved by Newton iteration and Keller continuation (as explained in section 4.2) to find an approximation of the steady solution.

For the numerical approximation of the eigenvalues, the linearized perturbation equations are discretized, and thus the values of the steady solution are needed only at specific locations (the grid points) and the computed approximations are used. That is, the linearization is made about the approximate solution. Thus, upon discretization, the partial differential eigenvalue problem becomes a generalized matrix eigenvalue problem.

4.2. Solution techniques. We are interested in computing the steady solution for a wide range of parameter values. To do this, we implement pseudoarclength continuation with the Keller correction condition (see, e.g., [10]) and use a Newton method to solve the resulting equations. If a solution is known for a particular set of parameter values, then this method can be used effectively to follow solutions as a parameter is varied, i.e., to find a solution curve (with respect to the parameter).

Here, we know that for $\Delta T = 0$ the trivial solution satisfies the equations for u, ξ, and T. Thus, for ΔT small, the trivial solution is a reasonable prediction of the solution, and Newton's method is used for the correction. In pseudoarclength continuation, the parameter is considered as an unknown, and initial guesses of the solution are found by following the tangent, or a secant line approximation, to the solution curve. Increments are made approximately along the solution curve, and not by incrementing the parameter. The Keller condition ensures that the corrections to the initial guesses occur approximately perpendicularly to the tangent. This method is particularly useful because it is able to compute solutions along the solution curve even when there is a limit point on the curve, i.e., when the solution curve turns back on itself. In practice, the evaluation of the Jacobian is expensive, and therefore, in order to reduce the number of Jacobian evaluations, we use a quasi-Newton method in which the Jacobian is not updated on each iteration.

The generalized matrix eigenvalue problem that results from the discretization of (3.5) is solved in MATLAB using the implicitly restarted Arnoldi method [18], which is a memory-efficient iterative method for finding a specified number of the largest eigenvalues. A generalized Cayley transformation [10] is made so that the Arnoldi iteration finds the eigenvalues of interest. In particular, the generalized Cayley transformation

$$\mathbf{C}(\mathbf{L}, \mathbf{A}) = (\mathbf{L} - \alpha_1 \mathbf{A})^{-1} (\mathbf{L} - \alpha_2 \mathbf{A}) \tag{4.1}$$

maps eigenvalues λ of the generalized matrix eigenvalue problem $\lambda \mathbf{A} v = \mathbf{L} v$ to eigenvalues σ of the transformed matrix $\mathbf{C}(\mathbf{L}, \mathbf{A})$ such that the eigenvalues λ with $\text{Real}(\lambda) > (\alpha_1 + \alpha_2)/2$ are mapped to the eigenvalues σ with $|\sigma| > 1$, where α_1 and α_2 are the real parameters of the Cayley transformation. The parameters of the transformation can be chosen to improve convergence properties. The matrix $\mathbf{C}(\mathbf{L}, \mathbf{A})$ does not have to be formed explicitly, because the Arnoldi iteration only requires matrix-vector products involving $\mathbf{C}(\mathbf{L}, \mathbf{A})$ [18]. Thus, the full sparseness properties of $\mathbf{L}$ and $\mathbf{A}$ can be exploited, and computer memory requirements can be reduced.

5. Existence and stability results. The specific values of the parameters that are used in the computations are listed in Table 1. While these parameters remain fixed, we vary the gap width R (and thus $r_b \equiv r_a + R$), differential heating ΔT, and the rate of rotation Ω, and compute the azimuthal (or zonal) fluid velocity u, the stream function ξ, and temperature deviation T.

For $\Delta T = 0$, there is no movement of the fluid. However, for all $\Delta T > 0$ fluid motion is induced. For small values of $\Delta T > 0$ a single convection cell develops, which we call a Hadley cell. An example of such a solution for gap width $R = 3.4$ and rotation rate $\Omega = 0.01$ is plotted in Figure 2. In the figure, the stream function $\xi = \xi(r, \theta)$, the azimuthal (zonal) velocity $u = u(r, \theta)$, and the temperature deviation from the temperature on the inner

Table 1

The parameters of the spherical shell and fluid used in the computations. See section 2 for definitions of the symbols.

r_a	10	cm
ν	$1.01e^{-2}$	$\mathrm{cm}^2/\mathrm{sec}$
κ	$1.41e^{-3}$	$\mathrm{cm}^2/\mathrm{sec}$
α	$2.06e^{-4}$	$1/^\circ$ C
ρ_0	0.998	$\mathrm{gm}/\mathrm{cm}^3$
T_r	20.0	$^\circ$C
g	980	$\mathrm{cm}/\mathrm{sec}^2$

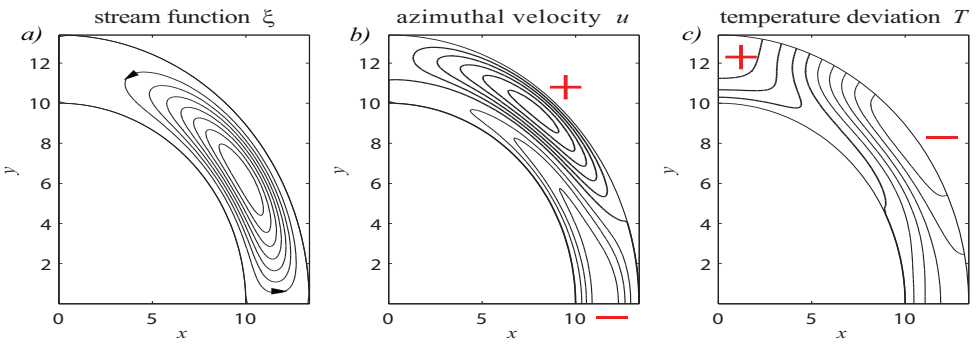


Figure 2. *An example of a single-cell circulation pattern observed for heating parameter $\Delta T = 0.0016$, gap width $R = 3.4$, and rotation rate $\Omega = 0.01$.* (a) *The stream function ξ—the flow tends to follow the contours;* (b) *the azimuthal (or zonal) velocity u; and* (c) *the temperature deviation T from the temperature prescribed on the lower boundary.*

boundary $T = T(r, \theta)$ are plotted on the unscaled domain $r_a \le r \le r_b$, $0 \le \theta \le \pi/2$. The figure represents a cross section of the solution at an arbitrary value of the azimuthal variable φ. The solution corresponding to the full equations (2.2)–(2.4) is obtained by rotating about the polar axis and reflecting across the equator. The "+" and "−" indicate the contours corresponding to positive and negative values of the functions, respectively. Contours of the stream function ξ and the azimuthal velocity that intersect the inner or outer boundary necessarily correspond to zeros of the function, while contours of the temperature deviation T that intersect the inner boundary correspond to zeros. The polar velocity v and the radial velocity w can be found from the stream function ξ using (3.1). In particular, the component of $\mathbf{u}$ that lies within a meridional plane is tangential to the contours represented in Figure 2(a), and thus the flow tends to follow these contours. More specifically, streamlines of the flow are restricted to lie on isosurfaces of ξ. The arrows in Figure 2(a) indicate the direction of the flow along the contours. In particular, in this flow, the fluid rises at the equator and falls at the pole. For this figure and all others that follow, we have taken $N = 40$.

The eigenvalues with the ten largest real parts associated with the axisymmetric eigenfunctions, as well as those associated with the eigenfunctions for each wave number m between 1 and 8, are approximated using the techniques described in section 3. It is found that all eigenvalues that are computed have negative real part. For the larger values of the wave

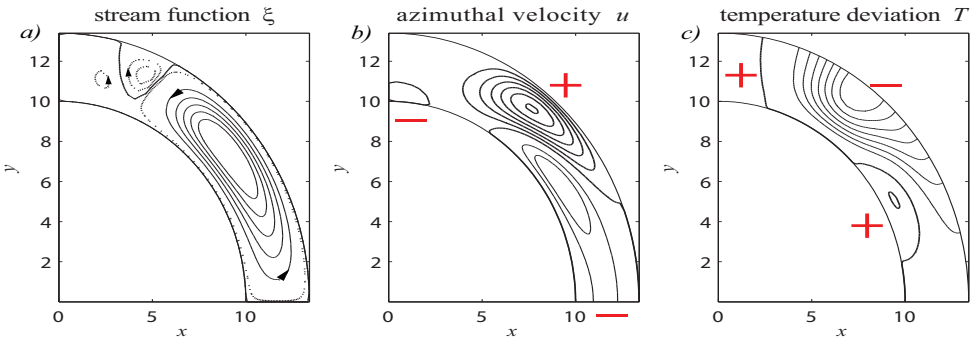


Figure 3. *An example of a three-cell circulation pattern observed for heating parameter* $\Delta T = 0.0036$*, gap width* $R = 3.4$*, and rotation rate* $\Omega = 0.01$*.* (a) *The stream function* ξ*—the flow tends to follow the contours;* (b) *the azimuthal (or zonal) velocity* u*; and* (c) *the temperature deviation* T *from the temperature prescribed on the lower boundary. In* (a)*, the dashed lines represent contours at* $1/20$ *of the interval of the solid contours.*

number m, the eigenvalues become more negative as the wave number is increased. Thus we expect that all eigenvalues associated with all wave numbers have negative real parts, and we conclude that the circulation pattern in Figure 2 is a linearly stable solution of the full three-dimensional equations.

If the heating parameter ΔT is increased further while keeping the gap width R fixed, a transition is observed. First, the flow passes through an intermediate stage, in which the stream function near the pole flattens. Then as ΔT is increased further, a three-cell pattern develops (see Figure 3). The three-cell pattern resembles the zonally (azimuthally) averaged circulation pattern observed in Earth's atmosphere, with a strong cell close to the equator (the Hadley cell), a weaker counter-rotating cell in the midlatitude (sometimes called the Ferrel cell), and finally an even weaker cell near the pole with the same direction of rotation as the Hadley cell [28]. Distinct differences between this pattern and that observed in Earth's atmosphere is that the equatorial cell extends to higher latitude than the Hadley cell of the atmosphere, and the middle cell does not extend to the inner sphere as does the corresponding cell of the atmosphere. However, as the differential heating ΔT is increased further, the middle cell does extend to the inner surface. The azimuthal velocity u is also similar to the azimuthally averaged azimuthal velocity observed in Earth's atmosphere, except that in the atmosphere the jet stream occurs at a somewhat lower latitude, and the negative velocity near the surface does not extend as far from the equator [28]. It is found that this solution is linearly stable to all axisymmetric and nonaxisymmetric perturbations considered, although stability to nonaxisymmetric perturbations is lost as the differential heating ΔT is increased.

It is of particular interest that, although there is clearly a transition in the flow pattern as ΔT is increased, there is no point at which the solution is neutrally stable; i.e., no eigenvalue crosses the imaginary axis. This is not entirely unexpected. Other studies in which flow transitions in systems with a lack of symmetry were investigated have revealed such behavior; for example, see [7, 23, 15, 29, 27]. Although such transitions have been observed in the absence of a corresponding nearby bifurcation (e.g., [7, 23]), they may be induced by a broken pitchfork bifurcation or a perturbed hysteresis bifurcation (see below). Therefore, we search for such a mechanism.

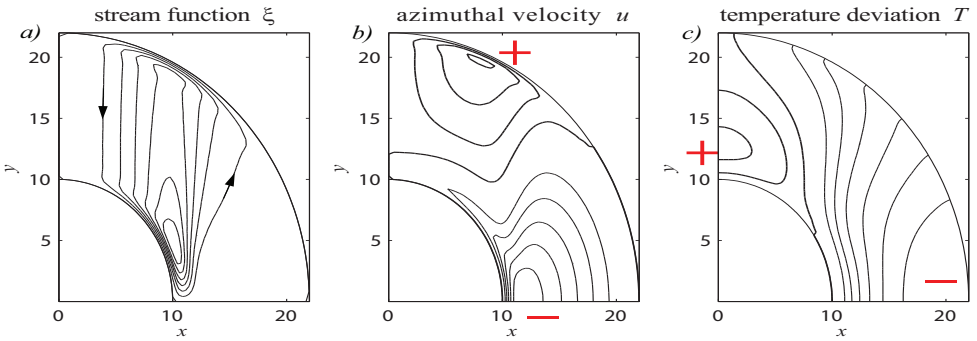


Figure 4. *An example of a single-cell circulation pattern observed for heating parameter* $\Delta T = 0.014$*, gap width* $R = 12$*, and rotation rate* $\Omega = 0.1$*.* (a) *The stream function* ξ*—the flow tends to follow the contours;* (b) *the azimuthal (or zonal) velocity* u*; and* (c) *the temperature deviation* T *from the temperature prescribed on the lower boundary.*

Although our full system has $SO(2) \times \mathbb{Z}_2$ symmetry, we are looking for solutions that preserve this symmetry, so we have used the symmetry to simplify the equations for u, ξ, and T. The simplified equations possess none of these symmetries. The symmetries of a system determine the types of bifurcation that are likely to occur, i.e., which types are generic for the system. For equations with no symmetry, saddle-node and Hopf bifurcations are generic and are expected to be observed if a real eigenvalue crosses through zero or a complex pair of eigenvalues crosses the imaginary axis, respectively, as a single parameter is varied.

The presence of a saddle-node bifurcation in itself cannot explain the observed transition, which occurs without a zero eigenvalue. However, such a transition may occur near an organizing center or bifurcation of codimension 2. This implies that it will be necessary to vary a second parameter to find it. The types of codimension-2 bifurcation points that could occur generically in a system like ours are the broken pitchfork bifurcation [9, 15, 29] and the hysteresis (or cusp) bifurcation [27] as described above and in [9, 16].

In order to clarify the origin of the observed transition, it is useful to explore the solutions at larger gap width and rate of rotation. If the gap width is increased to $R = 12$ and the rotation rate to $\Omega = 0.1$, then, when the heating ΔT is small enough, the linearly stable one-cell pattern is maintained, with a slight distortion of the stream function near the outer boundary and an increase in the retrograde velocity near the equator. An example is plotted in Figure 4. As the heating parameter ΔT is increased, there is a transition to a two-cell pattern (Figure 5), and again this transition occurs without an eigenvalue corresponding to an axisymmetric eigenfunction crossing the imaginary axis.

It is observed that as the gap width R and the rotation rate Ω are increased, the steady solutions become less stable to nonaxisymmetric perturbations. As a result, although this two-cell pattern is stable to axisymmetric perturbations, it is linearly unstable to nonaxisymmetric perturbations. The loss of stability to the nonaxisymmetric perturbations occurs as a Hopf bifurcation, and thus the bifurcating solution is expected to be a rotating wave with azimuthal wave number m. This implies that this two-cell solution will not be physically observable directly. However, because the loss of stability occurs near the one-cell to two-cell transition, it is possible that the rotating wave will inherit the θ and r dependence of the two-cell solution;

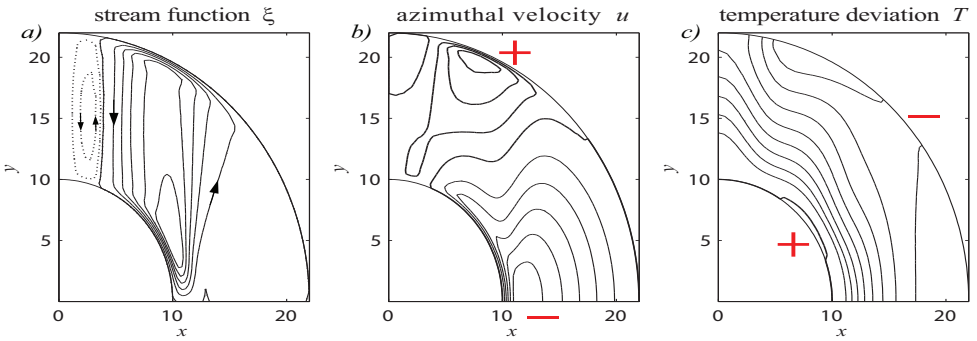


Figure 5. *An example of a two-cell circulation pattern observed for heating parameter $\Delta T = 0.016$, gap width $R = 12$, and rotation rate $\Omega = 0.1$.* (a) *The stream function ξ—the flow tends to follow the contours;* (b) *the azimuthal (or zonal) velocity u; and* (c) *the temperature deviation T from the temperature prescribed on the lower boundary. In* (a)*, the dashed lines represent contours at* 1/5 *of the value of the solid contours.*

i.e., an azimuthal average of the rotating wave will have a two-cell structure. We do not show this here, as it is secondary to our purpose for presenting this example. Regardless of the stability with respect to nonaxisymmetric perturbations, the two-cell pattern is a linearly stable solution of the axisymmetric equations, and it forms in qualitatively the same manner as the transition that is observed at lower R and Ω (see below for further discussion). Thus, an investigation of the transition to this flow, in the axisymmetric equations, will provide a mechanism for the transition. Below we consider solutions and stability only with respect to the axisymmetric equations.

As mentioned above, the transition at $R = 12$ and $\Omega = 0.1$ occurs without an eigenvalue corresponding to an axisymmetric eigenfunction crossing the imaginary axis. Because the eigenvalue with largest real part is real, this implies that there is no parameter value at which there is a zero eigenvalue. In Figure 6, we plot the real value of the (axisymmetric) eigenvalue with largest real part as a function of the heating parameter ΔT, and we plot a "bifurcation diagram" that indicates how the solutions change as ΔT changes, where the vertical axis is the L^2-norm of the stream function ξ. In the figure it can be seen that as the heating parameter ΔT increases, the value of the eigenvalue increases until it reaches a maximum that is negative, at which point it begins to decrease without ever reaching zero. Heuristically, it is observed that the development of the two-cell pattern begins to occur near this maximum.

If the gap width is increased to $R = 16.2$, then for small values of ΔT a linearly stable one-cell pattern is again observed; see Figure 7. A gap width of $R = 16.2$ corresponds to an aspect ratio $\eta = R/r_a = 1.35$. Although we are not necessarily interested in flows at large aspect ratio, they will help to explain the transition that is observed for smaller gap width.

As we increase the differential heating ΔT, again we see a transition from the one-cell to a two-cell pattern, shown in Figure 8. However, in this case, a real eigenvalue does cross the imaginary axis. In Figure 9, we plot both the real part of the eigenvalue with largest real part, as a function of ΔT, and the corresponding bifurcation diagram. The crossing of the imaginary axis by a real eigenvalue corresponds to a saddle-node bifurcation, also referred to as a limit point or fold. As the solution curve is followed past the bifurcation point, the solution becomes linearly unstable, and the value of ΔT begins to decrease. For a short interval,

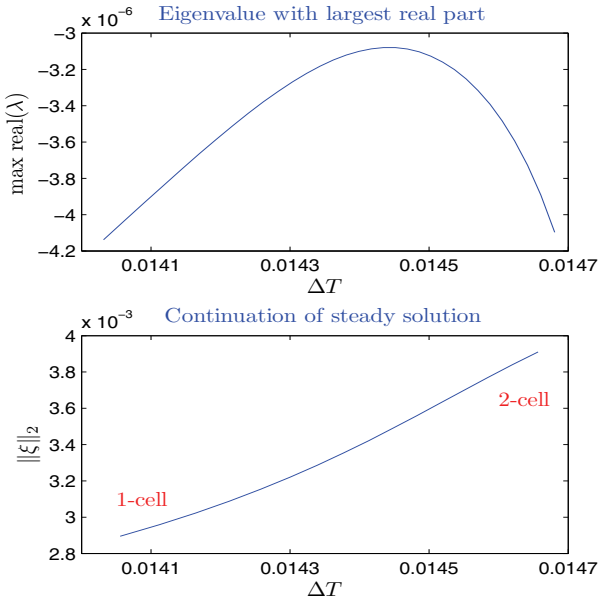


Figure 6. *Results for $R = 12$, $\Omega = 0.1$. (top) Real part of eigenvalue with largest real part versus ΔT, and (bottom) bifurcation diagram in ΔT; the vertical axis represents the L^2-norm of the stream function ξ.*

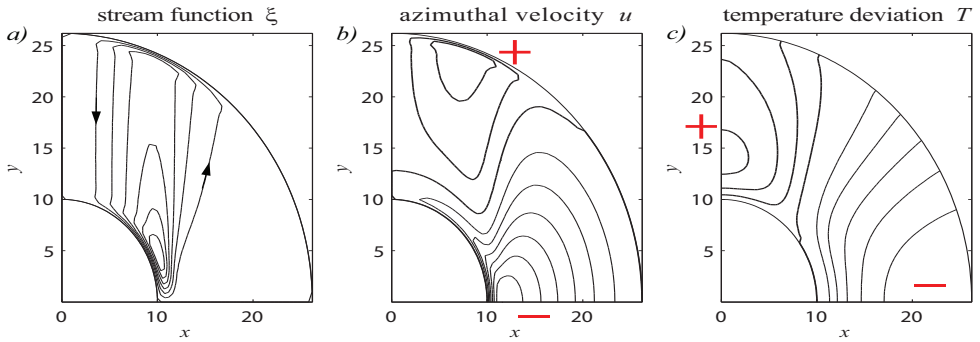


Figure 7. *An example of a single-cell circulation pattern observed for heating parameter $\Delta T = 0.013$, gap width $R = 16.2$, and rotation rate $\Omega = 0.1$.* (a) *The stream function ξ—the flow tends to follow the contours;* (b) *the azimuthal (or zonal) velocity u; and* (c) *the temperature deviation T from the temperature prescribed on the lower boundary.*

the real eigenvalue increases until it reaches a maximum, after which it begins to decrease. Subsequently, it again crosses the imaginary axis, and a second saddle-node bifurcation is observed. As we follow the curve past this point, the solution returns to being linearly stable, and ΔT once again begins to increase. This pair of saddle-node bifurcations results in an S-shaped bifurcation diagram. As we follow the solution curve, from the lower part of the S to the upper part of the S, the real part of the eigenvalue with largest real part traces out the loop seen in Figure 9. This form of solution curve results in a classical mechanism for hysteresis, as seen in the hysteresis loop shown in Figure 1.

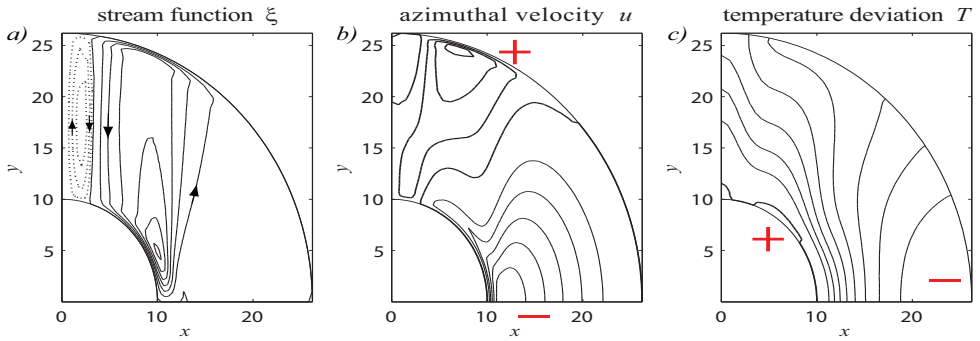


Figure 8. *An example of a two-cell circulation pattern observed for heating parameter* $\Delta T = 0.015$*, gap width* $R = 16.2$*, and rotation rate* $\Omega = 0.1$*.* (a) *The stream function* ξ*—the flow tends to follow the contours;* (b) *the azimuthal (or zonal) velocity* u*; and* (c) *the temperature deviation* T *from the temperature prescribed on the lower boundary. In* (a)*, the dashed lines represent contours at* $1/5$ *of the value of the solid contours.*

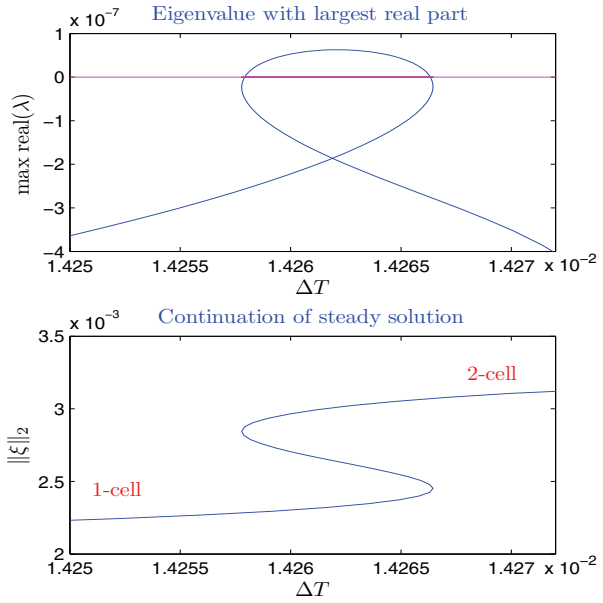


Figure 9. *Results for* $R = 16.2$*. (top) Real part of eigenvalue with largest real part versus* ΔT*, and (bottom) bifurcation diagram in* ΔT*; the vertical axis represents the* L^2*-norm of the stream function* ξ*.*

6. Hysteresis bifurcation. The behavior observed in Figures 6 and 9 is indicative of a hysteresis (or cusp) bifurcation point. The typical behavior of a nonlinear system near a hysteresis point is as shown in Figure 1. It can be expected that at some critical value of the gap width $R = R_c$ a one-dimensional bifurcation diagram in ΔT will have a single vertical tangent, with slopes on either side being positive. For values of $R < R_c$ we expect no bifurcations as ΔT is increased (as in Figure 6), while for $R > R_c$ we expect a pair of saddle-node bifurcations (as in Figure 9). This behavior can be seen in Figure 1, where these different one-dimensional bifurcation diagrams in ΔT may be obtained as slices through M, taken with different values of constant R. Mathematically, a hysteresis or cusp bifurcation

point is determined by the following three conditions [16]:

1. There exists a steady solution.
2. There is a simple zero eigenvalue of the linearization of the equations about the steady solution.
3. The coefficient of the second order term of the normal form equations on the center manifold vanishes.

Therefore, in order to demonstrate that the observed behavior is indeed generated by a hysteresis bifurcation, it is necessary to show that each of the above three conditions is satisfied. Although it may seem that this would be a daunting task, in fact these conditions can be verified by an explicit calculation that can be performed numerically as follows; see [10, 16]. To elucidate these conditions and the means that we use to compute the hysteresis point, we write the equations for ξ, u, and T in the abstract form:

$$\dot{U} = LU + N(U,U), \tag{6.1}$$

where

$$U = \begin{pmatrix} \xi \\ u \\ T \end{pmatrix}$$

is the dependent variable, L is the linear operator such that LU is the linear part of the equations, $N(U,U)$ is the nonlinear part of the equations, and the dot represents differentiation with respect to time. Note that the nonlinear part N has only quadratic terms (from the Navier–Stokes equation), and thus we may write it in the bilinear form $N(U,U)$.

Assume that for some critical values of the parameters $(\Delta T, R) = (\Delta T_c, R_c)$ there is a steady solution U_0 of (6.1); i.e., U_0 satisfies

$$LU_0 + N(U_0,U_0) = 0. \tag{6.2}$$

Assume also that at $(\Delta T, R) = (\Delta T_c, R_c)$ the linearization L_0 about the steady solution U_0 has a simple zero eigenvalue $\lambda_0 = 0$, while all other eigenvalues have negative real part, where L_0 is given by

$$L_0 V = LV + N(V,U_0) + N(U_0,V). \tag{6.3}$$

That is, we have

$$L_0 \Psi = 0, \tag{6.4}$$

where Ψ is the eigenfunction corresponding to the zero eigenvalue.

Under certain conditions on L_0, the dependent variable U can be written in the form

$$U = w\Psi + \Phi, \tag{6.5}$$

where $w \in \mathbb{R}$ and thus $w\Psi \in \operatorname{span}\{\Psi\}$, and $\Phi \in E_s$. Here E_s is called the stable subspace and is the space spanned by all eigenfunctions corresponding to eigenvalues with negative real part.

If we write U as in (6.5), then under certain technical conditions a center manifold and normal form reduction can be performed on (6.1) to obtain the equation on the center manifold in normal form

$$\dot{w} = \beta_1 + \beta_2 w + a w^2 + c w^3 + O(w^4), \tag{6.6}$$

where a and c are coefficients of the normal form and β_1 and β_2 are unfolding parameters that are in general functions of the parameters ΔT and R. It can be shown that if $c \neq 0$, then neglecting the terms of $O(w^4)$ does not change the qualitative features of the solutions.

The center manifold and normal form theories state that for $(\Delta T, R)$ near $(\Delta T_c, R_c)$ and when the solutions are in some sense small, the dynamics of (6.1) can be deduced from (6.6). In particular, solutions of (6.6) are in one-to-one correspondence with those of (6.1).

Formulas for the coefficients of the normal form equation can be derived by performing a center manifold and normal form reduction in the general case [10, 16]. In particular, the coefficient of the second order term is given by

$$a = 1/2 \left\langle \Psi^*, N(\Psi, \Psi) \right\rangle, \tag{6.7}$$

where Ψ is the eigenfunction corresponding to λ_0, Ψ^* is the corresponding adjoint eigenfunction corresponding to λ_0, and

$$\langle U, V \rangle = \iint U \cdot V d\mathbf{r} \tag{6.8}$$

is the inner product on the domain.

For $a = 0$, a hysteresis bifurcation occurs when $\beta_1 = \beta_2 = 0$. For $\beta_2/c > 0$, there is a single solution to (6.6) for all β_1. For $\beta_2/c < 0$, there is a region in the two-dimensional parameter space (β_1, β_2) in which there are three solutions. The borders of this region are given asymptotically by the two curves

$$\beta_1 = \pm \frac{2}{3} \sqrt{\frac{-\beta_2}{c}} \beta_2. \tag{6.9}$$

As β_2 approaches zero, these two curves approach each other and meet in a cusp at $\beta_1 = \beta_2 = 0$. This is the origin of the name "cusp bifurcation."

In order to show that a hysteresis or cusp bifurcation occurs in the model, we need to show that the three aforementioned conditions are satisfied. That is, we must find parameter values $(\Delta T, R) = (\Delta T_c, R_c)$ such that the following three equations are satisfied:

$$L U_0 + N(U_0, U_0) = 0, \tag{6.10}$$

$$L_0 V = 0, \qquad \langle V, V \rangle = 1, \tag{6.11}$$

$$a = 1/2 \left\langle \Psi^*, N(\Psi, \Psi) \right\rangle = 0, \tag{6.12}$$

where L_0 is given by (6.3).

These equations have the unfortunate property that for some values of $\Delta T \neq \Delta T_c$, $R \neq R_c$, L_0 will not be singular, and thus (6.11) will not have a solution for any V. Therefore, it will

be convenient to use the following defining system [10]:

$$LU_0 + N(U_0, U_0) = 0, \tag{6.13}$$

$$g = 0, \tag{6.14}$$

$$g' = 0, \tag{6.15}$$

where g and g' are scalars given by

$$L_0 V + gB = 0, \qquad \langle C, V\rangle = 1, \tag{6.16}$$

$$L_0 V' + g'B = -N(V, V), \qquad \langle C, V'\rangle = 0, \tag{6.17}$$

and where B is not in the range of L_0, and C is not in the range of the adjoint operator L_0^*, which is defined by

$$\langle L_0 U, V\rangle = \langle U, L_0^* V\rangle$$

for all U and V.

Each of the three equations (6.13)–(6.15) corresponds to one of the hysteresis point defining conditions. Specifically, U_0 is a steady solution of (6.1) when (6.13) is satisfied. If we set $g = 0$ in (6.16) and there is a solution, then L_0 has a zero eigenvalue with corresponding eigenfunction V. Thus, the second hysteresis condition is satisfied when (6.14) is satisfied. In this case, when L_0 is not singular, there will still be a solution of (6.16), namely one for which $g \neq 0$. This is assured by choosing B not in the range of L_0. If we set $g' = 0$ in (6.17), and there is a solution, then we have that $N(V, V)$ must be in the range of L_0. If we also have $g = 0$, then from above we have that V is the eigenfunction of L_0 with zero eigenvalue; i.e., $V = \Psi$. Thus, if $N(V, V)$ is in the range of L_0, then, by the "Fredholm alternative" property of the adjoint, $N(V, V)$ must be orthogonal to the eigenfunction of L_0^* corresponding to the zero eigenvalue; i.e., we must have $\langle \Psi^*, N(\Psi, \Psi)\rangle = 0$, and thus the second order coefficient of the normal form vanishes. When L_0 is nonsingular, (6.17) has a solution regardless of g'. However, there is also the possibility that close to the cusp there are parameter values such that L_0 is singular, while $N(V, V)$ is not in the range of L_0; this occurs at the saddle-node bifurcations. Thus, there will be no solution of (6.17) for $g' = 0$. However, because B is not in the range of L_0, we are assured a solution with $g' \neq 0$.

In order to make the equations linear in V and V', we can choose the normalization conditions as shown. Solutions are assured when C is chosen to be not in the range of the adjoint operator L_0^*. In practice, because the kernel of L_0 is only one-dimensional, it is easy to choose B and C with the required properties.

It can be proved that when the system (6.13)–(6.15) has a solution, then not only is the nondegeneracy condition for a cusp bifurcation satisfied, but also the transversality condition [10]. Thus, we are assured that we have found a cusp bifurcation.

Upon discretization of (6.13)–(6.15) and (6.16)–(6.17) on an $N \times N$ grid, we obtain a system of nonlinear algebraic equations. For various values of N, we find that there are critical parameter values $(\Delta T, R) = (\Delta T_c, R_c)$ such that there are a U_0, g, and g' that satisfy the discretized system. Results are listed in Table 2. Although for small N it appears that we are not in the asymptotic range, the results for a higher value of N provide evidence of

Table 2

Critical parameter values at which the hysteresis bifurcation occurs, for various values of N. The results provide evidence of convergence.

N	ΔT	R
40	0.0143	16.0
80	0.0155	24.2
120	0.0163	27.8
160	0.0165	28.0

convergence. The large variation in R between the results at $N = 40$ and $N = 80$ is possibly caused by the nonlinear dependence of the equations on R; see (2.9)–(2.13). In particular, for large R, we expect that a large change in R is required to produce even a small change in behavior of the solutions. This could also be an indication that the boundary layers are not resolved sufficiently at low resolution to put us in the asymptotic range of the convergence. Regardless, the results provide evidence that a hysteresis bifurcation does in fact exist in the model with cusp point $(\Delta T_c, R_c) \approx (0.017, 28)$, and thus the lower resolution captures the correct qualitative behavior.

There is strong evidence that the transition from the one-cell to the two-cell pattern, which is observed at gap width $R = 12$ as the differential heating ΔT is increased, is associated with this cusp bifurcation. We also postulate that the transition from the one-cell to the three-cell pattern, which is observed at gap width $R = 3.4$ as the differential heating ΔT is increased, not only is the same mechanism, but is associated with the same bifurcation. This is evident if we consider the transition from the one-cell pattern for a sequence of gap widths R and rotation rates Ω that decrease from $R = 12$, $\Omega = 0.1$ to $R = 3.4$, $\Omega = 0.01$. For all transitions in this sequence, regardless of whether the transition results in a two-cell or three-cell pattern, the eigenvalue with largest real part behaves in the same manner (as shown in Figure 6) with only small quantitative changes. Furthermore, in all cases, the stream function component of the eigenfunctions corresponding to the eigenvalue with largest real part has a two-cell structure. The azimuthal velocity component of the eigenfunctions also shows little variation. There is nothing to indicate that there is another bifurcation that is taking place. That is, all solutions discussed above lie on a single solution manifold that is folded at the cusp point; see the manifold M in Figure 1. In [27], such a manifold is shown to connect qualitatively different solutions in a model of rotating convection in a cylinder with centrifugal buoyancy.

Furthermore, there is a smooth variation of the qualitative behavior as R is varied. For large R, the transition from a one-cell to two-cell pattern begins with a flattening of the stream function ξ near the pole, corresponding to a decrease in fluid velocity in this region, and is followed by the formation of the counter-rotating cell near the pole. The new cell is first observed as a small cell adjacent to the pole and grows as the differential heating ΔT is increased. Changes in the rotation rate Ω affect the stability of the solution to nonaxisymmetric perturbations but do not affect the qualitative features of the transition, and therefore, here, we refer only to changes in gap width R. If R is decreased, the flattening becomes more pronounced before the second cell is observed. When the cell is observed, it grows more quickly with ΔT than when the gap width is larger. If the differential heating is increased sufficiently, a transition from the two-cell pattern to a three-cell pattern is observed. For a

yet smaller gap width (e.g., $R = 3.4$), again the transition begins with a flattening. However, in this case, the second cell does not first appear adjacent to the pole but peeks out a small distance from the pole. This transitional stage is not in essence a two-cell pattern, because between the new cell and the pole there is a very weak (almost quiescent) region in which the fluid rotates in the same sense as the large cell near the equator.

The development of the two-cell pattern is easily explained in terms of the lowest order dynamics of the cusp bifurcation. To lowest order, the solution U to the perturbation equations will be given by $w\Psi$, where $w \in \mathbb{R}$ and Ψ is the eigenfunction corresponding to the eigenvalue with largest real part; see (6.5). Thus, to first order, the solution to the axisymmetric equations will be $w\Psi$ plus the solution about which we have linearized, i.e., the one-cell solution. Thus, because the stream function component of the eigenfunction Ψ has a two-cell structure, we expect that, to lowest order, the solution will also develop a two-cell structure. The development of the three-cell pattern in this context is not as easily explained. However, it may be possible that insight could be gained from a higher order computation.

We have already pointed out that the variation of the rotation rate Ω has relatively smaller qualitative effect on the solutions of the axisymmetric equations. Indeed, a cusp is observed if the rotation is held fixed at $\Omega = 0.01$ and only the gap width R is increased. We choose not to present this example because, in this case, an additional saddle-node bifurcation occurs at values of the differential heating slightly larger than where the cusp is observed. As the rotation rate is increased, this bifurcation moves further away from the cusp, and thus the example for $\Omega = 0.1$ more clearly indicates the origin of the observed transitions.

7. Discussion and conclusions. This work has shown that a Boussinesq fluid in a rotating spherical shell, differentially heated on its inner surface, can exhibit a variety of stable rotationally symmetric flow patterns. Distinctive features of these flow patterns include a Hadley cell with a flow pattern much like the Hadley cells of Earth, and a high azimuthal velocity jet stream located at high altitudes and midlatitudes, much like the jet stream in each hemisphere of Earth. For small values of the differential heating parameter $\Delta T > 0$, the Hadley cell exists and extends from equator to pole. For larger values of ΔT, first one then two or more additional convection cells may form between the Hadley cell and the pole in the mid- and polar latitudes. The first cell to form after the Hadley cell exhibits counter-rotating flow (westerly winds near the inner surface), and the third cell has Hadley-like rotation (easterly winds near the polar surface). The observed transitions are distinctly related to the spherical geometry of the system. This is evident because, in all models of differentially heated fluids in domains with cylindrical geometry, such transitions that do not break the rotational symmetry occur only at very high differential heating.

The features of the transitions are not affected by moderate changes in the rate of rotation Ω, although the solutions become less stable to nonaxisymmetric perturbations as Ω is increased. However, changes in the gap width R can induce significant changes. Mathematical analysis of the axisymmetric model demonstrates that it possesses a codimension-2 hysteresis (or cusp) bifurcation for a critical choice of the parameters $(R, \Delta T) = (R_c, \Delta T_c)$. In a neighborhood of this hysteresis point, for larger R, there exists a region of bistability in which two different states of the system are both linearly stable solutions of the axisymmetric equations. These two stable states are separated by a third unstable state. For such R (fixed), there exists an interval of values of ΔT exhibiting a hysteresis loop, as illustrated in Figure 1.

At each end of this interval, a small change in ΔT can cause a transition in the state of the system, to a qualitatively different flow pattern, e.g., one with a different number of convection cells.

The results of this paper lead to more questions than answers and will form the basis of much future work. In the model, changes in the Hadley cell have no influence on the temperature difference parameter ΔT. This is not the case in a real planetary system, where convection cells are known to act as "thermal conveyor belts." In the case of small ΔT and a large Hadley cell extending from equator to pole, this conveyor belt would have the effect of warming the polar region and cooling the equatorial region, thus *reducing* the temperature difference ΔT. This can be seen in the temperature deviation plots of Figures 2 and 4, in which the gradient of the temperature deviation is essentially opposite to the imposed differential heating. Therefore, the thermal conveyor belt function of a large Hadley cell enhances its persistence. On the other hand, if ΔT increases (for some other reason) to a value where the single Hadley cell is replaced by multiple cells, this would curtail the thermal conveyor belt acting from equator to pole. As a result, the polar regions would cool relative to the equatorial regions and ΔT would increase, pushing the system further to the right along the bifurcation curve. We conjecture that this feedback mechanism, from the convection cell flow back to the temperature difference ΔT, is present in real planetary atmospheres and implies a modification of the predictions of our model. The effect is most easily stated with reference to Figure 1: the interval of bistability in ΔT would lengthen and the cusp point would move to smaller values of R. In other words, the net effect of this thermal feedback would be to increase both the likelihood and the amplitude of the hysteresis behavior demonstrated in the model.

In addition, the model should be reconsidered to take into account the fact that the atmosphere of the earth is a strongly stratified compressible fluid, with properties very different from water. The vertical motion of a strongly stratified fluid is inhibited, thus causing an elongation (in θ) of the cells. The cells in a Boussinesq fluid typically have an aspect ratio close to 1, which implies that R must be rather large in order to see only three cells. In a strongly stratified fluid, a much smaller R (for a similar ΔT) would be sufficient to see a similar number of cells. Therefore, we predict that the phenomena exhibited in this model could exist in the atmosphere of a planet such as Earth, but for much smaller aspect ratios R/r_a. Thus, both the thermal feedback and the stratified fluid properties of real atmospheres, which have been neglected in this simple model, can be expected to amplify rather than inhibit the hysteresis mechanism demonstrated for the model.

In further work the model could be extended to physical dimensions on the scale of a planet such as Earth. Another future goal is an analysis of the nonaxisymmetric bifurcations that would lead to rotating waves, as well as analyses of spherical shell models that break the north-south reflectional symmetry or satisfy different boundary conditions.

Acknowledgments. The authors would like to thank Wayne Nagata, Martin Golubitsky, and Laurette Tuckerman for helpful discussions, and would like to thank the referees for their helpful comments.

REFERENCES

[1] D. J. Acheson, *Elementary Fluid Dynamics*, Oxf. Appl. Math. Comput. Sci., Clarendon Press, Oxford, UK, 1990.

[2] C. D. Andereck and F. Hayot, *Ordered and Turbulent Patterns in Taylor–Couette Flow*, NATO Adv. Sci. Inst. Ser. B Phys. 297, Plenum, New York, 1992.

[3] P. Beltrame, V. Travnikov, M. Gellert, and C. Egbers, *GEOFLOW: Simulation of convection in a spherical shell under central force field*, Nonlinear Process. Geophys., 13 (2006), pp. 413–423.

[4] S. Chandrasekhar, *Hydrodynamic and Hydromagnetic Stability*, 2nd ed., Dover Publications, New York, 1981 (first published 1961, Oxford University Press).

[5] P. Chossat, *Bifurcation and stability of convective flows in a rotating or not rotating spherical shell*, SIAM J. Appl. Math., 37 (1979), pp. 624–647.

[6] P. Chossat and G. Iooss, *The Couette–Taylor Problem*, Springer-Verlag, New York, 1994.

[7] M. P. Escudier, *Observations of the flow produced in a cylindrical container by a rotating endwall*, Exp. Fluids, 2 (1984), pp. 189–196.

[8] M. Golubitsky and W. F. Langford, *Pattern formation and bistability in flows between counterrotating cylinders*, Phys. D, 32 (1988), pp. 362–392.

[9] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory*, Appl. Math. Sci. 51, Springer-Verlag, New York, 1985.

[10] W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, SIAM, Philadelphia, 2000.

[11] J. E. Hart, G. A. Glatzmaier, and J. Toomre, *Space-laboratory and numerical simulations of thermal convection in a rotating hemispherical shell with radial gravity*, J. Fluid Mech., 173 (1986), pp. 519–544.

[12] R. Hide and P. J. Mason, *Sloping convection in a rotating fluid*, Adv. Geophys., 24 (1975), pp. 47–100.

[13] P. Hignett, A. A. White, R. D. Carter, W. D. N. Jackson, and R. M. Small, *A comparison of laboratory measurements and numerical simulations of baroclinic wave flows in a rotating cylindrical annulus*, Quart. J. Roy. Meteorol. Soc., 111 (1985), pp. 131–154.

[14] R. Hollerbach, *Instabilities of the Stewartson layer Part* 1. *The dependence on the sign of* R_o, J. Fluid Mech., 492 (2003), pp. 289–302.

[15] A. Juel, A. G. Darbyshire, and T. Mullin, *The effect of noise on pitchfork and Hopf bifurcations*, Proc. Roy. Soc. London Ser. A, 453 (1997), pp. 2627–2647.

[16] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 3rd ed., Springer-Verlag, New York, 2004.

[17] W. F. Langford, R. Tagg, E. J. Kostelich, H. L. Swinney, and M. Golubitsky, *Primary instabilities and criticality in flow between counter-rotating cylinders*, Phys. Fluids, 31 (1988), pp. 776–785.

[18] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environ. Tools 6, SIAM, Philadelphia, 1998.

[19] V. Lesueur, A. Abouelainine, A. Mangeney, and P. Drossart, *Geostrophic motions of a Boussinesq fluid in a thick rotating spherical shell*, Geophys. Astrophys. Fluid Dynam., 91 (1999), pp. 1–43.

[20] G. M. Lewis and W. Nagata, *Linear stability analysis for the differentially heated rotating annulus*, Geophys. Astrophys. Fluid Dynam., 98 (2004), pp. 129–152.

[21] G. M. Lewis and W. Nagata, *Double Hopf bifurcations in the differentially heated rotating annulus*, SIAM J. Appl. Math., 63 (2003), pp. 1029–1055.

[22] L. Li, P. Zhang, X. Liao, and K. Zhang, *Multiplicity of nonlinear thermal convection in a spherical shell*, Phys. Rev. E, 71 (2005), paper 016301.

[23] J. M. Lopez, *Axisymmetric vortex breakdown Part* 1. *Confined swirling flow*, J. Fluid Mech., 221 (1990), pp. 533–552.

[24] E. N. Lorenz, *The Nature and Theory of the General Circulation of the Atmosphere*, World Meteorological Society, Geneva, Switzerland, 1967.

[25] P. S. Marcus and L. S. Tuckerman, *Simulation of flow between concentric rotating spheres. Part* 1. *Steady states*, J. Fluid Mech., 185 (1987), pp. 1–30.

[26] P. S. Marcus and L. S. Tuckerman, *Simulation of flow between concentric rotating spheres. Part* 2. *Transitions*, J. Fluid Mech., 185 (1987), pp. 31–65.

[27] F. Marques, I. Mercader, O. Batiste, and J. M. Lopez, *Centrifugal effects in rotating convection: Axisymmetric states and three-dimensional instabilities*, J. Fluid Mech., 580 (2007), pp. 303–318.
[28] J. P. Peixoto and A. H. Oort, *Physics of Climate*, American Institute of Physics, New York, 1992.
[29] A. M. Rucklidge and A. R. Champneys, *Boundary effects and the onset of Taylor vortices*, Phys. D, 191 (2004), pp. 282–296.
[30] G. P. Williams, *Baroclinic annulus waves*, J. Fluid Mech., 49 (1971), pp. 417–449.
[31] T. Yanagisawa and Y. Yamagishi, *Rayley–Bénard convection in a spherical shell with infinite Prandtl number at high Rayleigh number*, J. Earth Sim., 4 (2005), pp. 11–17.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1445–1476
© 2008 Society for Industrial and Applied Mathematics

Stable Synchrony in Globally Coupled Integrate-and-Fire Oscillators*

Yu-Chuan Chang† and Jonq Juang†

Abstract. A model of integrate-and-fire oscillators is studied. In the special case of identical oscillators, the model was first proposed and analyzed by Mirollo and Strogatz [*SIAM J. Appl. Math.*, 50 (1990), pp. 1645–1662]. We assume, as in Mirollo and Strogatz's model, that each oscillator x_i evolves according to a map f_i. Our main results are to demonstrate that the concavity structure of f_i plays an important role in determining whether Peskin's second conjecture holds true. Specifically, the following statements are proved. First, the system of convex oscillators (i.e., $f_i'' < 0$ for all i), in general, synchronizes when the oscillators are not quite identical. Second, the system of a certain class of concave oscillators (i.e., $f_i'' > 0$ for all i) will not achieve synchrony for initial conditions in a set of positive measure when the oscillators are nearly identical. Third, the system of concave oscillators may achieve synchrony under certain sufficient conditions, provided that the oscillators are not quite nonidentical and that its concavity is small.

Key words. stable synchrony, nonidentical oscillators, integrate-and-fire, concavity

AMS subject classifications. 92A09, 34C15, 58F40

DOI. 10.1137/070709220

1. Introduction.

Large assemblies of oscillator units can spontaneously evolve to a state of large scale organization. Synchronization is the best known phenomenon of this kind, where after some transient regime a coherent oscillatory activity of the set of oscillators emerges. This interesting phenomenon is quite common in many different disciplines such as engineering [62], physics [15, 35, 51], chemistry [36], as well as biology [61]. For example, in southeastern fireflies, thousands of individuals gathered on trees may flash in unison. Other examples of biological oscillators are the rhythmic activity of cells of the heart pacemaker [29, 40, 43, 55], of cells of the pancreas [48, 49], and of neural networks [9, 13, 20, 43, 45, 50]. Synchronization of oscillators has been studied in both phase-coupled models [3, 4, 5, 6, 11, 16, 17, 18, 19, 30, 33, 37, 38, 39, 42, 44, 52, 53, 55, 56, 57, 60, 58, 63], where the interaction between the oscillators is smooth and continuous in time, and pulse-coupled models [1, 7, 10, 12, 23, 24, 25, 27, 28, 31, 32, 36, 41, 46, 47, 57, 59], where the membrane voltage is discontinuously reset to a fixed value once it reaches a certain threshold. It should be noted that pulse-coupled models are of greater relevance for neuroscience applications since synaptic coupling is often spike mediated.

This paper deals with a population of integrate-and-fire oscillators with all-to-all pulse coupling. We begin with describing Peskin's model of n integrate-and-fire oscillators. Let the state of the ith oscillator be denoted by x_i, where x_i are subject to the dynamics $\frac{dx_i}{dt} =$

*Received by the editors November 25, 2007; accepted for publication (in revised form) by D. Terman July 17, 2008; published electronically December 3, 2008.
http://www.siam.org/journals/siads/7-4/70922.html
†Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan, R.O.C. (shamrock.am94g@nctu.edu.tw, jjuang@math.nctu.edu.tw).

$-r_i x_i + I_i$, $0 \le x_i \le 1$, $i = 1, 2, \ldots, n$, with input $I_i > 0$, a normalized threshold 1, and leakiness $r_i \ge 0$. When $x_i = 1$, the ith oscillator fires and x_i jumps back to zero. As a consequence of the firing of the ith oscillator, the activation of any other oscillator j is incremented by the coupling $\omega_{i,j}$. Should no confusion arise, we write $\omega_{i,j}$ as ω_{ij}. This model was later generalized by Mirollo and Strogatz [41]. It was assumed that the state variable x_i evolves according to a map f_i. When x_i reaches the threshold, the oscillator fires and x_i jumps back instantly to zero, and the activation of any other oscillator j is incremented by the positive coupling ω_{ji}. Specifically, x_i evolve according to $x_i = f_i(\phi_i)$, where $f_i : [0,1] \to [0,1]$ is smooth and strictly increasing, i.e., $f_i' > 0$ on $(0,1)$. Here ϕ_i is a phase variable so that (i) $\frac{d\phi_i}{dt} = \frac{1}{T_i}$, where T_i is the cycle period for oscillator x_i when evolving freely; (ii) $\phi_i = 0$ when the oscillator is at its lowest state $x_i = 0$; and (iii) $\phi_i \equiv 1$ at the end of cycle when the oscillator reaches the threshold $x_i = 1$. Therefore, f_i satisfy $f_i(0) = 0$, $f_i(1) = 1$. These maps f_i are to be called evolution maps. The inverses of f_i are to be denoted by g_i. If $f_i \equiv f$, $T_i \equiv T$, and $\omega_{ij} \equiv \omega$ for all i, j, then the corresponding system is called identical. Otherwise, it is called nonidentical. To describe the dynamics of the model, let $\Phi^0 = (\phi_1^0, \phi_2^0, \ldots, \phi_n^0) \in \mathbb{R}^n$ be the initial condition of the oscillators. Here $0 = \phi_1^0 \le \phi_2^0 \le \cdots \le \phi_n^0 < 1$. Further, $\Phi^k = (\phi_{k_1}^k, \phi_{k_2}^k, \ldots, \phi_{k_n}^k)$, where $0 = \phi_{k_1}^k \le \phi_{k_2}^k \le \cdots \le \phi_{k_n}^k < 1$, is the state of n oscillators after the kth firing. Denote by $V_k(\Phi^0)$ the set of the indexes of oscillators reaching threshold simultaneously and thus firing the kth time at the same instance. After the $(k-1)$th firing, there will be at least one oscillator ready to fire at the next instance. Such an index set $V_k(\Phi^0)$ of the next firing oscillators is called the trigger set with respect to the initial condition Φ^0 at the kth stage. Let $U_k(\Phi^0)$ be the index set of oscillators which reach the threshold at the kth stage. Note that $U_k(\Phi^0) \supset V_k(\Phi^0)$. Hence, $U_k(\Phi^0)$ may contain the index of the oscillators which reach the threshold after receiving activations from other oscillators in $V_k(\Phi^0)$. Such a set $U_k(\Phi^0)$ is to be termed the spike set with respect to the initial condition Φ^0 at the kth stage. The terms for sets U_k and V_k were first used in [57]. Should no confusion arise, we shall write $V_k(\Phi^0)$ and $U_k(\Phi^0)$ as V_k and U_k, respectively. Immediately after the first firing, the resulting state $\Phi^1 = (\phi_{1_1}^1, \phi_{1_2}^1, \ldots, \phi_{1_n}^1)$, $0 = \phi_{1_1}^1 \le \phi_{1_2}^1 \le \cdots \le \phi_{1_n}^1 < 1$, is given by

$$\begin{aligned} \phi_{1_\ell}^1 &= g_{1_\ell}\left(f_{1_\ell}\left(\frac{T_{i_0}}{T_{1_\ell}}(1 - \phi_{i_0}^0) + \phi_{1_\ell}^0 \right) + \sum_{j \in U_1} \omega_{1_\ell, j} \right) \\ &=: g_{1_\ell}(f_{1_\ell}(\delta_{1_\ell}) + \omega_{1_\ell}), \qquad i_0 \in V_1 \text{ and } 1_\ell \in \{1, 2, \ldots, n\} - U_1 =: S_n - U_1. \end{aligned} \tag{1.1}$$

Note that the first firing consists of firings due to some oscillators reaching threshold simultaneously as well as any other oscillators then reaching threshold due to chain reaction of the earlier firings that are infinitesimally apart. All those chains of firings can be lumped into one set of "simultaneously firing" oscillators. The states $\Phi^k = (\phi_{k_1}^k, \phi_{k_2}^k, \ldots, \phi_{k_n}^k)$ of n oscillators after the kth firing can then be defined accordingly. If the cardinality of the spike set U_k, $k = 1, 2, \ldots, n$, is one, then we shall say that the system of n oscillators undergoes one *whole* cycle of firings or no *absorption* occurs for the system of n oscillators within one cycle of firings. For Peskin's model, $f_i(\phi) = \frac{I_i}{r_i}(1 - e^{-r_i T_i \phi})$ and $T_i = \ln(\frac{I_i}{I_i - r_i})/r_i$. Peskin conjectured that, first, for identical oscillators, the system approaches a state in which all oscillators are firing synchronously for almost all initial conditions and that, second, this remains true even when the oscillators are not quite identical. The first part of the conjecture was essentially proved

by Mirollo and Strogatz [41] with convex oscillators (i.e., $f_i'' < 0$). The second part of Peskin's conjecture was verified by Urbanczik and Senn [57] with flat oscillators (i.e., $f_i'' \equiv 0$). The key feature in those proofs relies on the nonconcavity of the evolution functions f_i. However, Bottani [8] numerically showed that even concave oscillators (i.e., $f_i'' > 0$) can synchronize, provided that the concavity is not too large. The purpose of this paper is two-fold. First, we prove the second part of Peskin's conjecture for the system of convex oscillators. Second, we prove Bottani's numerical results and more. Specifically, we shall show that for the system of n "identical" concave oscillators, no synchronization occurs for initial values in a set of positive measure, provided that $n = 3$ or n is even or phase responding curve $h(x) = g(f(x) + \omega)$ is concave upward. That is to say, in general, concave oscillators may synchronize for almost all initial conditions only if the concavity of the evolution maps is small. Indeed, we prove that the imbalance between the speeds and/or coupling strengths of the oscillators induces the synchronization of the system, provided that the concavity of the evolution maps is sufficiently small.

Since the work of Mirollo and Strogatz, current research into pulse-coupled or integrate-and-fire oscillators has become motivated by more elaborate questions (see, e.g., [25, 32, 47]). There have been many papers [7, 13, 25, 26, 39, 45, 46, 47] discussing those more advanced and complicated models. Some progress has also been made for more realistic biophysical models such as oscillators subject to small noise [36], constant delays [21], or a finite duration of synaptic response [2, 14, 22, 26].

We conclude this introductory section by mentioning the organization of the paper. Section 2 is devoted to the stability conditions for systems of two or more oscillators. In section 3, we derive the absorption conditions for systems of two or more oscillators. In particular, the necessary and sufficient condition for the absorption of two oscillators is given. This, in turn, provides some insight into the role that concavity of the evolution maps plays in determining the absorption process for systems of more than two oscillators. Some sufficient conditions for the absorption conditions for systems of more than two oscillators are derived. The main results of the paper are also recorded in this section.

2. Stable partial and full synchrony. Before beginning the analysis, we give an intuitive account of the way that synchrony develops as the system evolves: oscillators begin to clump together in "groups" that fire at the same time. For nonidentical oscillators, such groups of oscillators when they reach partial/full synchrony may break up again as the system continues to evolve. Consequently, it is desirable to find *stability conditions* for which a group of oscillators reaching the threshold at the same time will remain coordinated in the future. Such stable partial synchrony then gives rise to a positive feedback process, and thereby tends to grow by "absorbing" other oscillators. Absorptions reduce the number of groups until ultimately only one group remains—at that point the population is synchronized. The scenario above was first pointed out for a different system by Winfree [60], and the phrase "absorption" was coined by Mirollo and Strogatz [41]. With the characteristic of constant speed and equal coupling strengths, the system of identical oscillators always has the stability conditions satisfied. In this section, we shall derive stability conditions. The absorption conditions of the system are to be derived in section 3.

Unless otherwise stated, throughout this paper, the system of oscillators under consideration is either one of two types: convex or concave oscillators.

2.1. Stability conditions for two oscillators. We begin with the study of the system of two oscillators, which provides some insight as to why the system may or may not synchronize. The stability condition for two oscillators is to be derived in this subsection. To this end, we first need certain common properties shared by f and its inverse g.

Lemma 2.1. *Let $h_i : [0,1] \to [0,1]$ be smooth and strictly increasing maps with $h_i(0) = 0$ and $h_i(1) = 1$. Moreover, we assume that h_i have no inflection points and that $\lim_{x\to 0^+} xh_i'(1-x) = 0$ and $\lim_{x\to 0^+} xh_i'(x) = 0$. For each i, let two points, $A = (a_1, a_2)$ and $B = (b_1, b_2)$, be on $y = h_i(x)$ with $b_1 - a_1 \geq \omega_{\min}$. Here $\omega_{\min}$, the minimum of coupling strength, is defined to be*

$$\omega_{\min} = \min_{i,j} \omega_{ij}. \tag{2.1a}$$

Let m_h and M_h be, respectively, the minimum and maximum slope of the secant to h_i with the difference in x being at least $\omega_{\min}$. They are, respectively, defined as follows:

$$m_h = \min_i \left\{ \min \left\{ \frac{h_i(\omega_{\min})}{\omega_{\min}}, \frac{1 - h_i(1-\omega_{\min})}{\omega_{\min}} \right\} \right\} \tag{2.1b}$$

and

$$M_h = \max_i \left\{ \max \left\{ \frac{h_i(\omega_{\min})}{\omega_{\min}}, \frac{1 - h_i(1-\omega_{\min})}{\omega_{\min}} \right\} \right\}. \tag{2.1c}$$

Then

$$M_h \geq \frac{h_i(b_1) - h_i(a_1)}{b_1 - a_1} \geq m_h, \qquad m_h \leq 1 \text{ and } M_h \geq 1. \tag{2.2}$$

The equalities hold only if $b_1 - a_1 = \omega_{\min}$ and $a_1 = 0$ or $b_1 = 1$.

Proof. We illustrate only the case that $h_i''(x) > 0$ on $(0,1)$. Clearly, $\frac{h_i(a+x)-h_i(a)}{x} \geq \frac{h_i(x)}{x}$ for any $a \geq 0$, $x > 0$, and $1 \geq a + x \geq 0$. Moreover, $\frac{h_i(x)}{x}$ is increasing and bounded above by 1, and $\frac{1-h_i(1-x)}{x}$ is decreasing and bounded below by 1. Consequently, $M_h \geq \frac{1-h_i(1-\omega_{\min})}{\omega_{\min}} \geq \frac{h_i(b_1)-h_i(a_1)}{b_1-a_1} \geq \frac{h_i(\omega_{\min})}{\omega_{\min}} \geq m_h$. ∎

Remark 2.1.

1. *The geometric and physical meanings of m_h and M_h can be roughly interpreted as follows. Let the difference of two points in the vertical axis be the sum $\sum \omega_{ij}$ of certain coupling strengths due to the firings of certain oscillators; then the resulting difference in h is no smaller than $m_h \sum \omega_{ij}$ and no better than $M_h \sum \omega_{ij}$. See Figure 1.*
2. *Let $\omega_{\max} = \max_{i,j} \omega_{ij}$. An immediate application to Lemma 2.1 and Remark 2.1.1 is the following interpretation of the meaning of the quantities $M_g\omega_{\max}$ and $m_g\omega_{\min}$.*
   (a) *If an oscillator is within the distance $m_g\omega_{\min}$ of the threshold, then it will reach the threshold whenever it receives an activation jump due to the firings of other oscillators. On the other hand, if an oscillator is at least $M_g\omega_{\max}$ away from the threshold, then it will not reach the threshold whenever it receives an activation jump due to a single firing of another oscillator.*

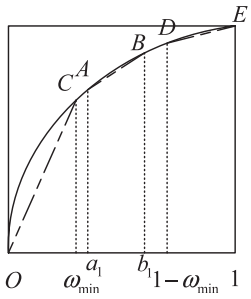


Figure 1. *Points O, C, A, B, D, and E are on a convex map $y = h(x)$. In this situation, $m_{\overline{OC}}$ is defined as the slope of $\overline{OC} = M_h$ and $m_{\overline{DE}} = m_h$. The assertions of Lemma 2.1 can easily be seen from the figure.*

(b) *If the ith oscillator has just received an impulse of strength ω_{ij} at x from the jth oscillator, then its phase jump, $g_i(f_i(x)+\omega_{ij})-x$, is at least $m_g\omega_{\min}$ and at most $M_g\omega_{\max}$ away from the origin.*

Theorem 2.2. *Let*

$$t_{\max} = \max_{i,j} \frac{T_i}{T_j}, \quad \Delta T = t_{\max} - 1, \quad \text{and} \quad \omega_{\min} = \min_{i,j} \omega_{ij}. \tag{2.3}$$

Suppose that f_i satisfy the same assumption as those maps h_i in Lemma 2.1. Let

$$m_g\omega_{\min} \geq \Delta T. \tag{2.4}$$

Then the system of two oscillators is stable.

Proof. Let $\Phi^0 = (\phi_1^0 = 0, \phi_2^0 = 0) \in \mathbb{R}^2$. We may assume that ϕ_2^0 has a greater speed $\frac{1}{T_2}$ and, hence, is the one that first reaches the threshold. Thus, $\phi_1^1 = g_1(f_1(\frac{T_2}{T_1})+\omega_{12})$. Therefore, $\phi_1^1 < 1$ if and only if $\frac{1-g_1(1-\omega_{12})}{\omega_{12}}\omega_{12} < 1 - \frac{T_2}{T_1}$. If $f_i''(x) > 0$, or equivalently, $g_i''(x) < 0$, and (2.4) holds, then we conclude, via (2.2), that $\phi_1^1 \geq 1$. Consequently, the assertion of the theorem holds. Suppose that $f_i''(x) < 0$, or equivalently, $g_i''(x) > 0$, and that (2.4) is satisfied. Then $\frac{1-g_1(1-\omega_{12})}{\omega_{12}}\omega_{12} \geq \frac{g_1(\omega_{12})}{\omega_{12}}\omega_{12} \geq m_g\omega_{\min} \geq \Delta T \geq 1 - \frac{T_2}{T_1}$. We have completed the proof of the theorem. ∎

The quantity ΔT is the phase difference between the fastest and slowest oscillators when evolving freely from their lowest state 0 toward the threshold. Therefore, if (2.4) holds, then two oscillators will remain firing synchronously according to Remark 2.1.2(a). To derive the stability condition and the absorption condition of the system, we make use of Lemma 2.1. From here on, we shall consider only the evolution maps that cannot turn "too" sharply at both ends. That is, the evolution maps f_i under consideration have the property that $\lim_{x\to 0^+} xf_i'(1-x) = 0$ and $\lim_{x\to 0^+} xf_i'(x) = 0$. It should be noted that each of the inverses of maps f_i cannot turn too sharply at both ends either.

2.2. Stable partial synchrony for n oscillators. To derive stable partial synchrony for n oscillators, we first need to derive conditions to exclude the possibility that one oscillator will run "too fast." The following proposition gives conditions that will prevent any oscillator from running too fast.

Proposition 2.3. *Let h_i be given as in Lemma 2.1, and let Δh and $\Delta\omega$ be given as follows:*

$$\Delta h = \max_{i,j} \max_{0\le\phi\le 1} |h_i(\phi) - h_j(\phi)| \tag{2.5a}$$

and

$$\Delta\omega = \max_{\substack{i,j\\ i\neq j}} \max_{T} \left(\sum_{\ell\in T} |\omega_{i\ell} - \omega_{j\ell}| \right), \tag{2.5b}$$

where $T \subset S_n - \{i,j\}$. If $n = 2$, then $\sum_{\ell\in T} |\omega_{i\ell} - \omega_{j\ell}|$ is to be interpreted as $|\omega_{ij} - \omega_{ji}|$.

1. *Let $\Phi^0 = (\phi_1^0, \dots, \phi_n^0)$ with ϕ_1^0 just reaching the threshold and being reset to zero. Assume $U_{k'}$, $k' = 1, 2, \dots, k$, are mutually exclusive and that $1, i \in S_n - \bigcup_{k'=1}^{k} U_{k'}$ with $\phi_i^0 \neq 0$. Suppose*

$$\begin{aligned} m_g^2 m_f \omega_{\min} &\ge \left(\sum_{j=0}^{k-1} \frac{1}{(m_f m_g)^j} \right) (M_g \Delta\omega + \Delta g + M_g(M_f(\Delta T + 1)\Delta T + \Delta f)) \\ &=: \left(\sum_{j=0}^{k-1} \frac{1}{(m_f m_g)^j} \right) \Delta. \end{aligned} \tag{2.6a}$$

   Then $\phi_i^{k'} \ge \phi_1^{k'}$, $k' = 1, 2 \dots, k$.
2. *Let*

$$m_g^2 m_f \omega_{\min} \ge \left(\sum_{j=0}^{n-1} \frac{1}{(m_f m_g)^j} \right) \Delta. \tag{2.6b}$$

   Suppose an oscillator has just reached the threshold. Then such an oscillator will not reach the threshold again until every other oscillator does. Moreover, suppose that the system of n oscillators undergoes one whole cycle of firings. Let the resulting phase of the system of oscillators be $\Phi^n = (\phi_{i_1}^n, \phi_{i_2}^n, \dots, \phi_{i_n}^n)$. Then the firing order for the next cycle with respect to the new initial condition Φ^n is preserved. That is, $\phi_{i_{k_2}}$ fires no earlier than $\phi_{i_{k_1}}$ does whenever $k_1 > k_2$.
3. *Let ϕ_i^m and ϕ_j^m be any two oscillators with $\phi_i^m = \phi_j^m < 1$ and $i, j \notin U_{m+1}$. Then the quantity Δ represents the maximum phase difference between these two oscillators after the next firing. That is, $|\phi_i^{m+1} - \phi_j^{m+1}| < \Delta$.*

Proof. Let δ_i and ω_i be given as in (1.1). Applying the mean value theorem, we get that

$$\begin{aligned} f_i(\delta_i) - f_i(\delta_1) &= f_i'(\xi) \left((1 - \phi_{i_0}^0) \frac{T_{i_0}}{T_i} \left(1 - \frac{T_i}{T_1} \right) + \phi_i^0 \right) \\ &\ge f_i'(\xi) \left((1 - \phi_{i_0}^0) \frac{T_{i_0}}{T_i} \left(1 - \frac{T_i}{T_1} \right) + g(\omega_{i1}) \right) \\ &\ge m_f m_g \omega_{\min} - M_f t_{\max} \Delta T. \end{aligned} \tag{2.7a}$$

Here $f_i'(\xi) = \frac{f_i(\delta_i)-f_i(\delta_1)}{\delta_i-\delta_1}$. The assumption that ϕ_1^0 just reach the threshold and Lemma 2.1 have been used to justify the inequalities in (2.7a). Using (2.5a), (2.5b), (2.6a), and (2.7a), we get that

$$\begin{aligned}\phi_i^1 - \phi_1^1 &= [g_i(f_i(\delta_i)+\omega_i) - g_i(f_i(\delta_i)+\omega_1)] + [g_i(f_i(\delta_i)+\omega_1) - g_i(f_i(\delta_1)+\omega_1)] \\ &\quad + [g_i(f_i(\delta_1)+\omega_1) - g_i(f_1(\delta_1)+\omega_1)] + [g_i(f_1(\delta_1)+\omega_1) - g_1(f_1(\delta_1)+\omega_1)] \\ &\geq \left(\sum_{j=1}^{k-1} \frac{1}{(m_f m_g)^j}\right)\Delta. \qquad \text{(2.7b)}\end{aligned}$$

Inductively, we have that

$$\begin{aligned}&\phi_i^{k'} - \phi_1^{k'} \geq \left(\sum_{j=k'}^{k-1} \frac{1}{(m_f m_g)^j}\right)\Delta, \quad k' = 1, 2, \ldots, k-1, \\ &\text{and } \phi_i^k - \phi_1^k \geq 0, \qquad \text{(2.8)}\end{aligned}$$

and the first part of the proposition follows. It should be noted that on the induction part, ϕ_i^0 in (2.7a) is to be replaced by $\phi_i^{k'-1} - \phi_1^{k'-1}$. Other parts of the estimates remain the same. Let Φ^0 be given. Suppose that the second assertion of the proposition were false. Then there exists a pair of indexes (i, j) such that the ith oscillator is the first oscillator reaching the threshold and the jth oscillator is the index of the first nonzero state oscillator that is outrun by the ith oscillator. To save notation, let the resulting phase state when the ith oscillator reaches the threshold be reset as ϕ_1^0, and the old index j be reset as j again. That is, ϕ_1^0 has just arrived at the threshold. Let k be the number of firings needed for ϕ_1^0 to reach the threshold. From how the indexes of 1 and j are chosen, we conclude that $k \leq n-1$ and that the spike sets associated with those firings are mutually disjoint. It follows from the first part of the proposition that if $\phi_1^k \geq 1$, then $\phi_i^k \geq \phi_1^k \geq 1$, a contradiction. We have just completed the proof of the first assertion of the second part of the proposition, and the second assertion of the second part of the proposition follows. To complete the proof of the last assertion of the proposition, we see that $\phi_i^{m+1} - \phi_j^{m+1}$ can be similarly expressed as those in (2.7b). The corresponding four terms in the brackets of (2.7b) are, respectively, bounded by $M_g\Delta\omega$, $M_g M_f t_{\max}\Delta T$, $M_g\Delta f$, and Δg. ■

We are now ready to state the stability conditions for synchrony.

Theorem 2.4. *Assume that the following stability condition holds:*

$$m_g^2 m_f \omega_{\min} \geq \max\left\{\sum_{j=0}^{n-1} \frac{1}{(m_f m_g)^j}, \sum_{j=0}^{n-2} (M_f M_g)^j\right\}\Delta. \qquad \text{(2.9)}$$

Then any group of oscillators which reaches the threshold simultaneously at some point will keep doing so in the future.

Proof. Let the ith and the jth oscillators be any two oscillators in the group spiking synchronously. Now reset both oscillators as $\phi_1^0 = \phi_2^0 = 0$. Suppose $1 \in U_{k+1}$ and $2 \notin \bigcup_{k'=1}^{k+1} U_{k'}$. It then follows from Proposition 2.3.2 that $U_{k'}$, $k' = 0, 1, \ldots, k+1$, are mutually

disjoint and that $k \le n-2$. Following from Proposition 2.3.3, we conclude that $|\phi_1^1 - \phi_2^1| \le \Delta$ and, inductively, $|\phi_1^k - \phi_2^k| \le (\sum_{j=0}^{k-1} \frac{1}{(M_f M_g)^j})\Delta$. Since $\phi_2^{k+1} = g_2(f_2(\frac{T_1}{T_2}(1-\phi_1^k)+\phi_2^k) + \sum_{\ell \in U_{k+1}} \omega_{2\ell})$, the index 2 being not in the set $\bigcup_{k'=1}^{k+1} U_{k'}$ implies that $\frac{T_1}{T_2}(1-\phi_1^k)+\phi_2^k < g_2(1-\sum_{\ell \in U_{k+1}} \omega_{2\ell})$. Upon using (2.2), we conclude that $m_g \omega_{\min} \le g_2'(\xi) \sum_{\ell \in U_{k+1}} \omega_{2\ell} < 1 - \frac{T_1}{T_2}(1-\phi_1^k) - \phi_2^k \le \Delta T + (\sum_{j=0}^{k-1} M_f M_g)\Delta \le (\sum_{j=0}^{n-2} M_f M_g)\Delta$, a contradiction to (2.9). ∎

Each of the terms in (2.9) can be verified analytically. Moreover, the inequality in (2.9) gives a measurement as to how not quite identical the system can be to get the stability condition. Roughly speaking, stability condition (2.9) amounts to saying that the total "weighted" measurements in how "nearly" identical the system is should be less than the minimum of the coupling strengths of the oscillators. In particular, the system of identical oscillators is always stable.

3. Absorption conditions.

In this section, we shall derive the conditions for which the absorption process of the system will forge ahead. In fact, we will show that the absorption process always occurs for a system of convex oscillators satisfying stability condition (2.9). On the other hand, the absorption process generally will not occur for a "nearly" identical system of concave oscillators. However, for a system of concave oscillators whose concavity is small, the absorption process is made possible by inducing an imbalance between the speeds and coupling strengths of the oscillators.

3.1. Absorption conditions for two oscillators.

We begin with the study of two oscillators. Let $\Phi^0 = (\phi_1^0, \phi_2^0)$ with $0 \le \phi_1^0 < \phi_2^0 < 1$. Assume that $U_1 = \{2\}$ and $U_2 = \{1\}$. Letting $\phi_2^0 = \phi$, the return map $R_2(\phi)$ is defined to be ϕ_2^2, the phase of the second oscillator immediately after the second firing. Specifically,

$$\phi_1^1 = g_1\left(f_1\left(\frac{T_2}{T_1}(1-\phi_2^0)+\phi_1^0\right)+\omega_{12}\right) =: h_1(\phi), \tag{3.1a}$$

$$\phi_2^2 = g_2\left(f_2\left(\frac{T_1}{T_2}(1-\phi_1^1)\right)+\omega_{21}\right) =: h_2(\phi_1^1), \tag{3.1b}$$

$$\phi_2^2 = h_2 h_1(\phi) =: R_2(\phi). \tag{3.1c}$$

Define the absorption map $A_2(\phi)$ as

$$A_2(\phi) = R_2(\phi) - \phi. \tag{3.1d}$$

The domain of the return map is the set of points for which $U_1 = \{2\}$ and $U_2 = \{1\}$. That is, no absorption occurs within one cycle of the firings whenever the initial values are in the domain of the return map. Now, $U_1 = \{2\}$ if and only if

$$\phi_2^0 > \ell_{12}, \quad \text{where } \ell_{ij} =: 1 - \frac{T_i}{T_j} g_i(1-\omega_{ij}), \tag{3.2a}$$

and $U_2 = \{1\}$ if and only if

$$\phi_1^1 > \ell_{21}. \tag{3.2b}$$

It should be noted that the positivity of ℓ_{ij} can be guaranteed by (2.4). The inequalities (3.2a) and (3.2b) amount to saying that there are limitations as to how close ϕ_2^0 can be to $0(=\phi_1^0)$ and $1(=\phi_1^0)$, respectively. To see why the second observation holds true, let

$$\gamma_{ij} = g_j(\omega_{ji}) - \ell_{ij}. \tag{3.3}$$

Note first that (3.2b) is equivalent to

$$f_1\left(\frac{T_2}{T_1}(1-\phi_2^0)+\phi_1^0\right)+\omega_{12} > f_1(\ell_{21}). \tag{3.4a}$$

If

$$\omega_{12} > f_1(\ell_{21}) \quad \text{or, equivalently,} \quad \gamma_{21} > 0, \tag{3.4b}$$

then ϕ_2^0 can be taken arbitrarily close to 1 from the left and (3.4a) still be satisfied. On the other hand, if $\gamma_{21} < 0$, then ϕ_2^0 cannot get too close to 1. In fact, $\phi_2^0 < h_1^{-1}(\ell_{21}) < 1$. Thus, the sign of γ_{21} determines how close ϕ_2^0 can be to 1 and therefore determines what is the boundary of the domain of the return map at the right end, which, in turn, influences the direction of the flow of the return map near the boundary of the domain. Such direction of the flow then determines whether the absorption process for the system of concave oscillators is to occur. (See Proposition 3.3.) We next show that for "nearly" identical oscillators the signs of γ_{ij} are determined by the concavity structure of the evolution maps.

Lemma 3.1. *Let ∇g be a measurement for the concavity of g_i, which is defined as follows:*

$$\nabla g = \min_i \left| \frac{g_i(\omega_{\max}) + g_i(1-\omega_{\max}) - 1}{\omega_{\max}} \right|. \tag{3.5}$$

Let $\widetilde{\Delta}\omega = \max_{i\neq j} |\omega_{ij} - \omega_{ji}|$. Assume that (2.4) and the following inequality, which is to be called the nearly identical condition, hold:

$$\omega_{\min}\nabla g > \Delta g + M_g\widetilde{\Delta}\omega + \Delta T. \tag{3.6}$$

Then $\gamma_{ij} < 0$ (resp., > 0) for all $i \neq j$, provided that $f_i'' < 0$ (resp., > 0) for all i.

Proof. Let $\widetilde{h}(x) =: \frac{h(x)+h(1-x)-1}{x}$. Here h is a map satisfying the assumptions of the maps given in Lemma 2.1. Then $\widetilde{h}(x)$ is increasing (resp., decreasing) on $(0,1)$, provided that $h''(x) > 0$ (resp., < 0). To see this, we have that $\widetilde{h}'(x) = \frac{x(h'(x)-h'(1-x))-(h(x)+h(1-x)-1)}{x^2} =: \frac{\widetilde{h_1}(x)}{x^2}$ and $\widetilde{h_1}'(x) = x(h''(x)+h''(1-x)) > 0$. Therefore, $\lim_{x\to 0^+}\widetilde{h_1}(x) = 0$, and so $\widetilde{h}(x)$ is increasing on $(0,1)$. The case for $h''(x) < 0$ can be similarly obtained. It is also clear that $\widetilde{h}(x) \leq 0$ (resp., ≥ 0) whenever $h''(x) > 0$ (resp., < 0). Consequently,

$$|-1+g_1(\omega_{12})+g_1(1-\omega_{12})| = \left|\frac{-1+g_1(\omega_{12})+g_1(1-\omega_{12})}{\omega_{12}}\omega_{12}\right| \leq \nabla g\omega_{\min}.$$

Suppose (3.6) holds. Then

$$\begin{aligned}\gamma_{ij} &= -1 + g_j(\omega_{ji}) + g_j(1-\omega_{ji}) + g_i(1-\omega_{ji}) - g_j(1-\omega_{ji}) \\ &\quad + g_i(1-\omega_{ij}) - g_i(1-\omega_{ji}) + \left(\frac{T_i}{T_j}-1\right)g_i(1-\omega_{ij}) < 0 \quad (\text{resp., } > 0),\end{aligned} \tag{3.7}$$

provided that $f_i''(x) < 0$ (resp., > 0), and the assertions of the lemma now follow. ∎

Remark 3.1.

1. *The consequences of Lemma 3.1 give that if the system of two oscillators is "nearly" identical in the sense that (3.6) are satisfied, then the domain of the absorption map A_2 is $(\frac{T_1}{T_2}\phi_1^0+\ell_{12}, h_1^{-1}(\ell_{21}))$ (resp., $(\frac{T_1}{T_2}\phi_1^0+\ell_{12}, 1)$), provided that $f_i'' < 0$ (resp., $f_i'' > 0$) for all i.*
2. *If ϕ_2^0 is not in the domain of the absorption map, then the two oscillators must fire simultaneously within one cycle of the firings. The corresponding system then will stay firing synchronously, provided that stability condition (2.4) is satisfied.*

The domain and monotonicity of the absorption map A_2 play an important role in determining whether the system is to forge ahead in the absorption process. The following lemma shows that the monotonicity of the absorption map depends on the concavity structure of f.

Lemma 3.2. *$\frac{\partial A_2}{\partial \phi} > 0$ (resp., < 0) on its domain, provided that $f_i'' < 0$ (resp., > 0) on $[0,1]$ for all i.*

Proof. We illustrate only the case that $f_i'' < 0$. The other cases can be similarly obtained. Applying the chain rule, we get

$$\begin{aligned}\frac{\partial R_2}{\partial \phi} &= \frac{\partial \phi_2^2}{\partial \phi} \\ &= g_2'\left(f_2\left(\frac{T_1}{T_2}(1-\phi_1^1)\right)+\omega_{21}\right) f_2'\left(\frac{T_1}{T_2}(1-\phi_1^1)\right) \\ &\quad \cdot g_1'\left(f_1\left(\frac{T_2}{T_1}(1-\phi_2^0)+\phi_1^0\right)+\omega_{12}\right) f_1'\left(\frac{T_2}{T_1}(1-\phi_2^0)+\phi_1^0\right).\end{aligned}$$

Using the facts that $g_i'' > 0$ and $g_i'(f_i(x)), f_i'(x) = 1$, $i = 1, 2$, we see immediately that $\frac{\partial R_2}{\partial \phi} > 1$, and hence $\frac{\partial A_2}{\partial \phi} > 0$. ∎

Proposition 3.3. *Assume that (2.4) is satisfied. Then the following statements hold:*

1. *Let (3.6) hold or $\gamma_{21} < 0$. Then $R_2(\phi)$ has a repelling fixed point, provided that $f_i'' < 0$ for all i. If $\gamma_{12} - \frac{T_1}{T_2}\phi_1^0 > 0$, then $R_2(\phi) - \phi > 0$ for all ϕ in its domain.*
2. *If $f_i'' > 0$ for all i and $\gamma_{21} < 0$, then $R_2(\phi) - \phi > 0$ for all ϕ in its domain.*
3. *Let $f_i'' > 0$ for all i. Assume that (3.6) holds. If $\phi_1^0 < \frac{T_1}{T_2}\gamma_{12}$, then $R_2(\phi)$ has a stable fixed point. If $\phi_1^0 > \frac{T_2}{T_1}\gamma_{12}$, then $R_2(\phi) - \phi < 0$ for all ϕ in its domain.*

Proof. Let $\phi = \frac{T_1}{T_2}\phi_1^0 + \ell_{12}$. Then

$$A_2(\phi) = \gamma_{12} - \frac{T_1}{T_2}\phi_1^0. \tag{3.8a}$$

Thus $A_2(\phi) < 0$, provided that $\gamma_{12} < 0$. On the other hand,

$$A_2(h_1^{-1}(\ell_{21})) = h_2(\ell_{21}) - h_1^{-1}(\ell_{21}) = 1 - h_1^{-1}(\ell_{21}) > 0, \tag{3.8b}$$

and the first part of the proposition now follows. The second part of the proposition is a direct consequence of Lemma 3.1, Lemma 3.2, and (3.8a). To complete the last part of the proposition, it remains to show that $A_2(1) < 0$ or, equivalently, $f_2(\frac{T_1}{T_2}(1 - g_1(\omega_{12}))) + \omega_{21} < 1$

or, equivalently, $\gamma_{12} > 0$, which follows from Lemma 3.1. We have just completed the proof of the proposition. ∎

Theorem 3.4.

1. *Assume that (2.4) holds. Then we have the following:*
   (a) *The system of two convex oscillators, in general, fires synchronously. Specifically, if $\gamma_{21} > 0$, then the synchrony of the system occurs for all initial values. Otherwise, that is, if $\gamma_{21} \le 0$, it synchronizes for almost all initial values. Consequently, for such a system, stability alone implies synchronization.*
   (b) *The system of two concave oscillators converges for all initial values to synchronous firing if and only if*

$$\gamma_{21} < 0 \quad \textit{or} \quad \gamma_{12} < 0. \tag{3.9}$$

   The inequalities in (3.9) are to be called the absorption condition for the system of two concave oscillators.
2. *Assume that (2.4) and (3.6) hold. Let $\phi_1^0 = 0$. Then the system of two concave oscillators will settle into a fixed nonfiring state if and only if ϕ_2^0 is in the domain of the absorption map A_2, that is, if $\ell_{12} < \phi_2^0 < 1$.*

Proof. To discuss synchrony for the system of two oscillators, we may just assume $\phi_1^0 = 0$. The statement 1(a) now follows from Proposition 3.3.1. The statement 2 follows easily from Proposition 3.3.3 and Lemma 3.1. It remains to prove statement 1(b). Consider the worst possible cases: (i) $\gamma_{21} < 0$ and $\gamma_{12} > 0$ or (ii) $\gamma_{21} > 0$ and $\gamma_{12} < 0$. The system will achieve synchronization at finite time for all initial conditions. To see this, we consider the case (ii). Let $\Phi^0 = (\phi_1^0, \phi_2^0)$ with $0 \le \phi_1^0 < \phi_2^0 < 1$. Then either Φ^1 is in synchrony or $\Phi^1 = (\phi_2^1, \phi_1^1)$ with $0 = \phi_2^1 < \phi_1^1 < 1$. Consequently, if no synchrony is achieved after the first firing, then the return map R_2 with respect to the initial phase state Φ^0 has a stable fixed point, while the return map R_2 with respect to the initial phase state Φ^1 has the property that $R_2(\phi) - \phi > 0$. However, the latter case will win out because it takes ϕ_1^1 finite time to reach the threshold and it takes ϕ_2^0 infinite time to reach the fixed point. On the other hand, if both γ_{21} and γ_{12} are nonnegative, then the corresponding return map has a stable fixed point. ∎

For the system of two convex oscillators, the associated return map is (volume) expanding; i.e., there exists some $r > 1$ such that $|A_2(\phi) - A_2(\overline{\phi})| > r|\phi - \overline{\phi}|$ for all $\phi \ne \overline{\phi}$ in the domain. Thus, the absorption is bound to happen except for the initial value being the fixed point of the absorption map. The sign of γ_{12} (or γ_{21}) then plays the role of determining whether the absorption map has a (repelling) fixed point or not. On the other hand, for the system of concave oscillators, the corresponding return map is (volume) contracting. If the flow of the return map at both ends of the domain points inward, which is the case for a nearly identical system (see Proposition 3.3.3), then its return map has a stable fixed point. As a result, the corresponding system converges to a nonfiring state. To make the system of concave oscillators fire synchronously, the flow of the return map at both ends has to point in the same direction, which in turn makes the absorption process go forward. The above scenario occurs whenever there is a certain degree of imbalance between oscillators (i.e., $\gamma_{12} < 0$ or $\gamma_{21} < 0$). To see this, note that $\gamma_{12} < 0$ is equivalent to $g_2(\omega_{21}) + \frac{T_1}{T_2} g_1(1 - \omega_{12}) < 1$. For identical concave oscillators, the inequality above will not be satisfied. Thus, to drive such a system into synchrony, the variations in the speed and/or the coupling strength cannot be too small.

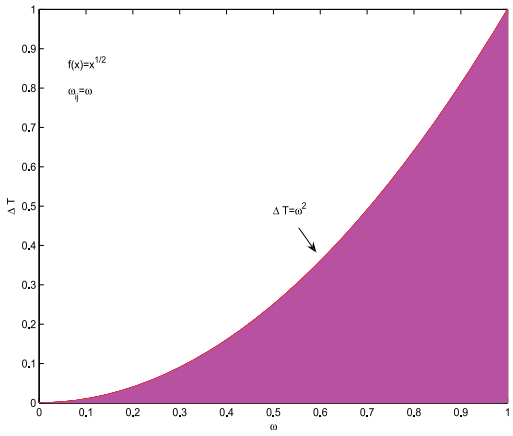


Figure 2. *The shaded area is the set of parameters satisfying* (3.10).

3.2. Feasible parameter and examples. For practical purposes, we consider how feasible it is to verify those stability and absorption conditions. Some numerical results are also provided to support the validity of the theorem. To simplify our calculations, we consider the following three cases: (i) $f_i(x) = \sqrt{x}$, $g_i(x) = x^2$, and $\omega_{12} = \omega_{21} = \omega$; (ii) $f_i(x) = x^2$, $g_i(x) = \sqrt{x}$, and $\omega_{12} = \omega_{21} = \omega$; (iii) $f_i(x) = x^2$, $g_i(x) = \sqrt{x}$, and $T_1 = T_2$.

Case (i): Since $m_g = \omega$, (2.4) becomes

$$\omega^2 \geq \Delta T. \tag{3.10}$$

In the $\omega - \Delta T$ plane, the equality in (3.10) is a parabola. As shown in Theorem 3.4, no absorption condition is needed to achieve synchrony for the system considered here. By choosing parameters randomly from the feasible region (see Figure 2), the numerical results (see Figure 3) indeed support our theory.

Case (ii): For case (ii), if (3.6) is satisfied, then no absorption occurs. Thus, the system in general will not fire synchronously unless ϕ_2^0 is too close to $\phi_1^0 = 0$. To see this, note that $\nabla g = \frac{\sqrt{\omega}+\sqrt{1-\omega}-1}{\omega}$, $m_g = \frac{1-\sqrt{1-\omega}}{\omega}$, and $\Delta g = \Delta\omega = 0$. The stability condition and (3.6) for the associated system then reduce to

$$(1 - \sqrt{1-\omega}) \geq \Delta T \tag{3.11}$$

and

$$\sqrt{\omega} + \sqrt{1-\omega} - 1 > \Delta T, \tag{3.12}$$

respectively. The feasible parameters region in the $\omega - \Delta T$ plane is nonempty (see Figure 4). Picking parameters from this region, we see, via Figure 5, that if $0 \leq \phi_2^0 < \ell_{12}$, then each of the corresponding systems will fire synchronously. Otherwise, they will settle into a nonfiring state. In fact, we choose various sets of parameters from different locations of the region, and all the corresponding systems behave as predicted in Theorem 3.4.2 (see Figure 5).

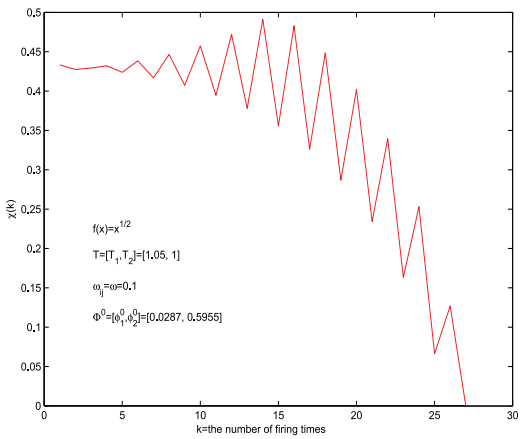


Figure 3. *The evolution of the synchronization order parameter $\chi(k)$ is defined as the sum of the minimum distances between any two oscillators at the kth stage* $= \sum_{i=1}^{n} \sum_{j=i+1}^{n} d(\phi_i^k, \phi_j^k)$, *where* $d(x, y) = \min(|x - y|, |x - y + 1|, |x - y - 1|)$. *If* $\chi(k) = 0$ *for some large* k, *then the system fires synchronously at finite time. If* $\lim_{k\to\infty} \chi(k) = 0$, *then the system fires synchronously eventually or asymptotically.*

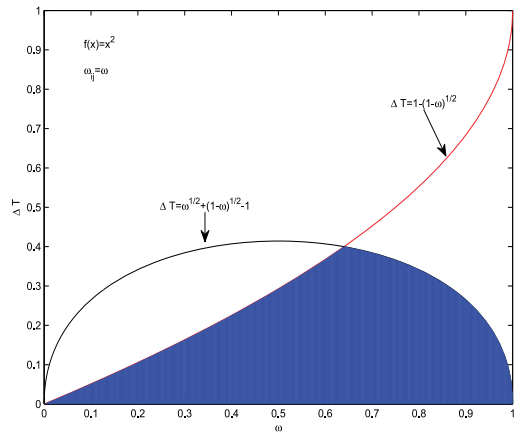


Figure 4. *The shaded area is the set of parameters satisfying* (3.11) *and* (3.12).

Case (iii): The absorption condition studied here is (3.9). Since $\Delta T = 0$, the stability condition is automatically satisfied. Moreover, (3.9) becomes

$$(\omega_{12})^{\frac{1}{2}} + (1 - \omega_{21})^{\frac{1}{2}} < 1. \tag{3.13}$$

The feasible parameters region in the ω_{21}-ω_{12} plane, as given in Figure 6, shows the "imbalance" between parameters ω_{12} and ω_{21}. The numerical results, as demonstrated in Figure 7, also support our theory.

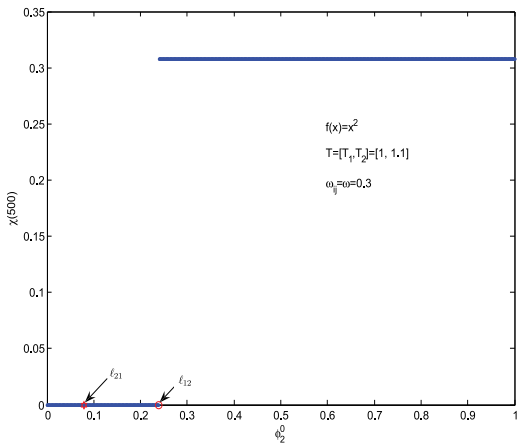


Figure 5. *Choosing parameters T_i and ω from the shaded part in Figure 4, we see that after 500 firings, the synchronization order parameter $\chi(500)$ is a step function with respect to the initial state ϕ_2^0. As predicted, if $\ell_{12} < \phi_2^0 < 1$, then the system settles into a nonfiring state. Otherwise, it fires synchronously.*

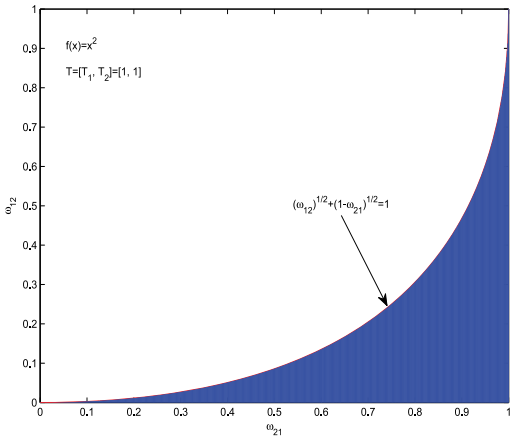


Figure 6. *The shaded area is the stability region for case* (iii).

3.3. Absorption conditions. To understand the absorption process of a system of more than two oscillators, we begin with defining the return map, which was originally defined in [41]. Throughout this section, we shall assume that stability condition (2.9) holds. Unlike the system of two oscillators, the corresponding return map under study in this section is now a high-dimensional map. Let the system of n oscillators undergo one whole cycle of firings. Assume that the resulting phase is denoted by $(\phi_1^0 = 0, \phi_2^0, \ldots, \phi_n^0)$. Let $\Phi^0 = (\phi_2^0, \ldots, \phi_n^0)$. Then the return map $R_n : \text{Domain}(R_n) =: A_n \subset \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ is defined to be

$$R_n(\Phi^0) = \Phi^n = (\phi_2^n, \phi_3^n, \ldots, \phi_n^n) =: (r_{2,n}(\Phi^0), \ldots, r_{n,n}(\Phi^0)). \tag{3.14a}$$

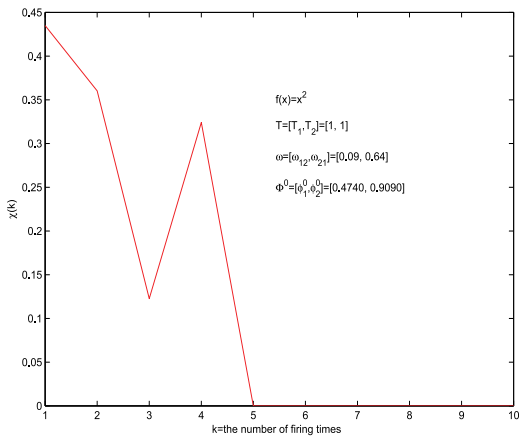


Figure 7. *Let the speed of the oscillators be* 1. *Pick the parameters* ω_{ij} *from the shaded region in Figure* 6. *The synchronization order parameter* $\chi(k)$ *reaches zero after* 5 *firings. The imbalance of parameters in activation gives the synchrony of the system.*

It should be noted, via Proposition 2.3.2, that the maps in (3.14a), (3.14b) are well defined. Moreover,

$$R_n(\Phi^0) = H_n \cdots H_2 H_1(\Phi^0), \tag{3.14b}$$

where

$$H_i = \tau_i \Sigma(\Phi). \tag{3.14c}$$

Here $\Phi = (\phi_2, \phi_3, \ldots, \phi_n)$,

$$\begin{aligned}\Sigma(\Phi) &= (\sigma_2, \sigma_3, \ldots, \sigma_n)\\ &=: \left(\frac{T_n}{T_1}(1-\phi_n), \frac{T_n}{T_2}(1-\phi_n)+\phi_2, \ldots, \frac{T_n}{T_{n-1}}(1-\phi_n)+\phi_{n-1}\right),\end{aligned}$$

and

$$\tau_i(\sigma_2, \sigma_3, \ldots, \sigma_n) = (g_1(f_1(\sigma_2)+\omega_{1,n}), \ldots, g_{n-1}(f_{n-1}(\sigma_{n-1})+\omega_{n-1,n})).$$

Note that we have implicitly relabeled the oscillators, so each of the image vectors $H_i(\Phi)$ represents the phases of the oscillators $1, 2, \ldots, n-1$. That is, the original oscillator 1 has become 2, oscillator 2 has become 3, ..., and oscillator n has become oscillator 1. It also follows from Proposition 2.3.3, Remark 2.1.2(b), and stability condition (2.9) that domain $(R_n) \subset S$, where $S = \{\Phi^0 = (\phi_2^0, \ldots, \phi_n^0) \in \mathbb{R}^{n-1} : 0 < \phi_2^0 < \phi_3^0 < \cdots < \phi_n^0 < 1\}$. In fact, the domain of the return map R_n is the set of points in S so that the spike sets $U_i = \{n-i+1\}$, $i = 1, 2, \ldots, n$. Having such spike sets is equivalent to the following inequalities:

$$\phi_{n-i+1}^{i-1} - \frac{T_{n-i}}{T_{n-i+1}}\phi_{n-i}^{i-1} > \ell_{n-i,\,n-i+1}, \qquad i = 1, 2, \ldots, n, \tag{3.15}$$

where $\ell_{n-i,\,n-i+1}$ are defined as in (3.2a) and T_0, $\ell_{0,1}$, and ϕ_0 are interpreted as T_n, $\ell_{n,1}$, and ϕ_n, respectively. Consequently, the domain A_n of the return map is

$$A_n = \{\Phi^0 \subset S : \text{the inequalities in (3.15) hold}\}. \tag{3.16a}$$

Since A_n is the finite intersection of open sets, it is open. Moreover, the domain A_k of H_k is the set of initial points satisfying the inequalities in (3.15) for $i = 1, 2, \ldots, k$. So A_i is the set of initial values that will have at least i firings before an absorption occurs. Then

$$A = \bigcap_{i=1}^{\infty} A_i = \text{the set of initial values that live forever without any absorptions.} \tag{3.16b}$$

We next state some properties of the return map $R_n : A_n \to S$. The first assertion of the theorem below is essentially due to Mirollo and Strogatz (see Theorem 3.1 of [41]).

Theorem 3.5. *Assume that stability condition (2.9) holds for a system of n oscillators. The following hold true:*

1. *Let $f_i'' < 0$ for all i. Then R_n is volume-expanding on A_n. Consequently, the set A has Lebesgue measure zero.*
2. *Let $f_i'' > 0$ for all i. Then R_n is volume-contracting on A_n.*

Proof. To prove the first assertion of the theorem, it suffices to show that the Jacobian determinant of R_n has absolute value greater than one. From (3.14b) and (3.14c) and the definitions of τ_i and Σ, $\det(DR_n) = \prod_{i=1}^n \det(DH_i) = \prod_{i=1}^n \det(D\tau_i)\det(D\Sigma)$. The map Σ is affine and satisfies $\sigma^n = I$, so $\det(D\Sigma) = \pm 1$. Note that $D\tau_i$ is a diagonal matrix; thus it is easily seen that $\det D\Sigma > 1$ under the assumption that each of the evolution maps is convex. Hence $|\det(DR)| > 1$. The arguments for proving the second assertion of the theorem are similar to those of the first. ∎

Since the return map of the system of convex oscillators is volume-expanding, the set of initial values that live forever without any absorptions has measure zero. Hence, it is the nature of the system of convex oscillators to grow by absorbing other oscillators. On the other hand, if the flow of the return map of the system of concave oscillators near the boundary of the domain points inward, such as that of identical concave oscillators, then the system converges to a fixed point, which is a nonfiring state. Hence, to break such a natural tendency of the system one has to introduce some imbalance between the parameters so as to make the direction of the flow point outward near a certain portion of the boundary, as in the case for two oscillators, where a necessary and sufficient condition has been established. Due to the technical difficulty of this, only sufficient conditions are established for systems of more than two oscillators. Such a result is stated in the following.

Theorem 3.6. *Let the number of concave oscillators under consideration be no less than three. Assume the following absorption condition, which is to say that the imbalance measurement is greater than or equal to the concavity of the inverse of the evolution maps:*

$$\frac{M_g}{m_g} \le \max_{0\le i\le n-1} \left(\frac{T_i \omega_{i,i+1}}{T_{i+1}\omega_{i+1,i}} \right). \tag{3.17}$$

Suppose that (2.9) and (3.17) hold. Then the absorption of the system must occur.

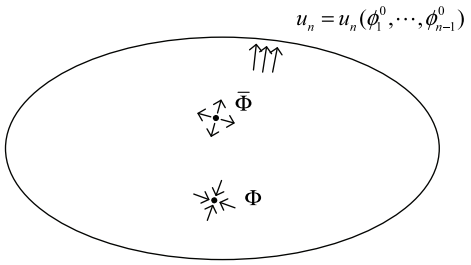


Figure 8. *A visualization of the claim in Step 3 of Theorem 3.6.*

Proof. Let $\max_{0\le i\le n-1} \frac{T_i\omega_{i,i+1}}{T_{i+1}\omega_{i+1,i}} = \frac{T_m\omega_{m,m+1}}{T_{m+1}\omega_{m+1,m}}$ for some m. Suppose that no absorption occurs after the first $(n-m)$ firings; then we may relabel oscillators so that the indexes $m+1,\dots,n,1,\dots,m$ of the oscillators become $1,2,\dots,n$, respectively. We then may assume that $\max_{0\le i\le n-1} \frac{T_i\omega_{i,i+1}}{T_{i+1}\omega_{i+1,i}} = \frac{T_n\omega_{n,1}}{T_1\omega_{1,n}}$. The proof the theorem then breaks into three steps. The first part is to prove that sufficient condition (3.17) is so given that the inequality in (3.15) with $i = n$ is violated whenever ϕ_n^0 is sufficiently close to 1 from the left. Consequently, if the system is to undergo one whole cycle of firings, ϕ_n^0 must stay away from 1. That is, $\phi_n^0 < u_n = u_n(\phi_1^0,\dots,\phi_{n-1}^0) < 1$ for some u_n depending on $\phi_1^0,\dots,\phi_{n-1}^0$ and being away from 1. Here $u_n = u_n(\phi_1^0,\dots,\phi_{n-1}^0)$ is a portion of the boundary of the domain of the return map described by $\phi_1^{n-1} - \frac{T_n}{T_1}\phi_n^{n-1} = \ell_{n,1}$. The second step of the proof is to show that the direction of the flow points outward to the boundary whenever ϕ_n^0 is sufficiently close to u_n from the left. Finally, to complete the proof the theorem, we need to show that the return map has no periodic points.

Step 1. Let ϕ_n^0 be sufficiently close to 1 from the left so that $\phi_1^1 - \phi_n^1 (= 0) < M_g\omega_{\min}$. We have used Lemma 2.1 to ensure that the above assertion can be done. Note that each of $g_i(f_i(\phi)+\omega)-\phi$, the phase jump at ϕ, is decreasing in ϕ. Hence, the phase jump is greater when the phase position ϕ is closer to the origin. Upon using Lemma 2.1, we conclude that

$$\begin{aligned}\phi_1^{n-1} - \frac{T_n}{T_1}\phi_n^{n-1} &< \phi_1^1 - \phi_n^1 + \left(1 - \frac{T_n}{T_1}\right)\\ &< M_g\omega_{1,n} + \left(1 - \frac{T_n}{T_1}\right) \le \frac{T_n}{T_1} m_g\omega_{n,1} + \left(1 - \frac{T_n}{T_1}\right) < \ell_{n,1}.\end{aligned}$$

We just proved that the boundary of the domain of the return map cannot get arbitrarily close to $\phi_n^0 = 1$. Note that if $n = 2$, then $\phi_2^1 = 0$, and so the first inequality above is not necessarily true.

Step 2. Suppose that ϕ_n^0 is close to u_n. Then $\phi_1^{n-1} - \frac{T_n}{T_1}\phi_n^{n-1}$ is close to $\ell_{n,1}$. Consequently, ϕ_n^n is close to 1. Therefore, $\phi_n^n > \phi_n^0$ whenever ϕ_n^0 is sufficiently close to u_n.

Step 3. Since R_n is volume-contracting, any of its periodic points, if one exists, must be stable. Assume, to the contrary, that there exists a periodic point Φ with period k. Let $\overline{R} = R_n^k$. Then Φ becomes a stable fixed point of $\overline{R}$. Moreover, the direction of the flow under $\overline{R}$ near the boundary of the domain still points outward. Consequently, there must exist a unstable fixed point $\bar{\Phi}$ of $\overline{R}$, a contradiction (see Figure 8).

Using Steps 1–3, we conclude that the direction of the flow of the return map points outward to the boundary u_n. Hence, the absorption must occur. We have just completed the proof of the theorem. ∎

From the proof of the above theorem as well as that of Theorem 3.4.1(b), it is easily concluded that for the system of concave oscillators to undergo the absorption process, the domain of the return map contains only the points for which their ϕ_n^0's must stay away from 1. This, in turn, makes the direction of flow near the boundary $u_n = u_n(\phi_1^0, \ldots, \phi_{n-1}^0)$ point outward. While the best possible condition to ensure such a scenario for the system of two concave oscillators can be obtained, it is not clear whether the condition that $\min_{1\le i\le n} \gamma_{i-1,i} < 0$ (here $\gamma_{0,1}$ is to be interpreted as $\gamma_{n,1}$) is the best absorption condition for the system of more than two oscillators. Nevertheless, if the concavity of a system is small, then the inequalities in (3.17) can be satisfied by inducing an imbalance between the speeds and weights of oscillators, which will be demonstrated in Proposition 3.7.

We next discuss the dynamics under iteration of the absorption maps. Assume an initial value Φ^0, not necessarily in the domain of the return map. Suppose after initial firings that the system forms k partially synchronous groups. Let the ith group, $1 \le i \le k$, contain k_i oscillators, where $\sum_{i=1}^{k} k_i = n$, and let these be treated as one new oscillator, denoted by $\overline{\phi}_i$. Clearly, when oscillator $\overline{\phi}_i$ is firing, the activation of each oscillator ϕ_j in the $(i+1)$th synchronous group, where $(\sum_{\ell=1}^{i} k_\ell) + 1 \le j \le \sum_{\ell=1}^{i+1} k_\ell =: \sigma_{i+1}$, is incremented by the positive coupling $\sum_{k=\sigma_{i-1}+1}^{\sigma_i} \omega_{jk} =: \widetilde{\omega}_{ji}$. For each j, $\sigma_{i-1} + 1 \le j \le \sigma_i$, we may define $\widetilde{\omega}_{ji+1}$ similarly. Since the ith and $(i+1)$th synchronous groups may contain more than one oscillator, the new cycle periods $\overline{T}_i$ and $\overline{T}_{i+1}$ of the new oscillators $\overline{\phi}_i$ and $\overline{\phi}_{i+1}$ are chosen as the minimum cycle periods among the oscillators in each group, i.e., $\overline{T}_i = \min_{\sigma_{i-1}+1\le i\le\sigma_i} T_i$ and $\overline{T}_{i+1} = \min_{\sigma_i+1\le i\le\sigma_{i+1}} T_i$. That is, the speed of each group is chosen to be the fastest speed among oscillators in the group. With $\overline{T}_i$ and $\overline{T}_j$ now being fixed, the corresponding new coupling strengths $\overline{\omega}_{i,i+1}$ and $\overline{\omega}_{i+1,i}$ are so chosen that

$$\max_{\sigma_{i-1}+1\le\ell\le\sigma_i} \left(\max_{\sigma_i+1\le j\le\sigma_{i+1}} \frac{\overline{T}_i \widetilde{\omega}_{\ell,i+1}}{\overline{T}_{i+1} \widetilde{\omega}_{j,i}} \right) = \frac{\overline{T}_i \overline{\omega}_{i,i+1}}{\overline{T}_{i+1} \overline{\omega}_{i+1,i}}. \tag{3.18}$$

The idea for such choices is to make the inequality (3.17) as easy as possible to satisfy. Due to the presence of the stability condition, we are allowed to make such choices. For these newly formed synchronous groups to continue their absorption process, we need to further assume that for any permissible set $\{k, k_1, k_2, \ldots, k_k\}$, where $2 < k \le n$ and $\sum_{i=1}^{k} k_i = n$,

$$\frac{M_g}{m_g} \le \max_{0\le i\le k-1} \left(\frac{\overline{T}_i \overline{\omega}_{i,i+1}}{\overline{T}_{i+1} \overline{\omega}_{i+1,i}} \right). \tag{3.19}$$

The right-hand side of the inequality above is to be called the imbalance measurement for the system of more than two oscillators. Note that the quantity $\frac{M_g}{m_g}$ is a measurement for the concavity of g. The closer $\frac{M_g}{m_g}$ is to 1, the more flat the g is. With such an absorption condition, the system continues to grow by absorption until it reaches full synchrony or reduces to two synchronous groups of oscillators. To ensure that these two synchronous groups continue to grow by absorption, we need to have a modified absorption condition for these two groups.

To this end, we assume that the first group consists of old oscillators $\phi_{\ell_1}, \dots, \phi_{\ell_2}$, where $1 \le \ell_1 < \ell_2 < n$ or $1 < \ell_1 < \ell_2 \le n$, while the second group contains the remaining oscillators. Then the parameters in γ_{12} and γ_{21}, as given in (3.3), need to be updated as well. Let $N_1 = \{\ell_1, \dots, \ell_2\}$ and $N_2 = \{1, 2, \dots, n\} - N_1$. Set $\widetilde{\omega}_{j1} = \sum_{i \in N_1} \omega_{ji}$, $j \in N_2$, and $\widetilde{\omega}_{j2} = \sum_{i \in N_2} \omega_{ji}$, $j \in N_1$. Define the new cycle periods of groups N_1 and N_2 to be the minimum cycle periods among the oscillators in each group. Denote such new periods by $\overline{T}_1$ and $\overline{T}_2$. Let

$$\gamma_{12}(\ell_1, \ell_2) = \min_{\substack{i \in N_1 \\ j \in N_2}} \left(g_j(\widetilde{\omega}_{j1}) - 1 + \frac{\overline{T}_1}{\overline{T}_2} g_i(1 - \widetilde{\omega}_{i2}) \right) \tag{3.20a}$$

and

$$\gamma_{21}(\ell_1, \ell_2) = \min_{\substack{j \in N_1 \\ i \in N_2}} \left(g_j(\widetilde{\omega}_{j2}) - 1 + \frac{\overline{T}_2}{\overline{T}_1} g_i(1 - \widetilde{\omega}_{i1}) \right). \tag{3.20b}$$

Then the absorption condition for any two sizes of synchronous groups of oscillators is

$$\min \left\{ \max_{\ell_1, \ell_2} \gamma_{12}(\ell_1, \ell_2), \max_{\ell_1, \ell_2} \gamma_{21}(\ell_1, \ell_2) \right\} < 0. \tag{3.20c}$$

The left-hand side of the inequality in (3.20c) is to be called the imbalance measurement for the system of two oscillators. With those absorption conditions on hand, one would expect the full synchrony of the system. The drawback of absorption conditions (3.19) and (3.20c) is that when n is large, there are enormously many cases needing to be checked. As a consequence, the question of nonemptiness of the set of parameters satisfying the constraints (3.19) and (3.20c) has to be addressed.

Proposition 3.7. *Let the coupling strengths $\omega_{ij} (= \omega)$ of a system of n oscillators all be equal. Let the period cycles of oscillators all be different. Assume that $\omega < \frac{2}{n}$ and that*

$$\frac{[\frac{n}{3}] + 1}{[\frac{n}{3}]} > t_{\max}. \tag{3.21}$$

Then the absorption conditions (3.19) *and* (3.20c) *are satisfied, provided that the concavity of the evolution maps is sufficiently small.*

Proof. With the speed of oscillators being all different, $t_{\max} > 1$. Suppose that the absorption occurs after the initial firings. Assume that the system evolves into k, $k > 2$, synchronous groups with sizes of groups being $k_1, k_2, \dots$, and k_k. If $k_1 = k_2 = \cdots = k_k$, then the system continues to grow by absorption, provided that $\frac{M_g}{m_g}$ is sufficiently close to 1. Suppose that the sizes of k synchronous groups are not all equal. Then there must exist an index i for which $\frac{\overline{\omega}_{i,i+1}}{\overline{\omega}_{i+1,i}} \ge ([\frac{n}{3}] + 1)/[\frac{n}{3}] > t_{\max}$. Here $[x]$ is the greatest integer that is equal to or less than x. Consequently, the imbalance measurement for this system is greater than one. The system then must reach full synchrony or reduce to the system of two synchronous groups, provided that the concavity of the evolution maps is small. In the case of the latter,

we assume that the sizes of these two groups N_1 and N_2 are ℓ and $n-\ell$, respectively, and let $f_i(x) = x$ for all i. Then (3.20c) reduces to

$$\gamma_{12}(\ell, n-\ell) = \ell\omega - \left(\frac{\overline{T}_i}{\overline{T}_j}\right)(n-\ell)\omega + \frac{\overline{T}_i}{\overline{T}_j} - 1, \qquad i \in N_1,\ j \in N_2,$$
$$\gamma_{21}(\ell, n-\ell) = (n-\ell)\omega - \left(\frac{\overline{T}_j}{\overline{T}_i}\right)\ell\omega + \frac{\overline{T}_j}{\overline{T}_i} - 1, \qquad i \in N_1,\ j \in N_2.$$

If n is even and $\ell = n - \ell$, then

$$\gamma_{12}(\ell, n-\ell) = \left(\frac{\overline{T}_i}{\overline{T}_j} - 1\right)\left(1 - \frac{n\omega}{2}\right) \quad \text{and} \quad \gamma_{21}(\ell, n-\ell) = \left(\frac{\overline{T}_j}{\overline{T}_i} - 1\right)\left(1 - \frac{n\omega}{2}\right).$$

Since $t_{\max} > 1$, either $\gamma_{12}(\ell, n-\ell)$ or $\gamma_{21}(\ell, n-\ell)$ is negative. If $n - \ell = \ell + \ell_1$, where $\ell_1 \geq 1$, then $\gamma_{12}(\ell, n-\ell) = (\frac{\overline{T}_i}{\overline{T}_j} - 1)(1 - \ell\omega) - \frac{\overline{T}_i}{\overline{T}_j}\ell_1\omega$. Suppose $\frac{\overline{T}_i}{\overline{T}_j} \leq 1$. Then $\gamma_{12}(\ell, n-\ell) < 0$. If $\frac{\overline{T}_i}{\overline{T}_j} > 1$, then $\gamma_{12}(\ell, n-\ell) < \Delta T - \omega_{\min} \leq 0$. The last inequality is justified by stability condition (2.4). The case that $\ell = (n-\ell) + \ell_1$, where $\ell_1 \geq 1$, can be similarly addressed. Therefore, the remaining two synchronous groups will achieve full synchrony, provided that the concavity of the evolution maps is small. ■

The result of the proposition supports the numerical observation of Bottani [8]. We next define phase responding function $h(x)$ and phase difference function $D(x)$. Both functions are helpful in determining the direction of the flow of the system near the boundary of the return map whenever the number of oscillators is greater than three. Assume that an oscillator receives an activation ω at x. Let the resulting phase $g(f(x) + \omega)$ be denoted by $h(x)$, and define $D(x)$ as

$$D(x) = h(x + a) - h(x). \tag{3.22}$$

Here $a > 0$ is a constant.

Proposition 3.8.

1. *Consider an identical system of three concave oscillators. That is, $f_i \equiv f$, $g_i \equiv g$, $T_i \equiv T$, and $\omega_{ij} = \omega$. Then the direction of the flow near the boundary of the domain of the return map points inward.*
2. *Suppose $h''(x) > 0$. Then $D(x)$ is increasing in x.*
3. *Consider an identical system of n concave oscillators. If $h''(x) > 0$, then*

$$\phi_n^0 - \phi_{n-1}^0 < \phi_n^n - \phi_{n-1}^{n-1} \tag{3.23}$$

whenever $\phi_n^0 - \phi_{n-1}^0$ is sufficiently close to $\ell_{ij} = 1 - g(1-\omega)$ from the left. Consequently, the direction of the flow of the system points inward near the boundary of the domain of the return map.

Proof. The boundary of the domain of the return map consists of three pieces of curves Γ_1, Γ_2, and Γ_3 defined by $\phi_{n-i+1}^{i-1} - \phi_{n-i}^{i-1} = 1 - g(1-\omega)$, $i = 1, 2, 3$, respectively. To prove the first part of the proposition, it suffices to show that for $i = 1, 2, 3$

$$\phi_{n-i+1}^{i-1} - \phi_{n-i}^{i-1} < \phi_{n-i+1}^{n+i-1} - \phi_{n-i}^{n+i-1} \tag{3.24}$$

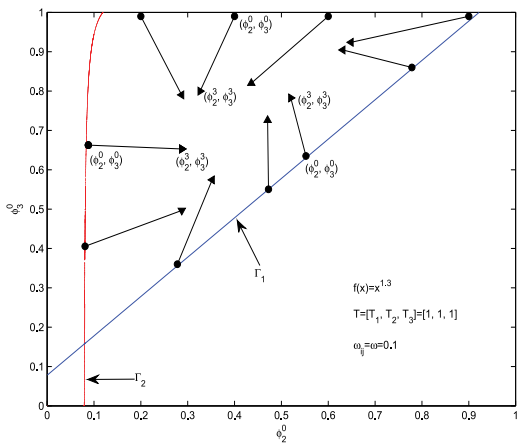


Figure 9. *The initial position and ending position of each arrow are* (ϕ_2^0, ϕ_3^0) *and* (ϕ_2^3, ϕ_3^3)*, respectively. The direction of the flow near the boundary of the domain of the return map indeed points inward as predicted.*

whenever $\phi_{n-i+1}^{i-1} - \phi_{n-i}^{i-1}$ are sufficiently close to $1-g(1-\omega)$. The inequalities in (3.24) amount to saying that $R(\phi_2^0, \phi_3^0) = (\phi_2^3, \phi_3^3)$ are moving further away from their respective boundaries whenever (ϕ_2^0, ϕ_3^0) are near Γ_1, Γ_2, and Γ_3, respectively (see Figure 9). To this end, we first prove that Γ_3 can be interpreted as $\phi_3^0 = 1$. For any $\phi_3^0 < 1$, we have that $\phi_1^1 > g(\omega)$. And so, for any $\phi_2^0 + 1 - g(1-\omega) < \phi_3^0 < 1$, we see that $\phi_1^2 - \phi_3^2 = \phi_1^1 - (\phi_3^2 - (\phi_1^2 - \phi_1^1)) > g(\omega) - (g(\omega) - (1 - g(1-\omega))) = 1 - g(1-\omega)$. We have used the fact that the phase jump function $h(x) - x$ is decreasing to justify the above inequality. Hence, Γ_3 can be interpreted as claimed. Now, if $\phi_3^0 - \phi_2^0 \approx (1 - g(1-\omega))^-$, then $\phi_2^1 \approx 1^-$, and so $\phi_3^2 - \phi_2^2 \approx (g(\omega))^+$. Here $\phi_2^2 = 0$. Consequently, $\phi_3^3 - \phi_2^3 = \phi_3^2 - ((\phi_2^3 - \phi_2^2) - (\phi_3^3 - \phi_3^2)) > g(\omega) - (g(\omega) - (1 - g(1-\omega))) = 1 - g(1-\omega)$. To prove (3.24) for $i = 1$, it remains to show that there exists an $\varepsilon > 0$ such that $\phi_3^3 - \phi_2^3 = 1 - g(1-\omega) + \varepsilon$ whenever (ϕ_2^0, ϕ_3^0) is near the boundary of Γ_1. To prove this, we need to make sure that $R(\phi_2^0, \phi_3^0)$ stay away from $1 - g(1-\omega)$ whenever (ϕ_2^0, ϕ_3^0) are near $\Gamma_1 \cap \Gamma_2$ and $\Gamma_1 \cap \Gamma_3$ (see Figure 9). Suppose that (ϕ_2^0, ϕ_3^0) is near the boundaries of Γ_1 and Γ_2. Then $\phi_1^2 \approx 1^-$. Thus, $\phi_3^3 - \phi_3^2 \approx g(2\omega) - g(\omega) = 1 - g(1-\omega) + \varepsilon$, where $\varepsilon > 0$. Similarly, if (ϕ_2^0, ϕ_3^0) is near the boundaries of Γ_1 and Γ_3, $\phi_3^3 - \phi_2^3$ is also bounded away from $1 - g(1-\omega)$. Hence, $\phi_3^3 - \phi_2^3$ is bounded away from $1 - g(1-\omega)$ whenever (ϕ_2^0, ϕ_3^0) is near the boundary of Γ_1. Similarly, one can prove that (3.24) holds for $i = 2, 3$. We have completed the first assertion of the proposition. The second assertion of the proposition is obvious. Suppose $\phi_n^0 - \phi_{n-1}^0 \approx (1 - g(1-\omega))^-$. Then $\phi_{n-1}^1 \approx 1^-$. Since $D(x)$ is increasing in x, $\phi_n^2 - \phi_{n-1}^2 \geq h(\omega) - h(0) = g(\omega)$. Inductively, we see that

$$\phi_n^n - \phi_{n-1}^n \geq g((n-1)\omega) - g((n-2)\omega) > 1 - g(1-\omega).$$

The second assertion of the proposition has been used repeatedly to justify the first inequality above. The second inequality above follows from (2.2). Therefore, (3.23) holds whenever ϕ_n^0 is sufficiently close to $1 - g(1-\omega)$. Hence, the direction of the flow of the system near the

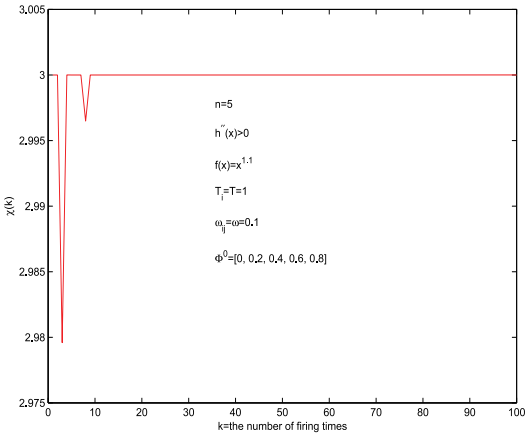


Figure 10. *For the choice of f, its phase responding function $h(x)$ is concave upward. The system in general does not synchronize as predicted in Proposition 3.8.3 and Theorem 3.9.2(b).*

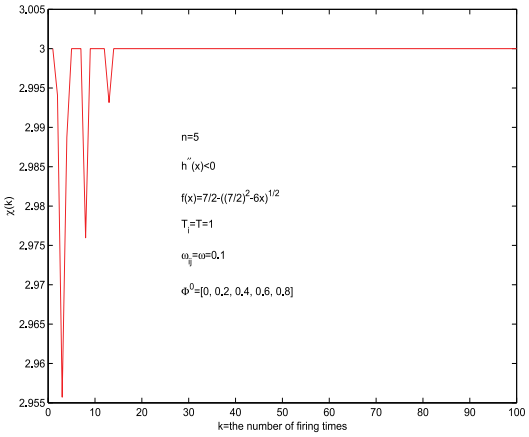


Figure 11. *For the choice of f, its phase responding function $h(x)$ is concave downward. Nevertheless, the system in general does not synchronize either.*

piece of boundary defined by $\phi_n^0 - \phi_{n-1}^0 = 1 - g(1-\omega)$ points inward. Similarly,

$$\phi_{n-i+1}^{i-1} - \phi_{n-i}^{i-1} < \phi_{n-i+1}^{n+i-1} - \phi_{n-i}^{n+i-1}, \qquad i = 2, \ldots, n,$$

whenever $\phi_{n-i+1}^{i-1} - \phi_{n-i}^{i-1}$ is close to $1 - g(1-\omega)$. We have just completed the proof of the proposition. ■

Two questions naturally arise from the proposition above. First, is the restriction $h''(x) > 0$ necessary for the validity of the second assertion of Proposition 3.8? Second, what kind of evolution maps with $f'' > 0$ satisfy the constraint $h''(x) > 0$? For the first question, we expect that the answer should be no (see Figures 10 and 11). However, we are unable to prove this.

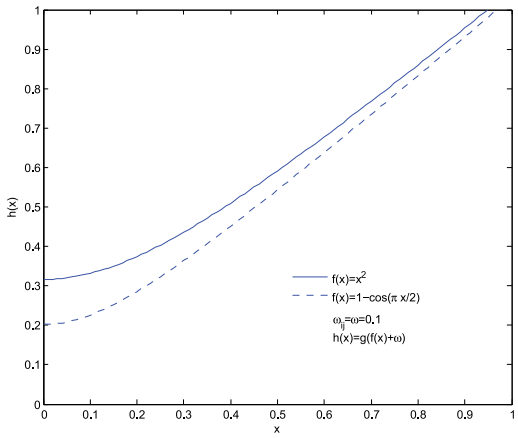


Figure 12. *Two graphs of $h(x)$ with two different f's are shown above. Their graphs are all concave upward.*

For the second question, we see in Figure 12, via the help of the computer, that $f(x) = x^r$, $r > 1$, and $f(x) = 1 - \cos(\pi x/2)$ satisfy $h''(x) > 0$.

We are now ready to state the main result of the paper.

Theorem 3.9.

1. *Suppose that stability condition (2.9) holds. Then the system of convex oscillators will achieve synchrony for all initial values, except possibly for those in a set of measure zero. In particular, the system of identical convex oscillators is to fire synchronously for all initial values, expect for those in a set of measure zero.*
2. (a) *The identical system with an even number of concave oscillators will not achieve full synchrony for certain initial values in a set of positive measure.*
   (b) *Suppose that the phase responding function $h(x)$ is concave upward. Then the identical system of concave oscillators will not synchronize for all initial values in the domain of its return map.*
   (c) *The identical system of three concave oscillators will not synchronize for all initial values in the domain of its return map.*
3. *Suppose that stability condition (2.9) and the absorption conditions (3.19) and (3.20c) are satisfied. Then the system of concave oscillators will achieve synchrony for all initial values.*

Proof. As shown in Theorem 3.5.1, the natural tendency of the system of convex oscillators is to grow by absorption regardless of their coupling strengths and speeds. Therefore, the system will continue to grow by absorption even though we need to update the new coupling strengths and speeds at each stage. The assertion of the first part of the theorem now follows. The third assertion of the theorem is now obvious. It remains to prove the second assertion of the theorem. To this end, let the number of oscillators be $2k$, and let ω and T be the constant coupling strength and constant cycle period, respectively. Pick $\Phi^0 = (\phi_1^0, \phi_2^0, \ldots, \phi_n^0)$

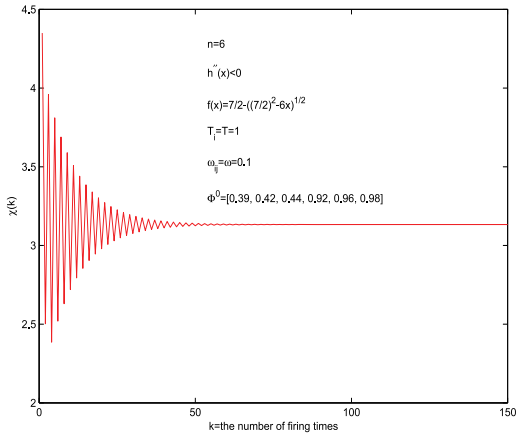


Figure 13. *The choice of f as above has the properties that $f''(x) > 0$ and $h''(x) < 0$. Since the number of oscillators chosen in this case is even, the numerical result demonstrated as above is consistent with the result of Theorem 3.9.2(a).*

to satisfy that

$$\phi_j^0 \in (1 - m_g\omega, 1),\ j = k+1, \dots, n, \text{ and} \tag{3.25}$$
$$\phi_1^0, \dots, \phi_k^0 \in (M_g(k+1)\omega - m_g\omega,\ M_g(k+1)\omega).$$

It then follows from Remark 2.1.2(a) that the system will reduce to two synchronous groups after initial firings. In fact, the first group contains oscillators $\phi_1^1, \dots, \phi_k^1$. The new coupling strengths for these two groups are equal. Denote by $\widetilde{\gamma}_{21}$ and $\widetilde{\gamma}_{12}$ the new corresponding γ_{21} and γ_{12}, respectively. Then $\widetilde{\gamma}_{21} = \widetilde{\gamma}_{12} = g(k\omega) + g(1 - k\omega) - 1 > 0$. Therefore, such a set of the initial values, which has a positive measure, will converge to a nonfiring state (see Figure 13). The assertions in 2(b) and 2(c) are now direct consequences of Proposition 3.8. We have just completed the second part of the theorem. ∎

The numerical stimulation suggests that a "nearly" identical system of any number of oscillators in general will not synchronize with or without the requirement that the phase responding curve be concave upward. Such a conjecture remains to be completed.

3.4. Examples and discussion. For the illustration of Theorem 3.9, the following three cases of systems of three oscillators are considered: (i) $f_i(x) = \sqrt{x}$, $\omega_{ij} = \omega$; (ii) $f_i(x) = x^{1.3}$ or $f_i(x) = \frac{7}{2} - \sqrt{(\frac{7}{2})^2 - 6x}$, $T_i = T$, and $\omega_{ij} = \omega$; (iii) $f_i(x) = x^r$, where $r > 1$, and $\omega_{ij} = \omega$.

Case (i): For this case, $m_g = \omega$, $m_f = \frac{1-\sqrt{1-\omega}}{\omega}$, $M_g = 2 - \omega$, and $M_f = \frac{1}{\sqrt{\omega}}$. Moreover, we have that $\frac{1}{m_g m_f} \geq M_g M_f$. Thus, as $n = 3$, equation (2.9) becomes

$$\frac{m_g^4 m_f^3 \omega}{(1 + m_f m_g + m_f^2 m_g^2) M_f M_g} \geq \Delta T(\Delta T + 1). \tag{3.26}$$

The corresponding feasible parameters region in $\omega - \Delta T$ is plotted in Figure 14. In the

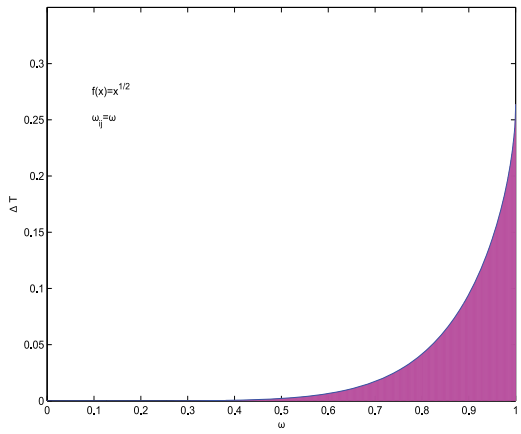


Figure 14. *The shaded part of the region is the set of parameters* $(\omega, \Delta T)$ *satisfying stability condition* (3.26).

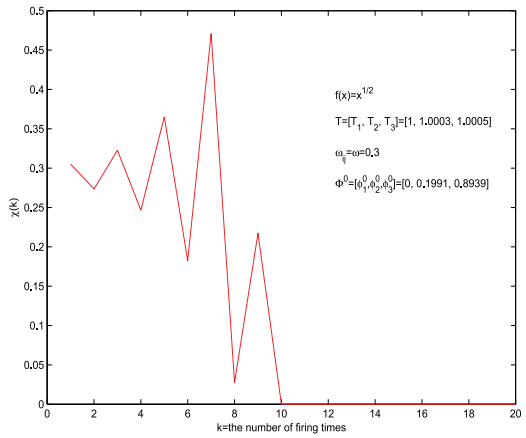


Figure 15. $\chi(t)$*, the synchronization order parameter, is defined in Figure* 3*. The parameters* ω_{ij} *and* T_i *are chosen so as to be from the stability region, Figure* 14*. With initial state being given as above, the system reaches full synchrony in* 10 *firings.*

numerical simulations, we pick randomly more than 20 sets of parameters with various sets of initial values; all the numerical results suggest the synchrony of the system. One such set of parameters and initial values and its corresponding numerical results are recorded in Figure 15.

Case (ii): The identical system is considered here. Let the number of oscillators be three. Figures 16 and 17 give the set of initial values not reaching synchrony, which contains the domain of the return map. Γ_3 is interpreted as $\phi_3^0 = 1$.

Case (iii): The case under consideration is the system of concave oscillators satisfying stability condition (2.9) and a modified absorption condition, which is stronger but easier to

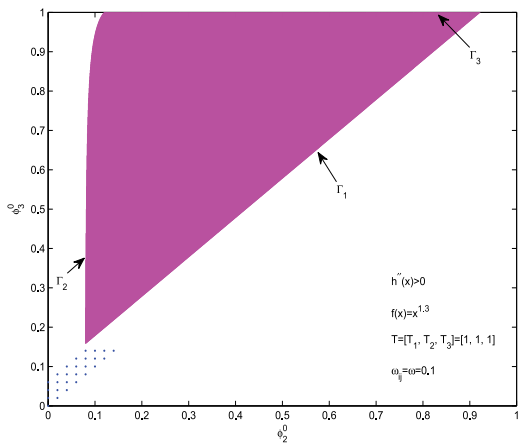


Figure 16. *The set of initial values reaching synchrony numerically is denoted by the dotted region. The points not in the dotted region, including the shaded region, will not acquire synchrony. In fact, the shaded region is the domain of the return map. This figure is consistent with the assertion of Theorems* 3.9.2(b) *and* 3.9.2(c).

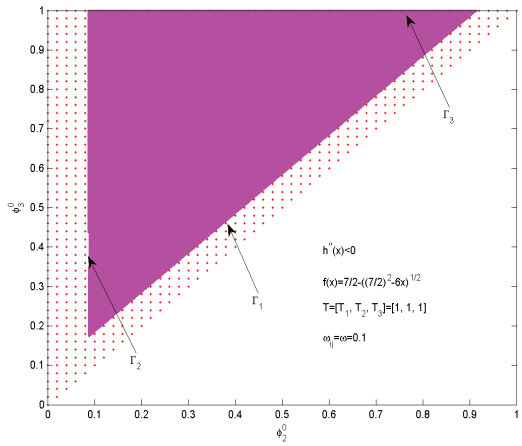
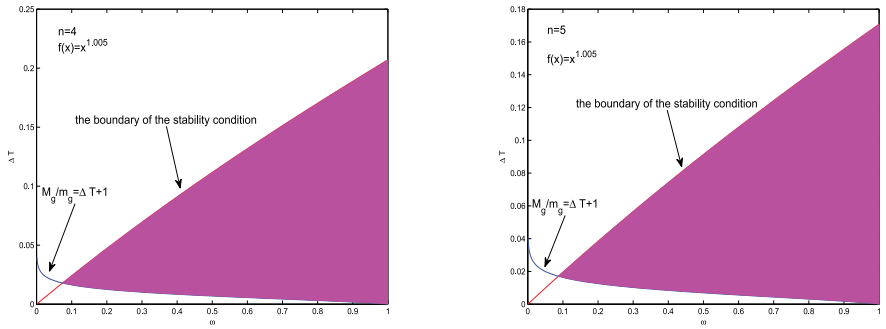


Figure 17. *The set of initial values reaching synchrony numerically is denoted by the dotted region. The points not in the dotted region, including the shaded region, will not acquire synchrony. In fact, the shaded region is the domain of the return map. This figure supports the assertion of Theorem* 3.9.2(c).

verify. Specifically, we consider the following absorption condition:

$$\frac{n}{n-2} > t_{\max} = \Delta T + 1 > \frac{M_g}{m_g}. \tag{3.27}$$

With such a stronger condition, the system will achieve full synchrony or reduce to two synchronous groups. However, in the case of the latter, to acquire full synchrony, the concavity of the evolution maps is still required to be sufficiently small. Numerically, we have that the



(a) The shaded part above is the region satisfied by both stability condition (2.9) and absorption conditions (3.27) for $n = 4$.

(b) The shaded part above is the region satisfied by both stability condition (2.9) and absorption conditions (3.27) for $n = 5$.

Figure 18.

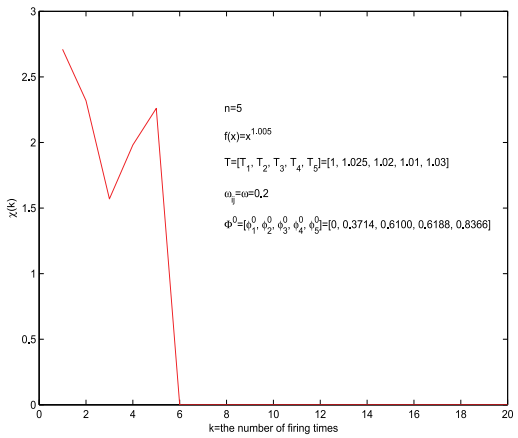


Figure 19. *Let $f(x) = x^{1.005}$. $\chi(t)$, the synchronization order parameter, is defined in Figure 3. The parameters ω_{ij} and T_i are so chosen to be in the stability region, Figure 18. Note that 5 is a prime number. Hence, when absorption occurs, the system breaks into a number of synchronous groups with their sizes being not all equal. Such an imbalance in coupling strength speeds the process of full synchrony. With initial state being given as above, the system reaches full synchrony in 6 firings.*

line $\left\{(\omega, \Delta T) : \Delta T = \frac{[\frac{n}{3}]+1}{[\frac{n}{3}]}\right\}$ does not intersect with the boundary of the stability condition. The parameter regions in the $\omega - \Delta T$ space satisfying (2.9) and (3.27) are, respectively, shown in the shaded regions in Figure 18(a) and (b). Picking the parameters from these regions, we see, in Figures 19 and 20, that the systems of both five and four concave oscillators reach full synchrony after a number of firings, provided that the concavity of the evolution maps is small. It should be mentioned that if n is a prime number, whenever the absorption occurs the system will acquire full synchrony in a short period of time. In this scenario, the imbalance in coupling strengths for the newly formed system is significant. In fact, it needs only six

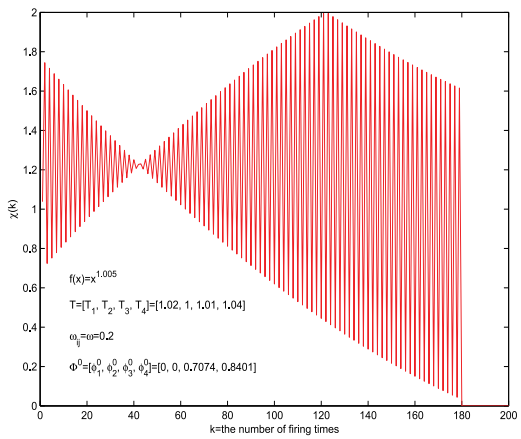


Figure 20. *With the evolution map, parameters, and initial state being given as above, the system reaches full synchrony in* 180 *firings. The reason that it takes so long for the system to synchronize is because* n *is an even number. When the absorption occurs, each of the synchronous groups may still have equal coupling strengths. Consequently, it takes longer for the system to synchronize since the imbalance in speed is insignificant.*

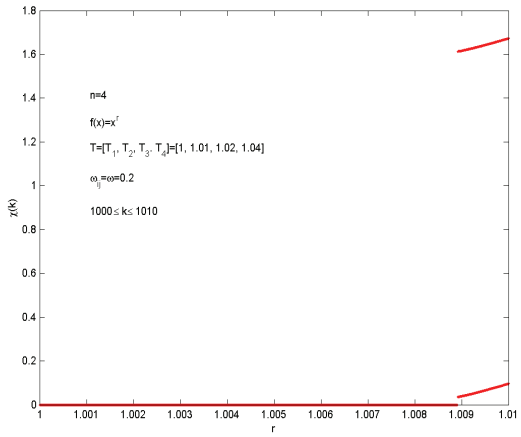


Figure 21. *The horizontal axis is the exponent of the evolution map of the form* $f(x) = x^r$, $r > 1$. *We plot* $\chi(k)$, $1000 \le k \le 1010$, *on the vertical axis.* 1000 *firings are needed to determine whether the corresponding system will achieve full synchrony or not. From the computer simulation, we see that the system will reach full synchrony, provided that* r *is roughly less than* 1.009.

firings to achieve full synchrony for $n = 5$. As for $n = 4$, the number of firings is 180 to secure synchrony. (See Figures 19 and 20, respectively.) To further support the validity of Proposition 3.7, we consider the evolution maps of the form $f(x) = x^r$, $r > 1$. The smaller r is, the smaller its concavity is. Treat r as a bifurcation parameter; Figures 21–23 show how we determine the smallest r that will make its corresponding system synchronize with various

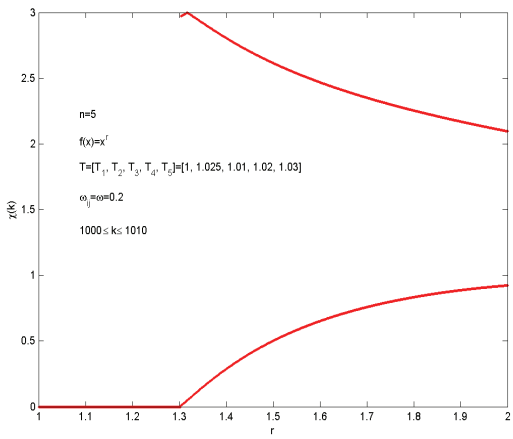


Figure 22. *Since* 5 *is a prime number, the imbalance measurement is "relatively" large if and when the system reduces to two synchronous groups. Therefore, the system is allowed to have a "larger" concavity at* $r \approx 1.3$.

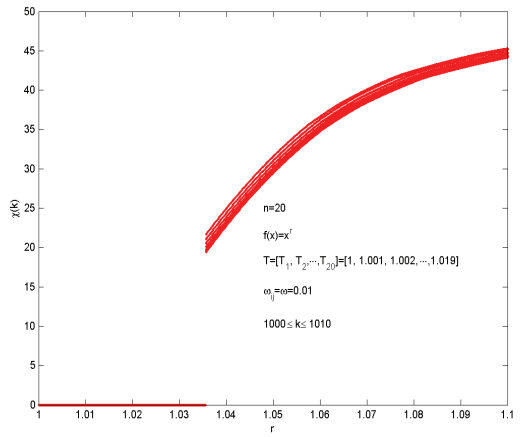


Figure 23. *With a greater number of oscillators present in the system, the computer simulation is consistent with the theory predicted in Proposition* 3.7.

choices of sizes of oscillators.

In conclusion, we prove stable synchrony for an integrate-and-fire model provided by Mirollo and Strogatz. Our results include the proof of Peskin's second conjecture. The next question is whether the results obtained here can be generalized to higher dimensional oscillators such as conductance-based models of neurons and/or phase-coupled networks via phase-response curves (see, e.g., [25] and the work cited therein). Note that the system presented here is just a special case for the phase-response curves approach. Nevertheless, the key ingredients for proving the full synchrony for those more current and advanced models should remain the same even though new technical difficulties might arise. For instance,

we still need to derive stability conditions so that the nonidentical system behaves like the identical system. We also need to have some kind of absorption conditions. For example, if the underlining model is dissipative, i.e., its time T-map decreases volume for all $T > 0$, then the natural tendency of the system would be to settle into a nonfiring state unless the direction of the flow of the "associated" return map points outward. If, on the other hand, the underlining model is volume-expanding, then the absorption process of the system tends to occur. It is certainly worthwhile to work on those problems.

Acknowledgments. The authors would like to thank the editor and referees for their helpful comments.

REFERENCES

[1] L. F. Abbott, *A network of oscillators*, J. Phys. A, 23 (1990), pp. 3835–3859.
[2] L. F. Abbott and C. van Vreeswijk, *Asynchronous states in networks of pulse-coupled oscillators*, Phys. Rev. E, 48 (1993), pp. 1483–1490.
[3] V. N. Belykh, N. N. Verichev, L. J. Kocarev, and L. O. Chua, *Chua's Circuit: A Paradigm for Chaos*, World Scientific, Singapore, 1993.
[4] V. N. Belykh, I. V. Belykh, K. V. Nevidin, and M. Hasler, *Hierarchy and stability of partially synchronous oscillations of diffusively coupled dynamical systems*, Phys. Rev. E, 62 (2000), pp. 6332–6345.
[5] V. N. Belykh, I. V. Belykh, K. V. Nevidin, and M. Hasler, *Cluster synchronization in three-dimensional lattices of diffusively coupled oscillators*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 13 (2003), pp. 755–799.
[6] V. N. Belykh, I. V. Belykh, and M. Hasler, *Connection graph stability method for synchronized coupled chaotic systems*, Phys. D, 195 (2004), pp. 159–187.
[7] I. Belykh, E. de Lange, and M. Hasler, *Synchronization of bursting neurons: What matters in the network topology*, Phys. Rev. Lett., 94 (2005), paper 188101.
[8] S. Bottani, *Pulse-coupled relaxation oscillators: From biological synchronization to self-organized criticality*, Phys. Rev. Lett., 74 (1995), pp. 4189–4192.
[9] A. Brailove, *The dynamics of two pulse-coupled relaxation oscillators*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 2 (1992), pp. 341–352.
[10] P. C. Bressloff and S. Coombes, *Synchrony in an array of integrate-and-fire neurons with dendritic structure*, Phys. Rev. Lett., 78 (1997), pp. 4665–4668.
[11] P. C. Bressloff and S. Coombes, *Desynchronization, mode locking, and bursting in strongly coupled integrate-and-fire oscillators*, Phys. Rev. Lett., 81 (1998), pp. 2168–2171.
[12] P. C. Bressloff and S. Coombes, *Symmetry and phase-locking in a ring of pulse-coupled oscillators with distributed delays*, Phys. D, 126 (1999), pp. 99–122.
[13] J. Buck, *Synchronous rhythmic flashing of fireflies.* II, Quart. Rev. Biol., 63 (1988), pp. 265–289.
[14] C. Chow, *Phase-locking in weakly heterogeneous neuronal networks*, Phys. D, 118 (1998), pp. 343–370.
[15] H. Daido, *Lower critical dimension for populations of oscillators with randomly distributed frequencies: A renormalization-group analysis*, Phys. Rev. Lett., 61 (1988), pp. 231–234.
[16] H. Daido, *Intrinsic fluctuation and its critical scaling in a class of populations of oscillators with distributed frequencies*, Progr. Theoret. Phys., 81 (1989), pp. 727–731.
[17] H. Daido, *Intrinsic fluctuations and a phase transition in a class of large populations of interacting oscillators*, J. Statist. Phys., 60 (1990), pp. 753–800.
[18] G. B. Ermentrout and N. Kopell, *Frequency plateaus in a chain of weakly coupled oscillators*, SIAM J. Math. Anal., 15 (1984), pp. 215–237.
[19] G. B. Ermentrout, *Synchronization in a pool of mutually coupled oscillators with random frequencies*, J. Math. Biol., 22 (1985), pp. 1–9.
[20] G. Ermentrout, *An adaptive model for synchrony in the firefly Pteroptyx malaccae*, J. Math. Biol., 29 (1991), pp. 571–585.

[21] U. Ernst, K. Pawelzik, and T. Geisel, *Delay-induced multistable synchronization of biological oscillators*, Phys. Rev. E, 57 (1998), pp. 2150–2162.
[22] W. Gerstner, J. L. van Hemmen, and J. D. Cowan, *What matters in neuronal locking*, Neural Computation, 8 (1996), pp. 1653–1676.
[23] W. Gerstner and W. M. Kistler, *Spiking Neuron Models. Single Neurons, Populations, Plasticity*, Cambridge University Press, Cambridge, UK, 2002.
[24] W. Gerstner, R. Ritz, and J. L. van Hemmen, *A biologically motivated and analytically soluble model of collective oscillations in the cortex:* I. *Theory of weak locking*, Biolog. Cybernet., 68 (1993), pp. 363–374.
[25] P. Goel and B. Ermentrout, *Synchrony, stability, and firing patterns in pulse-coupled oscillators*, Phys. D, 163 (2002), pp. 191–216.
[26] D. Hansel, G. Mato, and C. Meunier, *Synchrony in excitatory neural networks*, Neural Computation, 7 (1995), pp. 307–337.
[27] D. Hansel and G. Mato, *Existence and stability of persistent states in large neuronal networks*, Phys. Rev. Lett., 86 (2001), pp. 4175–4178.
[28] E. Izhikevich, *Class* 1 *neural excitability, conventional synapses, weakly connected networks, and mathematical foundations of pulse-coupled models*, IEEE Trans. Neural Networks, 10 (1999), pp. 499–507.
[29] J. Jalife, *Mutual entrainment and electrical coupling as mechanisms for synchronous firing of rabbit sinoatrial pacemaker cells*, J. Physiol., 356 (1984), pp. 221–243.
[30] J. Juang, C. L. Li, and Y. H. Liang, *Global synchronization in lattices of coupled chaotic systems*, Chaos, 17 (2007), paper 033111.
[31] J. Juang and Y.-H. Liang, *Synchronous chaos in coupled map lattices with general connectivity topology*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 755–765.
[32] N. Kopell and G. B. Ermentrout, *Mechanisms of phase-locking and frequency control in pairs of coupled neural oscillators*, in Handbook of Dynamical Systems, Vol. 3, Towards Applications, B. Fiedler, G. Iooss, and N. Kopell, eds., Elsevier, New York, 2000.
[33] Y. Kuramoto, *Self-entrainment of a population of coupled non-linear oscillators*, in International Symposium on Mathematical Problems in Theoretical Physics, Lecture Notes in Physics 39, H. Araki, ed., Springer, Berlin, 1975, pp. 420–422.
[34] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, New York, 1984.
[35] Y. Kuramoto and I. Nishikawa, *Statistical macrodynamics of large dynamical systems. Case of a phase transition in oscillator communities*, J. Statist. Phys., 49 (1987), pp. 596–605.
[36] Y. Kuramoto, *Collective synchronization of pulse-coupled oscillators and excitable units*, Phys. D, 50 (1991), pp. 15–30.
[37] J. Lü, X. Yu, and G. Chen, *Chaos synchronization of general complex dynamical networks*, Phys. A, 334 (2004), pp. 281–302.
[38] J. Lü, X. Yu, and G. Chen, *A time-varying complex dynamical network model and its controlling synchronization criteria*, IEEE Trans. Automat. Control, 50 (2005), pp. 841–846.
[39] P. C. Matthews and S. H. Strogatz, *Phase diagram for the collective behavior of limit-cycle oscillators*, Phys. Rev. Lett., 65 (1990), pp. 1701–1704.
[40] D. C. Michaels, E. P. Matyas, and J. Jalife, *Mechanisms of sinoatrial pacemaker synchronization: A new hypothesis*, Circulation Res., 61 (1987), pp. 704–714.
[41] R. E. Mirollo and S. H. Strogatz, *Synchronization of pulse-coupled biological oscillators*, SIAM J. Appl. Math., 50 (1990), pp. 1645–1662.
[42] T. Nishikawa, A. E. Motter, Y. C. Lai, and F. C. Hoppensteadt, *Heterogeneity in oscillator networks: Are smaller worlds easier to synchronize?*, Phys. Rev. Lett., 91 (2003), paper 014101.
[43] C. S. Peskin, *Mathematical Aspects of Heart Physiology*, Courant Institute of Mathematical Sciences, New York University, New York, 1975.
[44] A. Pogromsky and H. Nijmeijer, *Cooperative oscillatory behavior of mutually coupled dynamical systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 48 (2001), pp. 152–162.
[45] J. Rubin and D. Terman, *Geometric analysis of population rhythms in synaptically coupled neuronal networks*, Neural Computation, 12 (2000), pp. 597–645.
[46] J. Rubin and D. Terman, *Analysis of clustered firing patterns in synaptically coupled networks of oscillators*, J. Math. Biol., 41 (2000), pp. 513–545.

[47] J. Rubin and D. Terman, *Synchronized activity and loss of synchrony among heterogeneous conditional oscillators*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 146–174.
[48] A. Sherman, J. Rinzel, and J. Keizer, *Emergence of organized bursting in clusters of pancreatic beta-cells by channel sharing*, Biophys. J., 54 (1988), pp. 411–425.
[49] A. Sherman and J. Rinzel, *Collective properties of insulin secreting cells*, in Cell to Cell Signalling: From Experiments to Theoretical Models, A. Goldbeter, ed., Academic Press, London, 1989, pp. 61–75.
[50] S. Strogatz, *Norbert Wieners brain waves*, in Frontiers in Mathematical Biology, Lecture Notes in Biomathematics 100, Springer, Berlin, 1994, pp. 122–138.
[51] S. H. Strogatz, C. M. Marcus, R. M. Westervelt, and R. E. Mirollo, *Simple model of collective transport with phase slippage*, Phys. Rev. Lett., 61 (1988), pp. 2380–2383.
[52] S. H. Strogatz and R. E. Mirollo, *Phase-locking and critical phenomena in lattices of coupled nonlinear oscillators with random intrinsic frequencies*, Phys. D, 31 (1988), pp. 143–168.
[53] S. H. Strogatz and R. E. Mirollo, *Collective synchronisation in lattices of nonlinear oscillators with randomness*, J. Phys. A, 21 (1988), pp. L699–L705.
[54] D. Terman, N. Kopell, and A. Bose, *Dynamics of two mutually coupled slow inhibitory neurons*, Phys. D, 117 (1998), pp. 241–275.
[55] V. Torre, *A theory of synchronization of heart pace-maker cells*, J. Theoret. Biol., 61 (1976), pp. 55–71.
[56] M. Tsodyks, I. Mitkov, and H. Sompolinsky, *Pattern of synchrony in inhomogeneous networks of oscillators with pulse interactions*, Phys. Rev. Lett., 71 (1993), pp. 1280–1283.
[57] R. Urbanczik and W. Senn, *Similar nonleaky integrate-and-fire neurons with instantaneous couplings always synchronize*, SIAM J. Appl. Math., 61 (2000), pp. 1143–1155.
[58] C. van Vreeswijk, L. Abbott, and G. Ermentrout, *When inhibition not excitation synchronizes neural firing*, J. Comp. Neurosci., 1 (1994), pp. 313–321.
[59] C. van Vreeswijk, *Partially synchronized states in networks of pulse-coupled neurons*, Phys. Rev. E, 54 (1996), pp. 5522–5537.
[60] A. T. Winfree, *Biological rhythms and the behavior of populations of coupled oscillators*, J. Theoret. Biol., 16 (1967), pp. 15–42.
[61] A. T. Winfree, *The Geometry of Biological Time*, Springer-Verlag, New York, 1980.
[62] C. W. Wu, *Synchronization in Coupled Chaotic Circuits and Systems*, World Sci. Ser. Nonlinear Sci. Ser. A Monogr. Treatises 41, World Scientific, Singapore, 2002.
[63] J. Yang, G. Hu, and J. Xiao, *Chaos synchronization in coupled chaotic oscillators with multiple positive Lyapunov exponents*, Phys. Rev. Lett., 80 (2003), pp. 496–499.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1477–1506
© 2008 Society for Industrial and Applied Mathematics

Algorithms for Rigorous Entropy Bounds and Symbolic Dynamics*

Sarah Day†, Rafael Frongillo‡, and Rodrigo Treviño§

Abstract. The aim of this paper is to introduce a method for computing rigorous lower bounds for topological entropy. The topological entropy of a dynamical system measures the number of trajectories that separate in finite time and quantifies the complexity of the system. Our method relies on extending existing computational Conley index techniques for constructing semiconjugate symbolic dynamical systems. Besides offering a description of the dynamics, the constructed symbol system allows for the computation of a lower bound for the topological entropy of the original system. Our overall goal is to construct symbolic dynamics that yield a high lower bound for entropy. The method described in this paper is algorithmic and, although it is computational, yields mathematically rigorous results. For illustration, we apply the method to the Hénon map, where we compute a rigorous lower bound of 0.4320 for topological entropy.

Key words. topological entropy, symbolic dynamics, Conley index, Hénon map, computer-assisted proof

AMS subject classifications. 37B10, 37B40, 37B30, 37C25, 37M99

DOI. 10.1137/070688080

1. Introduction.

There has been a significant increase in computer-assisted proofs in dynamical systems in the past ten years. Many of these studies use topological techniques and carry at heart ideas introduced by Conley [Con78] as well as extensions derived from them. Conley's ideas, which were generalizations of Morse's theory for gradient-like flows, have spawned two computational approaches for studying complicated dynamics in discrete dynamical systems. The first is the method of *correctly aligned windows* (also known as the method of *covering relations*), which traces its roots to work on *windows* introduced by Easton in [Eas75]. In this paper, we exploit the algorithmic nature of a second approach that relies on the more general tools of Conley index theory. While many of the algorithms for this approach were introduced in earlier works (see, e.g., [Szy95], [DJM04], [Day03] and references therein), it was necessary to develop additional techniques and algorithms for this project. In particular, we describe in section 3.1 extended techniques for locating a region of the domain to be used for computations, and present in section 3.2 a newly developed automated procedure for taking a computed Conley index and producing an appropriate representative symbolic dynamical system.

We use the computational techniques based on Conley index theory to build a semiconjugacy from a map $f : S \to S$, $S \subset \mathbb{R}^n$, to a symbolic dynamical system and obtain a

*Received by the editors April 12, 2007; accepted for publication (in revised form) by H. Kokubu July 24, 2008; published electronically December 3, 2008. This work was made possible through the support of the Cornell University Summer 2006 REU program.
http://www.siam.org/journals/siads/7-4/68808.html
†Department of Mathematics, The College of William and Mary, P. O. Box 8795, Williamsburg, VA 23187-8795 (sday@math.wm.edu).
‡Computer Science Division, University of California at Berkeley, Berkeley, CA 94720 (raf@cs.berkeley.edu).
§Department of Mathematics, University of Maryland, College Park, MD 20742-4015 (rodrigo@math.umd.edu).

corresponding lower bound on the topological entropy (one measure of complexity) for the system. Since the symbols we use to construct the symbolic dynamics correspond to disjoint regions of the phase space $\mathbb{R}^n$, one benefit of this approach is that the symbolic dynamics offers a description of the dynamics on S, including information about the location of points along trajectories in S. Furthermore, the symbolic dynamics acts as a lower bound (via the semiconjugacy) for the dynamics of f on S; for any trajectory in the symbolic system there is at least one corresponding trajectory of f in S. It follows, as stated in Theorem 2.7, that the topological entropy of the symbolic system is a lower bound for the topological entropy of f. Since our goal is to compute a high lower bound, our approach relies on trying to maximize the complexity of the constructed symbolic system. We discuss our main approach for maximizing the complexity of the constructed system in section 3.

Topological entropy is a measurement that many have studied (see, e.g., [NBGM08], [AAC90], [ACE+87], [Col02], [Gal02]) using a variety of techniques. We see as the two main strengths in our approach the automation of our techniques and their independence from the typical constraint that stable and unstable manifolds are one-dimensional and restricted to the plane. Indeed, results in [DJM04] lead to entropy bounds for the infinite-dimensional Kot–Schaffer map in a way similar to the work described here, and in future work we plan to apply the automated techniques introduced in section 3 to this map to improve the bounds. In this paper, we apply our approach to the well-studied Hénon map in order to obtain results to compare with previous work in this area. We use our automated computational approach based on the ideas outlined above to construct a semiconjugacy between the dynamics on an appropriate subset of the Hénon attractor and a constructed symbolic dynamical system. Based on this construction, we give a rigorous lower bound of 0.4320 on the topological entropy of the Hénon map in Theorem 4.2. Section 4 also contains a comparison of this sample result with other work in this area.

This paper is organized in the following way. In section 2, we review the necessary background from dynamical systems and computational Conley index theory. Section 3 contains a detailed description of our extensions of this work to produce automated procedures for constructing complicated semiconjugate symbolic dynamics. Finally, in section 4 we apply these procedures to give sample results for the Hénon system.

2. Background.

In this section we review some basic definitions and ideas from dynamical systems theory and computational Conley index theory. We will state definitions and theorems which are relevant to our work, and refer the reader to [Rob95], [Con78], [MM02] and references therein for further development and details.

2.1. Symbolic dynamics and topological entropy.

Let $f : X \to X \subset \mathbb{R}^n$ be a continuous map. We will focus on maps that exhibit complicated dynamics on a compact subset $S \subset X$. Because the study of such maps and sets can be very complicated, they are often studied via a representation on a symbol space, giving rise to *symbolic dynamics*.

We focus on symbolic dynamics in the form of *subshifts of finite type*. Fix an integer $m \geq 2$, and let T be an $m \times m$ matrix with entries $t_{ij} \in \{0, 1\}$. The corresponding symbol space is

$$\Sigma_T := \{\mathbf{s} = (s_0 s_1 \ldots) \mid t_{s_k s_{k+1}} = 1 \text{ for all } k\}.$$

Although the matrix T is often referred to as the *adjacency matrix* in graph theory literature,

we will refer to T as the *symbol transition matrix* since it captures the allowed or admissible "transitions" between symbols. Finally, we define the *shift map* $\sigma_T : \Sigma_T \to \Sigma_T$ by

$$\sigma_T(\mathbf{s}) := \mathbf{s}', \quad \text{where } s'_i = s_{i+1}.$$

In this framework, (Σ_T, σ_T) is called a *subshift of finite type*, denoting both that we are working with only a finite list of (m) symbols and that only a subset of the set of all sequences on these m symbols is allowed by the symbol transition matrix T.

It is important to note that for an appropriate choice of metric on Σ_T, σ_T is a continuous map and $\sigma_T : \Sigma_T \to \Sigma_T$ is a dynamical system (see, e.g., [Rob95]). Subshifts of finite type are particularly nice in that dynamical objects of interest are often readily identifiable. For example, if one is looking for a period n orbit, then one checks that there is a symbol sequence $\mathbf{s}^* = (s_0, s_1, \ldots) \in \Sigma_T$ such that $s_{i+n} = s_i$ for all $i = 0, 1, \ldots$. If we view T as an adjacency matrix defining a directed graph, then $\mathbf{s}^*$ corresponds to an n-cycle, or cycle of length n, in the graph. For clarity, we include the following definition of the terms *cycle* and *simple cycle*.

Definition 2.1. *A* path *of length n in the directed graph G is a sequence of vertices $v_0, v_1, \ldots, v_n$ such that each pair (v_i, v_{i+1}) is an edge in G. If, in addition, $v_0 = v_n$, then $v_0, v_1, \ldots, v_n$ is a* cycle *of length n. Finally, a cycle $v_0, v_1, \ldots, v_n$ is a* simple cycle *provided that it contains no shorter cycles, namely, $v_i = v_j$ with $i \neq j$ if and only if $i, j \in \{0, n\}$.*

While subshifts of finite type and symbolic dynamical systems in general are nice to work with from a mathematical point of view, many interesting dynamical systems do not come in this form. Instead, as mentioned above, we may seek to represent a more general discrete dynamical system by symbolic dynamics. This representation often comes in the form of a *topological conjugacy* or *topological semiconjugacy*.

Definition 2.2. *A continuous map $\rho : X \to Y$ is a* topological semiconjugacy *between $f : X \to X$ and $g : Y \to Y$ if $\rho \circ f = g \circ \rho$ and ρ is surjective (onto). If, in addition, ρ is injective (one-to-one), then ρ is a* topological conjugacy.

Topological conjugacies preserve many properties of dynamical systems. One such example is the following theorem. (For more details, see [Dev89].)

Theorem 2.3. *Let ρ be a topological conjugacy between $f : X \to X$ and $g : Y \to Y$. Then $y \in Y$ is a periodic point of period n under g (i.e., $g^n(y) = y$) if and only if $\rho^{-1}(y)$ is a periodic point of period n under f.*

If f is topologically conjugate to a subshift of finite type, then we have a convenient list of trajectories of f given by the subshift. Indeed, in this case, the topological conjugacy acts as a coordinate transformation of the original system onto a decipherable (symbolic) system. In practice, such a complete description may be beyond our reach, and we instead construct topological semiconjugacies to appropriate subshifts of finite type. As illustrated by Theorem 2.7, these semiconjugate subshift systems offer lower bounds for the complexity of the dynamics of the original system.

One way to quantify the complexity of a given dynamical system is to compute its *topological entropy*. The following is based on Bowen's definition of topological entropy in [Bow71].

Definition 2.4. *Let $f : X \to X$ be a continuous map. A set $W \subset X$ is called (n, ϵ, f)-separated if for any two different points $x, y \in W$ there is an integer j with $0 \leq j < n$ so that the distance between $f^j(x)$ and $f^j(y)$ is greater than ϵ. Let $s(n, \epsilon, f)$ be the maximum*

cardinality of any (n, ϵ, f)*-separated set. The* topological entropy *of* f *is the number*

$$h_{top}(f) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{\log(s(n, \epsilon, f))}{n}. \tag{2.1}$$

As a measurement of chaos, we say that a map f for which $h_{top}(f) > 0$ is chaotic, and, if $h_{top}(f) > h_{top}(g)$, then f is "more chaotic" than g.

Once again, we can turn to symbolic dynamics in order to perform concrete computations.

Theorem 2.5 (see Robinson [Rob95]). *Let* T *be a symbol transition matrix, and let* $\sigma_T : \Sigma_T \to \Sigma_T$ *be the associated subshift of finite type. Then*

$$h_{top}(\sigma_T) = \log(sp(T)),$$

where $sp(T)$ *is the spectral radius of* T*.*

In essence, (n, ϵ, σ_T)-separation is encoded in the representation of the system and may be computed directly from the symbol transition matrix T.

Computing the topological entropy of a system not given as a subshift proves to be more challenging. In this setting and from a computational perspective, (2.1) may appear daunting. For one thing, sensitive dependence on initial conditions, a property commonly associated with chaotic systems, makes careful, precise measurements of (n, ϵ, f)-separation for large n and small ϵ difficult if not impossible. One technique for dealing with this problem is to focus on computing periodic points up to some cut-off period N rather than length N segments of general trajectories. The problem of finding periodic points may be reduced to finding fixed points for a sufficiently high iterate of the map, and two different periodic orbits of period n are necessarily (n, ϵ, f)-separated for sufficiently small ϵ. One then checks that

$$\left\{ \frac{\log(\#\{\text{periodic points of period } n\})}{n} \right\}_{n \leq N}$$

appears to be converging. Galias employed this approach in his study of the Hénon map in [Gal01]. The question now becomes, "is N sufficiently large to yield a good approximation for topological entropy?" This leads us to a second fundamental obstacle to a mathematically rigorous computational approach—the need to obtain asymptotic measurements in both n and ϵ. In Theorem 2.7 we use a special construction of a semiconjugacy to a subshift system to overcome these difficulties and to compute a rigorous lower bound.

This construction relies on tools from *Conley index theory* discussed in section 2.2. We use these tools to build the subshift system with the *itinerary function* serving as the semiconjugacy linking the systems.

Definition 2.6. *Suppose that* $N \subset X$ *may be decomposed into* $m < \infty$ *disjoint, closed subsets* $(N = \cup_{i=1,\ldots,m} N_i)$*. Let* S *be the maximal invariant set in* N *(i.e.,* S *is the largest set such that* $S \subset N$ *and* $f(S) = S$*). Then* $f^j(S) \subset N$ *for all* $j = 0, 1, \ldots$*. Finally, let* T *be the* $m \times m$ *symbol transition matrix given by*

$$t_{ij} = \begin{cases} 1 & \text{if } f(S \cap N_j) \cap N_i \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

The itinerary function $\rho : S \to \Sigma_T$ *is given by* $\rho(x) = s_0 s_1 \ldots$, *where* $s_j = i$ *for* $f^j(x) \in N_i$. *This function is continuous under the appropriate choice of metrics. (See* [Dev89], [Rob95] *for more details.)*

Finally, the following theorem allows us to use this semiconjugacy to obtain a lower bound for the topological entropy of the system under study.

Theorem 2.7. *Suppose that the itinerary function* ρ *is a semiconjugacy from* $f : S \to S$ *to* $\sigma_T : \Sigma_T \to \Sigma_T$ *for some* $S \subset X$ *and subshift of finite type* (σ_T, Σ_T) *with symbol transition matrix* T. *Then*

$$h_{top}(f) \geq \log(sp(T)) = h_{top}(\sigma_T),$$

where $sp(T)$ *is the spectral radius of* T.

Proof. Let $d(N_i, N_j) := \min_{x \in N_i,\, y \in N_j} d(x, y) > 0$ be the minimal distance between the two disjoint, closed sets N_i and N_j. Since there are only a finite number of these sets, $\epsilon_* := \min_{1 \leq i \neq j \leq m} d(N_i, N_j) > 0$.

For $\mathbf{s} = (s_0, s_1, \ldots) \in \Sigma_T$, call the sequence of n symbols, $B_n := (s_0, \ldots, s_{n-1})$, an *admissible* n*-block under* T. For each admissible n-block $B_n = (s_0, \ldots, s_{n-1})$, choose $x_{B_n} \in S$ such that $\rho(x_{B_n}) = (s_0, s_1, \ldots, s_{n-1}, s_n, \ldots) \in \Sigma_T$. Such a point exists in S since ρ maps S onto Σ_T. Furthermore, for $\epsilon < \epsilon_*$, the points chosen in S corresponding to two different admissible n-blocks must be (n, ϵ, f)-separated since, within n iterates, their itineraries carry them to two disjoint subsets of $S \cap N$, separated by a distance of at least ϵ_*.

We now have that for $\epsilon < \epsilon_*$, $s(n, \epsilon, f) \geq \#\{\text{admissible } n\text{-blocks under } T\}$. The asymptotic size of the set of admissible n-blocks may be computed as follows (see Theorem 1.9(b) in [Rob95]), to obtain the desired result:

$$\begin{aligned} h_{top}(f) &= \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{\log(s(n, \epsilon, f))}{n} \\ &\geq \limsup_{n \to \infty} \frac{\log(\#\{\text{admissible } n\text{-blocks under } T\})}{n} \\ &= \log(sp(T)) \\ &= h_{top}(\sigma_T). \qquad \blacksquare \end{aligned} \tag{2.2}$$

Thus, we may bound the topological entropy of a map f from below by finding a semiconjugacy from f to an appropriate subshift of finite type. The higher the spectral radius of the symbol transition matrix T, the better the lower bound we achieve for the topological entropy of the original system.

2.2. Conley index theory. We now present some of the topological tools used to build the subshift of finite type required for Theorem 2.7. These tools are based on Conley index theory for which we now give definitions, facts, and theorems which are relevant to our work. A discussion of the implementation of these ideas in a computational framework follows in section 2.4.

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous map. A *trajectory through* $x \in \mathbb{R}^n$ is a sequence

$$\gamma_x := (\ldots, x_{-1}, x_0, x_1, \ldots) \tag{2.3}$$

such that $x_0 = x$ and $x_{n+1} = f(x_n)$ for all $n \in \mathbb{Z}$. We now define the *invariant set relative to* $N \subset \mathbb{R}^n$ as

$$Inv(N, f) := \{x \in N \mid \text{there exists a trajectory } \gamma_x \text{ with } \gamma_x \subset N\}. \tag{2.4}$$

One example of a relative invariant set is the domain $S = Inv(N, f)$ on which we defined the itinerary function ρ in Definition 2.6.

We are now ready to present some of the basic structures in Conley index theory.

Definition 2.8. *A compact set $N \subset \mathbb{R}^n$ is an* isolating neighborhood *if*

$$Inv(N, f) \subset int(N), \tag{2.5}$$

where $int(N)$ denotes the interior of N. S is an isolated invariant set *if $S = Inv(N, f)$ for some isolating neighborhood N.*

We use the next two definitions to encode the dynamics on an isolating neighborhood.

Definition 2.9. *Let $P = (P_1, P_0)$ be a pair of compact sets with $P_0 \subset P_1 \subset X$. The map induced on the pointed quotient space $(P_1/P_0, [P_0])$ is*

$$f_P(x) := \begin{cases} f(x) & \text{if } x, f(x) \in P_1 \setminus P_0, \\ [P_0] & \text{otherwise.} \end{cases} \tag{2.6}$$

Definition 2.10 (see [RS88]). *The pair of compact sets $P = (P_1, P_0)$ with $P_0 \subset P_1 \subset X$ is an* index pair *for f, provided that*

1. *the induced map, f_P, is continuous,*
2. *$\overline{P_1 \setminus P_0}$, the closure of $P_1 \setminus P_0$, is an isolating neighborhood.*

In this case, we say that P is an index pair for the isolated invariant set $S = Inv(\overline{P_1 \setminus P_0}, f)$.

The following definition is required for the definition of the Conley index.

Definition 2.11. *Two group homomorphisms, $\phi : G \to G$ and $\psi : G' \to G'$ on abelian groups G and G', are* shift equivalent *if there exist group homomorphisms $r : G \to G'$ and $s : G' \to G$ and a constant $m \in \mathbb{N}$ (referred to as the "lag") such that*

$$r \circ \phi = \psi \circ r, \quad s \circ \psi = \phi \circ s, \quad r \circ s = \psi^m, \quad \text{and} \quad s \circ r = \phi^m.$$

The shift equivalence class of ϕ, *denoted $[\phi]_s$, is the set of all homomorphisms ψ such that ψ is shift equivalent to ϕ.*

Definition 2.12. *Let $P = (P_1, P_0)$ be an index pair for the isolated invariant set $S = Inv(\overline{P_1 \setminus P_0}, f)$, and let $f_{P*} : H_*(P_1, P_0) \to H_*(P_1, P_0)$ be the map induced on the relative homology groups $H_*(P_1, P_0)$ from the map f_P. The* Conley index of S *is the shift equivalence class of f_{P*},*

$$Con(S, f) := [f_{P*}]_s. \tag{2.7}$$

The Conley index for the isolated invariant set S given in Definition 2.12 is well defined; namely, every isolated invariant set has an index pair, and the corresponding shift equivalence class remains invariant under different choices for this index pair (see, e.g., [MM02]).

So far we have passed from continuous maps to induced maps on relative homology. Our overall goal, however, is to describe the dynamics of our original map. Here we present

measurements based on the map on homology that may give us information about the original map. The first theorem is Ważewski's principle in the context of Conley index theory.

Theorem 2.13. *If* $Con(S, f) \neq [0]_s$, *then* $S \neq \emptyset$.

By requiring additional structure in the isolating neighborhood N of S, we can use a modification of Theorem 2.13 to study finer structure in S.

Corollary 2.14. *Let* $N \subset X$ *be the union of disjoint compact sets* $N_1, \ldots, N_m$, *and let* $S := Inv(N, f)$ *be the isolated invariant set relative to* N. *Let*

$$S' = Inv(N_1, f_{N_n} \circ \cdots \circ f_{N_1}) \subset S,$$

where f_{N_i} *denotes the restriction of the map* f *to the region* N_i. *If*

$$Con(S', f_{N_n} \circ \cdots \circ f_{N_1}) \neq [0]_s, \tag{2.8}$$

then $S' \neq \emptyset$. *More specifically, there exists a point in* S *whose trajectory under* f *travels through the regions* $N_1, \ldots, N_n$ *in the prescribed order.*

We here note that, given the hypotheses of Corollary 2.14, there is a nice technique for obtaining the index of S' given the computed index map f_{P*}, where $P = (P_1, P_0)$ is an index pair with $N = \overline{P_1 \setminus P_0}$. Using an approach developed by Szymczak in [Szy95], we set

$$f_P^{ij}(x) := \begin{cases} f(x) & \text{if } x \in N_i \text{ and } f(x) \in N_j, \\ [P_0] & \text{otherwise.} \end{cases} \tag{2.9}$$

Then $f_{P*}^{ij} : H_*(P_1, P_0 \cup (\cup_{l \neq i} N_l)) \to H_*(P_1, P_0 \cup (\cup_{l \neq j} N_l))$. Given f_{Pk} in matrix form representing the linear map on $H_k(P_1, P_0)$, we may label the columns/rows by location of the associated relative homology generators in the subgroups $H_k(P_1, P_0 \cup (\cup_{l \neq 1} N_l)), \ldots,$ $H_k(P_1, P_0 \cup (\cup_{l \neq n} N_l))$. To simplify notation, we say that generator g is in region N_i if $g \in H_k(P_1, P_0 \cup (\cup_{l \neq i} N_l))$. Then f_{Pk}^{ij} is the $n_j \times n_i$ submatrix with n_i columns corresponding to the n_i generators in N_i and n_j rows corresponding to the n_j generators in N_j. Furthermore, $(P_1, P_0 \cup (\cup_{l \neq 1} N_l))$ is an index pair for the isolated invariant set $S' = Inv(N_1, f_{N_n} \circ \cdots \circ f_{N_1})$ with index map $f_{P*}^{n1} \circ \cdots \circ f_{P*}^{12} : H_*(P_1, P_0 \cup (\cup_{l \neq 1} N_l)) \to H_*(P_1, P_0 \cup (\cup_{l \neq 1} N_l))$. Therefore,

$$Con(S', f_{N_n} \circ \cdots \circ f_{N_1}) = [f_{P*}^{n1} \circ \cdots \circ f_{P*}^{12}]_s. \tag{2.10}$$

Since the more general problem of determining whether the linear map $f_{Pk} : H_k(P_1, P_0) \to H_k(P_1, P_0)$ is not shift equivalent to 0 may be difficult, we here focus on a computable sufficient condition. Trace is preserved by shift equivalence, and we adopt the notation

$$\mathrm{tr}_k(Con(S, f)) := \mathrm{tr}(f_{Pk}),$$

where $\mathrm{tr}(f_{Pk})$ denotes the trace of the linear map $f_{Pk} : H_k(P_1, P_0) \to H_k(P_1, P_0)$. Then if $\mathrm{tr}_k(Con(S, f)) \neq 0$ for some k, $Con(S, f) \neq [0]_s$.

Corollary 2.15. *If* $\mathrm{tr}_k(Con(S', f_{N_{i_n}} \circ \cdots \circ f_{N_{i_1}})) \neq 0$ *for some* k, *then there exists* $x \in S'$ *with* $\rho(x) = i_1 i_2 \ldots i_n i_1 i_2 \ldots i_n \ldots$.

Taking this approach, we are close to showing something stronger, namely that there is a periodic orbit under f with the corresponding cyclic symbol sequence. This stronger statement relies on computing the Lefschetz number.

Definition 2.16. *Let S be an isolated invariant set. The* Lefschetz number of S *is defined as*

$$L(S,f) := \sum_k (-1)^k \operatorname{tr}(f_{Pk}), \tag{2.11}$$

where $P = (P_1, P_0)$ is an index pair for S.

The Lefschetz number is essential to the following theorem and its corollary.

Theorem 2.17. *Let S be an isolated invariant set. If*

$$L(S,f) \neq 0, \tag{2.12}$$

then S contains a fixed point.

For a proof, see [Szy96]. As before, a refinement of the approach allows us to study symbolic dynamics.

Corollary 2.18. *Let $N \subset X$ be the finite union of disjoint compact sets $N_1, \ldots, N_m$, and let $S := Inv(N, f)$. Let $S' = Inv(N_1, f_{N_n} \circ \cdots \circ f_{N_1}) \subset S$, where f_{N_i} denotes the map f restricted to the region N_i. If*

$$L(S', f_{N_n} \circ \cdots \circ f_{N_1}) \neq 0, \tag{2.13}$$

then $f_{N_n} \circ \cdots \circ f_{N_1}$ contains a fixed point in S' that corresponds to a periodic point of period n in S that under f travels through the regions $N_1, \ldots, N_n$ in order.

In what follows, we will develop algorithms based on Corollary 2.15 to construct and verify symbolic dynamics. However, in the special case where the index map f_{P*} is nontrivial on exactly one level (as occurs with the Hénon map), we may use Corollary 2.18 to show that the constructed semiconjugate symbolic system has the added stronger property that every periodic orbit in the symbolic system corresponds to a periodic orbit in the original system of the same period.

2.3. Multivalued and combinatorial maps. Now that we have the relevant tools from Conley index theory, we can begin applying them algorithmically to extract information about the dynamical system $f : X \to X$. In this section, we describe the construction of a combinatorial representation of f. The first step is to define a *multivalued map* F that will be used to incorporate bounded errors in the representation.

Definition 2.19. *The* multivalued map $F : X \rightrightarrows X$ *is a map from X to its power set; i.e., for all $x \in X$, $F(x) \subset X$. If, for a continuous single-valued map $f : X \to X$, $f(x) \in F(x)$ and $F(x)$ is acyclic (i.e., has the homology of a point) for all $x \in X$, then f is a* continuous selector of F, *and F is an* enclosure *of f.*

In what follows, we discuss how to construct an enclosure of the map under study. The purpose of the enclosure is to incorporate round-off and other errors that occur in computations. This construction requires rigorous small error bounds in order to create an enclosure whose images are not so large as to obscure all relevant dynamics. Given an appropriate enclosure, the topological tools from section 2.2 may be used to uncover dynamics of the underlying map. Furthermore, there are algorithms for both the construction of the enclosure and the computation of the Conley index. These algorithms require a further step—discretizing the domain in order to store it in the computer as a finite list of objects.

We begin by using the subdivision procedure implemented in the software package *GAIO* [DFJ01] to create a grid $\mathcal{G}$ on a compact (rectangular) region in X. In practice, the region chosen for representation is usually determined either experimentally through nonrigorous numerical simulations or analytically given a special structure or symmetry for the system (e.g., a compact attracting region). We partition a specified rectangular set $W = \prod_{k=1}^{n} [x_k^-, x_k^+] \subset \mathbb{R}^n$ into a *cubical grid*

$$\mathcal{G}^{(d)} := \left\{ \prod_{k=1}^{n} \left[x_k^- + \frac{i_k r_k}{2^d},\, x_k^- + \frac{(i_k+1) r_k}{2^d} \right] \;\middle|\; i_k \in \left\{ 0, \ldots, 2^d - 1 \right\} \right\},$$

where $r_k = x_k^+ - x_k^-$ is the radius of W in the kth coordinate and the depth d is a nonnegative integer. We call an element of the grid, $B = \prod_{k=1}^{n} \left[x_k^- + \frac{i_k r_k}{2^d},\, x_k^- + \frac{(i_k+1) r_k}{2^d} \right]$, a *box*. For a collection of boxes, $G \subset \mathcal{G} = \mathcal{G}^{(d)}$, define the *topological realization* of G as $|G| := \cup_{B \in G} B \subset \mathbb{R}^n$.

Constructing a useful *combinatorial enclosure* involves bounding all round-off and other errors. In our study of the Hénon map in section 4, we construct a combinatorial enclosure by computing images of $f(|G|)$ using interval arithmetic software. This produces a bounding box, $\tilde{f}(|G|)$, for the image $f(|G|)$, which is then intersected with the grid $\mathcal{G}$ to produce the combinatorial enclosure image

$$\mathcal{F}(G) := \{ G' \in \mathcal{G} : |G'| \cap \tilde{f}(|G|) \neq \emptyset \}.$$

This combinatorial enclosure, $\mathcal{F} : \mathcal{G} \rightrightarrows \mathcal{G}$, yields an enclosure $F = |\mathcal{F}|$ of f in the following way: define $|\mathcal{F}| : W \rightrightarrows W$, where $W = \cup_{G \in \mathcal{G}} |G|$,

$$|\mathcal{F}|(x) := \bigcup_{G \in \mathcal{G} \text{ with } x \in |G|} |\mathcal{F}(G)|. \tag{2.14}$$

More importantly, efficient algorithms exist for computing isolating neighborhoods, index pairs, and Conley indices for f from an appropriate combinatorial enclosure $\mathcal{F}$ of f.

2.4. Computational Conley index theory. Now we give algorithms for computing the isolating neighborhoods, index pairs, and Conley indices first introduced in section 2.2 in the setting of combinatorial enclosures.

Definition 2.20. *A* combinatorial trajectory *of a combinatorial enclosure $\mathcal{F}$ through $G \in \mathcal{G}$ is a bi-infinite sequence $\gamma_G = (\ldots, G_{-1}, G_0, G_1, \ldots)$ with $G_0 = G$, $G_n \in \mathcal{G}$, and $G_{n+1} \in \mathcal{F}(G_n)$ for all $n \in \mathbb{Z}$.*

Definition 2.21. *The* combinatorial invariant set *in $\mathcal{N} \subset \mathcal{G}$ for a combinatorial enclosure $\mathcal{F}$ is*

$$Inv(\mathcal{N}, \mathcal{F}) := \{ G \in \mathcal{G} : \textit{there exists a trajectory } \gamma_G \subset \mathcal{N} \}.$$

Definition 2.22. *The* combinatorial neighborhood *of $\mathcal{B} \subset \mathcal{G}$ is*

$$o(\mathcal{B}) := \{ G \in \mathcal{G} : |G| \cap |\mathcal{B}| \neq \emptyset \}.$$

This set, $|o(\mathcal{B})|$, sometimes referred to as a *one box neighborhood of $\mathcal{B}$ in $\mathcal{G}$*, is the smallest representable neighborhood of $|B|$ in the grid $\mathcal{G}$.

While there are different characterizations of isolation in the setting of combinatorial enclosures, we chose the following for this work.

Definition 2.23. *If*

$$o(Inv(\mathcal{N},\mathcal{F})) \subset \mathcal{N},$$

then $\mathcal{N} \subset \mathcal{G}$ *is a* combinatorial isolating neighborhood *under* $\mathcal{F}$.

Note that, by construction, the topological realization $|\mathcal{N}|$ of a combinatorial isolating neighborhood $\mathcal{N}$ under $\mathcal{F}$ is an isolating neighborhood under any continuous selector $f \in |\mathcal{F}|$. This definition is stronger than what is actually required to guarantee isolation on the topological level. It is, however, the definition that we will use in this work and is computable using the following approach.

Let $\mathcal{S} \subset \mathcal{G}$. Set $\mathcal{N} = \mathcal{S}$, and let $o(\mathcal{N})$ be the combinatorial neighborhood of $\mathcal{N}$ in $\mathcal{G}$. If $Inv(o(\mathcal{N}),\mathcal{F}) = \mathcal{N}$, then $\mathcal{N}$ is isolated under $\mathcal{F}$. If not, set $\mathcal{N} := Inv(o(\mathcal{N}),\mathcal{F})$ and repeat the above procedure. In this way, we grow the set $\mathcal{N}$ until either the isolation condition is met or the set grows to intersect the boundary of $\mathcal{G}$, in which case the algorithm fails to locate an isolating neighborhood in $\mathcal{G}$. This procedure is outlined in more detail in the following algorithm from [DJM04].

Algorithm 1 (Grow isolating neighborhood).

```
INPUT:  grid 𝒢, combinatorial enclosure ℱ on 𝒢, set 𝒮 ⊂ 𝒢
OUTPUT: a combinatorial isolating neighborhood 𝒩 containing 𝒮
        or 𝒩 = ∅ if the isolation condition is not met
𝒩 = grow_isolating_neighborhood(𝒢, ℱ, 𝒮)
    𝒢_boundary := {G ∈ 𝒢 : |G| ∩ ∂|𝒢| ≠ ∅};
    𝒩 := 𝒮;
    while Inv(o(𝒩),ℱ) ⊄ 𝒩 and 𝒩 ∩ 𝒢_boundary = ∅,
        𝒩 := Inv(o(𝒩),ℱ);
    end
    if 𝒩 ∩ 𝒢_boundary = ∅, return 𝒩;
    else return ∅;
    end
```

Once we have an isolating neighborhood for f, our next goal is to compute a corresponding index pair. The following definition of a *combinatorial index pair* again emphasizes our goal of using the combinatorial enclosure to compute structures for f.

Definition 2.24. *A pair* $\mathcal{P} = (\mathcal{P}_1,\mathcal{P}_0)$ *of cubical sets is a* combinatorial index pair *for a combinatorial enclosure* $\mathcal{F}$ *if the corresponding topological realization* $P = (P_1,P_0)$*, where* $P_i := |\mathcal{P}_i|$*, is an index pair for any continuous selector* $f \in |\mathcal{F}|$*. Namely,* $\overline{P_1 \setminus P_0} = |\mathcal{P}_1 \setminus \mathcal{P}_0|$ *is an isolating neighborhood under* f*, and the map* f_P*, as defined in Definition 2.9, is continuous.*

The following algorithm produces a combinatorial index pair associated with a combinatorial isolating neighborhood produced via Algorithm 1. While there are other algorithms for producing combinatorial index pairs, this algorithm works well with later index computations. For more details, see the description of *modified combinatorial index pairs* in [Day03].

Algorithm 2 (Build index pair).

```
INPUT:  grid 𝒢, combinatorial enclosure ℱ on 𝒢,
        combinatorial isolating neighborhood 𝒩 produced by Algorithm 1
```

```
OUTPUT: combinatorial index pair P = (P_1, P_0) with P_1 \ P_0 = N
[P_1, P_0] = build_index_pair(G, F, N)
    P_0 := ∅;
    New := F(N) ∩ o(N) \ N;
    while New ≠ ∅
        P_0 := P_0 ∪ New;
        New := (F(P_0) ∩ o(N)) \ P_0;
    end
    P_1 := N ∪ P_0;
    return [P_1, P_0];
```

We now have an isolating neighborhood $|\mathcal{N}|$ and corresponding index pair $P := (|\mathcal{P}_1|, |\mathcal{P}_0|)$ for f. What remains in computing the Conley index for the associated isolated invariant set, $S := Inv(|\mathcal{N}|, f)$, is to compute the map $f_{P*} : H_*(|\mathcal{P}_1|, |\mathcal{P}_0|) \to H_*(|\mathcal{P}_1|, |\mathcal{P}_0|)$. Once again, the combinatorial enclosure offers the appropriate computational framework, and we use the software program *homcubes* in [Pil98] to compute f_{P*}. This step is outlined in Algorithm 3.

Algorithm 3 (Compute index map).

```
INPUT:  grid G, combinatorial enclosure F on G,
        combinatorial index pair P = (P_1, P_0) produced by Algorithm 2
OUTPUT: relative homology groups H_*(|P_1|, |P_0|),
        the induced index map f_{P*} : H_*(|P_1|, |P_0|) → H_*(|P_1|, |P_0|),
        and the induced submaps {f^{ij}_{Pk}} on connected components
[f_{P*} H_*(|P_1|, |P_0|) {f^{ij}_{Pk}}] = compute_index_map(G, F, P_1, P_0)
    Q_1 = F(P_1);
    Q_0 = F(P_0);
    [f_{P*} H_*(|P_1|, |P_0|) {f^{ij}_{Pk}}] := homcubes(P_1, P_0, Q_1, Q_0, F);
    return [f_{P*} H_*(|P_1|, |P_0|) {f^{ij}_{Pk}}];
```

Algorithm 3 produces a sequence of matrices for the maps $f_{P0}, f_{P1}, \ldots, f_{Pn}$, where n is the dimension of the phase space X. For $k > n$, $f_{Pk} = 0$. The associated Conley index is $Con_*(S) = [f_{P*}]_s$ for $S := Inv(|\mathcal{P}_1 \setminus \mathcal{P}_0|, f)$. The submaps $f^{ij}_{Pk} : H_k(|\mathcal{P}_1|, |\mathcal{P}_0| \cup_{l \neq i} |\mathcal{N}_l|) \to H_k(|\mathcal{P}_1|, |\mathcal{P}_0| \cup_{l \neq j} |\mathcal{N}_l|)$, where $|\mathcal{N}_1|, \ldots, |\mathcal{N}_n|$ are the connected components of $|\mathcal{P}_1 \setminus \mathcal{P}_0|$, are given as submatrices of f_{Pk}. These are the maps required for Corollaries 2.14, 2.15, and 2.18. In the following section, we describe an algorithmic procedure for using this index information to construct the appropriate subshift of finite type.

3. Constructing and verifying complicated symbolic dynamics. Given $f : X \to X$, the general method we adopt for computing a lower bound on topological entropy consists of the following steps:

- constructing a fixed cubical grid $\mathcal{G}$ on a subset of X and a combinatorial enclosure $\mathcal{F}$ of f on $\mathcal{G}$ (section 2.3),
- locating a region of interest $\mathcal{S}$ in $\mathcal{G}$ (section 3.1),
- computing the associated Conley index (section 2.4),
- constructing semiconjugate symbolic dynamics (section 3.2),
- using the constructed symbolic dynamical system to compute a lower bound on the topological entropy of f (Theorem 2.7).

While many steps of this general procedure have been carried out in previous work (e.g., [DJM04] for the first four steps, and [Gal01] for the last step), we here seek to uncover far more complicated symbolic dynamics. This requires a more automated approach based on setting verifiable conditions for uncovering and proving the existence of cyclic symbolic dynamics and ignoring or giving up on the verification of dynamics that does not satisfy these conditions. Along these lines, we now give algorithms for locating a region of interest (section 3.1) and processing the resulting index information (section 3.2) that allow us to uncover more complicated dynamics than previously found using related techniques. This improved procedure produces the entropy bounds presented in section 4.

3.1. Locating a region of interest. We now turn to the second task in this list—that of locating the region of the grid where we will attempt to compute interesting symbolic dynamics. More specifically, the set that we are calling the *region of interest* will serve as the input, $\mathcal{S}$, for Algorithm 1. We show three different methods for locating regions of interest for the Hénon map in sections 4.1, 4.2, and 4.3. In this section, we focus on the method that, of these three, both is general (i.e., is not restricted to studies of the Hénon map) and yields high entropy bounds. This is the method followed in section 4.2. The first step in this approach is similar in spirit to the work of Cvitanović and others in using periodic orbits of low periods to approximate chaotic attractors. We begin by finding short cycles in the combinatorial enclosure (directed graph) $\mathcal{F}$. These short cycles correspond to possible periodic orbits of low period for f. We then add a level of complexity by searching for paths in the directed graph between these short cycles. From a dynamics point of view, these paths represent possible mixing between the periodic regions.

We construct a list of short cycles in $\mathcal{G}$ by setting ourselves a computational parameter `Max_Cycle_Length` $\in \mathbb{Z}^+$ and locating the cycles in $\mathcal{F}$ of length k with $1 \leq k \leq$ `Max_Cycle_Length`. These cycles are nonzero entries on the diagonal of $\mathcal{F}$ (when viewed as a transition matrix) raised to the kth power. The corresponding computed vertices in $\mathcal{F}$ are the regions in $\mathcal{G}$ that may contain period k points of f. Starting with $\mathcal{S} = \emptyset$, we begin adding the short cycles to $\mathcal{S}$ one by one, starting with the shortest. Just before adding a cycle to $\mathcal{S}$, we grow its isolating neighborhood using Algorithm 1 and then check that this neighborhood does not intersect the isolating neighborhood of the current collection. This corresponds to a possible increase in the number of symbols and/or the number of admissible transitions between symbols in the resulting constructed symbolic system and may eventually lead to a higher entropy bound. If this condition is not met, we do not add the cycle but move to the next cycle in the list, continuing until the list is exhausted. We next use breadth first search (BFS) to find shortest path, pairwise connections between the short cycles in $\mathcal{S}$ and add these connecting paths to $\mathcal{S}$. This procedure is outlined in Algorithm 4.

Algorithm 4 (Locating region of interest/joining low cycles).

```
INPUT:      grid 𝒢, combinatorial enclosure ℱ on 𝒢,
            computational parameter Max_Cycle_Length
OUTPUT:     region of interest 𝒮 ⊂ 𝒢
𝒮 = find_and_connect_low_cycles(𝒢, ℱ, Max_Cycle_Length)
   𝒮 = ∅;
   𝒩 = ∅;
```

```
for n = 1 : Max_Cycle_Length,
     for each length n cycle c in F,
          N_c = grow_isolating_neighborhood(G, F, c);
          if N_c ∩ N = ∅,
               S = S ∪ c;
               N = grow_isolating_neighborhood(G, F, N ∪ c);
          end
     end
end
S_c := S;
for each vertex v_i ∈ S_c,
     for each vertex v_j ∈ S_c,
          γ = shortest_path in F from v_i to v_j in G\o(o(S));
          S := S ∪ γ;
     end
end
return S;
```

Here, we explicitly compute cycles with lengths up to `Max_Cycle_Length`, which in practice is small. However, we obtain many new cycles by adding pairwise connections between the computed cycles. This allows us to uncover complicated dynamics without having to explicitly search for the long cycles that correspond to periodic orbits of high period. As illustration, Figure 1 depicts a subshift of finite type constructed from a region of interest consisting of a length 2 cycle, two length 4 cycles, and pairwise shortest connecting paths between these three objects. Note that the resulting subshift system contains infinitely many cycles (of lengths 5, 8, 10, and higher) and positive topological entropy.

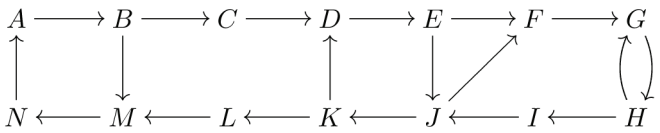


Figure 1. *Symbol transition graph constructed from a* 2*-cycle, two* 4*-cycles, and pairwise connections.*

While effective in computation of entropy bounds for the Hénon map (see section 4.2), this approach for the construction of the region of interest, $\mathcal{S}$, could be improved. Given a fixed combinatorial enclosure $\mathcal{F}$ on a grid $\mathcal{G}$, one goal would be to optimize the construction of $\mathcal{S}$ in order to produce a subshift of finite type with the highest possible entropy. As a first step along these lines, the relationship between the entropy bound and the maximal cycle length used in Algorithm 4 in a study of the Hénon map is depicted in Figure 5. In addition, there is a clear trade-off between refining the grid in order to find, isolate, and connect more low period cycles to produce a higher bound and the associated increase in computational cost. (The effect of refining the grid on increasing the bound is illustrated for the Hénon map in Figure 6.) Another improvement to these techniques related to this balance would involve making the computation of $\mathcal{G}$, and therefore $\mathcal{S}$, adaptive. The goal here would be to refine the grid in areas where new low-period periodic orbits and connec-

tions may be uncovered without having to recompute structures in the remainder of the space.

3.2. Processing index information. Recall that our goal is to compute complicated symbolic dynamics. If we are successful in locating an appropriate region of interest in the domain (one approach is described in section 3.1), the corresponding Conley index computed by the algorithms described in section 2.4 is given as a large matrix representing the map induced on an index pair consisting of many disjoint components.

From this index map, we wish to find a symbol transition matrix T such that f is semiconjugate to the subshift on Σ_T. We first use some properties of shift equivalence to simplify the computed index map. We then construct T from a collection of cycles, called *verified cycles*, that satisfy the hypotheses of Corollary 2.15.

3.2.1. Removing transient generators. We begin our processing of the index map $f_{P*} : H_*(P_1, P_0) \to H_*(P_1, P_0)$ by removing generators from $H_*(P_1, P_0)$ that do not correspond to asymptotic invariant behavior. More specifically, we utilize the fact that the Conley index, $\mathrm{Con}_*(S, f)$, is the shift equivalence class of f_{P*} to construct a new representative of the class obtained by removing generators $\alpha \in H_k(P_1, P_0)$ such that $f_{Pk}^l(\alpha) = 0$ or $\alpha \notin f_{Pk}^l(H_k(|\mathcal{P}_1|, |\mathcal{P}_0|))$ for some $l \in \mathbb{Z}$.

Note that since we are considering continuous maps f on $\mathbb{R}^n$, $f_{Pk} : H_k(P_1, P_0) \to H_k(P_1, P_0)$ are linear maps on (finite) vector spaces. We therefore choose to think of f_{Pk} as a square matrix. Suppose that f_{Pk} is similar to a matrix A, i.e., $f_{Pk} = B^{-1}AB$ for some invertible matrix B. Then, by setting $r = B$, $s = B^{-1}$, and $m = 0$ in Definition 2.11, we see that $[f_{Pk}]_s = [A]_s$. In what follows, B will be an appropriate reordering of the basis so that A takes on the block lower-triangular form required for the following theorem.

Theorem 3.1. *Let*

$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & 0 \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

be a 3×3 *block lower-triangular matrix, with square matrices* A_{ii} *on the diagonal. If* $A_{11}^l = 0$ *and* $A_{33}^l = 0$ *for some* l*, then* A *is shift equivalent to* A_{22}*.*

Proof. For $i = 1, 2, 3$, let $n_i \times n_i$ be the size of the square matrix A_{ii}, and define projection and inclusion maps respectively as follows:

$$\pi = \begin{bmatrix} 0_{n_{22} \times n_{11}} & I_{n_{22}} & 0_{n_{22} \times n_{33}} \end{bmatrix}$$

and

$$\iota = \pi^\top.$$

One can check that the maps $R := \pi A^l$ and $S := A^l \iota$ satisfy the conditions stated in Definition 2.11 to give the desired shift equivalence between A and A_{22} with lag constant $m = 2l$. ∎

The motivation for the previous theorem was to find a simpler representative for the shift equivalence class of f_{Pk}. This relies on finding a reordering of the basis for f_{Pk} that yields a similar matrix A satisfying the hypotheses of Theorem 3.1. In order to use existing efficient algorithms, we now turn to a graph interpretation of the $l \times l$ matrix f_{Pk}. More specifically,

we consider the directed graph $G = (V, E)$ with vertices $1, \dots, l$ and edges $(j, i) \in E$ if and only if $f_{Pk}(i, j) \neq 0$. Let

$$V_3 := \{v \in V \mid \text{any path starting at } v \text{ has length less than } l\}, \tag{3.1}$$

$$V_1 := \{v \in V \setminus V_3 \mid \text{any path ending at } v \text{ has length less than } l\}, \tag{3.2}$$

and

$$V_2 := V \setminus (V_1 \cup V_3). \tag{3.3}$$

Note that since there are l vertices, V_1 is the set of vertices that are not connected to cycles in backward time and V_3 is the set of all vertices that are not connected to cycles in forward time. The following two lemmas show that the partition $\{V_1, V_2, V_3\}$ of the vertex set V is useful for finding zeros in the matrix f_{Pk}.

Lemma 3.2. *The submatrix $f_{Pk}(V_1, V_2 \cup V_3)$ of f_{Pk} corresponding to the rows indexed by V_1 and columns indexed by $V_2 \cup V_3$ is the zero matrix of the appropriate size.*

Proof. Suppose that $f_{Pk}(w, v) \neq 0$ for $w \in V_1$ and $v \in V_2 \cup V_3$. Then (v, w) is an edge in the associated directed graph G. Since v is not in V_1, there exists a path $v_1, \dots, v_l, v$ in G. Then $v_1, \dots, v_l, v, w$ is a length $l + 1$ path in G, contradicting our assumption that $w \in V_1$. ∎

Lemma 3.3. *The submatrix $f_{Pk}(V_2, V_3)$ of f_{Pk} corresponding to the rows indexed by V_2 and columns indexed by V_3 is the zero matrix of the appropriate size.*

Proof. Suppose that $f_{Pk}(w, v) \neq 0$ for $w \in V_2$ and $v \in V_3$. Then (v, w) is an edge in the associated directed graph G. Since w is not in V_3, there exists a path $w, v_1, \dots, v_l$ in G. Then $v, w, v_1, \dots, v_l$ must also be a path in G, contradicting our assumption that $v \in V_3$. ∎

We have now shown that if we reorder the basis by listing the basis elements in V_1, followed by those in V_2, followed by those in V_3, we obtain the following block form (with rows and columns labeled by location in the specified sets):

$$f_{Pk} \sim A = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \end{array} \begin{array}{c} \begin{array}{ccc} V_1 & V_2 & V_3 \end{array} \\ \begin{pmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & 0 \\ A_{31} & A_{32} & A_{33} \end{pmatrix} \end{array}.$$

What remains to show in order to use Theorem 3.1 is the following lemma.

Lemma 3.4. *The two matrices A_{11}^l and A_{33}^l are zero matrices of the appropriate sizes.*

Proof. We obtained the block lower-triangular matrix A by a reordering of the basis for the matrix f_{Pk}. Therefore, the associated directed graph G_A for A is the directed graph G with relabeled vertices. With a slight abuse of notation, we consider again the subsets V_1, V_2, V_3 in G_A to be the sets satisfying (3.2), (3.3), and (3.1), respectively. Interpreting nonzero entries of A to be weights on the corresponding edges, we may use powers of A to study paths in G_A. More specifically, $A^l(i, j) \neq 0$ implies that there exists a length l path from vertex j to vertex i in G_A (see, e.g., [Die05]). Now suppose that $A^l(i, j) = A_{11}^l(i, j) \neq 0$ for some $i, j \in V_1$. Then, by the above argument, there exists a length l path in G_A that ends at a vertex in V_1. This contradicts (3.2). Therefore, $A_{11}^l = 0$. A similar argument shows that $A_{33}^l(i, j) = 0$ for all $i, j \in V_3$. ∎

We now have that f_{Pk} is similar (and hence shift equivalent) to A which is shift equivalent to $\tilde{f}_{Pk} := A_{22}$ by Lemma 3.4 and Theorem 3.1. Therefore, we may take $\tilde{f}_{Pk}$ to be the new, possibly smaller representative of the Conley index

$$\mathrm{Con}(S, f) = [f_{Pk}]_s = [\tilde{f}_{Pk}]_s.$$

This procedure is outlined in Algorithm 5. Here, algorithms based on depth or breadth first search may be used to efficiently compute the required sets V_1, V_2, and V_3. As we will show in section 4 this technique may give a drastic decrease in the size of the representative index map.

Algorithm 5 (Remove transient generators).

INPUT: square matrix f_{Pk}
OUTPUT: shift equivalent (square) matrix $\tilde{f}_{Pk}$
$\tilde{f}_{Pk} =$ remove_transient_generators(f_{Pk})
$G = (V, E)$ is the directed graph associated with f_{Pk};
$V_3 = \{v \in V \mid$ any path starting at v is finite$\}$;
$V_1 = \{v \in V \mid$ any path ending at v is finite$\}$;
$V_2 = V \setminus (V_1 \cup V_3)$;
$\tilde{f}_{Pk} = f_{Pk}(V_2, V_2)$;
return $\tilde{f}_{Pk}$;

3.2.2. Cycle verification. We now automate a procedure for using Conley index computations to construct a semiconjugate subshift of finite type. As described in Theorem 3.6 below, we construct the subshift system from a collection of cycles that are *verified* using Corollary 2.15. As will be seen in section 4, the automation of this procedure becomes necessary as we build increasingly complicated subshifts of finite type. In particular, building a subshift system containing infinitely many periodic orbits may, in principle, lead to an infinite list of computations to verify that the hypotheses of Corollary 2.15 hold for each cycle. In the following approach, we present an algorithm which uses a finite list of computations to verify a possibly infinite set of cycles.

Given an index pair $P = (P_1, P_0)$, we begin by labeling each of the (m) disjoint regions of the isolating neighborhood $N := \overline{P_1 \setminus P_0}$. Let $N = \cup_{i=1}^{m} N_i$ with N_i closed and nonempty and $N_i \cap N_j = \emptyset$ for all $i \neq j$. By construction, each N_i has a corresponding cubical representation $\mathcal{N}_i \subset \mathcal{N}$. Recall that the associated itinerary function ρ is defined by $\rho(x) = (s_0 s_1 \ldots)$ with $s_j = i$ if $f^j(x) \in N_i$. Let $\tilde{T}$ be the matrix of admissible transitions between the regions N_i allowed by $\mathcal{F}$. More specifically, $\tilde{T}$ is the $m \times m$ matrix with entries

$$t_{ij} = \begin{cases} 1 & \text{if } \mathcal{F}(\mathcal{N}_j) \cap \mathcal{N}_i \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Then $\rho : \tilde{S} \to \Sigma_{\tilde{T}}$, where $\tilde{S} := Inv(N, f)$ and $\Sigma_{\tilde{T}}$ and $\sigma_{\tilde{T}} : \Sigma_{\tilde{T}} \to \Sigma_{\tilde{T}}$ are the subshift of finite type defined in section 2. As previously discussed, $\rho : \tilde{S} \to \Sigma_{\tilde{T}}$ may not be surjective, and hence $\sigma_{\tilde{T}} : \Sigma_{\tilde{T}} \to \Sigma_{\tilde{T}}$ may not be semiconjugate to $f : \tilde{S} \to \tilde{S}$. We will now construct a subshift system, $\sigma_T : \Sigma_T \to \Sigma_T$, with $\Sigma_T \subset \Sigma_{\tilde{T}}$, that we prove is semiconjugate via ρ to $f : S \to S$ with $S \subset \tilde{S}$.

Let $G = (V, E)$ be the directed graph associated with the symbol transition graph $\tilde{T}$ (viewed as an adjacency matrix). More specifically, the vertices are named for the regions N_i with $V = \{1, 2, \dots, m\}$, and the edge set $E := \{(j, i) \in V \times V \mid t_{ij} = 1\}$ represents the admissible transitions between regions. In our approach, we begin by removing all paths in G that are not contained in cycles. These paths correspond to dynamics that we will not check using index theory. A practical way to perform this step is to remove edges and vertices not contained in the strongly connected components (SCC) of G. We will now study Conley indices for periodic symbol sequences in $\Sigma_{\tilde{T}}$ represented by cycles in G.

As discussed in section 2.2, we consider restricted index maps

$$f_{Pk}^{ij} : H_k(P_1, P_0 \cup (\cup_{l \neq i} N_l)) \to H_k(P_1, P_0 \cup (\cup_{l \neq j} N_l)).$$

To do this, we first group the generators of $H_k(P_1, P_0)$ remaining after running Algorithm 5 by region. Again, thinking of f_{Pk} as a matrix with rows and columns corresponding to the generators of $H_k(P_1, P_0)$, we have that

$$f_{Pk}^{ij} := f_{Pk}(g_{N_j}, g_{N_i}), \tag{3.5}$$

where g_{N_i} are the (row/column) indices of generators in $H_k(P_1, P_0 \cup \bigcup_{l \neq i} N_l) \subset H_k(P_1, P_0)$. Here, f_{Pk}^{ij} is as an $n_j \times n_i$ matrix, where n_i and n_j are the number of generators in regions N_i and N_j, respectively. To simplify notation, for a path $p = (s_1, s_2, \dots, s_n)$, let

$$f_{Pk}^{p} := f_{Pk}^{s_{n-1}s_n} \circ \cdots \circ f_{Pk}^{s_1 s_2}. \tag{3.6}$$

Definition 3.5. *We say that a cycle $c = (s_1, s_2, \dots, s_n, s_1)$ in G is* verified *if, for some k,*

$$\operatorname{tr}(f_{Pk}^{c}) = \operatorname{tr}_k \operatorname{Con}(S', f_{N_{s_n}} \circ \cdots \circ f_{N_{s_1}}) \neq 0,$$

where $S' := Inv(N_{s_1}, f|_{N_{s_n} \to N_{s_1}} \circ \cdots \circ f|_{N_{s_1} \to N_{s_2}})$. Note that, by Corollary 2.15, $\rho^{-1}(\mathbf{s}) \neq \emptyset$, where $\mathbf{s} = (s_1 s_2 \dots s_n s_1 \dots s_n \dots)$ is the periodic symbol sequence corresponding to the verified cycle.

Before discussing our automated approach for verifying cycles, we give the following theorem to serve as motivation for this work.

Theorem 3.6. *Let Σ_T be the space of symbol sequences with symbol transition matrix T, and let $\operatorname{Per}(\Sigma_T)$ be the set of periodic symbol sequences in Σ_T under σ_T. Suppose that $\Sigma_T = \overline{\operatorname{Per}(\Sigma_T)}$ and for each $\mathbf{s} = (s_1 \dots s_n s_1 \dots s_n \dots) \in \operatorname{Per}(\Sigma_T)$ the corresponding cycle $c = (s_1, s_2, \dots, s_n, s_1)$ in G has been verified according to Definition 3.5. Then the itinerary function ρ is a semiconjugacy between $f : S \to S$ and $\sigma_T : \Sigma_T \to \Sigma_T$, where $S := \rho^{-1}(\Sigma_T) \subset \tilde{S}$.*

Proof. The itinerary function $\rho : \tilde{S} \to \Sigma_T$ is continuous and $\rho \circ f = \sigma_T \circ \rho$ (see section 2 and references therein). Furthermore, since each cycle in G corresponding to a periodic symbol sequence in Σ_T has been verified according to Definition 3.5, ρ maps onto $\operatorname{Per}(\Sigma_T)$. Since ρ is continuous, $\tilde{S} := Inv(N, f)$ is compact, and Σ_T is Hausdorff, ρ must map onto the closure of the set of of periodic symbol sequences, $\overline{\operatorname{Per}(\Sigma_T)} = \Sigma_T$. Therefore, $\rho : S \to \Sigma_T$ is a semiconjugacy. ∎

The list of cycles that may be verified according to Definition 3.5 relies implicitly on the form of f_P and, more specifically, on f_{Pk}^{ij} for $k = 0, 1, 2, \dots$ and $i, j \in \{1, \dots, m\}$. For the

examples studied in section 4, the homology maps f_{Pk} are trivial for all $k \neq 1$. Therefore, for these examples we fix $k = 1$, as any other choice will necessarily lead to a zero trace and failure to verify all cycles. For different systems, there may be more flexibility in the choice of k. Given a fixed k, the question of how the list of verified cycles relies on choices of i and j is far more subtle. We begin this discussion by fixing k and considering the case where each region contains exactly one homology generator ($n_i = 1$ for all $i = 1, \dots, m$). We will then discuss the more difficult case where some regions have multiple homology generators.

Note that if there is only one generator per region, then f_{Pk}^{ij} is a scalar for all admissible transitions $\tilde{t}_{ji} = 1$ in $\tilde{T}$. In this case, if $f_{Pk}^{ij} \neq 0$ for all admissible transitions, then for any admissible periodic symbol sequence $\mathbf{s} = (s_1 s_2 \dots s_n s_1 s_2 \dots s_n \dots)$ with corresponding cycle $c = (s_1 s_2 \dots s_n s_1)$, $\operatorname{tr}(f_{Pk}^c) \neq 0$, and therefore all cycles in G are verified. If, on the other hand, $f_{Pk}^{ij} = 0$ for some admissible transition, then any cycle c with edge (i, j) will have $\operatorname{tr}(f_{Pk}^c) = 0$ and cannot be verified using this approach. In this case, we remove this transition from the set of admissible transitions by removing the edge (i, j) from G and, correspondingly, by setting $t_{ji} = 0$ in T. In essence, cycle verification computations in the setting where there is exactly one homology generator per component boil down to a (finite) check that entries in f_{Pk} corresponding to admissible transitions in T are nonzero.

If there are regions that contain more than one generator of homology, then these computations become more complicated. In what follows, we will systematically process the cycles in G. In the first phase of the procedure, we process paths and cycles in $\mathcal{G}$ in an attempt verify cycles. Alternatively, one can think about identifying all cycles that may not be verified by our approach. Along these lines, we will label certain cycles as *unverifiable* and certain paths as *unconcatenable*. From these, we will identify a collection of edges that need to be removed from the graph so that all remaining cycles are verified cycles. Note that in what follows, labeling a path *unconcatenable* does not mean that cycles containing this path may not be verified according to Definition 3.5, only that we may not verify some such cycles using our prescribed list of finite computations. Let `Max_Iter` be a nonnegative integer that will serve as a computational parameter.

Definition 3.7. *Define the* edge set for a path $p = (v_0, \dots, v_n)$ *to be*

$$E(p) := \{(v_i, v_{i+1}) \in E(G) \mid i = 0, \dots, n-1\}$$

and the length of p *to be* $|p| = n$. *Consider a cycle* $c = (s, v_2, v_3, \dots, v_{n-1}, s)$ *starting and ending at vertex* s. *If* $\operatorname{tr}(f_{Pk}^c) = 0$, *then* c *is* unverifiable. *(See also Definition 3.5.)*

A path $p = (s, v_2, v_3, \dots, v_{n-1}, t)$ *from* s *to* t *of length* $|p| \leq$ `Max_Iter` *is* unconcatenable *if* $f_{Pk}^p = \mathbf{0}$.

For a path $p = (s, v_2, v_3, \dots, v_{n-1}, t)$ *from* s *to* t *of length* $|p| =$ `Max_Iter`, p *is* concatenable *if there exists a path* p' *from* s *to* t *with* $|p'| <$ `Max_Iter`, $E(p') \subseteq E(p)$, *and* $f_{Pk}^p = \alpha f_{Pk}^{p'} \neq \mathbf{0}$ *for some* $\alpha \neq 0$. *If no such path* p' *exists, then* p *is* unconcatenable.

Finally, an edge set E *is* prohibited *(at computational parameter* `Max_Iter`*) if at least one of the following holds:*

1. *there exists an unverifiable cycle* c *with* $|c| \leq$ `Max_Iter` *and* $E(c) \subseteq E$,
2. *there exists an unconcatenable path* p *with* $E(p) \subseteq E$.

Lemma 3.8. *If c is a cycle whose edge set $E(c)$ is not prohibited, then c is a verified cycle.*

Proof. Suppose that $|c| \leq$ `Max_Iter`. Since $E(c)$ is not prohibited, c must be a verified cycle.

Next, notice that in the natural partial ordering on edge sets, if E' is prohibited, then so is E for any E containing E'. Therefore, $E(c)$ must not contain any prohibited subsets. If $|c| >$ `Max_Iter`, then c is the concatenation of two paths, p_1 and p_2, where $|p_2| =$ `Max_Iter`. We will use the notation p_1p_2 to denote the concatenation of paths p_1 and p_2. Label the start/end vertices s_1, t_1 and s_2, t_2 of p_1 and p_2, respectively. Note that $s_1 = t_2$ and $t_1 = s_2$ by construction. Since $E(p_2) \subseteq E(c)$ is not prohibited, there exists a path p_2' from s_2 to t_2 with $E(p_2') \subseteq E(p_2)$, $|p_2'| <$ `Max_Iter`, and $f_{Pk}^{p_2} = \alpha f_{Pk}^{p_2'}$ for some $\alpha \neq 0$. Therefore,

$$\begin{aligned} f_{Pk}^{c} &= f_{Pk}^{p_2} f_{Pk}^{p_1} \\ &= \alpha f_{Pk}^{p_2'} f_{Pk}^{p_1} \\ &= \alpha f_{Pk}^{c'}, \end{aligned}$$

where $c' = p_1p_2'$ is a cycle with $E(c') = E(p_1) \cup E(p_2') \subseteq E(c)$ and length $|c'| \leq |c| - 1$. Continuing this process, we obtain a cycle $\tilde{c}$ with $|\tilde{c}| \leq$ `Max_Iter`, $E(\tilde{c}) \subseteq E(c)$, and $f_{Pk}^{c} = \tilde{\alpha} f_{Pk}^{\tilde{c}}$ for some $\tilde{\alpha} \neq 0$. Since $E(\tilde{c})$ cannot be prohibited, $\tilde{c}$ must be verifiable and

$$\operatorname{tr}(f_{Pk}^{c}) = \operatorname{tr}(\tilde{\alpha} f_{Pk}^{\tilde{c}}) = \tilde{\alpha} \operatorname{tr}(f_{Pk}^{\tilde{c}}) \neq 0.$$

Therefore, c is a verified cycle. ∎

Lemma 3.8 and Theorem 3.6 provide an outline of our approach for constructing the semiconjugate system. By Lemma 3.8, we know that all cycles that do not have prohibited edge sets are verified cycles and may be used to construct the semiconjugate system according to Theorem 3.6. In practice, we use the prohibited edge sets to identify a collection of edges to be removed from G, resulting in the desired semiconjugate system.

We now give an outline of our procedure for locating prohibited edge sets by collecting and testing appropriate matrix products along paths in G. The algorithm outputs a collection of minimal prohibited edge sets $\mathcal{B}$; that is, for any prohibited edge set E, there is a prohibited edge set $E' \in \mathcal{B}$ with $E' \subseteq E$.

Algorithm 6 (Find prohibited edge sets).

```
INPUT:   graph G, index map f_Pk, computational parameter Max_Iter;
OUTPUT:  list of minimal prohibited edge sets B
B = find_prohibited_edge_sets(G, {f_Pk}, Max_Iter)
  B = ∅;
  for all s,t ∈ V(G), E ⊂ E(G), set all M(s,t,E,k) = ∅;
  for all (s,t) ∈ E(G),
    if s == t and tr(f_Pk^st) == 0, B = B ∪ {(s,t)};
    else M(s,t,{(s,t)},1) = {f_Pk^st};
    end
  end
  for k = 1...Max_Iter,
    for s,t ∈ V(G), E ⊆ E(G), M ∈ M(s,t,E,k),
```

```
      for (t,u) ∈ E(G),
        E' = E ∪ (t,u);
        M' = f^{tu}_{Pk} M;
        if (M' == 0) or (s == u and tr(M') == 0),
          B = B ∪ {E'};
          set M(s',t',E'',ℓ) = ∅ for all s',t' ∈ V(G), E' ⊆ E'' ⊆ E(G), ℓ ≤ k;
        else if ∄M'' ∈ ⋃_{E''⊆E', ℓ<k} M(s,t,E'',ℓ), with M'' == αM' for some α ≠ 0,
          M(s,t,E',k+1) = M(s,t,E',k+1) ∪ {M'};
        end
      end
    end
  end
  B = B ∪ {E ⊂ E(G) | M(s,t,E,Max_Iter) ≠ ∅ and E minimal};
  return B;
```

In practice, it is more efficient to apply Algorithm 6 only to a subgraph of G that captures the behavior of the system in the regions with multiple homology generators. More specifically, we first study G restricted to the vertices for multiple generator regions and the neighboring single generator regions. This allows us to take advantage of the fact that f^p_{Pk} is a scalar for all paths p starting and ending at vertices for single generator regions. By removing enough edges so that there are no remaining prohibited edge sets in the subgraph, we can reduce the check that cycles remaining in G are verified to a check that the maps f^{ij}_{Pk} between single generator regions are nonzero. This is the approach we adopt for the results described in section 4.

For all cycles c in G that do not contain any prohibited edge sets (listed in $\mathcal{B}$), c is a verified cycle by Lemma 3.8. What remains for the construction of a subgraph G' of verified cycles is to remove enough edges so that we no longer have any cycles with prohibited edge sets. Since our goal is to obtain a high lower bound for entropy, we will select one edge from each prohibited edge set so that the removal of these edges results in the semiconjugate symbolic system with highest entropy. Again, since the list of prohibited edge sets is finite (and each prohibited edge set is finite), the computation of optimal edges to remove is finite. Removing the edges yields a graph in which all cycles may be verified using Corollary 2.15. By Theorem 3.6, the corresponding adjacency matrix, T, defines a semiconjugate symbolic system.

The following is an outline of the procedure for breaking prohibited edge sets.

Algorithm 7 (Break prohibited edge sets).

```
INPUT:  graph G, a list of prohibited edge sets B,
OUTPUT: graph G' in which all cycles may be verified via Corollary 2.15
G' = break_prohibited_edge_sets(G, B)
  if B = ∅, return G' = G;
  h_max = −1;
  E_c = ∅;
  for each set {e_1, e_2, ..., e_N}, where e_i is an edge on the ith cycle in B,
    let G' be the subgraph of G obtained by removing edges e_1, e_2, ..., e_N;
```

```
    let T(G') be the adjacency matrix for G';
    h = log(sp(T(G')));
    if h > h_max,
      h_max = h;
      E_c = {e_1, e_2, ..., e_N};
    end
  end
  let G' be the subgraph of G obtained by removing the edges in E_c;
  return G';
```

Combining Algorithms 6 and 7, Theorem 3.6 guarantees that the following algorithm produces a symbol transition matrix T with $\sigma : \Sigma_T \to \Sigma_T$ semiconjugate to $f : S \to S$. Noting that $f_{Pk}^{ij} = \mathbf{0}$ will cause the verification procedure to fail for any cycle containing edge (i, j), we will start with a graph G on the same vertex set with the edge set $E = \{(i, j) \mid f_{Pk}^{ij} \neq \mathbf{0}\}$.

Algorithm 8 (Build subshift).

```
INPUT:      index map f_Pk : H_k(|P_1|,|P_0|) → H_k(|P_1|,|P_0|),
            computational parameter Max_Iter
OUTPUT:     symbol transition matrix T for semiconjugate subshift of
            finite type
T = build_subshift(f_Pk, H_k(|P_1|,|P_0|), Max_Iter)
  f_Pk = remove_transient_generators(f_Pk);
  set m to be the number of disjoint components of H_k(|P_1|,|P_0|);
  V = {1, ..., m};
  E = {(i,j) ∈ V × V | f_Pk^ij ≠ 0};
  G = G(V, E);
  G = SCC(G); (removes all edges not contained in cycles)
  U = find_prohibited_edge_sets(G, f_Pk, Max_Iter);
  G' = break_prohibited_edge_sets(G, U);
  T is the adjacency matrix for graph G';
  return T;
```

4. An example: The Hénon map. As illustration, we now apply our techniques to the Hénon map

$$h(x, y) = (1 + y - ax^2, bx) \tag{4.1}$$

at the classical parameters a = 1.4, b = 0.3. Since its first appearance in [Hén76], there has been extensive research on the Hénon map. The first result concerning a real description of the chaotic dynamics of the Hénon map is [MS80], where the existence of a transverse homoclinic point, and hence the existence of horseshoe dynamics, is proved. In [Szy97], Szymczak used Conley index theory to give a computer-assisted proof of the existence of periodic orbits of all periods except three and five. In [Gal02], Galias employed the method of covering relations (related to Easton's windows) to give a computer-assisted proof of the existence of an infinite number of homoclinic and heteroclinic trajectories. [Gal02] also contains a result which gives

a rigorous lower bound for the topological entropy of the map $h_{top}(h) \geq 0.4300$. In [NBGM08], Newhouse et al. use the planar structure of the Hénon map to compute $h_{top}(h) \geq 0.46469$, the highest lower bound on the entropy for Hénon at the classical parameter values currently reported.

For this work, we use the *GAIO* software package to construct grids, $\mathcal{G}^{(d)}$, at discretization depths $0 \leq d \leq 12$, on the initial box $[-1.425, 1.425] \times [-0.425, 0.425]$ (see section 2.3). We then use the interval arithmetic package INTLAB [Cse99] to compute a combinatorial enclosure, $\mathcal{H}$, on $\mathcal{G}^{(d)}$ as

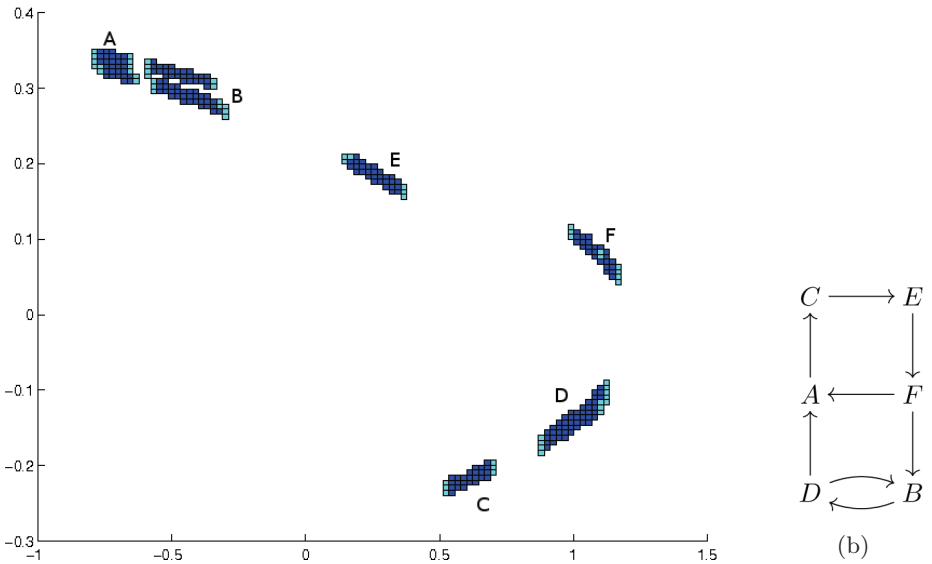
$$\mathcal{H}(I_1 \times I_2) = \{G \in \mathcal{G}^{(d)} \mid \tilde{h}(I_1, I_2) \cap G \neq \emptyset\},$$

where $I_1 \times I_2$ is an element in $\mathcal{G}^{(d)}$ in interval product notation and $\tilde{h}(I_1, I_2)$ denotes the rectangular image of $h(I_1, I_2)$ computed using (outward rounding) interval arithmetic. Finally, we use MATLAB scripts encoding the algorithms outlined throughout the paper to find and compute the required Conley index structures and subshifts of finite type. In the following sample results, we describe three different techniques for producing the region of interest $\mathcal{S} \subset \mathcal{G}^d$. Given $\mathcal{S}$, the main approach is the following.

Algorithm 9 (Main).

INPUT: grid $\mathcal{G}^d$, combinatorial enclosure $\mathcal{H}$ on $\mathcal{G}$,
region of interest $\mathcal{S}$, computational parameter Max_Iter
OUTPUT: lower bound on the topological entropy of h ENTROPY
ENTROPY = compute_entropy_lower_bound($\mathcal{G}^d$, $\mathcal{H}$, $\mathcal{S}$, Max_Iter)
  ENTROPY = 0;
  $\mathcal{N}$ = grow_isolating_neighborhood($\mathcal{S}$); (Algorithm 1)
  $[\mathcal{P}_1, \mathcal{P}_0]$ = build_index_pair($\mathcal{N}$); (Algorithm 2)
  f_{P*} = compute_index_map($\mathcal{P}_1$, $\mathcal{P}_0$, $\mathcal{H}$, $\mathcal{G}^d$); (Algorithm 3)
  for $k = 1 \ldots \dim(\mathcal{G}^d)$, with $f_{Pk} \neq 0$,
    T = build_subshift(f_{Pk}, $H_k(|\mathcal{P}_1|, |\mathcal{P}_0|)$, Max_Iter); (Algorithm 8)
    ENTROPY := max{ENTROPY, $\log(sp(T))$};
  end
  return ENTROPY;

4.1. Joining two short cycles. For purposes of illustration, we begin with a relatively simple example on the grid at depth $d = 7$. Although the resulting entropy lower bound, 0.2406, is small, this example provides us with matrices of reasonable sizes for depicting the results of various stages of the procedure. For this example, we locate a region of interest, $\mathcal{S}$, by searching the computed enclosure $\mathcal{H}$ on $\mathcal{G}^{(7)}$ for a cycle of length 2, a cycle of length 4, and shortest path connections from the 2-cycle to the 4-cycle and from the 4-cycle to the 2-cycle. $\mathcal{S}$ is the union of these four objects. Applying Algorithms 1 and 2 to $\mathcal{S}$ results in the index pair given in Figure 2.



(a) (b)

Figure 2. (a) *A combinatorial index pair,* $(\mathcal{P}_1, \mathcal{P}_0)$*, computed using Algorithms 1 and 2 for the Hénon map at depth* $d = 7$*. (*$\mathcal{P}_0$ *is the collection of boxes shown in cyan.)* (b) *The corresponding symbol transition graph produced by Algorithm 8.*

Theorem 4.1. *The topological entropy of the Hénon map* (4.1) *is bounded from below by* 0.2406.

Proof. The computed index map for the index pair depicted in Figure 2(a) is

$$h_{P,1} = \begin{array}{c} \\ A \\ B \\ B \\ B \\ B \\ C \\ D \\ E \\ F \\ F \end{array} \begin{array}{c} \begin{array}{cccccccccc} A & B & B & B & B & C & D & E & F & F \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}.$$

The rows and columns are labeled by location of the corresponding homology generator in the labeled regions of the isolating neighborhood (see Figure 2(a)). Applying Algorithm 5 for removing transient generators to $h_{P,1}$ produces the shift equivalent matrix

$$A = \begin{array}{c} \\ A \\ B \\ B \\ C \\ D \\ E \\ F \\ F \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & B & C & D & E & F & F \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}.$$

This is the matrix labeled A_{22} in Theorem 3.1 and is obtained by an appropriate reordering of the basis. Note that this algorithm removed two of the homology generators in region B and, therefore, reduced the size of the representative of the shift equivalence class/Conley index.

As an example computation, using Corollary 2.15 to verify the cycle (B, D, B), we check that

$$\begin{aligned} \mathrm{tr}_1(Con(S', f|_{D\to B} \circ f|_{B\to D})) &= \mathrm{tr}\left(h_{P,1}^{DB} h_{P,1}^{BD}\right) \\ &= \mathrm{tr}_1\left(\begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}\right) \\ &= \mathrm{tr}_1\left(\begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}\right) \\ &\neq 0. \end{aligned}$$

Running Algorithm 8 on A to verify a collection of cycles results in the construction of a semiconjugate subshift system with symbol transition matrix

$$T = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}.$$

The corresponding symbol transition graph is given in Figure 2(b). Since the log of the spectral radius of T is greater than 0.2406, the result follows from Theorem 2.7. ∎

4.2. Joining low cycles (Algorithm 4). We now focus on improving the bound by refining the grid and using Algorithm 4 to compute a more complicated region of interest.

This approach results in the following theorem.

Theorem 4.2. *The topological entropy of the Hénon map* (4.1) *is bounded from below by* 0.4320.

Outline of proof. Given the enclosure $\mathcal{H}$ on $\mathcal{G}^{(12)}$, we use Algorithm 4 with Max_Cycle_Length = 7 to produce the region of interest $\mathcal{S}$. We then follow Algorithm 9. The index pair for $\mathcal{S}$ appears in Figure 3. Algorithm 3 returns an index map on 1521 relative homology gener-

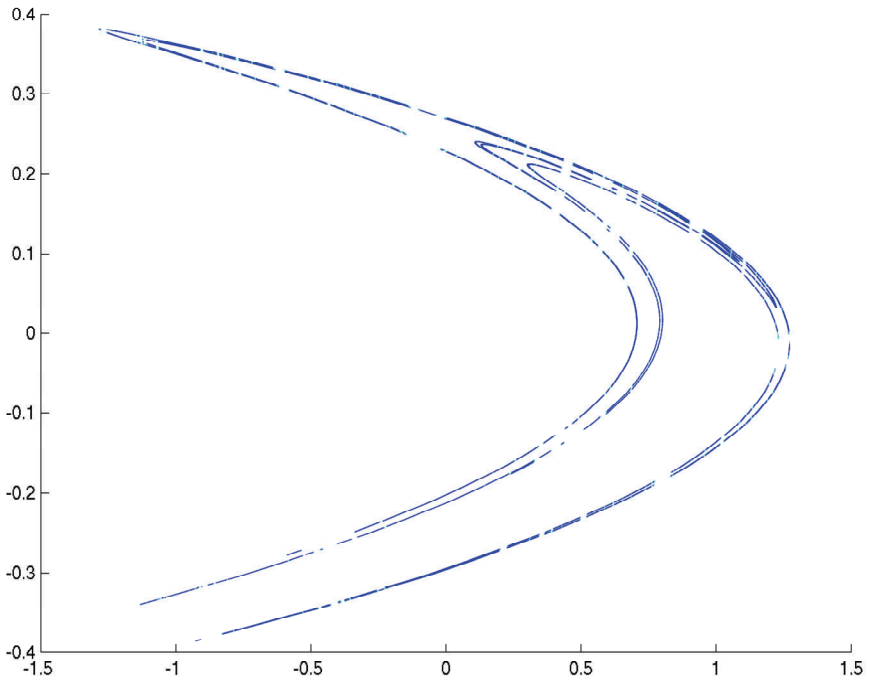


Figure 3. *The combinatorial index pair, $(\mathcal{P}_1, \mathcal{P}_0)$, constructed starting with Algorithm 4 for Theorem 4.2 at depth 12. ($\mathcal{P}_0$ is the collection of boxes shown in cyan.)*

ators. Algorithm 5 reduces this map to a shift equivalent map on 309 generators. Finally, Algorithm 8 produces a semiconjugate subshift of finite type with 247 symbols. The symbol transition matrix for the constructed subshift is depicted in Figure 4. The log of the spectral radius of T is bounded from below by 0.4320. The result then follows from Theorem 2.7. ■

For the above result computed on the grid $G^{(12)}$, we choose the maximal cycle length for Algorithm 4 to be Max_Cycle_Length = 7. This choice is made because choosing instead Max_Cycle_Length < 7 yields a lower bound than that given in Theorem 4.2, and choosing Max_Cycle_Length > 7 yields an entropy lower bound of 0. This behavior is depicted in Figure 5. The reason that choosing a large maximal cycle length leads to a 0 lower bound is that the corresponding isolating neighborhood produced by Algorithm 1 is a covering of the entire attractor, with corresponding trivial symbolic dynamics.

In principle, improving the bound requires only extra computational cost. Figure 6 shows the change in the computed entropy bound with increase in resolution of the grid (and corresponding increase in computational expense) for the Hénon map. The dip in the graph at depth 11 is of interest because, in general, we expect a monotonic increase in the computed entropy bound with increase in resolution of the grid. This nonmonotonic behavior indicates that our choice of region of interest, $\mathcal{S}$, in Algorithm 4 is indeed suboptimal. In fact, choosing $\mathcal{S}$ to be the boxes in $\mathcal{G}^{(11)}$ contained in the isolating neighborhood $\mathcal{N}$ returned by Algorithms 4 and 1 on $\mathcal{G}^{(10)}$ would yield the same entropy as that computed at depth 10, and so it is possible to compute a higher entropy bound at this resolution.

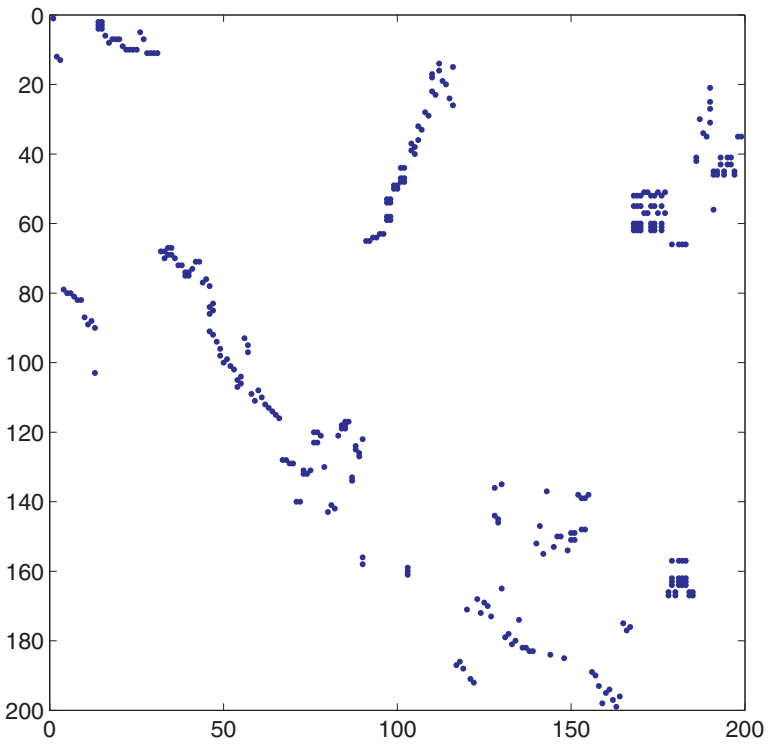


Figure 4. *A depiction of the nonzero entries of the* 247×247 *symbol transition matrix for the subshift of finite type constructed for Theorem* 4.2.

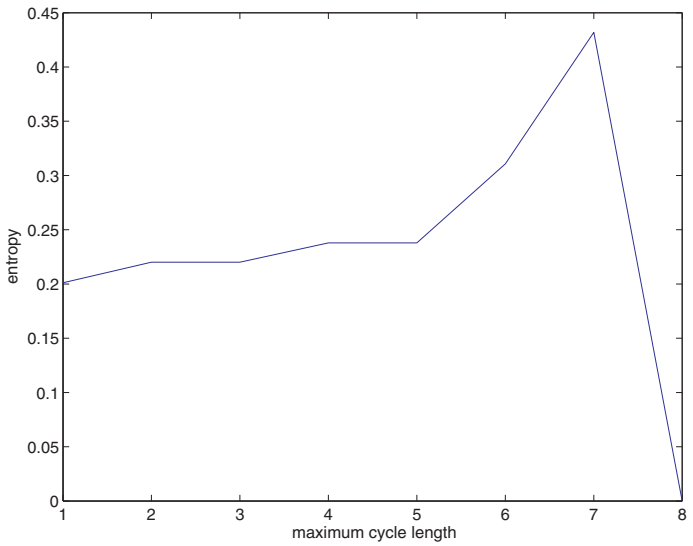


Figure 5. *Entropy lower bounds computed using Algorithm* 4 *for the Hénon map on grid* $\mathcal{G}^{(12)}$ *at varying maximal cycle lengths* N.

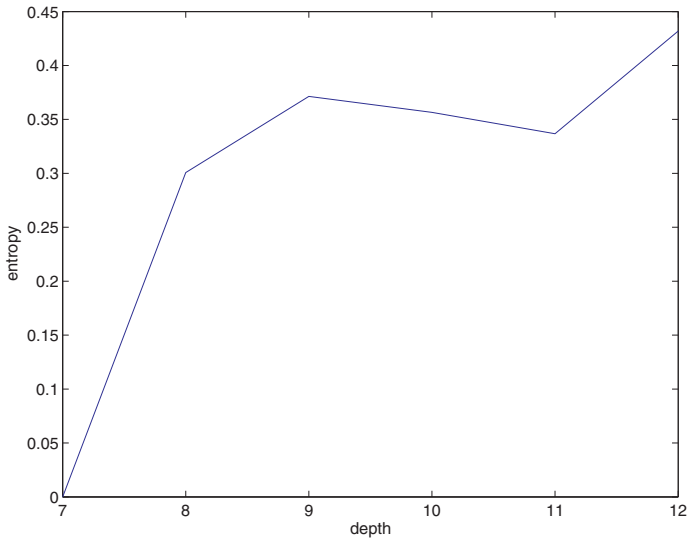


Figure 6. *Entropy lower bounds for the Hénon map computed on regions given by Algorithm 4 on grids $\mathcal{G}^{(d)}$ of varying depth d.*

4.3. Fold preimage removal. A priori knowledge of the Hénon map suggests another approach for constructing the region of interest $\mathcal{S}$. We notice that indices for cycles traveling too close to the "fold" of the attractor (at approximately $(1.2717, -0.0207)$) are necessarily trivial. Here, the Hénon map loses hyperbolicity, and the resulting induced map on homology maps the corresponding generator to zero. Out of curiosity, we now take the opposite approach of removing boxes from the covering of the attractor in an attempt to find an isolating neighborhood with interesting associated symbolic dynamics. Here we start with a box covering of the maximal invariant set (in this case, Hénon's strange attractor) and remove a small box neighborhood of the fold. We then remove a fixed number of preimages of this collection of boxes from the covering of the maximal invariant set. This procedure is outlined in Algorithm 10. From the resulting region of interest, we grow an isolating neighborhood and construct and verify symbolic dynamics as outlined in Algorithm 9.

Algorithm 10 (Fold preimage removal for constructing $\mathcal{S}$).

INPUT: grid $\mathcal{G}^d$, combinatorial enclosure $\mathcal{H}$ on $\mathcal{G}^d$,
region $\mathcal{N}_f^0 \subset \mathcal{G}^d$ containing the fold point,
computational parameter Max_Preimage_Iter
OUTPUT: region of interest $\mathcal{S}$

$\mathcal{S}$ = fold_preimage_removal($\mathcal{G}^d$, $\mathcal{H}$, Max_Preimage_Iter)
$\mathcal{N}_f = \mathcal{N}_f^0$;
$\mathcal{S} = \mathcal{G}^d \setminus \mathcal{N}_f$;
for $i = 1 \ldots$ Max_Preimage_Iter,
Fold_Iter $= \mathcal{H}^{-1}(\mathcal{N}_f)$;
$\mathcal{S} = \mathcal{S} \setminus \mathcal{N}_f$;
end
return $\mathcal{S}$;

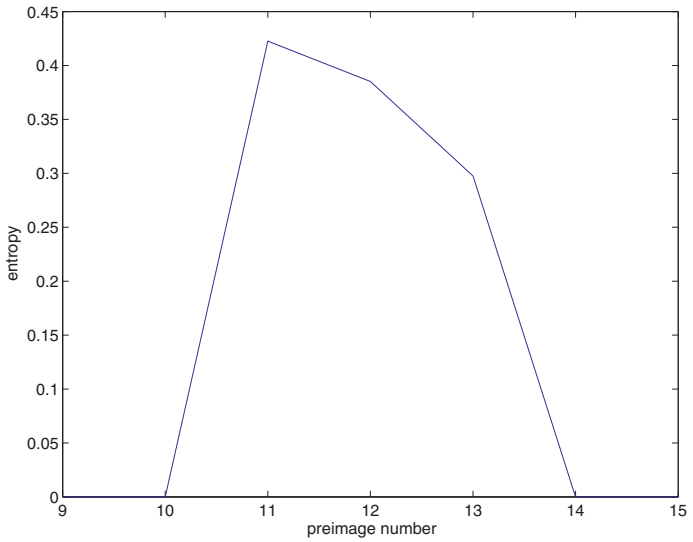


Figure 7. *Entropy lower bounds computed using the fold preimage removal technique for the Hénon map on grid* $\mathcal{G}^{(12)}$*. The horizontal axis gives the number,* `Max_Preimage_Iter`*, of preimages of the fold removed before growing the isolating neighborhood.*

Figure 7 depicts entropy bounds resulting from computations made starting with Algorithm 10 and various values of `Max_Preimage_Iter`. Removing too few preimages of the fold boxes (`Max_Preimage_Iter` small) does not yield interesting symbolic dynamics since we are unable to isolate this set at the given resolution. Removing too many preimages (`Max_Preimage_Iter` large) results in a subshift system consisting of disjoint cycles with 0 entropy. At depth 12, `Max_Preimage_Iter` $= 11$ provides the highest entropy bound, and this optimal constant increases at greater depths.

We obtain the following theorem by applying this third approach to the Hénon map.

Theorem 4.3 (fold and preimage removal). *The topological entropy of the Hénon map* (4.1) *is bounded from below by* 0.4225*.*

Outline of proof. Starting with a covering of the Hénon attractor by elements in $\mathcal{G}^{(12)}$, we use Algorithm 10 to remove `Max_Preimage_Iter` $= 11$ preimages (under $\mathcal{H}$) of $(1.2717, -0.0207) + [-0.04, 0.04] \times [-0.002, 0.002]$, a neighborhood of the "fold." We then use the resulting region of interest $\mathcal{S}$ together with $\mathcal{G}^{(12)}$, and $\mathcal{H}$ as the input for Algorithm 9. The computed index pair is shown in Figure 8. The homology map computed using Algorithm 3 is a map on 1281 generators of the first relative homology group. Algorithm 5 reduces the number of required generators to 191 by computing an appropriate shift equivalent index map. Finally, Algorithm 8 produces a topologically conjugate subshift on 129 symbols with topological entropy bounded from below by 0.4225. The result follows from Theorem 2.7. ∎

5. Concluding remarks.

We have described an automated, algorithmic method for studying the dynamics of a discrete dynamical system $f : X \to X$. The method not only constructs a semiconjugate subshift of finite type, but also uses this information to compute a rigorous lower bound on the topological entropy for the system. The essential ingredient to this ap-

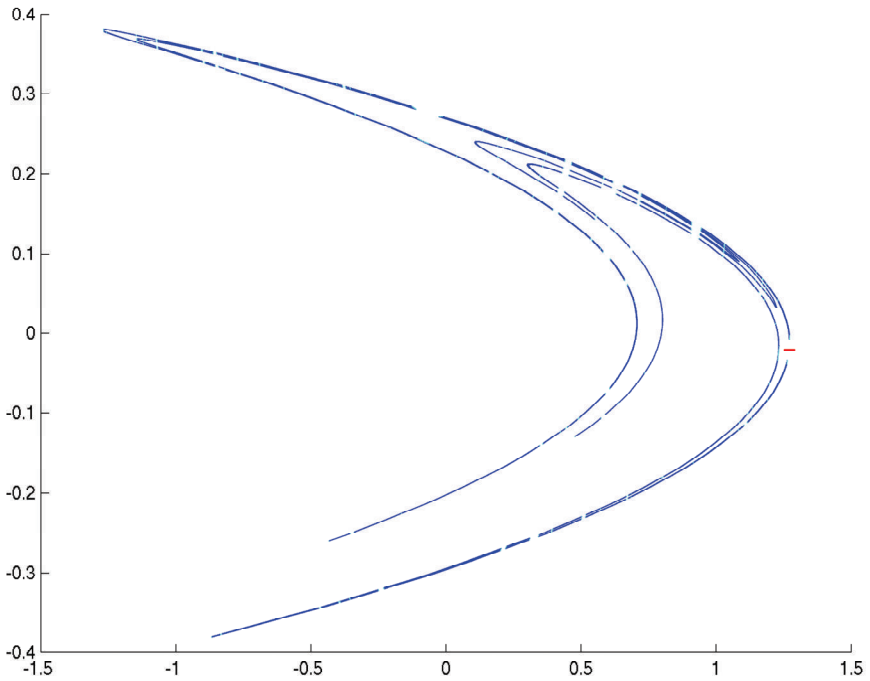


Figure 8. *The combinatorial index pair, $(\mathcal{P}_1, \mathcal{P}_0)$, constructed starting with Algorithm 10 at depth 12. ($\mathcal{P}_0$ is the collection of boxes shown in cyan.) The red rectangle shows the neighborhood of the fold point whose preimages were removed to construct the region of interest $\mathcal{S}$.*

proach is a computable "coarse" level of hyperbolicity in the map which is required to obtain a nontrivial Conley index. As the procedure stands, greater computational effort may be employed to improve the bounds. However, further analysis and optimization of the procedure described in section 3.1 for locating a region of interest should lead to even stronger results. A referee suggestion to consider more general sofic shifts rather than subshifts of finite type may also lead to the construction of semiconjugate symbolic dynamical systems with higher entropy.

The index processing techniques introduced in section 3.2 will enable further studies along these lines. As mentioned in the introduction, even infinite-dimensional systems may be studied in this manner. For such systems, it is necessary to incorporate both a dimension reduction for obtaining a computable system and analysis to overcome this reduction. These ideas are described in more detail in [DJM04] and would not, in principle, hinder entropy measurements of the type presented here.

Acknowledgments. The authors would like to thank Jim Wiseman and a referee for very helpful comments on the content and the structure of this paper.

REFERENCES

[AAC90] R. Artuso, E. Aurell, and P. Cvitanović, *Recycling of strange sets.* I. *Cycle expansions*, Nonlinearity, 3 (1990), pp. 325–359.

[ACE+87] D. Auerbach, P. Cvitanović, J.-P. Eckmann, G. Gunaratne, and I. Procaccia, *Exploring chaotic motion through periodic orbits*, Phys. Rev. Lett., 58 (1987), pp. 2387–2389.

[Bow71] R. Bowen, *Periodic points and measures for Axiom A diffeomorphisms*, Trans. Amer. Math. Soc., 154 (1971), pp. 377–397.

[Col02] P. Collins, *Symbolic dynamics from homoclinic tangles*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002), pp. 605–617.

[Con78] C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Reg. Conf. Ser. Math. 38, AMS, Providence, RI, 1978.

[Cse99] T. Csendes, ed., *INTLAB—INTerval LABoratory*, in Proceedings of the International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics, selected papers from the symposium (SCAN-98) held in Budapest, 1998, reprinted in Developments in Reliable Computing, Reliab. Comput. 5, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 213–357.

[Day03] S. Day, *A Rigorous Numerical Method in Infinite Dimensions*, Ph.D. thesis, School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 2003.

[Dev89] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, 2nd ed., Addison–Wesley Studies in Nonlinearity, Addison–Wesley (Advanced Book Program), Redwood City, CA, 1989.

[DFJ01] M. Dellnitz, G. Froyland, and O. Junge, *The algorithms behind GAIO-set oriented numerical methods for dynamical systems*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, Springer, Berlin, 2001, pp. 145–174.

[Die05] R. Diestel, *Graph Theory*, 3rd ed., Grad. Texts in Math. 173, Springer-Verlag, Berlin, 2005.

[DJM04] S. Day, O. Junge, and K. Mischaikow, *A rigorous numerical method for the global analysis of infinite-dimensional discrete dynamical systems*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 117–160.

[Eas75] R. W. Easton, *Isolating blocks and symbolic dynamics*, J. Differential Equations, 17 (1975), pp. 96–118.

[Gal01] Z. Galias, *Interval methods for rigorous investigations of periodic orbits*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 2427–2450.

[Gal02] Z. Galias, *Obtaining rigorous bounds for topological entropy for discrete time dynamical systems*, in Proceedings of the International Symposium on Nonlinear Theory and Its Applications, NOLTA'02 (Xi'an, PRC), 2002, pp. 619–622.

[Hén76] M. Hénon, *A two-dimensional mapping with a strange attractor*, Comm. Math. Phys., 50 (1976), pp. 69–77.

[MM02] K. Mischaikow and M. Mrozek, *Conley index*, in Handbook of Dynamical Systems, Vol. 2, North–Holland, Amsterdam, 2002, pp. 393–460.

[MS80] M. Misiurewicz and B. Szewc, *Existence of a homoclinic point for the Hénon map*, Comm. Math. Phys., 75 (1980), pp. 285–291.

[NBGM08] S. Newhouse, M. Berz, J. Grote, and K. Makino, *On the estimation of topological entropy on surfaces*, in Contemp. Math. 469, AMS, Providence, RI, 2008, pp. 243–270.

[Pil98] P. Pilarczyk, *Homology Computation-Software and Examples*, software guide, Jagiellonian University, 1998; http://www.im.uj.edu.pl/~vpilarczy/homology.htm.

[Rob95] C. Robinson, *Dynamical Systems. Stability, Symbolic Dynamics, and Chaos*, Stud. Adv. Math., CRC Press, Boca Raton, FL, 1995.

[RS88] J. W. Robbin and D. Salamon, *Dynamical systems, shape theory and the Conley index*, Ergodic Theory Dynam. Systems, 8 (1988), pp. 375–393.

[Szy95] A. Szymczak, *The Conley index for decompositions of isolated invariant sets*, Fund. Math., 148 (1995), pp. 71–90.

[Szy96] A. Szymczak, *The Conley index and symbolic dynamics*, Topology, 35 (1996), pp. 287–299.

[Szy97] A. Szymczak, *A combinatorial procedure for finding isolating neighbourhoods and index pairs*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1075–1088.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1507–1526
© 2008 Society for Industrial and Applied Mathematics

Asymptotic Expansions of I-V Relations via a Poisson–Nernst–Planck System*

Nicole Abaid†‡, Robert S. Eisenberg§, and Weishi Liu†

Abstract. We investigate higher order matched asymptotic expansions of a steady-state Poisson–Nernst–Planck (PNP) system with particular attention to the I-V relations of ion channels. Assuming that the Debye length is small relative to the diameter of the narrow channel, the PNP system can be viewed as a singularly perturbed system. Special structures of the zeroth order inner and outer systems make it possible to provide an *explicit* derivation of higher order terms in the asymptotic expansions. For the case of zero permanent charge, our results concerning the I-V relation for two oppositely charged ion species are (i) the first order correction to the zeroth order linear I-V relation is generally *quadratic* in V; (ii) when the electro-neutrality condition is enforced at both ends of the channel, there is NO first order correction, but the second order correction is *cubic* in V. Furthermore (Theorem 3.4), up to the second order, the cubic I-V relation has (except for a very degenerate case) three distinct real roots that correspond to the bistable structure in the FitzHugh–Nagumo simplification of the Hodgkin–Huxley model.

Key words. singular perturbation, matched asymptotic expansion, I-V relations

AMS subject classifications. 34E05, 34E13, 34B16, 92C35, 92C37

DOI. 10.1137/070691322

1. Introduction.

The Poisson–Nernst–Planck (PNP) systems are basic electro-diffusion equations modeling, for example, ion flow through membrane channels and transport of holes and electrons in semiconductors (see, for example, [3, 4, 5, 10, 21, 32, 39, 18, 19, 34, 35]). In the context of ion flow through a membrane channel, the flow of ions is driven by their concentration gradients and by the electric field modeled together by the Nernst–Planck continuity equations, and the electric field is in turn determined by the concentrations through the Poisson equation. The PNP system describes the current flow at low resolution; that is, it is an approximate description of the transport process [3] appropriate when channel selectivity (between different chemical species of ions) is not of importance, as, for example, in the numerous classical studies of the *gramicidin channel* [51] most recently reviewed in [1, 2] (see also [16, 36, 37, 11, 25, 51, 17, 20]). Derivation from a Langevin model of ionic diffusion [42] shows how correlations are approximated in PNP systems and suggests extensions of the PNP approach to deal with selectivity arising from excess chemical potentials [13, 14, 6]. The bio-

*Received by the editors May 11, 2007; accepted for publication (in revised form) by J. Keener August 11, 2008; published electronically December 3, 2008. The work is partially supported by NSF grant DMS-0406998, NIH grant NIGMS-076013-01, and University of Kansas General Research Fund allocation #2301158.
http://www.siam.org/journals/siads/7-4/69132.html
†Department of Mathematics, University of Kansas, Lawrence, KS 66045 (nabaid@math.ku.edu, wliu@math.ku.edu).
‡Current address: Department of Mechanical and Aerospace Engineering, Polytechnic Institute of New York University, Brooklyn, NY 11201 (nabaid01@students.poly.edu).
§Department of Molecular Biophysics and Physiology, Rush Medical Center, Chicago, IL 60612 (beisenbe@rush.edu).

logical properties of a channel called *permeation* can be described by the PNP equations. The biological properties called *selectivity* can be described by the extended PNP equations. Both permeation and selectivity are characterized by the current-voltage (I-V) relation measured experimentally under different ionic conditions.

The domain for the PNP system is a three-dimensional region including both the channel in the middle and the two baths at the ends of the tube. Thus there are some specifics regarding the domain geometry: the middle part of the domain representing the channel is much more narrow than the two ends that represent the two baths. A reasonable model of the domain would be a tubular-like region for the channel and two widely open conical regions for the baths. Another structure of the channel is its permanent charge, which is highly concentrated at the neck (center) of the channel. To capture the essence of the three-dimensional dynamics, one-dimensional PNP systems have been introduced with the key geometry ($h(x)$ in (1.1)) of the three-dimensional domain encoded in the equations. In particular, the following one-dimensional (steady-state) PNP system for n types of ion species was suggested by Nonner and Eisenberg in [32] and is derived in [42] and analyzed in [29] under the assumption that the characteristic radius of the channel is much smaller than its length:

$$\begin{aligned} &\frac{\epsilon^2}{h(x)}\frac{d}{dx}\left(h(x)\frac{d}{dx}\phi\right) = -\sum_{j=1}^{n}\alpha_j c_j - Q(x), \\ &\frac{dJ_k}{dx} = 0, \quad h(x)\frac{dc_k}{dx} + \alpha_k c_k h(x)\frac{d\phi}{dx} = -J_k, \quad k = 1, 2, \dots, n. \end{aligned} \tag{1.1}$$

Here the channel is normalized from $x = 0$ to $x = 1$; $h(x)$ is the scaled area function of the cross section of the channel at location x; $Q(x)$ is the permanent charge density; $\epsilon^2 = \lambda/r$, where λ is the Debye length and r is the characteristic radius of the channel; ϕ is the electric potential; and, for each $k = 1, 2, \dots, n$, c_k is the concentration, α_k is the valence, and J_k is the flux density (scaled by the diffusion constant) of the kth ion species. Note that J_k's are constant since we are considering the steady-state PNP system. The Debye length is customarily computed from the concentrations of ions on one side of the channel or the other, but we note for completeness that in some conditions the concentration of ions in the channel itself is quite different from that in the baths, and so the "local" Debye length (describing the inside of the channel) is quite different as well.

The baths are macroscopic regions in which the concentration of charges is nearly constant (because the dimensions of the reservoirs are macroscopic and so the total number of charges is hardly changed by the flows) and electrical potentials are nearly constant too. It is then natural to impose the following boundary conditions:

$$\phi(0) = V, \quad c_k(0) = L_k > 0; \quad \phi(1) = 0, \quad c_k(1) = R_k > 0. \tag{1.2}$$

Here V is the electric potential in the bath at the left end relative to that at the right end, and, for $k = 1, 2, \dots, n$, L_k and R_k are the concentrations of the kth ion species in the left and right baths, respectively.

In our one-dimensional setting, we model the region by a finite interval (normalized to $[0, 1]$). In this way, the interval can be partitioned into several subintervals so that the permanent charge takes large (in magnitude) values in the middle subintervals and small values away

from middle (zero near 0 and 1). There are other choices for the one-dimensional domain. For example, in [44], the whole real line is taken instead of a finite interval, and the boundary conditions are imposed at infinity. These two scalings correspond to different interpretations of relative ratios of different lengths involved in the biological problem. The mathematical treatment would be the same except at the boundaries. At this moment, it is not clear to us which interprets the biology better. Since the types of channels are so rich, we suspect that one scaling is better for certain types of channels and the other is better for other types.

With the setting in (1.1) and (1.2), the *I-V relation* means the dependence of the current $I = \sum \alpha_k J_k$ on the voltage V for fixed L_k's and R_k's. We remark that, in general, the I-V relation is NOT unique (see [30, 38, 39, 45, 46, 10] for $Q \neq 0$ and see [28] even for $Q = 0$). Understanding the nonuniqueness issue is critical for ion channels because nonuniqueness might explain the gating behavior of single channels, which switch suddenly and stochastically from one current level to another [16, 40, 31, 15]. Nonuniqueness is directly related to the stability of the corresponding steady states. The study of the stability problem is beyond the scope of this work. In section 3, we will consider special cases where the I-V relation is indeed unique.

System (1.1) together with the boundary condition (1.2) will be treated as a singular boundary value problem with $\epsilon \ll 1$ as the singular parameter. In, e.g., [5, 12, 27], the zeroth order I-V relation is explicitly obtained for two types of ion species and $Q = 0$ (see [10, 28] for a treatment of general situations and [44, 43] for a treatment using a three-dimensional PNP model). The zeroth order I-V relation, in this case, is linear. Experimental data show clearly a nonlinear I-V relation. But there is not much discussion on the nature of the nonlinearity except the famous Goldman–Hodgkin–Katz (GHK) I-V equation, which has been used to analyze experimental data for nearly 60 years but does not include such useful and necessary parameters as the charge of the channel protein.

It is our goal in this paper to examine higher order asymptotic expansions of the I-V relation. In particular, we are interested in higher order corrections to the zeroth order I-V relation. To obtain higher order asymptotic expansions of the I-V relation requires higher order asymptotic expansion of ϕ and c_k's. Both the classical matched asymptotic expansion method and the geometric singular perturbation method work well for the zeroth order term (see [5, 12, 27]) at least for the special case mentioned above. An advantage of the geometric method is that it also provides a rigorous justification directly of the validity of the zeroth order terms. But, for higher order terms, a direct application of the geometric singular perturbation theory seems not to work. Roughly speaking, it is not clear how to fully incorporate the information on the zeroth order terms in deriving systems for higher order terms. In fact, some natural approaches for higher order terms result in singularly perturbed systems that do not have the so-called *slow manifolds*. We therefore take the classical matched asymptotic expansion approach for higher order terms.

It is known that higher order terms satisfy linear but nonautonomous and nonhomogeneous systems. The homogeneous parts of the linear systems are the same and are nothing but the linearizations of the zeroth order nonlinear system along the zeroth order (inner and outer) solutions. Also known is the fact that it is generally impossible to get explicit solutions of a linear nonautonomous system. A special feature of the problem at hand is that the zeroth order nonlinear system possesses a complete set of integrals, and each integral provides an

integral for the linearization (see Propositions 2.1 and 2.2). It is this feature that allows us to carry out a detailed asymptotic analysis. Theoretically, any order of the asymptotic expansion can be found. But there are several technical difficulties even for obtaining the zeroth order terms; that is, one has to solve some nonlinear algebraic systems that are too complicated to expect explicit solutions in general. As a starting point, we will restrict ourselves to the simplest case where $n = 2$, $h = 1$, $Q = 0$, $\alpha_1 = -\alpha_2 = 1$. From the biological point of view, the most restrictive part of this constraint is $Q = 0$, since many channels have significant permanent charge. However, many channels are analyzed without consideration of permanent charge. The classical channel gramicidin, which has been simulated much more than any other channel we know of, has often been approximated that way [2, 51, 36, 37]. Maltoporin—one of the few channels with known high resolution crystallographic structure—has insignificant permanent charge [50, 24, 7, 49, 41]. Mathematically, this simplification allows us to obtain exact asymptotic expansions and, from those, explicit qualitative properties of I-V relations can be derived. Furthermore, a perturbation argument will allow one to generalize the result to cases where Q is small. Also, as mentioned above, we believe that the approach will work for general cases, say, a piecewise constant Q for quantitative results, particularly with the help of numerics. (This is a project that we will pursue in the future.) It should be clearly understood that some biological phenomena of importance—e.g., selectivity between cations—will not appear until three ions are considered, perhaps with different diffusion coefficients for each ionic type. Readers interested in more general cases may carry out the analysis along the lines of this paper with some extra analysis (see remarks in section 4).

For the special case where $n = 2$, $\alpha_1 = -\alpha_2 = 1$, $h = 1$, and $Q = 0$, our results give a *definite* nonlinear characterization of the I-V relations up to the second order ($O(\epsilon^2)$). More precisely, the first order correction to the zeroth order linear I-V relation is generally *quadratic* in V given in (3.20), but, when the electro-neutrality condition is imposed at both ends of the channel, the first order correction is *zero*. The second order correction is *cubic* in V even with the electro-neutrality condition (see formula (3.26)). Furthermore, the coefficient of the cubic term is always *negative* except for a highly degenerate case (see Theorem 3.4). The importance of this negative sign is that, up to the second order, the cubic I-V function has three distinct real roots—this agrees qualitatively with the I-V relation adopted in the FitzHugh–Nagumo simplification of the Hodgkin–Huxley systems. The existence of three roots of the I-V relation is responsible for the bistable structure in the FitzHugh–Nagumo system and may be related to the instabilities in biological channels called single channel gating [16, 40, 31, 15] and to the instabilities seen by [47, 48] in abiotic nanopores that have fixed structure. It should be pointed out that our analysis via the PNP system describes current flow through a single channel, and the FitzHugh–Nagumo equations (like the Hodgkin–Huxley system) describe an ensemble of channels in a biological membrane. The current through an ensemble of channels is determined by the current through a single channel and the gating process that determines the number of open channels. The FitzHugh–Nagumo equations have not yet been applied to abiotic channels of fixed structure [47, 48] as far as we know.

This paper is organized as follows. In section 2, we derive outer and inner systems for each order in the asymptotic expansions for a general situation. Starting in section 3, we restrict ourselves to the special case and examine the outer and inner expansions and matching. A treatment for the zeroth order is included in subsection 3.1 for completeness, and the first order

expansions and matching are detailed in subsection 3.2. In subsection 3.3, the analysis for the second order is carried out under the electro-neutrality assumption. Possible generalizations of our analysis and some general remarks are discussed in section 4.

For the reader's convenience, we provide a table of some notation to be used in this paper:

$x = 0$: the left end of the channel;
$x = 1$: the right end of the channel;
$\phi(x)$: the electric potential over the channel;
$0 = \phi(1)$: the reference of the potential set at the right end;
$V = \phi(0)$: the relative potential at the left end;
$c_k(x)$: the concentration of the kth ion species over the channel;
$L_k = c_k(0)$: the concentration of the kth ion species at the left end;
$R_k = c_k(1)$: the concentration of the kth ion species at the right end;
$Q(x)$: the permanent charge of the channel;
ε: the singular parameter related to the Debye length;
J_{kj}: the jth order in ε of the kth flux density;
$J_k = \sum_j \varepsilon^j J_{kj}$: the flux density of the kth ion species;
$T_j = \sum_k J_{kj}$: the jth order in ε of the diffusion flux density;
$I_j = \sum_k \alpha_k J_{kj}$: the jth order in ε of the current density;
$T = \sum_j \varepsilon^j T_j$: the diffusion flux density;
$I = \sum_j \varepsilon^j I_j$: the current density;
I-V relation: the dependence of I on V for fixed L_k, R_k, and Q.

2. Systems for asymptotic expansions. In this section, we derive the outer and inner systems for the asymptotic expansions and describe the matching principle to be employed. The outer systems govern the dynamics of ion flows within the channel, and the inner systems determine the potential boundary layers representing the effects of boundary conditions coming from the bath conditions. The matching principle then provides the interaction between the boundary conditions and the internal dynamics. In this sense, the boundary layers model the channel-bath interfaces. The process is standard but we will point out two special structures, (2.4) and Proposition 2.1, that allow us to obtain explicit information.

2.1. Outer systems for each order. We assume Q is constant and look for outer expansion of the form, for $k = 1, 2, \ldots, n$,

$$\begin{aligned} \phi(x;\epsilon) &= \phi_0(x) + \epsilon\phi_1(x) + \epsilon^2\phi_2(x) + \cdots, \\ c_k(x;\epsilon) &= c_{k0}(x) + \epsilon c_{k1}(x) + \epsilon^2 c_{k2}(x) + \cdots, \\ J_k &= J_{k0} + \epsilon J_{k1} + \epsilon^2 J_{k2} + \cdots. \end{aligned} \tag{2.1}$$

Substituting (2.1) into (1.1) and denoting the derivatives with respect to x by overdots, with the convention that $\phi_{-1} = \phi_{-2} = 0$, $\delta_0 = 1$, and $\delta_j = 0$ for $j \neq 0$, the jth order system is

$$\begin{aligned} \ddot{\phi}_{j-2} + h^{-1}(x)h_x(x)\dot{\phi}_{j-2} &= -\sum_{l=1}^{n} \alpha_l c_{lj} - \delta_j Q, \\ \dot{c}_{kj} &= -\sum_{p+q=j} \alpha_k c_{kp}\dot{\phi}_q - h^{-1}(x)J_{kj}. \end{aligned} \tag{2.2}$$

Upon introducing $u_j = h(x)\dot{\phi}_j$, system (2.2) becomes

$$\begin{aligned} \dot{\phi}_{j-2} &= h^{-1}(x)u_{j-2}, \\ \dot{u}_{j-2} &= -h(x)\sum_{l=1}^{n}\alpha_l c_{lj} - \delta_j h(x)Q, \\ \dot{c}_{kj} &= -h^{-1}(x)\sum_{p+q=j}\alpha_k c_{kp}u_q - h^{-1}(x)J_{kj}. \end{aligned} \tag{2.3}$$

An observation is that the homogeneous part for $\mathcal{C}_j = (c_{1j}, \dots, c_{nj})^T$ is

$$\dot{\mathcal{C}}_j = -\frac{u_0(x)}{h(x)}\mathcal{D}\mathcal{C}_j = -\dot{\phi}_0(x)\mathcal{D}\mathcal{C}_j, \tag{2.4}$$

where $\mathcal{D} = \operatorname{diag}\{\alpha_1, \dots, \alpha_n\}$ is a diagonal matrix. Once $\phi_0(x)$ is found, system (2.4) can be simply integrated:

$$\mathcal{C}_j(x) = \operatorname{diag}\left\{e^{-\alpha_1(\phi_0(x)-\phi_0(0))}, \dots, e^{-\alpha_n(\phi_0(x)-\phi_0(0))}\right\}\mathcal{C}_j(0).$$

Hence, system (2.3) can be solved by the method of variation of parameters.

2.2. Inner systems for each order. There will be two sets of inner systems, one at the boundary $x = 0$ and the other at the boundary $x = 1$.

2.2.1. Inner systems at the boundary $x = 0$. At the boundary $x = 0$, in terms of the inner variable $\xi = x/\epsilon$, let $\Phi(\xi;\epsilon) = \phi(\epsilon\xi;\epsilon)$, $C_k(\xi;\epsilon) = c_k(\epsilon\xi;\epsilon)$. System (1.1) becomes, for $k = 1, 2, \dots, n$,

$$\begin{aligned} h^{-1}(\epsilon\xi)\frac{d}{d\xi}\left(h(\epsilon\xi)\frac{d}{d\xi}\Phi\right) &= -\sum_{l=1}^{n}\alpha_l C_l - Q, \quad \frac{dJ_k}{d\xi} = 0, \\ h(\epsilon\xi)\frac{dC_k}{d\xi} + \alpha_k C_k h(\epsilon\xi)\frac{d\Phi}{d\xi} &= -\epsilon J_k. \end{aligned} \tag{2.5}$$

We look for the inner expansion of the form

$$\begin{aligned} \Phi(\xi;\epsilon) &= \Phi_0(\xi) + \epsilon\Phi_1(\xi) + \epsilon^2\Phi_2(\xi) + \cdots, \\ C_k(\xi;\epsilon) &= C_{k0}(\xi) + \epsilon C_{k1}(\xi) + \epsilon^2 C_{k2}(\xi) + \cdots, \\ J_k &= J_{k0} + \epsilon J_{k1} + \epsilon^2 J_{k2} + \cdots. \end{aligned} \tag{2.6}$$

Set

$$h(\epsilon\xi) = \sum_{p=0}\epsilon^p r_p^0 \xi^p, \quad h^{-1}(\epsilon\xi) = \sum_{p=0}\epsilon^p s_p^0 \xi^p,$$

where

$$r_p^0 = \frac{1}{p!}\frac{d^p h}{dx^p}(0) \quad \text{and} \quad s_p^0 = \frac{1}{p!}\frac{d^p h^{-1}}{dx^p}(0).$$

Denote by primes the derivatives with respect to ξ, and substitute (2.6) into (2.5) to get, for each $j = 0, 1, \ldots,$

$$\begin{aligned} &\sum_{p+q+l=j} s_p^0 \xi^p (r_q^0 \xi^q \Phi_l')' = -\sum_{l=1}^{n} \alpha_l C_{lj} - \delta_j Q, \\ &C_{kj}' = -\sum_{p+q=j} \alpha_k C_{kp} \Phi_q' - \sum_{p+q=j-1} s_p^0 \xi^p J_{kq}. \end{aligned} \tag{2.7}$$

By introducing $U_j = a_0 \Phi_j'$, we recast system (2.7) as

$$\begin{aligned} &\Phi_j' = s_0^0 U_j, \quad U_j' = -r_0^0 \sum_{l=1}^{n} \alpha_l C_{lj} - \delta_j r_0^0 Q, \\ &C_{kj}' = -s_0^0 \sum_{p+q=j} \alpha_k C_{kp} U_q - \sum_{p+q=j-1} s_p^0 \xi^p J_{kq}. \end{aligned} \tag{2.8}$$

For $j = 0$, the system is

$$\Phi_0' = s_0^0 U_0, \quad U_0' = -r_0^0 \sum_{l=1}^{n} \alpha_l C_{l0} - r_0^0 Q, \quad C_{k0}' = -s_0^0 \alpha_k C_{k0} U_0, \tag{2.9}$$

and, for all $j \geq 1$, system (2.8) has the same homogeneous part that is the linearization of the zeroth order system (2.9).

A specific structure of system (2.9) is revealed in the following.

Proposition 2.1. *The zeroth order inner system (2.9) has a complete set of $(n+1)$ first integrals given by, for $k = 1, 2, \ldots, n$,*

$$H_k = C_{k0} e^{\alpha_k \Phi_0}, \quad H_{n+1} = \frac{s_0^0}{2 r_0^0} U_0^2 - \sum_{l=1}^{n} C_{l0} + Q \Phi_0.$$

Proof. This can be verified directly (see also [28]). ∎

A crucial result is given below. We believe it is known but did not find a reference in the literature.

Proposition 2.2. *Consider an autonomous system*

$$z' = f(z), \quad z \in \mathbf{R}^m. \tag{2.10}$$

For a solution $z_0(t)$ of (2.10), consider the linearization along $z_0(t)$:

$$Z' = Df(z_0(t)) Z, \quad Z \in \mathbf{R}^m. \tag{2.11}$$

If a C^2 function $H : \mathbf{R}^m \to \mathbf{R}$ is an integral of system (2.10) (that is, $H(z(t))$ is independent of t for any solution $z(t)$ of (2.10)), then $G(Z, t) = \langle \nabla H(z_0(t)), Z \rangle$ is an integral of the linear system (2.11) (that is, $G(Z(t), t)$ is independent of t for any solution $Z(t)$ of (2.11)).

Proof. Since H is an integral of (2.10), $\frac{d}{dt}H(z(t)) = 0$ for all t, or

$$\langle \nabla H(z), f(z)\rangle = \sum_j \partial_j H(z) f_j(z) = 0$$

for all $z \in \mathbf{R}^m$. For any $k = 1, 2, \ldots, m$, take the partial derivative with respect to z_k to get

$$\sum_j \left(\partial^2_{kj} H(z) f_j(z) + \partial_j H(z) \partial_k f_j(z)\right) = 0. \tag{2.12}$$

As a consequence of (2.12), a computation gives that $\frac{d}{dt}G(Z(t), t) = 0$. ∎

As noted, the homogeneous part of (2.8) for $j \geq 1$ is the linearization of the zeroth order system (2.9); one can combine Propositions 2.1 and 2.2 to derive a complete set of integrals for the homogeneous part of (2.8). An application of variation of parameters allows one to get a closed form for the solutions of (2.8). For the general case presented in this section, certain technical difficulties arise in evaluating integrals explicitly. This is the main reason that, in section 3, we will restrict our analysis to a simple case.

2.2.2. Inner systems at the boundary $x = 1$. At the boundary $x = 1$, we use the inner variable $\xi = (-1 + x)/\epsilon$. Set $\Psi(\xi; \epsilon) = \phi(1 + \epsilon\xi; \epsilon)$ and $D_k(\xi; \epsilon) = c_k(1 + \epsilon\xi; \epsilon)$. We will then look for the inner expansion of the form:

$$\begin{aligned} \Psi(\xi;\epsilon) &= \Psi_0(\xi) + \epsilon\Psi_1(\xi) + \epsilon^2\Psi_2(\xi) + \cdots, \\ D_k(\xi;\epsilon) &= D_{k0}(\xi) + \epsilon D_{k1}(\xi) + \epsilon^2 D_{k2}(\xi) + \cdots, \\ J_k &= J_{k0} + \epsilon J_{k1} + \epsilon^2 J_{k2} + \cdots. \end{aligned} \tag{2.13}$$

In the same way as that for the boundary $x = 0$, one has, for each j,

$$\begin{aligned} \sum_{p+q+l=j} s^1_p \xi^p (r^1_q \xi^q \Psi'_l)' &= -\sum_{l=1}^{n} \alpha_l D_{lj} - \delta_j Q, \\ D'_{kj} &= -\sum_{p+q=j} \alpha_k D_{kp} \Psi'_q - \sum_{p+q=j-1} s^1_p \xi^p J_{kq}, \end{aligned} \tag{2.14}$$

where

$$r^1_p = \frac{1}{p!}\frac{d^p h}{dx^p}(1) \quad \text{and} \quad s^1_p = \frac{1}{p!}\frac{d^p h^{-1}}{dx^p}(1).$$

By introducing $V_j = r^1_0 \Psi'_j$, we get

$$\begin{aligned} \Psi'_j &= s^1_0 V_j, \quad V'_j = -r^1_0 \sum_{l=1}^{n} \alpha_l D_{lj} - \delta_j r^1_0 Q, \\ D'_{kj} &= -s^1_0 \sum_{p+q=j} \alpha_k D_{kp} V_q - \sum_{p+q=j-1} s^1_p \xi^p J_{kq}. \end{aligned} \tag{2.15}$$

2.3. Asymptotic matching principle. To piece together the inner solution and outer solution, one needs matching principles. There are two mainstreams in matching. One is the method of *intermediate matching* of Kaplun and Lagerstrom and the other is the *asymptotic matching principle* of Van Dyke (see [8, 9, 22, 23, 26]). The method of intermediate matching is based rigorously on the so-called *extension theorems*, but in general the implementation is more complicated than that of the asymptotic matching principle. The asymptotic matching principle, with a suitable hypothesis, can be also rigorously justified. It turns out that, for the problem handled in this paper, the so-called *outer manifold* is normally hyperbolic [10, 27], and Van Dyke's principle of asymptotic matching is justified (see, for example, [33]). We will thus use the asymptotic matching principle for our matching purpose.

To state the principle of asymptotic matching, we recall the notion of *kth order expansion operators* E_x^k and E_ξ^k in [26]: if, in terms of the outer variable x, $g(x;\epsilon) = \sum_{j=0}^{\infty} \epsilon^j g_j(x)$ and, in terms of the inner variable $\xi = x/\epsilon$, $f(\xi;\epsilon) = \sum_{j=0}^{\infty} \epsilon^j f_j(\xi)$, then

$$E_x^k(g(x;\epsilon)) = \sum_{j=0}^{k} \epsilon^j g_j(x), \quad E_\xi^k(f(\xi;\epsilon)) = \sum_{j=0}^{k} \epsilon^j f_j(\xi).$$

To match f and g up to the kth order at $x = 0$, one needs to express both $E_x^k(g(x;\epsilon))$ and $E_\xi^k(f(\xi;\epsilon))$ in the same variable. For example, to express $E_x^k(g(x;\epsilon)) = \sum_{j=0}^{k} \epsilon^j g_j(x)$ in terms of ξ, one replaces x by $\epsilon\xi$ in $g_j(x)$ and rewrites the expansion, say,

$$E_x^k(g(x;\epsilon)) = \sum_{j=0}^{k} \epsilon^j g_j(\epsilon\xi) = \sum_{j=0}^{\infty} \epsilon^j h_j(\xi);$$

in particular,

$$E_\xi^k E_x^k(g(x;\epsilon)) = \sum_{j=0}^{k} \epsilon^j h_j(\xi).$$

The kth order asymptotic matching principle for f and g at $x = 0$ is (see, for example, [8, 26]), in terms of the inner variable ξ,

$$E_\xi^k(f) = E_\xi^k E_x^k(g); \quad \text{that is,} \quad f_j(\xi) = h_j(\xi) \quad \text{for} \quad j = 0, 1, \ldots, k. \tag{2.16}$$

3. Matched asymptotic expansion: Case study. In this section, we will derive the matched asymptotic expansions for the case where $n = 2$, $\alpha_1 = -\alpha_2 = 1$, $Q = 0$, and $h = 1$. The zeroth order I-V relation turns out to be linear in V. While the first order correction is quadratic in V in general, it is zero when both ends are electro-neutral ($L_1 = L_2$ and $R_1 = R_2$). For this reason, we also carried out the analysis for the second order terms of the I-V relation under the electro-neutrality conditions and found that the I-V relation is a cubic in V.

For convenience, we set

$$I_j = \sum \alpha_k J_{kj} \qquad \text{and} \qquad T_j = \sum J_{kj}. \tag{3.1}$$

We point out that our main interest in the I-V relation is to derive the asymptotic expansion $I = I_0 + \epsilon I_1 + \epsilon^2 I_2 + \cdots$.

3.1. Zeroth order I-V relation. The zeroth order has been obtained in [5] using the asymptotic expansion method and in [27] using the geometric singular perturbation method. Here we rederive the zeroth order terms explicitly. This is crucial for an explicit formulation of higher order terms in the asymptotic expansions.

3.1.1. Zeroth order outer solution. From (2.3), the zeroth order outer system reads

$$0 = c_{10} - c_{20}, \quad \dot{c}_{10} = -c_{10}\dot{\phi}_0 - J_{10}, \quad \dot{c}_{20} = c_{20}\dot{\phi}_0 - J_{20}. \tag{3.2}$$

It is easy to solve system (3.2) to deduce

$$c_{10}(x) = c_{20}(x) = \frac{a_0 - T_0 x}{2}, \quad \phi_0(x) = b_0 + \frac{I_0}{T_0}\ln|a_0 - T_0 x| \tag{3.3}$$

for some constants a_0 and b_0 to be determined through matching. Here $I_0 = J_{10} - J_{20}$ and $T_0 = J_{10} + J_{20}$ from (3.1).

3.1.2. Zeroth order inner solution. At the boundary $x = 0$, the zeroth order inner system (2.8) is

$$\begin{aligned} &\Phi_0'(\xi) = U_0, \quad U_0'(\xi) = -C_{10} + C_{20}, \\ &C_{10}'(\xi) = -C_{10}U_0, \quad C_{20}'(\xi) = C_{20}U_0. \end{aligned} \tag{3.4}$$

In this case, Proposition 2.1 reads as follows.

Proposition 3.1. *System* (3.4) *has three first integrals given by*

$$H_1 = C_{10}e^{\Phi_0}, \quad H_2 = C_{20}e^{-\Phi_0}, \quad H_3 = \frac{1}{2}U_0^2 - C_{10} - C_{20}.$$

One can then solve system (3.4) explicitly (see [5, 27]) to get

$$\begin{aligned} &\Phi_0 = V + \frac{1}{2}\ln\frac{L_1}{L_2} + \ln\left(\frac{1 + le^{-\sqrt{M}\xi}}{1 - le^{-\sqrt{M}\xi}}\right)^2, \qquad U_0 = -\frac{4l\sqrt{M}e^{-\sqrt{M}\xi}}{1 - l^2e^{-2\sqrt{M}\xi}}, \\ &C_{10} = \sqrt{L_1L_2}\left(\frac{1 - le^{-\sqrt{M}\xi}}{1 + le^{-\sqrt{M}\xi}}\right)^2, \qquad C_{20} = \sqrt{L_1L_2}\left(\frac{1 + le^{-\sqrt{M}\xi}}{1 - le^{-\sqrt{M}\xi}}\right)^2, \end{aligned} \tag{3.5}$$

where

$$M = 2\sqrt{L_1L_2}, \quad l = \frac{L_2^{1/4} - L_1^{1/4}}{L_2^{1/4} + L_1^{1/4}}.$$

Similarly, at the boundary $x = 1$ with $\xi = (x - 1)/\epsilon$, we have

$$\begin{aligned} &\Psi_0(\xi) = \frac{1}{2}\ln\frac{R_1}{R_2} + \ln\left(\frac{1 + re^{\sqrt{N}\xi}}{1 - re^{\sqrt{N}\xi}}\right)^2, \qquad V_0(\xi) = \frac{4r\sqrt{N}e^{\sqrt{N}\xi}}{1 - r^2e^{2\sqrt{N}\xi}}, \\ &D_{10}(\xi) = \sqrt{R_1R_2}\left(\frac{1 - re^{\sqrt{N}\xi}}{1 + re^{\sqrt{N}\xi}}\right)^2, \qquad D_{20}(\xi) = \sqrt{R_1R_2}\left(\frac{1 + re^{\sqrt{N}\xi}}{1 - re^{\sqrt{N}\xi}}\right)^2, \end{aligned} \tag{3.6}$$

where

$$N = 2\sqrt{R_1R_2}, \quad r = \frac{R_2^{1/4} - R_1^{1/4}}{R_2^{1/4} + R_1^{1/4}}.$$

3.1.3. Zeroth order matching. In view of the matching principle (2.16), the matching conditions at the boundary $x = 0$ are

$$E^0_\xi E^0_x(c_k) = E^0_\xi(C_k) \quad \text{and} \quad E^0_\xi E^0_x(\phi) = E^0_\xi(\Phi).$$

From (3.3) and (3.5), we get

$$E^0_\xi E^0_x(c_k) = \frac{a_0}{2}, \quad E^0_\xi E^0_x(\phi) = b_0 + \frac{I_0}{T_0}\ln a_0,$$
$$E^0_\xi(C_k) = \sqrt{L_1 L_2}, \quad E^0_\xi(\Phi) = V + \frac{1}{2}\ln\frac{L_1}{L_2}.$$

The matching gives

$$a_0 = 2\sqrt{L_1 L_2} \quad \text{and} \quad b_0 = V + \frac{1}{2}\ln\frac{L_1}{L_2} - \frac{I_0}{T_0}\ln(2\sqrt{L_1 L_2}). \tag{3.7}$$

Similarly, the matching at the boundary $x = 1$ requires

$$a_0 - T_0 = 2\sqrt{R_1 R_2} \quad \text{and} \quad b_0 + \frac{I_0}{T_0}\ln(a_0 - T_0) = \frac{1}{2}\ln\frac{R_1}{R_2}. \tag{3.8}$$

We deduce, from (3.7) and (3.8), that

$$\begin{aligned} T_0 &= 2\sqrt{L_1 L_2} - 2\sqrt{R_1 R_2}, \\ I_0 &= \frac{2(\sqrt{L_1 L_2} - \sqrt{R_1 R_2})(2V + \ln(L_1 R_2) - \ln(L_2 R_1))}{\ln(L_1 L_2) - \ln(R_1 R_2)}. \end{aligned} \tag{3.9}$$

In particular, at the zeroth order, the I-V relation $I_0 = I_0(V)$ is *linear* in V. When $L_1 = L_2 = L$ and $R_1 = R_2 = R$ (electro-neutrality condition at both ends of the channel) hold, we have

$$T_0 = 2(L - R), \quad I_0 = \frac{2(L - R)}{\ln L - \ln R}V. \tag{3.10}$$

Note also that, as $L \to R$, we have $T_0 \to 0$ and $I_0 \to 2RV$.

3.2. First order I-V relation. Since the higher order outer systems (2.3) and inner systems (2.8) and (2.15) are nonautonomous, one cannot solve them explicitly in general. The upshot for our problem is that they can actually be solved explicitly. For inner systems, the solvability is due to Propositions 2.2 and 3.1.

3.2.1. First order outer solution. From (2.3), the first order outer system is

$$c_{11} = c_{21}, \quad \dot{c}_{11} = -(c_{10}\dot{\phi}_1 + c_{11}\dot{\phi}_0) - J_{11}, \quad \dot{c}_{21} = (c_{20}\dot{\phi}_1 + c_{21}\dot{\phi}_0) - J_{21}. \tag{3.11}$$

Recall, from (3.1), that $T_1 = J_{11} + J_{21}$ and $I_1 = J_{11} - J_{21}$. Using (3.3) for $(\phi_0(x), c_{10}(x), c_{20}(x))$, one solves system (3.11) to get

$$\begin{aligned} c_{11}(x) = c_{21}(x) &= \frac{a_1 - T_1 x}{2}, \\ \phi_1(x) &= b_1 + \frac{T_0 I_1 - I_0 T_1}{T_0^2}\ln|a_0 - T_0 x| + \frac{I_0(a_1 T_0 - a_0 T_1)}{T_0^2(a_0 - T_0 x)} \end{aligned}$$

for some unknown constants a_1 and b_1.

3.2.2. First order inner solution. The first order inner system (2.8) at $x = 0$ is

$$\begin{aligned} &\Phi_1' = U_1, \quad U_1' = -(C_{11} - C_{21}), \\ &C_{11}' = -(C_{10}U_1 + C_{11}U_0) - J_{10}, \quad C_{21}' = (C_{20}U_1 + C_{21}U_0) - J_{20}. \end{aligned} \tag{3.12}$$

As an application of Propositions 2.2 and 3.1, we have the next result.

Proposition 3.2. *The homogeneous part of* (3.12) *has the following integrals:*

$$\begin{aligned} G_1^h &= C_{11}e^{\Phi_0} + C_{10}e^{\Phi_0}\Phi_1, \\ G_2^h &= C_{21}e^{-\Phi_0} - C_{20}e^{-\Phi_0}\Phi_1, \\ G_3^h &= U_0U_1 - C_{11} - C_{21}. \end{aligned}$$

The full system (3.12) *has the following integrals:*

$$\begin{aligned} G_1 &= C_{11}e^{\Phi_0} + C_{10}e^{\Phi_0}\Phi_1 + J_{10}F_1(\xi), \\ G_2 &= C_{21}e^{-\Phi_0} - C_{20}e^{-\Phi_0}\Phi_1 + J_{20}F_2(\xi), \\ G_3 &= U_0U_1 - C_{11} - C_{21} - T_0\xi, \end{aligned}$$

where

$$\begin{aligned} F_1(\xi) &= \int_0^\xi e^{\Phi_0(s)}ds = -\sqrt{\frac{L_1}{L_2}}\frac{e^V}{\sqrt{M}}\left(\frac{4}{1 - le^{-\sqrt{M}\xi}} - \frac{4}{1-l} - \sqrt{M}\xi\right), \\ F_2(\xi) &= \int_0^\xi e^{-\Phi_0(s)}ds = -\sqrt{\frac{L_2}{L_1}}\frac{e^{-V}}{\sqrt{M}}\left(\frac{4}{1 + le^{-\sqrt{M}\xi}} - \frac{4}{1+l} - \sqrt{M}\xi\right). \end{aligned}$$

Proof. The first statement for the homogeneous part of system (3.12) follows from Propositions 2.2 and 3.1. The extra terms in the second statement are obtained by adding trial functions and forcing the resulting functions to be integrals of (3.12). This leads to

$$F_1(\xi) = \int_0^\xi e^{\Phi_0(s)}ds \quad \text{and} \quad F_2(\xi) = \int_0^\xi e^{-\Phi_0(s)}ds.$$

Direct integrations with Φ_0 in (3.5) give the explicit expressions for $F_1(\xi)$ and $F_2(\xi)$ as claimed. ∎

One can then use the integrals to solve system (3.12). Note that we have the initial conditions $\Phi_1(0) = 0$ and $C_{11}(0) = C_{21}(0) = 0$, but $U_1(0)$ has to be determined via matching. One finds, after careful integrations,

$$\begin{aligned} U_1(\xi) &= \frac{U_0(0)U_1(0) - (C_{10} - C_{20})\Phi_1 - J_{10}e^{-\Phi_0}F_1 - J_{20}e^{\Phi_0}F_2 + T_0\xi}{U_0}, \\ \Phi_1(\xi) &= -\frac{4l(T_0 + lI_0)}{M^{3/2}(1+l)(1-l)} - \frac{I_0}{M}\xi \\ &\quad + \frac{1}{2\sqrt{M}}\left(U_1(0) + \frac{I_0 + lT_0}{(1+l)(1-l)M}\right)e^{\sqrt{M}\xi} + o(e^{-\sqrt{M}\xi}). \end{aligned} \tag{3.13}$$

The term involving $e^{\sqrt{M}\xi}$ should disappear due to matching. Thus,

$$U_1(0) = -\frac{I_0 + lT_0}{(1+l)(1-l)M}.$$

In summary, we have

$$\begin{aligned}
\Phi_1(\xi) &= -\frac{4l(T_0 + lI_0)}{M^{3/2}(1+l)(1-l)} - \frac{I_0}{M}\xi + o(e^{-\sqrt{M}\xi}),\\
C_{11}(\xi) &= -\frac{2l(I_0 + lT_0)}{\sqrt{M}(1+l)(1-l)} - \frac{T_0}{2}\xi + o(e^{-\sqrt{M}\xi}),\\
C_{21}(\xi) &= -\frac{2l(I_0 + lT_0)}{\sqrt{M}(1+l)(1-l)} - \frac{T_0}{2}\xi + o(e^{-\sqrt{M}\xi}).
\end{aligned} \tag{3.14}$$

Similarly, at $x = 1$ with $x - 1 = \epsilon\xi$, we have

$$\begin{aligned}
\Psi_1(\xi) &= -\frac{4r(T_0 + rI_0)}{N^{3/2}(1+r)(1-r)} - \frac{I_0}{N}\xi + o(e^{-\sqrt{N}\xi}),\\
D_{11}(\xi) &= -\frac{2r(I_0 + rT_0)}{\sqrt{N}(1+r)(1-r)} - \frac{T_0}{2}\xi + o(e^{-\sqrt{N}\xi}),\\
D_{21}(\xi) &= -\frac{2r(I_0 + rT_0)}{\sqrt{N}(1+r)(1-r)} - \frac{T_0}{2}\xi + o(e^{-\sqrt{N}\xi}).
\end{aligned} \tag{3.15}$$

3.2.3. First order matching. We first consider the matching near $x = 0$. For the inner expansion, we have, from (3.5) and (3.14), for $k = 1, 2$,

$$\begin{aligned}
E^1_\xi(\Phi) &= E^1_\xi(\Phi_0(\xi) + \epsilon\Phi_1(\xi))\\
&= V + \frac{1}{2}\ln\frac{L_1}{L_2} - \epsilon\left(\frac{4l(T_0 + lI_0)}{M^{3/2}(1+l)(1-l)} + \frac{I_0}{M}\xi\right),\\
E^1_\xi(C_k) &= E^1_\xi(C_{k0}(\xi) + \epsilon C_{k1}(\xi))\\
&= \sqrt{L_1L_2} - \epsilon\left(\frac{2l(I_0 + lT_0)}{\sqrt{M}(1+l)(1-l)} + \frac{T_0}{2}\xi\right).
\end{aligned} \tag{3.16}$$

On the other hand, for the outer expansion, we have

$$\begin{aligned}
E^1_x(\phi) = E^1_x(\phi_0(x) + \epsilon\phi_1(x)) &= b_0 + \frac{I_0}{T_0}\ln a_0 - \frac{I_0}{a_0}x\\
&\quad + \epsilon\left(b_1 + \frac{T_0I_1 - I_0T_1}{T_0^2}\ln a_0 + \frac{I_0(a_1T_0 - a_0T_1)}{T_0^2a_0} + O(x)\right),\\
E^1_x(c_k) = E^1_x(c_{k0}(x) + \epsilon c_{k1}(x)) &= \left(\frac{a_0}{2} - \frac{T_0}{2}x\right) + \epsilon\left(\frac{a_1}{2} - \frac{T_1}{2}x\right).
\end{aligned}$$

Therefore, in terms of the inner variable ξ,

$$
\begin{aligned}
E^1_\xi E^1_x(\phi) &= b_0 + \frac{I_0}{T_0}\ln a_0 \\
&\quad + \epsilon\left(b_1 + \frac{T_0 I_1 - I_0 T_1}{T_0^2}\ln a_0 + \frac{I_0(a_1 T_0 - a_0 T_1)}{T_0^2 a_0} - \frac{I_0}{a_0}\xi\right), \\
E^1_\xi E^1_x(c_k) &= \frac{a_0}{2} + \epsilon\left(\frac{a_1}{2} - \frac{T_0}{2}\xi\right).
\end{aligned} \tag{3.17}
$$

The matchings $E^1_\xi(\Phi) = E^1_\xi E^1_x(\phi)$ and $E^1_\xi(C_k) = E^1_\xi E^1_x(c_k)$ imply, from (3.16) and (3.17),

$$
\begin{aligned}
a_0 &= M = 2\sqrt{L_1 L_2}, \qquad b_0 = V + \frac{1}{2}\ln\frac{L_1}{L_2} - \frac{I_0}{T_0}\ln a_0, \\
a_1 &= -\frac{4l(I_0 + lT_0)}{\sqrt{M}(1+l)(1-l)}, \\
b_1 &= -\frac{I_0(a_1 T_0 - a_0 T_1)}{T_0^2 a_0} - \frac{T_0 I_1 - I_0 T_1}{T_0^2}\ln a_0 \\
&\quad - \frac{4l(T_0 + lI_0)}{M^{3/2}(1+l)(1-l)}.
\end{aligned} \tag{3.18}
$$

Note that the relation for a_0 in (3.18) is consistent since $M = 2\sqrt{L_1 L_2}$.

Similarly, the matching near $x = 1$ gives

$$
\begin{aligned}
a_0 &= N + T_0 = 2\sqrt{R_1 R_2} + T_0, \qquad b_0 = \frac{1}{2}\ln\frac{R_1}{R_2} - \frac{I_0}{T_0}\ln(a_0 - T_0), \\
a_1 &= T_1 - \frac{4r(I_0 + rT_0)}{\sqrt{N}(1+r)(1-r)}, \\
b_1 &= -\frac{I_0(a_1 T_0 - a_0 T_1)}{2\sqrt{R_1 R_2}T_0^2} - \frac{T_0 I_1 - I_0 T_1}{T_0^2}\ln(2\sqrt{R_1 R_2}) \\
&\quad - \frac{4r(T_0 + rI_0)}{N^{3/2}(1+r)(1-r)}.
\end{aligned} \tag{3.19}
$$

As expected, one recovers (3.9) for T_0 and I_0 from the two expressions in (3.18) and (3.19) for a_0 and b_0. Using the two expressions for a_1 and b_1, we get

$$
\begin{aligned}
T_1 &= \frac{4r(I_0 + rT_0)}{\sqrt{N}(1+r)(1-r)} - \frac{4l(I_0 + lT_0)}{\sqrt{M}(1+l)(1-l)}, \\
I_1 &= T_0^{-1} I_0 T_1 - \frac{T_0^{-1} I_0(a_1 T_0 - a_0 T_1)}{\ln(R_1 R_2) - \ln(L_1 L_2)}\left(\frac{1}{\sqrt{R_1 R_2}} - \frac{1}{\sqrt{L_1 L_2}}\right) \\
&\quad - \frac{4rT_0(T_0 + rI_0)}{\sqrt{R_1 R_2 N}(1+r)(1-r)} + \frac{4lT_0(T_0 + lI_0)}{\sqrt{L_1 L_2 M}(1+l)(1-l)}.
\end{aligned} \tag{3.20}
$$

Note that T_1 is linear in I_0 and hence is linear in V; I_1 is quadratic in I_0 and hence is also quadratic in V. That is, the first order correction to the zeroth order linear I-V relation is *quadratic* in V in general.

What is interesting and potentially important is that, when $L_1 = L_2$ and $R_1 = R_2$ (electro-neutrality), one deduces that $T_1 = I_1 = 0$. That is, under the electro-neutrality condition at both ends, there is NO first order correction for the I-V relation.

3.3. Second order I-V relation with electro-neutrality. We now assume the electro-neutrality condition $L_1 = L_2$ and $R_1 = R_2$ and examine the I-V correction at the second order $O(\epsilon^2)$.

3.3.1. Second order outer expansion. The second order outer system (2.3) is

$$\begin{aligned} \ddot{\phi}_0 &= -c_{12} + c_{22}, \\ \dot{c}_{12} &= -(c_{12}\dot{\phi}_0 + c_{11}\dot{\phi}_1 + c_{10}\dot{\phi}_2) - J_{12}, \\ \dot{c}_{22} &= (c_{22}\dot{\phi}_0 + c_{21}\dot{\phi}_1 + c_{20}\dot{\phi}_2) - J_{22}. \end{aligned} \tag{3.21}$$

Note that, under the assumption $L_1 = L_2$ and $R_1 = R_2$, we have

$$a_1 = b_1 = T_1 = I_1 = c_{11}(x) = c_{21}(x) = \phi_1(x) = 0.$$

Upon using (3.3), one solves system (3.21) to get

$$\begin{aligned} c_{12}(x) &= \frac{a_2 - T_2 x}{2} + \frac{I_0^2 + 2I_0T_0}{4(a_0 - T_0x)^2}, \\ c_{22}(x) &= \frac{a_2 - T_2 x}{2} + \frac{I_0^2 - 2I_0T_0}{4(a_0 - T_0x)^2}, \\ \phi_2(x) &= b_2 - \frac{2I_0T_0}{3(a_0 - T_0x)^3} + \frac{I_0^3}{6T_0(a_0 - T_0x)^3} + \frac{I_0(a_2T_0 - a_0T_2)}{T_0^2(a_0 - T_0x)} \\ &\quad - \frac{I_0T_2}{T_0^2}\ln|a_0 - T_0x| + \frac{I_2}{T_0}\ln|a_0 - T_0x|, \end{aligned} \tag{3.22}$$

where a_2 and b_2 are unknown constants.

3.3.2. Second order inner expansion. The second order inner system (2.8) at $x = 0$ is

$$\begin{aligned} \Phi_2' &= U_2, \quad U_2' = -(C_{12} - C_{22}), \\ C_{12}' &= -(C_{10}U_2 + C_{11}U_1 + C_{12}U_0) - J_{11}, \\ C_{22}' &= (C_{20}U_2 + C_{21}U_1 + C_{22}U_0) - J_{21}. \end{aligned} \tag{3.23}$$

Similarly to the first order inner system, we have the following claim.

Proposition 3.3. *System* (3.23) *has the following integrals:*

$$\begin{aligned} G_1 &= C_{12}e^{\Phi_0} + C_{10}e^{\Phi_0}\Phi_2 + J_{11}F_1(\xi) + F_{12}(\xi), \\ G_2 &= C_{22}e^{-\Phi_0} - C_{20}e^{-\Phi_0}\Phi_2 + J_{21}F_2(\xi) - F_{22}(\xi), \\ G_3 &= U_0U_2 - C_{12} - C_{22} - T_1\xi + \frac{1}{2}U_1^2, \end{aligned}$$

where $F_1(\xi)$ *and* $F_2(\xi)$ *are given in Proposition* 3.2 *and*

$$F_{12}(\xi) = \int_0^\xi C_{11}(s)U_1(s)e^{\Phi_0(s)}ds, \qquad F_{22}(\xi) = \int_0^\xi C_{21}(s)U_1(s)e^{-\Phi_0(s)}ds.$$

Using $L_1 = L_2 = L$ and $R_1 = R_2 = R$, we have, from (3.5) and (3.6), that

$$\begin{aligned} &\Phi_0(\xi) = V, \quad U_0(\xi) = 0, \quad C_{10}(\xi) = C_{20}(\xi) = L, \\ &\Psi_0(\xi) = 0, \quad V_0(\xi) = 0, \quad D_{10}(\xi) = D_{20}(\xi) = R, \end{aligned}$$

and, from (3.14) and (3.15), that

$$\begin{aligned} &\Phi_1(\xi) = -\frac{I_0}{2L}\xi, \quad U_1(\xi) = -\frac{I_0}{2L}, \quad C_{11}(\xi) = C_{21}(\xi) = -\frac{T_0}{2}\xi, \\ &\Psi_1(\xi) = -\frac{I_0}{2R}\xi, \quad V_1(\xi) = -\frac{I_0}{2R}, \quad D_{11}(\xi) = D_{21}(\xi) = -\frac{T_0}{2}\xi. \end{aligned}$$

Also,

$$J_{11} = J_{21} = 0, \quad F_{12}(\xi) = \frac{I_0T_0}{8L}e^V\xi^2, \quad F_{22}(\xi) = \frac{I_0T_0}{8L}e^{-V}\xi^2.$$

Applying the integrals in Proposition 3.3, we can solve (3.23) with $\Phi_2(0) = C_{12}(0) = C_{22}(0) = 0$ to get

$$\Phi_2(\xi) = \left(\frac{I_0T_0}{8L^3} - A\right)e^{-\sqrt{2L}\xi} + Ae^{\sqrt{2L}\xi} - \frac{I_0T_0}{8L^3} - \frac{I_0T_0}{8L^2}\xi^2.$$

The matching will force $A = 0$. Thus, for $\xi \geq 0$,

$$\begin{aligned} \Phi_2(\xi) &= \frac{I_0T_0}{8L^3}\left(e^{-\sqrt{2L}\xi} - 1\right) - \frac{I_0T_0}{8L^2}\xi^2, \\ C_{12}(\xi) &= -\frac{I_0T_0}{8L^2}\left(e^{-\sqrt{2L}\xi} - 1\right), \quad C_{22}(\xi) = \frac{I_0T_0}{8L^2}\left(e^{-\sqrt{2L}\xi} - 1\right). \end{aligned} \tag{3.24}$$

Similarly, at $x = 1$, the second order inner solution is, for $\xi \leq 0$,

$$\begin{aligned} \Psi_2(\xi) &= \frac{I_0T_0}{8R^3}\left(e^{\sqrt{2R}\xi} - 1\right) - \frac{I_0T_0}{8R^2}\xi^2, \\ D_{12}(\xi) &= -\frac{I_0T_0}{8R^2}\left(e^{\sqrt{2R}\xi} - 1\right), \quad D_{22}(\xi) = \frac{I_0T_0}{8R^2}\left(e^{\sqrt{2R}\xi} - 1\right). \end{aligned} \tag{3.25}$$

3.3.3. Second order matching. From (3.22), in terms of $\xi = x/\epsilon$, the outer expansion at $x = 0$ is

$$\begin{aligned} E_\xi^2E_x^2(\phi) = b_0 &+ \frac{I_0}{T_0}\ln a_0 - \epsilon\frac{I_0}{a_0}\xi \\ &+ \epsilon^2\left(b_2 - \frac{I_0(2T_0 - I_0)(2T_0 + I_0)}{6T_0a_0^3} + \frac{I_0(a_2T_0 - a_0T_2)}{T_0^2a_0}\right. \\ &\left. + \frac{T_0I_2 - I_0T_2}{T_0^2}\ln a_0 - \frac{I_0T_0}{2a_0^2}\xi^2\right), \\ E_\xi^2E_x^2(c_1) = \frac{a_0}{2} &- \epsilon\frac{T_0}{2}\xi + \epsilon^2\left(\frac{a_2}{2} + \frac{I_0^2 + 2T_0I_0}{4a_0^2}\right), \\ E_\xi^2E_x^2(c_2) = \frac{a_0}{2} &- \epsilon\frac{T_0}{2}\xi + \epsilon^2\left(\frac{a_2}{2} + \frac{I_0^2 - 2T_0I_0}{4a_0^2}\right), \end{aligned}$$

and, in terms of $\xi = (x-1)/\epsilon$, the outer expansion at $x = 1$ is

$$\begin{aligned} E_\xi^2 E_x^2(\phi) &= b_0 + \frac{I_0}{T_0}\ln(a_0 - T_0) - \epsilon\frac{I_0}{a_0 - T_0}\xi \\ &\quad + \epsilon^2\left(b_2 - \frac{I_0(2T_0 - I_0)(2T_0 + I_0)}{6T_0(a_0 - T_0)^3} + \frac{I_0(a_2T_0 - a_0T_2)}{T_0^2(a_0 - T_0)}\right. \\ &\quad \left. + \frac{T_0I_2 - I_0T_2}{T_0^2}\ln(a_0 - T_0) - \frac{I_0T_0}{2(a_0 - T_0)^2}\xi^2\right), \\ E_\xi^2 E_x^2(c_1) &= \frac{a_0 - T_0}{2} - \epsilon\frac{T_0}{2}\xi + \epsilon^2\left(\frac{a_2 - T_2}{2} + \frac{I_0^2 + 2T_0I_0}{4(a_0 - T_0)^2}\right), \\ E_\xi^2 E_x^2(c_2) &= \frac{a_0 - T_0}{2} - \epsilon\frac{T_0}{2}\xi + \epsilon^2\left(\frac{a_2 - T_2}{2} + \frac{I_0^2 - 2T_0I_0}{4(a_0 - T_0)^2}\right). \end{aligned}$$

From (3.24) and (3.25), the inner expansion at $x = 0$ is

$$\begin{aligned} E_\xi^2(\Phi) &= V - \epsilon\frac{I_0}{2L}\xi - \epsilon^2\left(\frac{I_0T_0}{8L^3} + \frac{I_0T_0}{8L^2}\xi^2\right), \\ E_\xi^2(C_1) &= L - \epsilon\frac{T_0}{2}\xi + \epsilon^2\frac{I_0T_0}{8L^2}, \quad E_\xi^2(C_2) = L - \epsilon\frac{T_0}{2}\xi - \epsilon^2\frac{I_0T_0}{8L^2}, \end{aligned}$$

and the inner expansion at $x = 1$ is

$$\begin{aligned} E_\xi^2(\Psi) &= -\epsilon\frac{I_0}{2R}\xi - \epsilon^2\left(\frac{I_0T_0}{8R^3} + \frac{I_0T_0}{8R^2}\xi^2\right), \\ E_\xi^2(D_1) &= R - \epsilon\frac{T_0}{2}\xi + \epsilon^2\frac{I_0T_0}{8R^2}, \quad E_\xi^2(D_2) = R - \epsilon\frac{T_0}{2}\xi - \epsilon^2\frac{I_0T_0}{8R^2}. \end{aligned}$$

The matchings at $x = 0$ and at $x = 1$ then give

$$\begin{aligned} T_2 &= \frac{(L-R)^3(L+R)}{2L^2R^2(\ln L - \ln R)^2}, \\ I_2 &= \frac{(L-R)^4(L^2 + LR + R^2)}{3L^3R^3(\ln L - \ln R)^2}V - \frac{(L-R)^3(L^3 - R^3)}{3L^3R^3(\ln L - \ln R)^4}\nu_0^3 \\ &\quad + \frac{(L-R)^2(L^2 - R^2)}{2L^2R^2(\ln L - \ln R)^3}V^3. \end{aligned} \tag{3.26}$$

In particular, the second order correction $I_2(V)$ to the zeroth order I-V relation $I_0(V)$ is *cubic* in V. Note also that, as $L \to R$, one finds that $T_2 \to 0$ and $I_2 \to 0$.

Theorem 3.4. *If $L \neq R$, then, up to the order of ϵ^2, the I-V relation $I = I(V)$ in (3.26) is a cubic function with three distinct real roots.*

Proof. From (3.10) and (3.26), up to $O(\epsilon^2)$, $I = f(L, R; \epsilon)V - \epsilon^2 g(L, R)V^3$, where

$$\begin{aligned} f(L, R; \epsilon) &= \frac{2(L-R)}{\ln L - \ln R} + \epsilon^2\frac{(L-R)^4(L^2 + LR + R^2)}{3L^3R^3(\ln L - \ln R)^2}, \\ g(L, R) &= \frac{(L-R)^3(L^3 - R^3)}{3L^3R^3(\ln L - \ln R)^4} - \frac{(L-R)^2(L^2 - R^2)}{2L^2R^2(\ln L - \ln R)^3}. \end{aligned}$$

It is easy to see that $f(L,R;\epsilon) > 0$ for $L \neq R$ (and $f(R,R;\epsilon) = 2R$). It remains to show that $g(L,R) > 0$ for $L \neq R$. Note that $g(L,R) = g(R,L)$. Thus, it suffices to show $g(L,R) > 0$ for $L > R$. Assume now $L > R$ and rewrite $g(L,R)$ as

$$g(L,R) = \frac{(L-R)^3}{6L^3R^3(\ln L - \ln R)^4} h(L,R),$$

where $h(L,R) = 2(L^3 - R^3) - 3LR(L+R)(\ln L - \ln R)$. To show $h(L,R) > 0$ for $L > R$, we fix R and treat $h(L) = h(L,R)$ as a function of L. Then, a direct computation gives $h(R) = h'(R) = h''(R) = 0$ but $h'''(L) > 0$ for all L. Therefore, $h(L) > 0$ for $L > R$. ■

4. Some remarks. We investigated higher order asymptotic expansion of the I-V relation for biological channels via a one-dimensional steady-state Poisson–Nernst–Planck system. For the case of two oppositely charged ion species and zero permanent charge, we obtained explicit information on the I-V relation up to the second order. In particular, we found that the zeroth order I-V relation is *linear*, the first order correction to the zeroth order I-V relation is generally *quadratic*, and, with the electro-neutrality condition at both ends of the channel, there is NO first order correction but the second order correction is *cubic*. Furthermore, up to the second order, the cubic I-V relation has three real roots (Theorem 3.4), which is potentially related to the cubic-like feature of the average I-V relation of a population of channels in the FitzHugh–Nagumo simplification of the Hodgkin–Huxley model.

For second order terms, we only treated the electro-neutrality case because this is a natural biological assumption and the first order correction to the zeroth order linear I-V relation is zero. This occurs in the special case when the permanent charge Q is zero. In general, a realistic assumption is that Q is piecewise constant. To treat this general situation of Q, one can follow our approach to first work on each subinterval where Q is constant (see [10] for the zeroth order case). In doing so, one cannot assume the electro-neutrality condition since it is known to hold only at the two baths ($x = 0$ and $x = 1$). Another direction to be further explored is the case where three or more ion species are involved in the channel. We believe our analysis can be extended to those cases.

Acknowledgment. The authors thank the referees for their valuable comments and suggestions that helped improve the manuscript.

REFERENCES

[1] T. W. Allen, O. S. Andersen, and B. Roux, *Molecular dynamics—Potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels*, Biophys. Chem., 124 (2006), pp. 251–267.

[2] O. S. Andersen, R. E. Koeppe II, and B. Roux, *Gramicidin channels: Versatile tools*, in Biological Membrane Ion Channels: Dynamics, Structure, and Applications (Biological and Medical Physics, Biomedical Engineering), S. H. Chung, O. S. Andersen, and V. Krishnamurthy, eds., Springer, New York, 2006, pp. 33–80.

[3] V. Barcilon, *Ion flow through narrow membrane channels: Part* I, SIAM J. Appl. Math., 52 (1992), pp. 1391–1404.

[4] V. Barcilon, D.-P. Chen, and R. S. Eisenberg, *Ion flow through narrow membrane channels: Part* II, SIAM J. Appl. Math., 52 (1992), pp. 1405–1425.

[5] V. Barcilon, D.-P. Chen, R. S. Eisenberg, and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Perturbation and simulation study*, SIAM J. Appl. Math., 57 (1997), pp. 631–648.
[6] M. Burger, R. S. Eisenberg, and H. W. Engl, *Inverse problems related to ion channel selectivity*, SIAM J. Appl. Math., 67 (2007), pp. 960–989.
[7] R. Dutzler, T. Schirmer, M. Karplus, and S. Fischer, *Translocation mechanism of long sugar chains across the maltoporin membrane channel*, Structure, 10 (2002), pp. 1273–1284.
[8] W. Eckhaus, *Asymptotic Analysis of Singular Perturbations*, Stud. Math. Appl. 9, North–Holland, Amsterdam, New York, 1979.
[9] W. Eckhaus, *Fundamental concepts of matching*, SIAM Rev., 36 (1994), pp. 431–439.
[10] B. Eisenberg and W. Liu, *Poisson–Nernst–Planck systems for ion channels with permanent charges*, SIAM J. Math. Anal., 38 (2007), pp. 1932–1966.
[11] R. Elber, D. Chen, D. Rojewska, and R. S. Eisenberg, *Sodium in gramicidin: An example of a permion*, Biophys. J., 68 (1995), pp. 906–924.
[12] D. Gillespie, *A Singular Perturbation Analysis of the Poisson-Nernst-Planck System: Applications to Ionic Channels*, Ph.D. dissertation, Department of Molecular Biophysics and Physiology, Rush University at Chicago, Chicago, IL, 1999.
[13] D. Gillespie, W. Nonner, and R. S. Eisenberg, *Coupling Poisson–Nernst–Planck and density functional theory to calculate ion flux*, J. Phys. Condens. Matter, 14 (2002), pp. 12129–12145.
[14] D. Gillespie, W. Nonner, and R. S. Eisenberg, *Density functional theory of charged, hard-sphere fluids*, Phys. Rev. E, 68 (2003), pp. 1–10.
[15] O. P. Hamill, A. Marty, E. Neher, B. Sakmann, and F. J. Sigworth, *Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches*, Pflugers Arch., 391 (1981), pp. 85–100.
[16] S. B. Hladky and D. A. Haydon, *Discreteness of conductance change in bimolecular lipid membranes in the presence of certain antibiotics*, Nature, 225 (1970), pp. 451–453.
[17] U. Hollerbach and R. Eisenberg, *Concentration-dependent shielding of electrostatic potentials inside the gramicidin A channel*, Langmuir, 18 (2002), pp. 3262–3631.
[18] J. W. Jerome, *Consistency of semiconductor modeling: An existence/stability analysis for the stationary Van Roosbroeck system*, SIAM J. Appl. Math., 45 (1985), pp. 565–590.
[19] J. W. Jerome and T. Kerkhoven, *A finite element approximation theory for the drift diffusion semiconductor model*, SIAM J. Numer. Anal., 28 (1991), pp. 403–422.
[20] P. C. Jordan, *Trial by ordeal: Ionic free energies in gramicidin*, Biophys. J., 83 (2002), pp. 1235–1236.
[21] J. Keener and J. Sneyd, *Mathematical Physiology*, Interdisciplinary Appl. Math. 8, Springer-Verlag, New York, 1998.
[22] J. Kevorkian and J. D. Cole, *Perturbation Methods in Applied Mathematics*, Appl. Math. Sci. 34, Springer-Verlag, New York, Berlin, 1981.
[23] J. Kevorkian and J. D. Cole, *Multiple Scale and Singular Perturbation Methods*, Appl. Math. Sci. 114, Springer-Verlag, New York, 1996.
[24] L. Kullman, M. Winterhalter, and S. M. Bezrukov, *Transport of maltodextrins through maltoporin: A single-channel study*, Biophys. J., 82 (2002), pp. 803–812.
[25] M. G. Kurnikova, R. D. Coalson, P. Graf, and A. Nitzan, *A lattice relaxation algorithm for 3D Poisson–Nernst–Planck theory with application to ion transport through the gramicidin A channel*, Biophys. J., 76 (1999), pp. 642–656.
[26] P. A. Lagerstrom, *Matched Asymptotic Expansions*, Springer-Verlag, New York, 1988.
[27] W. Liu, *Geometric singular perturbation approach to steady-state Poisson–Nernst–Planck systems*, SIAM J. Appl. Math., 65 (2005), pp. 754–766.
[28] W. Liu, *Steady-state Poisson-Nernst-Planck systems for ion channels with multiple ion species*, J. Differential Equations, submitted.
[29] W. Liu and B. Wang, *Poisson-Nernst-Planck Systems for narrow tubular-like membrane channels*, Trans. Amer. Math. Soc., submitted.
[30] M. S. Mock, *An example of nonuniqueness of stationary solutions in device models*, COMPEL, 1 (1982), pp. 165–174.
[31] E. Neher and B. Sakmann, *Single channel currents recorded from the membrane of denervated muscle fibers*, Nature, 260 (1976), pp. 799–802.

[32] W. Nonner and R. S. Eisenberg, *Ion permeation and glutamate residues linked by Poisson-Nernst-Planck theory in L-type calcium channels*, Biophys. J., 75 (1998), pp. 1287–1305.
[33] R. E. O'Malley, Jr., *Singular Perturbation Methods for Ordinary Differential Equations*, Appl. Math. Sci. 89, Springer-Verlag, New York, 1991.
[34] J.-H. Park and J. W. Jerome, *Qualitative properties of steady-state Poisson–Nernst–Planck systems: Mathematical study*, SIAM J. Appl. Math., 57 (1997), pp. 609–630.
[35] A. Peskoff, R. S. Eisenberg, and J. D. Cole, *Matched asymptotic expansions of the Green's function for the electric potential in an infinite cylindrical cell*, SIAM J. Appl. Math., 30 (1976), pp. 222–239.
[36] B. Roux and M. Karplus, *Ion transport in a gramicidin-like channel: Dynamics and mobility*, J. Phys. Chem., 95 (1991), pp. 4856–4868.
[37] B. Roux and M. Karplus, *Ion transport in a model gramicidin channel: Structure and thermodynamics*, Biophys. J., 59 (1991), pp. 961–981.
[38] I. Rubinstein, *Multiple steady states in one-dimensional electrodiffusion with local electroneutrality*, SIAM J. Appl. Math., 47 (1987), pp. 1076–1093.
[39] I. Rubinstein, *Electro-Diffusion of Ions*, SIAM Stud. Appl. Math. 11, SIAM, Philadelphia, 1990.
[40] B. Sakmann and E. Neher, *Single Channel Recording*, Plenum, New York, 1995.
[41] R. H. Schirmer, T. A. Keller, Y. F. W., and J. P. Rosenbusch, *Structural basis for sugar translocation through maltoporin channels at* 3.1 *resolution*, Science, 267 (1995), pp. 512–514.
[42] Z. Schuss, B. Nadler, and R. S. Eisenberg, *Derivation of Poisson and Nernst-Planck equations in a bath and channel from a molecular model*, Phys. Rev. E, 64 (2001), pp. 1–14.
[43] A. Singer and J. Norbury, *A Poisson-Nernst-Planck model for biological ion channels—An asymptotic analysis in a* 3*-D narrow funnel*, European J. Appl. Math., 19 (2008), pp. 541–560.
[44] A. Singer, D. Gillespie, J. Norbury, and R. S. Eisenberg, *Singular Perturbation Analysis of the Steady State Poisson-Nernst-Planck System: Applications to Ion Channels*, preprint.
[45] H. Steinrück, *Asymptotic analysis of the current-voltage curve of a pnpn semiconductor device*, IMA J. Appl. Math., 43 (1989), pp. 243–259.
[46] H. Steinrück, *A bifurcation analysis of the one-dimensional steady-state semiconductor device equations*, SIAM J. Appl. Math., 49 (1989), pp. 1102–1121.
[47] Z. Siwy, M. R. Powell, E. Kalman, R. D. Asumian, and R. S. Eisenberg, *Negative incremental resistance induced by calcium in asymmetric nanopores*, Nano Lett., 6 (2006), pp. 473–477.
[48] Z. Siwy, M. R. Powell, A. Petrov, E. Kalman, C. Trautmann, and R. S. Eisenberg, *Calcium-induced voltage gating in single conical nanopores*, Nano Lett., 6 (2006), pp. 1729–1734.
[49] J. Tang, N. Saint, J. Rosenbusch, and R. Eisenberg, *Permeation through single channels of maltoporin*, Biophys. J., 72 (1997), p. A108.
[50] P. Van Gelder, F. Dumas, I. Bartoldus, N. Saint, A. Prilipov, M. Winterhalter, Y. Wang, A. Philippsen, J. P. Rosenbusch, and T. Schirmer, *Sugar transport through maltoporin of Escherichia coli: Role of the greasy slide*, J. Bacteriology, 184 (2002), pp. 2994–2999.
[51] B. A. Wallace, ed., *Gramicidin and Related Ion Channel Forming Peptides*, John Wiley, New York, 1999.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1527–1557
© 2008 Society for Industrial and Applied Mathematics

Separatrix Splitting in 3D Volume-Preserving Maps*

Héctor E. Lomelí† and Rafael Ramírez-Ros‡

Abstract. We construct a family of integrable volume-preserving maps in $\mathbb{R}^3$ with a two-dimensional heteroclinic connection of spherical shape between two fixed points of saddle-focus type. In other contexts, such structures are called Hill's spherical vortices or spheromaks. We study the splitting of the separatrix under volume-preserving perturbations using a discrete version of the Melnikov method. First, we establish several properties under general perturbations. For instance, we bound the topological complexity of the primary heteroclinic set in terms of the degree of some polynomial perturbations. We also give a sufficient condition for the splitting of the separatrix under some entire perturbations. A broad range of polynomial perturbations verify this sufficient condition. Finally, we describe the shape and bifurcations of the primary heteroclinic set for a specific perturbation.

Key words. separatrix splitting, volume-preserving maps, primary heteroclinic set, Melnikov method, bifurcations

AMS subject classifications. 34C37, 34C23, 37C29, 33E20

DOI. 10.1137/080713173

1. Introduction.

A fundamental question in dynamical systems is the effect that small perturbations of a dynamical system cause on its unperturbed invariant sets. The most studied unperturbed invariant sets are tori and stable/unstable invariant manifolds of hyperbolic sets. Usually, the unperturbed dynamical system is integrable and has separatrices; that is, its stable and unstable invariant manifolds overlap. After a generic perturbation, the perturbed stable and unstable invariant manifolds intersect transversely, which gives rise to the onset of chaos, through the creation of Smale horseshoes. This phenomenon is known as the problem of splitting of separatrices. A widely used technique for detecting such intersections is the Melnikov method.

Our goal is to apply the Melnikov method to the splitting of separatrices in the discrete volume-preserving framework. Similar questions have been considered before. However, we believe this is the first time that detailed *analytical* results about the structure of the primary heteroclinic set and its bifurcations are established for *specific* maps. This represents a step forward with respect to previous works [23, 24], in which once a formula for the Melnikov function in terms of an infinite series is written down, the approach becomes mainly numerical, because of the technical difficulties that obstruct the analytical one. Here, we have overcome

*Received by the editors January 11, 2008; accepted for publication (in revised form) by J. Meiss July 22, 2008; published electronically December 10, 2008.
http://www.siam.org/journals/siads/7-4/71317.html
†Department of Mathematics, Instituto Tecnológico Autónomo de México, Mexico, DF 01000 (lomeli@itam.mx). Current address: Department of Mathematics, The University of Texas, Austin, TX 78712. This paper was finished while this author held a Research Scholar position at the University of Texas at Austin and was supported in part by Asociación Mexicana de Cultura.
‡Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain (Rafael.Ramirez@upc.edu). This author was supported in part by MCyT-FEDER grant MTM2006-00478.

some of these difficulties using basic tools: complex variable theory, quasi-elliptic functions, homology, and several algebraic tricks. Nevertheless, we have not been able to find an explicit expression of the Melnikov function in terms of elementary functions for any specific perturbation. In contrast, such explicit expressions (in terms of elliptic functions) have been known for almost twenty years in the discrete area-preserving setting [18, 13].

This study is interesting because volume-preserving maps are the simplest and most natural higher-dimensional versions of the much-studied class of area-preserving maps. The infinite-dimensional group of volume-preserving diffeomorphisms on $\mathbb{R}^3$ is at the core of the ambitious program to reformulate hydrodynamics [3]. Volume-preserving maps arise in a number of applications such as the study of the motion of Lagrangian tracers in incompressible fluids or of the structure of magnetic field lines [19, 20, 33, 30]. Experimental methods have only recently been developed that allow the visualization of particle trajectories in spatial fluids [28, 32].

Given a system with a heteroclinic connection between two hyperbolic fixed points, the Melnikov function computes the rate at which the distance between the manifolds changes with a perturbation. After the introduction of the Melnikov method for periodic perturbations of one-degree-of-freedom Hamiltonian systems, many different versions appeared, most of them in continuous settings (flows). For instance, there are versions for three-dimensional incompressible flows in [29, 6, 5].

There exist also discrete versions of this method, in which the Melnikov function is no longer an integral, but an infinite sum whose domain is the unperturbed connection. The first steps towards a discrete Melnikov theory were performed for area-preserving maps [17, 18, 13, 21], and next for symplectic maps [14], for twist maps [22], for general n-dimension diffeomorphisms [9, 7, 25], and for spatial billiard maps [15]. Finally, volume-preserving maps have been considered in [23, 24]. These papers deal with codimension-one heteroclinic connections between fixed points of saddle-focus type and between hyperbolic invariant circles. The current paper is a natural continuation and uses some of their ideas.

We shall construct a family of integrable volume-preserving maps $f : \mathbb{R}^3 \to \mathbb{R}^3$ with a two-dimensional heteroclinic connection between two fixed points. This family is derived from another family of integrable planar maps introduced by McMillan [27]. The same construction can be found in [23], but we have decided to study a completely new family to minimize the overlap with previous works. Additionally, the new family has a purely rational character, whereas the previous one contains trigonometric terms. In other words, our model can be used to explore numerically phenomena that are connected to the splitting of separatrices. In particular, numerical computations using a multiple-precision arithmetic are feasible.

The two-dimensional separatrix has a spherical shape, and the fixed points are its "north pole" p_+ and its "south pole" p_-. Our integrable maps depend on two parameters: a *characteristic exponent* $h > 0$ and a *frequency* $\omega \in \mathbb{T}$. These names refer to the fact that

$$\operatorname{spec}[Df(p_\pm)] = \left\{ \mathrm{e}^{\pm 2h}, \mathrm{e}^{\mp h + \mathrm{i}\omega}, \mathrm{e}^{\mp h - \mathrm{i}\omega} \right\}.$$

Thus, the fixed points are of saddle-focus type, the characteristic exponent measures the hyperbolicity of the map, and the frequency quantifies the rotation speed of the trajectories on the separatrix. There also exists a one-dimensional straight heteroclinic connection between

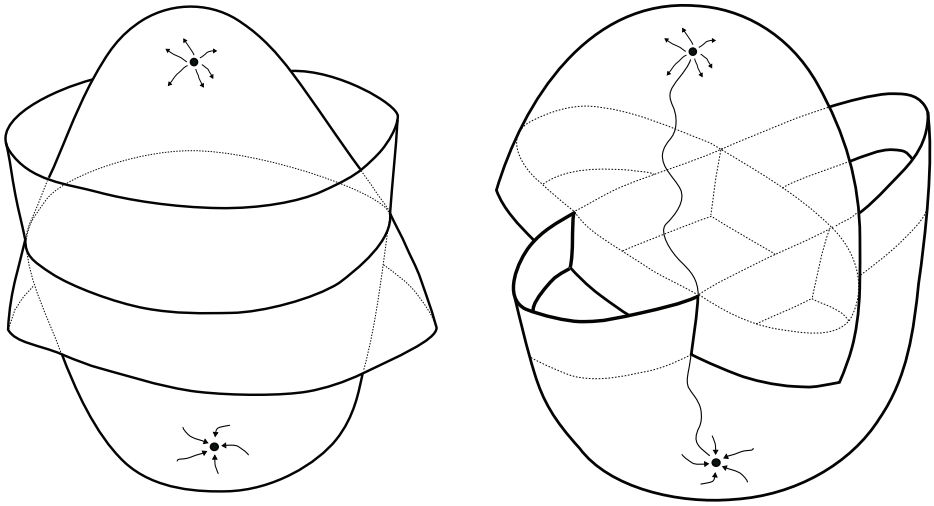


Figure 1. *In this figure we illustrate two possible intersections that appear as heteroclinic intersections of the stable and unstable manifolds that we will be considering. On the left, we have an* equatorial intersection. *On the right a* vertical intersection.

the fixed points. The same configuration appears in fluid dynamics under the name of *Hill's spherical vortex* or *bubble-type vortex breakdown* [33] and as a model for the magnetic field of stars. In a plasma physics context, the configuration is a confinement device that is called *spheromak* [26]. From a more theoretical point of view, we note that the integrable normal forms associated with families of volume-preserving flows with a *Hopf-zero singularity* have the same structure in the phase space [10].

In general, volume-preserving perturbations split the separatrix, but the perturbed stable and unstable manifolds still intersect along one-dimensional heteroclinic curves, which can be *vertical*, *equatorial*, or *bubble-type* ones. (This terminology is borrowed from [23].) Vertical curves are those heteroclinic intersections whose endpoints are both fixed points. Due to the rotational dynamics of the unperturbed map, these curves look like spirals connecting both poles when there is *swirl*; that is, when $\omega \neq 0$. On the contrary, equatorial and bubble-type curves are closed curves that do not approach the poles; in particular, they cannot appear in autonomous flows. The difference between equatorial and bubble-type curves is that the portions of the stable and unstable manifolds delimited by a bubble-type curve encircle a contractible region in $\mathbb{R}^3$; that is, a "bubble." See Remark 7 and Figure 1 for more details.

We shall describe the structure of the set of primary intersections under some perturbations. Roughly speaking, primary intersections are the sets of points where the stable and unstable manifolds "first" meet. In fact, primary intersections are the only intersections that can be followed by the perturbation. In the limit as $\epsilon \to 0$, they appear as zeroes of the Melnikov function. Therefore, nonprimary intersections are missed by standard Melnikov methods. See [23] for details.

First, we bound the topological complexity of the primary heteroclinic set in terms of the degree of volume-preserving polynomial perturbations of the form

$$f_\epsilon = (\mathrm{Id} + \epsilon\kappa) \circ f, \qquad \kappa(x, y, z) = (0, \alpha(x), \beta(x, y)).$$

In particular, it turns out that the primary heteroclinic set contains at most $2n$ vertical curves when $\alpha(x) \in \mathbb{R}_{n-1}[x]$ and $\beta(x,y) \in \mathbb{R}_n[x,y]$. Throughout this paper, we will use the notation $\mathbb{R}_m[x]$ for the set of degree-m polynomials in x with real coefficients and $\mathbb{R}_m[x,y]$ for the set of degree-m polynomials in x and y.

Next, we shall give a sufficient condition for the splitting of the separatrix under some entire perturbations. A broad range of polynomial perturbations verify this condition, for instance, those with $\kappa(x,y,z) = (0,0,\beta(x,y))$ for some even polynomial $\beta(x,y)$ of degree $4l+2$, provided that $\mathrm{e}^{4k\omega \mathrm{i}} \neq -1$ for $k = 1,\dots,2l+1$. In particular, nonresonant frequencies guarantee the breakdown of the unperturbed structure, which is in sharp contrast with some known principles in KAM (Kolmogorov–Arnold–Moser) theory.

Finally, we shall consider the perturbation with $\kappa(x,y,z) = (0,x,0)$. The primary heteroclinic set under this perturbation consists of four vertical curves for $\omega \neq \pm\pi/2$, whereas some heteroclinic bifurcations take place at $\omega = \pm\pi/2$. Unfortunately, we have found a complete proof of these facts only for $h \geq h_0 \approx 2.28$, but we conjecture, based on numerical experiments, that this picture holds for any $h > 0$. The previous upper bound on the number of vertical curves is optimal for this perturbation.

The proof of each analytical result is based on different tools. The bounds on the topological complexity follow from basic homology theory. The splitting result is obtained through the study of the complex singularities of the Melnikov function, an idea that goes back to Ziglin [35]. The part about bifurcations relies strongly on the fact that the Melnikov function can be expressed in terms of a quasi-elliptic function of order two. Additionally, each part has its own algebraic tricks.

We complete this introduction with a note on the organization of the paper. In section 2, we recall the Melnikov theory for volume-preserving maps. In section 3, we construct the family of integrable volume-preserving maps. In section 4, we derive an explicit expression for the Melnikov function associated with some volume-preserving perturbations. The next sections are devoted to bounding the topological complexity of the primary heteroclinic set and to establishing some sufficient conditions for the splitting. The study about the bifurcations of the primary heteroclinic set under the sample perturbation is contained in section 7. Some analytical details and numerical experiments are relegated to Appendices A and B, respectively.

2. The Melnikov theory for volume-preserving maps. In this section we shall briefly describe the Melnikov theory for volume-preserving maps developed in [23, 24, 25]. Let $f_\epsilon : \mathbb{R}^3 \to \mathbb{R}^3$ be a family of smooth volume-preserving maps such that the unperturbed map $f = f_0$ has two hyperbolic fixed points a and b whose stable and unstable invariant manifolds coincide, giving rise to a two-dimensional saddle connection $\Sigma = W^{\mathrm{u}}(a,f) \setminus \{a\} = W^{\mathrm{s}}(b,f) \setminus \{b\}$, where $W^{\mathrm{u}}(a,f)$ and $W^{\mathrm{s}}(b,f)$ denote the unstable invariant manifold of the point a and the stable invariant manifold of the point b, respectively. Both fixed points persist and remain hyperbolic for small ϵ. We want to study how the perturbed invariant manifolds $W^{\mathrm{u}}(a_\epsilon,f_\epsilon)$ and $W^{\mathrm{s}}(b_\epsilon,f_\epsilon)$ intersect.

Our goal is to describe the topology of the set of primary intersections $P_\epsilon \subset W^{\mathrm{u}}(a_\epsilon,f_\epsilon) \cap W^{\mathrm{s}}(b_\epsilon,f_\epsilon)$. We also try to elucidate when the separatrix Σ *splits* under the perturbation; that is, when there is no smooth family of saddle connections $\Sigma_\epsilon \subset W^{\mathrm{u}}(a_\epsilon,f_\epsilon) \cap W^{\mathrm{s}}(b_\epsilon,f_\epsilon)$ such that $\Sigma_0 = \Sigma$.

We recall some concepts that are going to be used below. A point ξ of a manifold Σ is a *regular point* of a smooth function $M : \Sigma \to \mathbb{R}$ when the differential form $\mathrm{d}M$ does not vanish at ξ, whereas $r \in \mathbb{R}$ is a *regular value* of M if every point in $M^{-1}(r)$ is a regular point. A zero of M is called *nondegenerate* when it is a regular point. If r is a regular value of M, then $M^{-1}(r)$ is a one-dimensional submanifold of Σ. On the contrary, $M^{-1}(r)$ can be much more complicated if 0 is a *singular value*, although its subset of regular points is also a one-dimensional submanifold of Σ. A diffeomorphism f is *symmetric* when there exists a diffeomorphism S such that $f \circ S = S \circ f$, and then S is called a *symmetry* of the map f. Analogously, f is *reversible* when there exists a diffeomorphism R such that $f \circ R = R \circ f^{-1}$, and then R is called a *reversor* of the map f and we denote by $\operatorname{Fix} R = \{\xi \in \mathbb{R}^3 : R(\xi) = \xi\}$ the set of its fixed points. These fixed points are called *symmetric* in the literature.

We collect in the following theorem the basic Melnikov-like results about this setup. See [23, 25].

Theorem 1. *Under the previous assumptions, there exists a smooth function $M : \Sigma \to \mathbb{R}$, called the* Melnikov function*, with the following properties:*

(i) *If ξ_0 is a nondegenerate zero of M, then $W^{\mathrm{u}}(a_\epsilon, f_\epsilon)$ and $W^{\mathrm{s}}(b_\epsilon, f_\epsilon)$ intersect transversely, for ϵ small enough, at a point $\xi_\epsilon = \xi_0 + \mathrm{O}(\epsilon) \in P_\epsilon$.*

(ii) *If 0 is a regular value of M, then the set of primary intersections P_ϵ is, for ϵ small enough, a one-dimensional submanifold of $\mathbb{R}^3$ such that $P_\epsilon = M^{-1}(0) + \mathrm{O}(\epsilon)$.*

(iii) *M is invariant by the unperturbed map: $M \circ f = M$.*

(iv) *If f_ϵ has a smooth family of*

1. *symmetries $S_\epsilon : \mathbb{R}^3 \to \mathbb{R}^3$ such that $S_0(\Sigma) = \Sigma$, then $M \circ S_0 = M$;*
2. *reversors $R_\epsilon : \mathbb{R}^3 \to \mathbb{R}^3$ such that $R_0(\Sigma) = \Sigma$, then $M \circ R_0 = -M$;*
3. *saddle connections $\Sigma_\epsilon \subset W^{\mathrm{u}}(a_\epsilon, f_\epsilon) \cap W^{\mathrm{s}}(b_\epsilon, f_\epsilon)$ with $\Sigma_0 = \Sigma$, then $M \equiv 0$.*

The Melnikov function is constructed in such a way that it measures the distance between the perturbed invariant manifolds $W^{\mathrm{u}}(a_\epsilon, f_\epsilon)$ and $W^{\mathrm{s}}(b_\epsilon, f_\epsilon)$ in first order. Because of this, the zero-level set $M^{-1}(0) \subset \Sigma$ is strongly related to the primary intersection set P_ϵ, and any change in its topology gives rise to some heteroclinic bifurcation. Further, a sufficient condition for the splitting of the separatrix is that the Melnikov function is not identically zero.

Symmetries, reversors, and symmetric heteroclinic points play an important role in the study of (primary) heteroclinic intersections; see [16]. For instance, the set of primary intersections is invariant by symmetries and reversors. Additionally, symmetric heteroclinic points persist under reversible perturbations. Concretely, if f_ϵ is R_ϵ-reversible and $\operatorname{Fix} R_0$ is a smooth curve that intersects transversely the saddle connection Σ at some point ξ_0, then there exists a unique point $\xi_\epsilon = \xi_0 + \mathrm{O}(\epsilon) \in P_\epsilon \cap \operatorname{Fix} R_\epsilon$.

Remark 1. If f_ϵ is not R_ϵ-reversible, but $f_\epsilon \circ R_\epsilon - R_\epsilon \circ f_\epsilon^{-1} = \mathrm{O}(\epsilon^2)$ and $R_0(\Sigma) = \Sigma$, then $M \circ R_0 = -M$. This has to do with the fact that the Melnikov function measures only first-order behaviors. We present an explicit example of this situation in Proposition 4.

In order to apply this theory, we must compute the Melnikov function. This is easier when the unperturbed map has a nondegenerate first integral $I : \mathbb{R}^3 \to \mathbb{R}$ and $f_\epsilon = (\mathrm{Id} + \epsilon\kappa) \circ f$ for some map $\kappa : \mathbb{R}^3 \to \mathbb{R}^3$. Under these assumptions, it is proved in [23, Lemma 8] that the Melnikov function is the absolutely convergent series

$$M = \sum_{k \in \mathbb{Z}} \langle \nabla I, \kappa \rangle \circ f^k. \tag{1}$$

This is the formula for Melnikov functions that we shall use in this paper.

We are interested only in perturbations that do not destroy the volume-preserving character of the unperturbed map f. This question has a simple answer: $f_\epsilon = (\mathrm{Id} + \epsilon\kappa) \circ f$ preserves volume if and only if the differential of the perturbation κ is nilpotent everywhere; see [23, Lemma 3]. This allows us to create simple examples of volume-preserving perturbations. For instance, we could take $\kappa(x, y, z) = (0, \alpha(x), \beta(x, y))$ or $\kappa(x, y, z) = (\gamma(y, z), \delta(z), 0)$ for any smooth functions $\alpha, \delta : \mathbb{R} \to \mathbb{R}$ and $\beta, \gamma : \mathbb{R}^2 \to \mathbb{R}$.

3. The maps. In this section we shall construct, following a methodology developed in [23], the perturbed volume-preserving maps f_ϵ that will be studied in the rest of the paper. As a starting point, we shall describe the unperturbed maps $f = f_0$ that form a family of integrable volume-preserving maps with a two-dimensional heteroclinic connection between a couple of hyperbolic fixed points. This integrable family is derived from another family of integrable planar standard-like maps introduced by McMillan [27].

Let $h > 0$ be the parameter of the family of planar standard-like maps. Then we consider the quantities $c = \cosh(h/2)$ and $s = \sinh(h/2)$, the rational transformation

$$z \mapsto \phi(z) = \frac{cz + s}{c + sz}, \tag{2}$$

and the area-preserving map

$$g(r, z) = \left(\phi\left(r + \phi^{-1}(z)\right) - z,\ r + \phi^{-1}(z)\right), \tag{3}$$

where $\phi^{-1}(z) = (cz - s)/(c - sz) = -\phi(-z)$ is the inverse transformation of (2). The phase portrait of this map is sketched in Figure 2. Its main dynamical properties are described in the next lemma.

Lemma 2. *The area-preserving map (3) has the following properties:*

(i) *The points $q_\pm = (0, \pm 1)$ are hyperbolic fixed points of g and*

$$\mathrm{spec}[Dg(q_\pm)] = \left\{\mathrm{e}^h, \mathrm{e}^{-h}\right\}.$$

(ii) *The function $J(r, z) = (c^2 - s^2z^2)r^2 + 2cs(z^2 - 1)r$ is a first integral, and the level $J^{-1}(0)$ contains two heteroclinic connections between the hyperbolic fixed points.*

(iii) *These heteroclinic connections are $\Gamma_0 = \left\{(r, z) \in \mathbb{R}^2 : r = 0,\ |z| < 1\right\}$ and*

$$\Gamma = \left\{(r, z) \in \mathbb{R}^2 : r = \phi(z) - \phi^{-1}(z) = \frac{2cs(1 - z^2)}{c^2 - s^2z^2},\ |z| < 1\right\}.$$

(iv) *The diffeomorphism $\gamma = (r, z) : \mathbb{R} \to \Gamma$, $z(t) = \tanh(t/2)$, $r(t) = z(t + h) - z(t - h)$ is a natural parametrization of the connection Γ; that is, $g(\gamma(t)) = \gamma(t + h)$.*

(v) *The map g is R-reversible and $R(\gamma(t)) = \gamma(-t)$, where $R(r, z) = (r, -z)$.*

Proof. It is a direct computation, so it is more enlightening to explain how these formulae are guessed. The canonical change of variables $(r, z) \mapsto (z, w = r + \phi^{-1}(z))$ transforms (3) into the planar standard-like map

$$\bar{g}(z, w) = (w, \psi(w) - z), \qquad \psi(w) = \phi(w) + \phi^{-1}(w) = \frac{2w}{c^2 - s^2w^2}$$

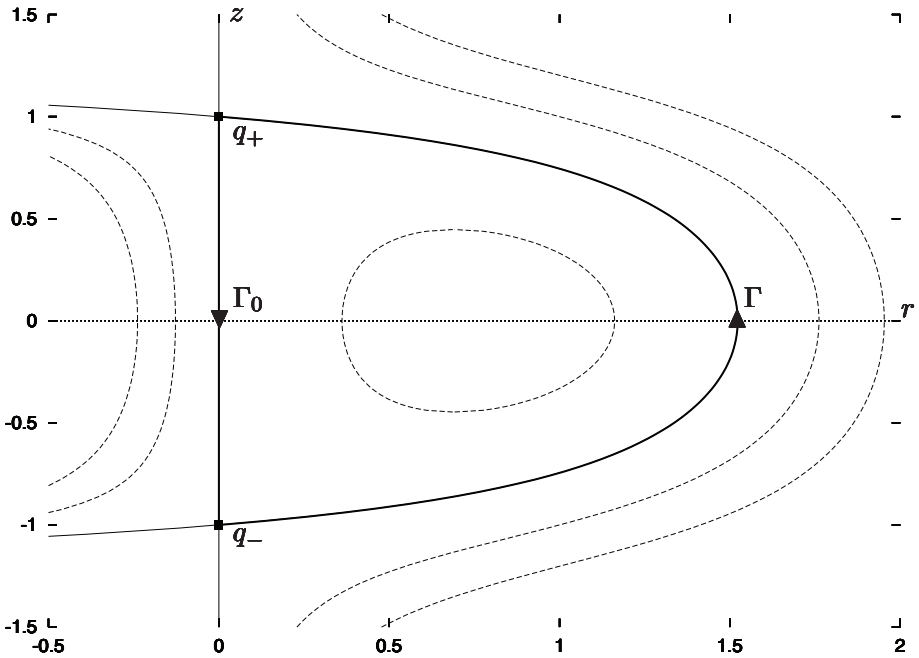


Figure 2. *The phase portrait of the area-preserving map (3) for $h = 2$. The solid squares denote the hyperbolic fixed points $q_\pm$. The thick lines denote the heteroclinic connections Γ_0 and Γ. The arrows denote the dynamics of the map on the connections.*

introduced by McMillan, which has similar properties [27, p. 232]. For instance,

$$\bar{J}(z,w) = s^2(z^2-1)(w^2-1) - (z-w)^2$$

is a known first integral of the McMillan map, and its zero-level set $\bar{J}^{-1}(0)$ contains two heteroclinic connections between the hyperbolic fixed points $\bar{q}_- = (-1,-1)$ and $\bar{q}_+ = (1,1)$. From the relation $\psi(z) = \phi(z) + \phi^{-1}(z)$, we get that

$$\bar{g}^k(z, \phi^{\pm 1}(z)) = \left(\phi^{\pm k}(z), \phi^{\pm(k+1)}(z)\right)$$

for all $k \in \mathbb{Z}$. Thus, using that $\phi : (-1,1) \to (-1,1)$ is a diffeomorphism such that $\lim_{k\to\pm\infty} \phi^k(z) = \pm 1$ for all $z \in (-1,1)$, we see that the heteroclinic connections are

$$\bar{\Gamma}_0 = \{w = \phi^{-1}(z)\}, \qquad \bar{\Gamma} = \{w = \phi(z)\}.$$

The change $(z,w) \mapsto (r = w - \phi^{-1}(z), z)$ transforms $\bar{\Gamma}_0$ into $\Gamma_0 = \{r = 0\}$ and $\bar{\Gamma}$ into $\Gamma = \{r = \phi(z) - \phi^{-1}(z)\}$. Finally, the natural parametrization follows from the relations $\phi(z(t)) = z(t+h)$ and $r(t) = \phi(z(t)) - \phi^{-1}(z(t)) = z(t+h) - z(t-h)$. ■

Next, we construct a volume-preserving map using the area-preserving map (3). The methodology consists, roughly speaking, of "rotating" the right half-plane $\{r > 0\}$ of Figure 2 around the vertical axis, using "canonical" cylindrical coordinates [23]. The map becomes fully three-dimensional if we introduce any nontrivial dynamics in the cylindrical angular variable

$\theta \in \mathbb{T} := \mathbb{R}/2\pi\mathbb{Z}$. For instance, a rigid rotation $\theta \mapsto \Theta = \theta + \omega$ suffices. See also Remark 2. The surface of revolution Σ obtained from the curve Γ is the two-dimensional heteroclinic connection we were looking for.

The construction would be a little obscure if we directly used the Cartesian coordinates (x, y, z). Hence, as an intermediate step, it is convenient to introduce the cylindrical angle $\theta \in \mathbb{T}$ and the cylindrical radius $\sqrt{2r} > 0$. That is, we will work with the canonical cylindrical coordinates (r, θ, z) defined by the relations

$$x = \sqrt{2r}\cos\theta, \qquad y = \sqrt{2r}\sin\theta, \qquad z = z. \tag{4}$$

The term canonical means that $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z = \mathrm{d}r \wedge \mathrm{d}\theta \wedge \mathrm{d}z$. Consider the map $(r, \theta, z) \mapsto (R, \Theta, Z)$, given by

$$\Theta = \theta + \omega, \qquad (R, Z) = g(r, z), \tag{5}$$

where g is the area-preserving map (3). This map preserves volume, since

$$\mathrm{d}X \wedge \mathrm{d}Y \wedge \mathrm{d}Z = -\mathrm{d}R \wedge \mathrm{d}Z \wedge \mathrm{d}\Theta = -\mathrm{d}r \wedge \mathrm{d}z \wedge \mathrm{d}\theta = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z.$$

Let

$$\rho(r, z) = \begin{cases} \sqrt{(\phi(r + \phi^{-1}(z)) - z)/r}, & r \neq 0, \\ \sqrt{\phi'(\phi^{-1}(z))}, & r = 0. \end{cases}$$

This function $\rho(r, z)$ is analytic at $r = 0$ for $|z| < c/s$. Using coordinates (4) in the map defined by (5), we get that, in Cartesian coordinates, the map that we want is

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = f \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos\omega & -\sin\omega & 0 \\ \sin\omega & \cos\omega & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \rho(r, z)x \\ \rho(r, z)y \\ r + \phi^{-1}(z) \end{pmatrix}, \tag{6}$$

where $r = (x^2 + y^2)/2$. We check, using formulation (6), that the map is well defined and analytic on $\{(0, 0, z) \in \mathbb{R}^3 : |z| < c/s\}$. This was not immediately clear from (4), since the change to cylindrical coordinates is singular at $r = 0$.

The map (6) is our unperturbed volume-preserving model. It depends on the *characteristic exponent* $h > 0$ and the *frequency* $\omega \in \mathbb{T}$. The characteristic exponent measures the hyperbolicity of the problem. In particular, numerical computations or analytical studies about separatrix splittings for small values of h will be hard, due to their exponential smallness. For instance, we have only been able to prove a conjecture presented in section 7 for $h \geq \log 16 - \log(\sqrt{113} - 9) \approx 2.282$.

The main dynamical properties of the integrable volume-preserving map (6) are described in the following lemma.

Lemma 3. *The volume-preserving map* (6) *has the following properties:*

(i) *The points* $p_\pm = (0, 0, \pm 1)$ *are hyperbolic fixed points of* f *such that*

$$\operatorname{spec}[Df(p_\pm)] = \left\{ \mathrm{e}^{\pm 2h}, \mathrm{e}^{\mp h + \mathrm{i}\omega}, \mathrm{e}^{\mp h - \mathrm{i}\omega} \right\}.$$

(ii) *The function* $I(x,y,z) = J((x^2+y^2)/2, z)$ *is a first integral of* f*, and the level* $I^{-1}(0)$ *contains two heteroclinic connections between the hyperbolic fixed points.*

(iii) *The heteroclinic connections are* $\Sigma_0 = \{(0,0,z) \in \mathbb{R}^3 : |z| < 1\}$ *and*

$$\Sigma = \left\{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 = \frac{4cs(1-z^2)}{c^2 - s^2 z^2},\ |z| < 1\right\}.$$

(iv) *The diffeomorphism* $\sigma : \mathbb{T} \times \mathbb{R} \to \Sigma$, $\sigma(\theta,t) = (x(\theta,t), y(\theta,t), z(t))$, *given by*

$$\left.\begin{array}{rcl} z(t) & = & \tanh(t/2), \\ r(t) & = & z(t+h) - z(t-h), \\ x(\theta,t) & = & \sqrt{2r(t)}\cos\theta, \\ y(\theta,t) & = & \sqrt{2r(t)}\sin\theta, \end{array}\right\} \tag{7}$$

is a natural parametrization of Σ*; that is,* $f(\sigma(\theta,t)) = \sigma(\theta+\omega, t+h)$.

(v) *The map* f *has the linear symmetry* $S(x,y,z) = (-x,-y,z)$ *and the involutive linear reversors* $R(x,y,z) = (x,-y,-z)$ *and* $T(x,y,z) = (-x,y,-z)$*. Additionally,* $S(\sigma(\theta,t)) = \sigma(\theta+\pi, t)$, $R(\sigma(\theta,t)) = \sigma(-\theta,-t)$, *and* $T(\sigma(\theta,t)) = \sigma(\pi-\theta,-t)$.

Proof. In the cylindrical coordinates (r,θ,z), the map f acts in the way described in (5). Therefore, these properties follow directly from the properties of the map g described in Lemma 2 and the fact that the involutions $\theta \mapsto -\theta$ and $\theta \mapsto \pi - \theta$ are reversors of the rigid rotations $\theta \mapsto \theta + \omega$. ■

Remark 2. We could have assumed that the frequency is not constant, but that it depends on the first integral: $\omega = \omega(I)$. In that case, since the expression of the Melnikov function only needs the values of the dynamics on the saddle connection Σ, only the value $\omega_0 = \omega(0)$ appears in the Melnikov computations.

The fixed sets of the reversors R and T are smooth curves. In fact, $\operatorname{Fix} R$ is the x-axis and $\operatorname{Fix} T$ is the y-axis. Additionally, each fixed set intersects the saddle connection transversely at a couple of opposite points, namely,

$$\Sigma \cap \operatorname{Fix} R = \{\xi^+, \xi^-\}, \qquad \Sigma \cap \operatorname{Fix} T = \{\zeta^+, \zeta^-\},$$

where $\xi^\pm = (\pm\eta, 0, 0)$, $\zeta^\pm = (0, \pm\eta, 0)$, and $\eta = 2\sqrt{s/c}$. Further, $\xi^+ = \sigma(0,0)$, $\xi^- = \sigma(\pi,0)$, $\zeta^+ = \sigma(\pi/2,0)$, and $\zeta^- = \sigma(3\pi/2,0)$. The question about when these symmetric heteroclinic points persist is answered in the following proposition.

Proposition 4. *Let* S, R, *and* T *be the symmetry and the reversors introduced in Lemma 3. Let* $\xi^\pm$ *and* $\zeta^\pm$ *be the symmetric heteroclinic points of the map (6) on the* x*-axis and* y*-axis, respectively. Let* P_ϵ *be the set of primary heteroclinic intersections of the perturbed map* $f_\epsilon = p_\epsilon \circ f$, *where* $p_\epsilon = \mathrm{Id} + \epsilon\kappa$ *and* $\kappa(x,y,z) = (0, \alpha(x), \beta(x,y))$.

(i) *If* $\alpha(x)$ *is odd and* $\beta(x,y)$ *is even, then* f_ϵ *is* S*-symmetric and* $S(P_\epsilon) = P_\epsilon$.

(ii) *If* $\beta(x,y)$ *is even in* y*, then* $f_\epsilon \circ R_\epsilon - R_\epsilon \circ f_\epsilon^{-1} = \mathrm{O}(\epsilon^2)$, *where* $R_\epsilon = p_\epsilon \circ R$*. If, in addition,* $\alpha(0) = 0$ *and* $\beta(0, \alpha(x)) = 0$, *then* f_ϵ *is* R_ϵ*-reversible,* $R_\epsilon(P_\epsilon) = P_\epsilon$, *and there exist points* $\xi_\epsilon^\pm = \xi^\pm + \mathrm{O}(\epsilon) \in P_\epsilon \cap \operatorname{Fix} R_\epsilon$.

(iii) *If* $\alpha(x)$ *is odd and* $\beta(x,y)$ *is even in* x*, then* $f_\epsilon \circ T_\epsilon - T_\epsilon \circ f_\epsilon^{-1} = \mathrm{O}(\epsilon^2)$, *where* $T_\epsilon = p_\epsilon \circ T$*. If, in addition,* $\alpha(0) = 0$ *and* $\beta(0, \alpha(x)) = 0$, *then* f_ϵ *is* T_ϵ*-reversible,* $T_\epsilon(P_\epsilon) = P_\epsilon$, *and there exist points* $\zeta_\epsilon^\pm = \zeta^\pm + \mathrm{O}(\epsilon) \in P_\epsilon \cap \operatorname{Fix} T_\epsilon$.

Proof. (i) If $\alpha(x)$ is odd and $\beta(x,y)$ is even, then $\kappa \circ S = S \circ \kappa$ and

$$p_\epsilon \circ S = (\mathrm{Id} + \epsilon\kappa) \circ S = S + \epsilon S \circ \kappa = S \circ (\mathrm{Id} + \epsilon\kappa) = S \circ p_\epsilon.$$

Therefore, $f_\epsilon \circ S = p_\epsilon \circ f \circ S = p_\epsilon \circ S \circ f = S \circ p_\epsilon \circ f = S \circ f_\epsilon$.

(ii) If $\beta(x,y)$ is even in y, then $\kappa \circ R = -R \circ \kappa$. Further, $p_\epsilon^{-1} = \mathrm{Id} - \epsilon\kappa + \mathrm{O}(\epsilon^2)$. Therefore, $R_\epsilon = (\mathrm{Id} + \epsilon\kappa) \circ R = R - \epsilon R \circ \kappa = R \circ (\mathrm{Id} - \epsilon\kappa) = R \circ p_\epsilon^{-1} + \mathrm{O}(\epsilon^2)$ and $f_\epsilon \circ R_\epsilon = p_\epsilon \circ f \circ R \circ p_\epsilon^{-1} + \mathrm{O}(\epsilon^2) = p_\epsilon \circ R \circ f^{-1} \circ p_\epsilon^{-1} + \mathrm{O}(\epsilon^2) = R_\epsilon \circ f_\epsilon^{-1} + \mathrm{O}(\epsilon^2)$.

If $\alpha(0) = 0$ and $\beta(0, \alpha(x)) = 0$, then $\kappa^2(x,y,z) = (0, \alpha(0), \beta(0, \alpha(x))) = 0$ and $p_\epsilon^{-1} = \mathrm{Id} - \epsilon\kappa$, so all the $\mathrm{O}(\epsilon^2)$ terms above vanish.

(iii) If $\alpha(x)$ is odd and $\beta(x,y)$ is even in x, then $\kappa \circ T = -T \circ \kappa$, so it suffices to replace R with T in the previous item. ∎

When R_ϵ and T_ϵ are true reversors, their fixed sets are

$$\begin{aligned} \operatorname{Fix} R_\epsilon &= \{(x,y,z) \in \mathbb{R}^3 : y = \epsilon\alpha(x)/2,\ z = \epsilon\beta(x, \epsilon\alpha(x)/2)/2\}, \\ \operatorname{Fix} T_\epsilon &= \{(x,y,z) \in \mathbb{R}^3 : x = 0,\ z = \epsilon\beta(0,y)/2\}, \end{aligned}$$

which are $\mathrm{O}(\epsilon)$-close to the x-axis and y-axis. (We have used that $\alpha(0) = 0$.)

4. The Melnikov function. Next, we want to derive an explicit expression for the Melnikov function associated with the volume-preserving perturbations

$$f_\epsilon = (\mathrm{Id} + \epsilon\kappa) \circ f, \qquad \kappa(x,y,z) = (0, \alpha(x), \beta(x,y)). \tag{8}$$

Other perturbations can also be studied. We do not aspire to be exhaustive.

If $\alpha(0) = \beta(0,0) = 0$, then $f_\epsilon(0,0,z) = f(0,0,z)$, and the one-dimensional heteroclinic connection Σ_0 is preserved under the perturbation (8). There is no similar persistence result for the two-dimensional heteroclinic connection Σ.

The first integral given in Lemma 3 is $I(x,y,z) = J(r,z)$, where $r = (x^2+y^2)/2$ and $J(r,z) = (c^2 - s^2z^2)r^2 + 2cs(z^2-1)r$. Additionally, $r = 2cs(1-z^2)/(c^2 - s^2z^2)$ on the saddle connection Σ. Finally, if the perturbation has the form (8), then the Melnikov function (1) can be written as

$$M : \Sigma \to \mathbb{R}, \qquad M(x,y,z) = \sum_{k \in \mathbb{Z}} m(x_k, y_k, z_k), \tag{9}$$

where $(x_k, y_k, z_k) = f^k(x,y,z)$ and

$$\begin{aligned} m(x,y,z) &= \langle \nabla I(x,y,z), \kappa(x,y,z) \rangle \\ &= \partial_y I(x,y,z)\alpha(x) + \partial_z I(x,y,z)\beta(x,y) \\ &= y\partial_r J(r,z)\alpha(x) + \partial_z J(r,z)\beta(x,y) \\ &= 2csy(1-z^2)\alpha(x) + 2szr(2c - sr)\beta(x,y). \end{aligned}$$

On the other hand, the natural parametrization $\sigma = (x,y,z) : \mathbb{T} \times \mathbb{R} \to \Sigma$ given in (7) provides a diffeomorphism between the saddle connection Σ and the cylinder $\mathbb{T} \times \mathbb{R}$, so that objects defined over Σ can be considered as depending on an angular variable $\theta \in \mathbb{T}$ and a

hyperbolic variable $t \in \mathbb{R}$. Henceforth, we will abuse the notation by not giving these objects new names. Thus, the Melnikov function (9) becomes

$$M : \mathbb{T} \times \mathbb{R} \to \mathbb{R}, \qquad M(\theta, t) = \sum_{k \in \mathbb{Z}} m(\theta + k\omega,\, t + kh), \tag{10}$$

where

$$\begin{aligned} m(\theta, t) &= \lambda(t) y(\theta, t) \alpha(x(\theta, t)) + \mu(t) \beta(x(\theta, t), y(\theta, t)) \\ &= \rho(t) \lambda(t) \alpha\big(\rho(t) \cos\theta\big) \sin\theta + \mu(t) \beta\big(\rho(t) \cos\theta, \rho(t) \sin\theta\big) \end{aligned} \tag{11}$$

and

$$\begin{aligned} r(t) &= z(t+h) - z(t-h) = \frac{2cs}{\cosh\left((t+h)/2\right) \cosh\left((t-h)/2\right)}, \\ \rho(t) &= \sqrt{2r(t)}, \\ \lambda(t) &= 2cs\left(1 - z(t)^2\right) = 4csz'(t) = \frac{2cs}{\cosh^2(t/2)}, \\ \mu(t) &= 2sz(t)r(t)\left(2c - sr(t)\right) = -4csr'(t). \end{aligned} \tag{12}$$

The rest of the paper deals with the computation and description of the zero-level set

$$Z = M^{-1}(0) = \{(\theta, t) \in \mathbb{T} \times \mathbb{R} : M(\theta, t) = 0\}$$

for several simple perturbations (8). In order to make that easier, we recall that the Melnikov function is invariant by the unperturbed map. In the current context, this implies that the Melnikov function M satisfies

$$M(\theta + \omega,\, t + h) = M(\theta, t) = M(\theta + 2\pi, t), \tag{13}$$

and therefore the zero-set Z is $(2\pi, 0)$ and (ω, h)-periodic. That is, if $(\theta^*, t^*) \in Z$, then $(\theta^* + \omega,\, t^* + h) \in Z$ and $(\theta^* + 2\pi, t^*) \in Z$.

A tilde will always denote the projection of a periodic object to the quotient torus,

$$\tilde{\tau}(\omega, h) = \tilde{\tau} := (\mathbb{T} \times \mathbb{R})/(\omega, h)\mathbb{Z} = \mathbb{R}^2/((2\pi, 0)\mathbb{Z} + (\omega, h)\mathbb{Z}), \tag{14}$$

which is diffeomorphic to the quotient of the saddle connection by the unperturbed map. The study of the projected set $\tilde{Z} = \tilde{M}^{-1}(0) \subset \tilde{\tau}$ is easier, because $\tilde{\tau}$ is compact.

The torus is represented in Figure 3 as the rectangle $[0, 2\pi] \times [0, h]$ with the appropriate identifications. We have not chosen the parallelogram shown in thin lines in that figure as the representation of the torus because its shape changes in ω, hindering posterior comparisons and the study of bifurcations that take place in ω.

Remark 3. Sometimes we will restrict ourselves to the case $\omega = 0$, which is called no-swirl in fluid dynamics. This is the simplest one, because then the quotient torus is a product: $\tilde{\tau}(0, h) = (\mathbb{R}/2\pi\mathbb{Z}) \times (\mathbb{R}/h\mathbb{Z})$, and the variable t is defined modulo h. Although the unperturbed map has just a two-dimensional dynamics for $\omega = 0$, it is still an interesting case—the perturbation will create a real three-dimensional dynamics.

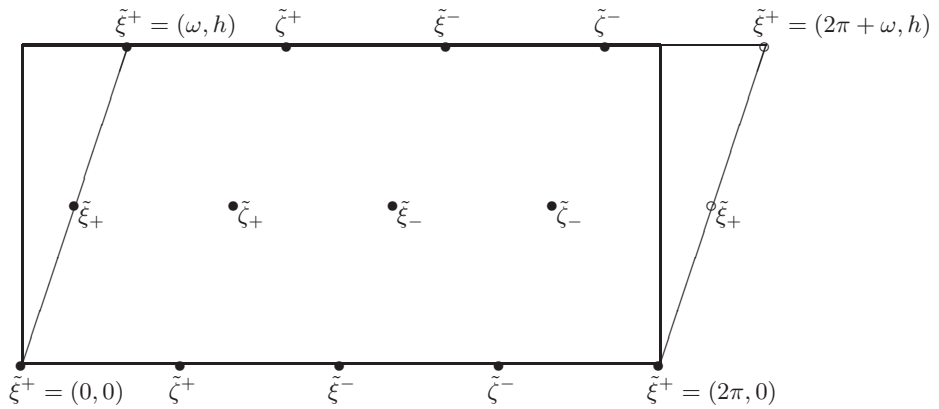


Figure 3. *A rectangular representation of the torus $\tilde{\tau}$ and the symmetric points described in Lemma 5 for $\omega = \pi/3$. Opposite sides of the rectangle are identified, although the identification of the horizontal ones is shifted by an amount equal to ω.*

Let us check that $\tilde{Z}$ contains at least eight *symmetric points* and has some useful symmetries and more periodicities when the perturbation preserves the symmetry and reversors of the unperturbed map. The symmetric points are shown in Figure 3.

Lemma 5. *Let $\tilde{Z}$ be the projection onto the torus (14) of the zero-level set of the Melnikov function (10).*

(i) *If $\alpha(x)$ is odd and $\beta(x, y)$ is even, then $\tilde{Z}$ is $(\pi, 0)$-periodic: $\tilde{Z} = \tilde{Z} + (\pi, 0)$.*
(ii) *If $\alpha(x)$ is even and $\beta(x, y)$ is odd, then $\tilde{Z}$ is $(\pi, 0)$-periodic: $\tilde{Z} = \tilde{Z} + (\pi, 0)$.*
(iii) *If $\beta(x, y)$ is even in y, then $\tilde{Z}$ contains (and is symmetric with regard to) the points $\tilde{\xi}^+ = (0, 0)$, $\tilde{\xi}^- = (\pi, 0)$, $\tilde{\xi}_+ = (\omega/2, h/2)$, and $\tilde{\xi}_- = (\pi + \omega/2, h/2)$.*
(iv) *If $\alpha(x)$ is odd and $\beta(x, y)$ is even in x, then $\tilde{Z}$ contains (and is symmetric with regard to) the points $\tilde{\zeta}^+ = (\pi/2, 0)$, $\tilde{\zeta}^- = (3\pi/2, 0)$, $\tilde{\zeta}_+ = (\pi/2 + \omega/2, h/2)$, and $\tilde{\zeta}_- = (3\pi/2 + \omega/2, h/2)$.*

Proof. We could write a geometric proof based on the geometric properties established in Proposition 4, but instead, we give a shorter analytic proof.

We consider the Melnikov function M as a function defined on the plane $\mathbb{R}^2$ with periods $(2\pi, 0)$ and (ω, h). Assume that M is odd with regard to a point $(\theta_0, t_0) \in \mathbb{R}^2$. Then $M^{-1}(0)$ contains (and is symmetric with regard to) the point (θ_0, t_0). But $M^{-1}(0)$ also contains (and is symmetric with regard to) the points $(\theta_0 + p, t_0 + q)$ for any semiperiod (p, q) of M, because

$$M(\theta_0 + p, t_0 + q) = -M(\theta_0 - p, t_0 - q) = -M(\theta_0 + p, t_0 + q).$$

The functions $r(t)$, $\rho(t)$, and $\lambda(t)$ given in (12) are even, whereas $\mu(t)$ is odd. Let m be the function defined in (11). Hence, the lemma is a consequence of the following observations:

(i) if $\alpha(x)$ is odd and $\beta(x, y)$ is even, then $m(\theta, t)$ is π-periodic in θ;
(ii) if $\alpha(x)$ is even and $\beta(x, y)$ is odd, then $m(\theta, t)$ is π-antiperiodic in θ;
(iii) if $beta(x, y)$ is even in y, then $m(\theta, t)$ is odd with regard to $(0, 0)$; and
(iv) if $\alpha(x)$ is odd and $\beta(x, y)$ is even in x, $m(\theta, t)$ is odd with regard to $(\pi/2, 0)$.

In this way, we conclude that the Melnikov function (10) verifies the conclusions of the lemma. ∎

Remark 4. One can obtain more symmetries or periodicities under more restrictive hypotheses. For instance, using basic trigonometric properties, one checks that Z is $(\pi/2, 0)$-periodic when $\kappa(x, y, z) = (0, x, 0)$ or when $\kappa(x, y, z) = (0, 0, \beta(x, y))$ for some $\beta(x, y)$ such that $\beta(x, y) = \beta(-y, x)$ or $\beta(x, y) = -\beta(-y, x)$.

Remark 5. It turns out that $\tilde{Z} \neq \emptyset$, even if the volume-preserving perturbation (8) has no symmetries. This has to do with the existence of an area form $\tilde{\eta}$ over the torus $\tilde{\tau}$ such that the integral of the two-form $\tilde{M}\tilde{\eta}$ vanishes. Therefore, the Melnikov function has to be zero at some points. This idea was used in [24]. We skip the details.

5. Bounds on the complexity of the primary heteroclinic set. First, we shall establish an upper bound on the cardinality of the horizontal sections of the zero-level set Z under polynomial perturbations of the form (8). These horizontal sections are defined as

$$Z_{t_0} = \{\theta \in \mathbb{T} : M(\theta, t_0) = 0\} = \{\theta \in \mathbb{T} : (\theta, t_0) \in Z\}, \qquad t_0 \in \mathbb{R}.$$

Proposition 6. *If $\alpha(x) \in \mathbb{R}_{n-1}[x]$ and $\beta(x, y) \in \mathbb{R}_n[x, y]$ for some integer $n \geq 1$, then either $Z_{t_0} = \mathbb{T}$ or $\#Z_{t_0} \leq 2n$.*

Proof. The function $m(\theta, t)$ given in (11) has the following simple forms under monomial perturbations. If $\alpha(x) = x^{i-1}$ and $\beta(x, y) = 0$, then $m(\theta, t) = (2r(t))^{i/2}\lambda(t)\cos^{i-1}\theta\sin\theta$, whereas if $\alpha(x) = 0$ and $\beta(x, y) = x^i y^j$, then $m(\theta, t) = (2r(t))^{(i+j)/2}\mu(t)\cos^i\theta\sin^j\theta$. Therefore, the Fourier expansion of $m(\theta, t)$ when $\alpha(x) \in \mathbb{R}_{n-1}[x]$ and $\beta(x, y) \in \mathbb{R}_n[x, y]$ has only the central $2n + 1$ harmonics. That is, $m(\theta, t) = \sum_{|j|\leq n} m_j(t)\mathrm{e}^{\mathrm{i}j\theta}$ for some coefficients $m_j(t)$. Thus the Fourier expansion of the Melnikov function (10) has the same form, since

$$\begin{aligned} M(\theta, t) &= \sum_{k\in\mathbb{Z}} m(\theta + k\omega,\, t + kh) \\ &= \sum_{|j|\leq n} \sum_{k\in\mathbb{Z}} m_j(t + kh)\mathrm{e}^{\mathrm{i}j(\theta+k\omega)} \\ &= \sum_{|j|\leq n} M_j(t)\mathrm{e}^{\mathrm{i}j\theta}, \end{aligned}$$

where $M_j(t) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{\mathrm{i}jk\omega} m_j(t + kh)$. To end the proof, it suffices to note that any nonzero trigonometric polynomial like $M_{t_0}(\theta) := M(\theta, t_0) = \sum_{|j|\leq n} M_j(t_0)\mathrm{e}^{\mathrm{i}j\theta}$ has at most $2n$ different roots in $\mathbb{T}$. ∎

If $\omega = 0$, there exists a similar bound for the cardinal of the vertical sections. This new bound is obtained by using some elementary facts of the theory of elliptic functions. We recall that a function is *elliptic* when it is meromorphic in the whole complex plane and has two complex periods that are independent over the reals. The *order* of a nonconstant elliptic function is the number of its poles (or zeroes), counted with multiplicity, that lie in a cell. A *cell* of an elliptic function with periods p_1 and p_2 is a parallelogram with vertexes s, $s + p_1$, $s + p_1 + p_2$, and $s + p_2$ such that its sides do not contain either zeroes or poles. For a general background on elliptic functions, we refer to [34].

We realized in Remark 3 that the Melnikov function $M(\theta, t)$ is h-periodic in the vertical coordinate t when the frequency ω is zero. In that case, the vertical sections of the projected

zero-level set $\tilde{Z} = \tilde{M}^{-1}(0) \subset \tilde{\tau}$ defined as

$$\tilde{Z}^{\theta_0} = \{t \in \mathbb{R}/h\mathbb{Z} : \tilde{M}(\theta_0, t) = 0\} = \{t \in \mathbb{R}/h\mathbb{Z} : (\theta_0, t) \in \tilde{Z}\}$$

are subsets of the quotient space $\mathbb{R}/h\mathbb{Z}$.

Proposition 7. *Assume that the perturbation (8) is polynomial and $\omega = 0$. Let $\tilde{Z}^{\theta_0}$ be any vertical section which does not cover the whole set $\mathbb{R}/h\mathbb{Z}$. Let $l, n \in \mathbb{N}$.*

(i) *If $\alpha(x) \in \mathbb{R}_{2l-1}[x]$ is odd and $\beta(x, y) = 0$, then $\#\tilde{Z}^{\theta_0} \leq l$.*

(ii) *If $\alpha(x) = 0$ and $\beta(x, y) \in \mathbb{R}_{2n}[x, y]$ is even, then $\#\tilde{Z}^{\theta_0} \leq n + 1$.*

(iii) *If $\alpha(x) \in \mathbb{R}_{2l-1}[x]$ is odd and $\beta(x, y) \in \mathbb{R}_{2n}[x, y]$ is even, then $\#\tilde{Z}^{\theta_0} \leq \max(l, n + 2)$.*

Proof. These three items a based on the following formulae. If $\alpha(x) = x^{2j-1}$ and $\beta(x, y) = 0$, then $m(\theta, t) = (2r(t))^j \lambda(t) \cos^{2j-1}\theta \sin\theta$. If $\alpha(x) = 0$ and $\beta(x, y) = x^i y^{2j-i}$, then $m(\theta, t) = (2r(t))^j \mu(t) \cos^i\theta \sin^{2j-i}\theta$. Hence, if the polynomial $\alpha(x) \in \mathbb{R}_{2l-1}[x]$ is odd and the polynomial $\beta(x, y) \in \mathbb{R}_{2n}[x, y]$ is even, the function $m(\theta, t)$ has the form

$$m(\theta, t) = \lambda(t) \sum_{j=1}^{l} a_j(\theta)(r(t))^j + \mu(t) \sum_{j=1}^{n} b_j(\theta)(r(t))^j$$

for some trigonometric polynomials $a_j(\theta)$ and $b_j(\theta)$. Since $\omega = 0$, we get that

$$\tilde{M}^{\theta_0}(t) := \tilde{M}(\theta_0, t) = \sum_{j=1}^{l} a_j(\theta_0) A_j(t) + \sum_{j=1}^{n} b_j(\theta_0) B_j(t), \tag{15}$$

where $A_j(t) = \sum_{k\in\mathbb{Z}} \lambda(t + kh)(r(t + kh))^j$ and $B_j(t) = \sum_{k\in\mathbb{Z}} \mu(t + kh)(r(t + kh))^j$.

The functions $z(t)$, $r(t)$, $\lambda(t)$, and $\mu(t)$ are 2πi-periodic and meromorphic in $\mathbb{C}$. On the one hand, the poles of $z(t)$ are the points in the set $\pi\mathrm{i} + 2\pi\mathrm{i}\mathbb{Z}$, all of them simple, and so $\lambda(t) = 4csz'(t)$ has the same poles, but they are double ones. On the other hand, the poles of $r(t)$ are the points in the sets $\pm h + \pi\mathrm{i} + 2\pi\mathrm{i}\mathbb{Z}$, all of them simple, and so $\mu(t) = -4csr'(t)$ has the same poles, but they are double ones.

Thus, $A_j(t)$, $B_j(t)$, and $\tilde{M}^{\theta_0}(t)$ are elliptic functions with periods h and 2πi. Their poles are the points in the set $\pi\mathrm{i} + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}$. Their orders are at most $\max(j, 2)$, $j + 2$, and $\max(l, n + 2)$, respectively. To end the common part of the proof, we note that $\tilde{M}^{\theta_0}(t)$ is nonconstant because $\tilde{Z}^{\theta_0} \neq \mathbb{R}/h\mathbb{Z}$.

(i) If $\beta(x, y) = 0$, the elliptic function (15) becomes $\tilde{M}^{\theta_0}(t) = \sum_{j=1}^{l} a_j(\theta_0) A_j(t)$, and its order is at most $\max(l, 2)$, so that it has at most $\max(l, 2)$ roots in a cell and $\#\tilde{Z}^{\theta_0} \leq \max(l, 2)$. We can substitute this last bound by $\#\tilde{Z}^{\theta_0} \leq l$ because if $\alpha(x) = x$ and $\omega = 0$, then either $\tilde{Z}^{\theta_0} = \mathbb{R}/h\mathbb{Z}$ or $\tilde{Z}^{\theta_0} = \emptyset$; see item (iv) of Theorem 17.

(ii) If $\alpha(x) = 0$, the elliptic function (15) becomes $\tilde{M}^{\theta_0}(t) = \sum_{j=1}^{n} b_j(\theta_0) B_j(t)$ and is odd, because $\mu(t)$ is odd and $r(t)$ is even. Its order is at most $n + 2$, but the rough bound $\#\tilde{Z}^{\theta_0} \leq n + 2$ can be improved using the symmetry. We get that $\tilde{M}^{\theta_0}(h/2 + \pi\mathrm{i}) = 0$, because

$$\tilde{M}^{\theta_0}(h/2 + \pi\mathrm{i}) = \tilde{M}^{\theta_0}(-h/2 + \pi\mathrm{i}) = \tilde{M}^{\theta_0}(-h/2 - \pi\mathrm{i}) = -\tilde{M}^{\theta_0}(h/2 + \pi\mathrm{i}).$$

This means that $\tilde{M}^{\theta_0}(t)$ has at most $n + 1$ real roots modulo h.

(iii) In this case, the bound $\#\tilde{Z}^{\theta_0} \leq \max(l, n+2)$ cannot be improved. ■

Remark 6. Proposition 6 also holds for the integrable trigonometric family of volume-preserving maps introduced in [23]. The proof does not requires any change. On the contrary, Proposition 7 cannot be directly translated into the trigonometric setting, because the complex singularities of the natural parametrization of the separatrix in that setting are more complicated.

Recall that in the introduction we had a brief discussion of the type of primary heteroclinic intersections that appear. In our case, by Theorem 1, it is enough to remember that primary intersections are the intersections that arise as the continuation of the nondegenerate zeroes of the Melnikov function.

Any vertical curve intersects the horizontal line $\{t = t_0\}$ in at least one point, so the number of vertical curves cannot be larger than the cardinal of the horizontal sections of the zero-level set. Therefore, as a by-product of Proposition 6, we get that there are at most $2n$ vertical curves when $\alpha(x) \in \mathbb{R}_{n-1}[x]$ and $\beta(x, y) \in \mathbb{R}_n[x, y]$. In fact, Propositions 6 and 7 have stronger consequences on the homology/homotopy classes of the heteroclinic intersections of the invariant manifolds.

We recall that $\tilde{Z}$ is the projection of the zero-level set $Z = M^{-1}(0)$ onto the torus $\tilde{\tau}$ defined in (14). Assume that 0 is a regular value of the Melnikov function. Then $\tilde{Z}$ is a submanifold of the torus, and its connected components are closed smooth curves. Therefore, once we have fixed an induced orientation on the torus, we can assign to each connected component $\tilde{\gamma}$ of $\tilde{Z}$ its homology class $[\tilde{\gamma}] \in H_1(\mathbb{T}^2) = \mathbb{Z}^2$.

In the case of the torus $\tilde{\tau}$, we will identify horizontal lines with the class $(1, 0)$ and vertical lines, generated by the vector (ω, h), with the class $(0, 1)$. Thus, $[\tilde{\gamma}] = (p, q) \in \mathbb{Z}^2$ means that $\tilde{\gamma}$ is a closed curve that wraps around the torus $|p|$ times in the horizontal direction and $|q|$ times in the vertical one. For instance, the set $\tilde{Z}$ has four connected components with homology class $(0, 1)$ or $(0, -1)$ in the subfigures 4(a)–4(d), whereas it has just two connected components with homology class $(1, -2)$ or $(-1, 2)$ in the subfigures 4(f)–4(i). Subfigure 4(e) is excluded because then 0 is a singular value of the Melnikov function.

Remark 7. With regard to the three types of heteroclinic curves mentioned in the introduction, we note that a connected component $\tilde{\gamma}$ such that $[\tilde{\gamma}] = (p, q)$ gives rise for ϵ small enough to vertical (resp., equatorial) (resp., bubble-type) curves when $q \neq 0$ (resp., $q = 0$ but $p \neq 0$) (resp., $p = q = 0$).

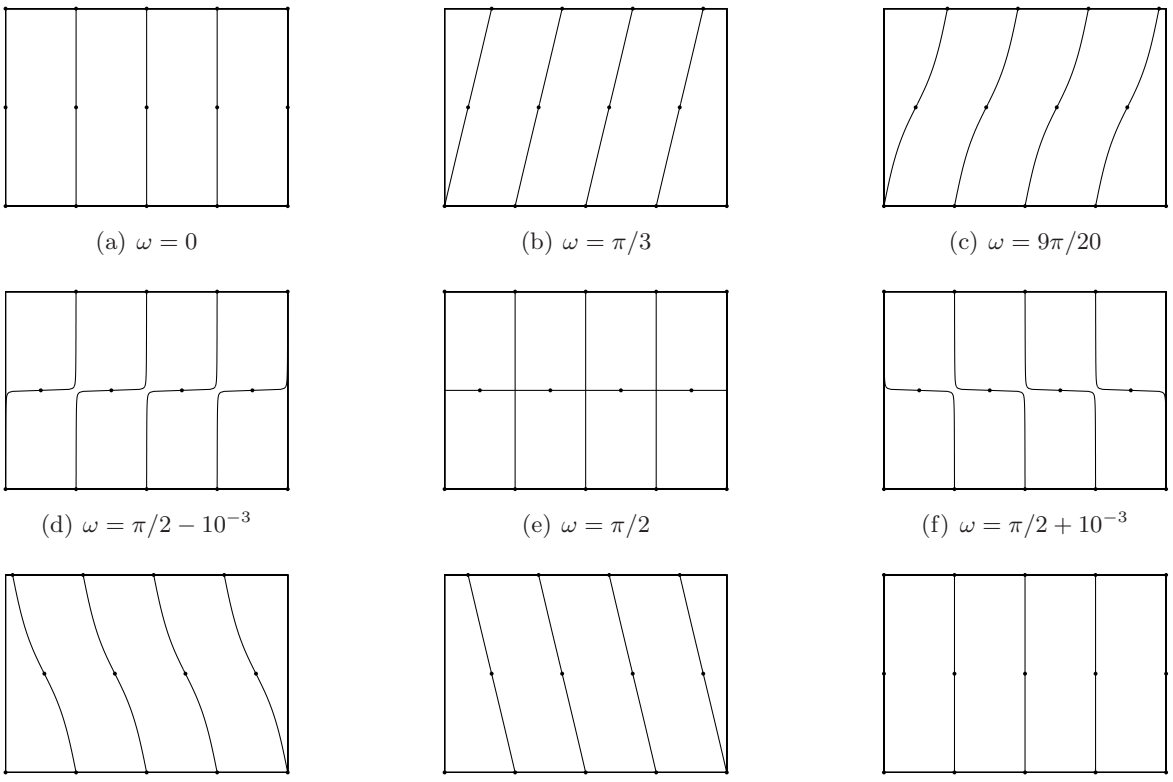
Remark 8. The first homology group and the first homotopy group (that is, the fundamental group) of a torus coincide: $H_1(\mathbb{T}^2) = \mathbb{Z}^2 = \pi_1(\mathbb{T}^2)$. Hence, we could use homotopy instead of singular homology.

We need the following result from Morse theory.

Lemma 8. *If $a \in \mathbb{R}$ is a regular value of a smooth function $f : X \to \mathbb{R}$ defined over a compact manifold X, the homology class of the level set $L_a = f^{-1}(a)$ is zero.*

Proof. It suffices to prove this for Morse functions, because Morse functions are dense and the homology class of a closed curve does not change under small perturbations.

Let a and b be two regular values of f such that $a < b$. Then L_a and L_b are the borders of the smooth manifold $f^{-1}([a, b])$, and so they have the same homology class. Let c be the maximum value of f. Since f is Morse, there exists a unique point $x \in X$ such that $f(x) = c$. This point is a nondegenerate maximum, and so if $\delta > 0$ is small enough, $c - \delta$ is a regular



(a) $\omega = 0$ (b) $\omega = \pi/3$ (c) $\omega = 9\pi/20$

(d) $\omega = \pi/2 - 10^{-3}$ (e) $\omega = \pi/2$ (f) $\omega = \pi/2 + 10^{-3}$

(g) $\omega = 11\pi/20$ (h) $\omega = 2\pi/3$ (i) $\omega = \pi$

Figure 4. *The only bifurcation of $\tilde{Z}$ in the range $0 \leq \omega \leq \pi$ under the perturbation $\kappa(x, y, z) = (0, x, 0)$ takes place at the singular frequency $\omega = \pi/2$. These pictures show this bifurcation for $h = 1$. The symmetric points move as the frequency ω varies.*

value and $L_{c-\delta}$ is just a small closed curve around x. Hence, $L_{c-\delta}$ is contractible, and its homology class is equal to zero. ■

The homology classes of the connected components of the projected zero-level set $\tilde{Z}$ are bounded in the following theorem. These bounds restrict the topological complexity of the primary heteroclinic set $P_\epsilon = Z + \mathrm{O}(\epsilon)$ for small values of ϵ.

Theorem 9. *Assume that 0 is a regular value of the Melnikov function associated with the perturbation (8). Let $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_r$ be connected components of the projected zero-level set $\tilde{Z} = \tilde{M}^{-1}(0)$. Let $[\tilde{\gamma}_1] = (p_1, q_1)$, $\ldots$, $[\tilde{\gamma}_r] = (p_r, q_r)$ be their homology classes.*

(i) *The homology of $\tilde{Z}$ is zero: $\sum_j [\tilde{\gamma}_j] = \sum_j (p_j, q_j) = (0, 0)$.*

(ii) *If $\alpha(x) \in \mathbb{R}_{n-1}[x]$ and $\beta(x, y) \in \mathbb{R}_n[x, y]$, then $2\,|q_j| \leq \sum_j |q_j| \leq 2n$.*

(iii) *Assume that $\omega = 0$. Let $l, n \in \mathbb{N}$. Then we have the following:*

1. *If $\alpha(x) \in \mathbb{R}_{2l-1}[x]$ is odd and $\beta(x, y) = 0$, then $2\,|p_j| \leq \sum_j |p_j| \leq l$.*
2. *If $\alpha(x) = 0$ and $\beta(x, y) \in \mathbb{R}_{2n}[x, y]$ is even, then $2\,|p_j| \leq \sum_j |p_j| \leq n + 1$.*
3. *If $\alpha(x) \in \mathbb{R}_{2l-1}[x]$ is odd and $\beta(x, y) \in \mathbb{R}_{2n}[x, y]$ is even, then $2\,|p_j| \leq \sum_j |p_j| \leq \max(l, n + 2)$.*

Proof. (i) It suffices to apply Lemma 8 to the projected function $\tilde{M} : \mathbb{T}^2 \to \mathbb{R}$.

(ii) We are under the hypotheses of Proposition 6, and 0 is a regular value of the Melnikov function, so there exists some $t_* \in \mathbb{R}$ such that $\#Z_{t_*} \leq 2n$. On the contrary, $Z_{t_0} = \mathbb{T}$ for all $t_0 \in \mathbb{R}$, and the Melnikov function should be identically zero.

Using that $\tilde{Z} = \tilde{\gamma}_1 \amalg \cdots \amalg \tilde{\gamma}_r$ and that each curve $\tilde{\gamma}_j$ wraps $|q_j|$ times in the vertical direction, we get that $\sum_j |q_j| \leq \sum_j \#\left(\tilde{\gamma}_j \cap (\mathbb{T} \times \{t_*\})\right) = \#Z_{t_*} \leq 2n$. Next, we obtain the bound $2\,|q_j| = |q_j| + \left|\sum_{i \neq j} q_i\right| \leq \sum_i |q_i| \leq 2n$ from the identity $\sum_j q_j = 0$.

(iii) This follows in a similar way, but from Proposition 7. ∎

6. Splitting of separatrices. In this section we shall present two theorems about the splitting of the separatrix. In the first one, we shall establish a sufficient condition for the splitting of the separatrix under some entire perturbations, whereas in the second one we find a broad class of polynomial perturbations that split the separatrix. The sufficient condition is obtained through the study of the complex singularities of the Melnikov function. To be more precise, if the Melnikov function can be analytically extended for complex values of its variables and this extension has some nonremovable singularity, then the original Melnikov function cannot be identically zero and the separatrix splits.

For simplicity, we have restricted our study to perturbations of the form

$$f_\epsilon = (\mathrm{Id} + \epsilon\kappa) \circ f, \qquad \kappa(x, y, z) = (0, 0, \beta(x, y)) \tag{16}$$

for some nonzero even entire function $\beta(x, y)$. The study is a bit more cumbersome when the entire perturbation has the more general form (8) with $\alpha(x)$ odd and $\beta(x, y)$ even. If $\alpha(x)$ is not odd or $\beta(x, y)$ is not even, our current technique does not work, because ramified singularities are harder to deal with than isolated ones.

Theorem 10. *Let $B_\theta : \mathbb{C} \to \mathbb{C}$ be the entire function*

$$B_\theta(r) = \int_0^r \beta(\sqrt{2s}\cos\theta, \sqrt{2s}\sin\theta)\mathrm{d}s. \tag{17}$$

Let $r(t) = z(t + h) - z(t - h)$ with $z(t) = \tanh(t/2)$. If the function

$$\delta_\theta(t) = \delta_\theta^+(t) + \delta_\theta^-(t), \qquad \delta_\theta^\pm(t) = B_{\theta \pm \omega}(r(t \pm h)) \tag{18}$$

has a nonremovable singularity at $t = \pi\mathrm{i}$ for some $\theta \in \mathbb{T}$, then the separatrix splits.

We note that $z(t)$ is meromorphic and its poles are the points in the set $\pi\mathrm{i} + 2\pi\mathrm{i}\mathbb{Z}$. Hence, since the function $B_\theta(r)$ is entire and nonzero, the compositions $\delta_\theta^+(t)$ and $\delta_\theta^-(t)$ always have a nonremovable singularity at the point $t = \pi\mathrm{i}$. Our sufficient condition for the splitting is that the sum $\delta_\theta^+(t) + \delta_\theta^-(t)$ be still singular at $t = \pi\mathrm{i}$, which is generic.

Proof. The function $B_\theta(r)$ is entire because the parity of the perturbation $\beta(x, y)$ cancels the square roots that appear in (17).

The first step is to rewrite the Melnikov function in a more convenient form. Using that $\alpha(x) = 0$ and the relation $\mu(t) = -4cs r'(t)$, the Melnikov function (10) has the form $M(\theta, t) = -4cs(\Delta_\theta)'(t)$, where

$$\Delta_\theta(t) = \sum_{k \in \mathbb{Z}} \delta_\theta^{[k]}(t), \qquad \delta_\theta^{[k]}(t) = B_{\theta + k\omega}(r(t + kh)).$$

Using that $\delta_\theta(t) = \delta_\theta^{[1]}(t) + \delta_\theta^{[-1]}(t)$, we shall prove that the series $\Delta_\theta(t)$—and hence, the Melnikov function—has a nonremovable singularity at $t = \pi\mathrm{i}$ for some $\theta \in \mathbb{T}$.

Since the function $B_\theta(r)$ is entire, the composition $\delta_\theta^{[k]}(t) = B_{\theta+k\omega}(r(t+kh))$ is analytic but at the poles of the meromorphic function $r(t+kh)$, which are the points in the sets $\pm h - kh + \pi\mathrm{i} + 2\pi\mathrm{i}\mathbb{Z}$. Thus, the difference

$$\Delta_\theta(t) - \delta_\theta(t) = \sum_{k\neq\pm 1} \delta_\theta^{[k]}(t)$$

is analytic at $t = \pi\mathrm{i}$ for any $\theta \in \mathbb{T}$. On the other hand, by hypothesis, $\delta_\theta(t)$ has a nonremovable singularity at $t = \pi\mathrm{i}$ for some $\theta \in \mathbb{T}$. ■

Next, we find some concrete perturbations of the form (16) that split the separatrix. For simplicity, we shall deal with perturbations such that the computation of the singular parts of the functions $\delta_\theta^\pm(t)$ defined in (18) around their singularity $t = \pi\mathrm{i}$ can be easily analyzed. Polynomial perturbations are a natural choice. We need the following notation for the statement of the result. Given any $\beta(x,y) \in \mathbb{R}_n[x,y]$, we shall denote by $\sum_{l=0}^n \beta_l(x,y)$ its decomposition as a sum of homogeneous polynomials. That is, $\beta_l(\rho x, \rho y) = \rho^l \beta_l(x,y)$ for all $\rho \in \mathbb{R}$. Let $R_\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ be the rotation

$$R_\varphi(x,y) = (x\cos\varphi - y\sin\varphi,\ x\sin\varphi + y\cos\varphi). \tag{19}$$

Proposition 11. *If $\beta(x,y) \in \mathbb{R}_{2n}[x,y]$ is even and $\beta_{2n} \circ R_{2\omega} \neq (-1)^n \beta_{2n}$, then the separatrix splits under the polynomial perturbation* (16).

Proof. The decomposition of the polynomial $\beta(x,y) \in \mathbb{R}_{2n}[x,y]$ has only even terms: $\beta(x,y) = \sum_{l=0}^n \beta_{2l}(x,y)$. Then the entire function $B_\theta : \mathbb{C} \to \mathbb{C}$ defined in (17) is the (not necessarily even) polynomial

$$B_\theta(r) = \sum_{l=0}^{n} \hat{B}_l(\theta) r^{l+1}, \qquad \hat{B}_l(\theta) = \frac{2^l \beta_{2l}(\cos\theta, \sin\theta)}{l+1}.$$

The point $t = \pi\mathrm{i}$ is a simple pole of the meromorphic function $z(t) = \tanh(t/2)$, and so it becomes a pole of order $n+1$ of the functions $\delta_\theta^+(t) = B_{\theta+\omega}(z(t+2h) - z(t))$ and $\delta_\theta^-(t) = B_{\theta-\omega}(z(t) - z(t-2h))$. In particular, there exist some Laurent coefficients $\hat{\delta}_1^\pm(\theta), \dots, \hat{\delta}_{n+1}^\pm(\theta)$ such that

$$\delta_\theta^\pm(t) = \frac{\hat{\delta}_{n+1}^\pm(\theta)}{(t-\pi\mathrm{i})^{n+1}} + \cdots + \frac{\hat{\delta}_1^\pm(\theta)}{t-\pi\mathrm{i}} + (\text{some analytic function at } t = \pi\mathrm{i}).$$

For instance, using that the residue of $z(t)$ at its poles is equal to 2, we get that the dominant Laurent coefficients are $\hat{\delta}_{n+1}^\pm(\theta) = (\mp 2)^{n+1} \hat{B}_n(\theta \pm \omega)$.

Finally, we note that if there exist some $\theta \in \mathbb{T}$ and some index $j = 1, \dots, n+1$ such that $\hat{\delta}_j^+(\theta) + \hat{\delta}_j^-(\theta) \neq 0$, then $\delta_\theta(t) = \delta_\theta^+(t) + \delta_\theta^-(t)$ has a nonremovable singularity at $t = \pi\mathrm{i}$ and the separatrix splits. The functional condition $\beta_{2n} \circ R_{2\omega} \neq (-1)^n \beta_{2n}$ is equivalent to the existence of some angle $\theta \in \mathbb{T}$ such that

$$\beta_{2n}(\cos(\theta+2\omega), \sin(\theta+2\omega)) \neq (-1)^n \beta_{2n}(\cos\theta, \sin\theta),$$

which is equivalent to the existence of θ such that $\hat{\delta}^+_{n+1}(\theta) + \hat{\delta}^-_{n+1}(\theta) \neq 0$. ■

Using this proposition, we shall obtain many polynomial perturbations that split the separatrix. To explain this, we introduce the complexified variables

$$z = x + y\mathrm{i}, \qquad \bar{z} = x - y\mathrm{i}. \tag{20}$$

In these variables, the functional equation $\beta_{2n} \circ R_{2\omega} = (-1)^n \beta_{2n}$ reads as $\tilde{\beta}_{2n} \circ \tilde{R}_{2\omega} = (-1)^n \tilde{\beta}_{2n}$. Here, $\tilde{R}_\varphi(z, \bar{z}) = (\mathrm{e}^{\varphi \mathrm{i}} z, \mathrm{e}^{-\varphi \mathrm{i}} \bar{z})$ and

$$\tilde{\beta}_{2n}(z, \bar{z}) = \sum_{k=-n}^{n} \tilde{\beta}^{[k]}_{2n} z^{n+k} \bar{z}^{n-k}$$

stand for the rotation (19) and the homogeneous polynomial $\beta_{2n}(x, y)$ in the complexified variables, respectively. The transformed polynomial $\tilde{\beta}_{2n}(z, \bar{z})$ is still a homogeneous polynomial of degree $2n$ because the change (20) is linear.

Lemma 12. *The functional equation $\beta_{2n} \circ R_{2\omega} = (-1)^n \beta_{2n}$ holds if and only if*

$$\tilde{\beta}^{[k]}_{2n} \left(\mathrm{e}^{4\omega k \mathrm{i}} - (-1)^n \right) = 0 \qquad \forall k = -n, \dots, n. \tag{21}$$

Proof. $\tilde{R}_{2\omega}(z, \bar{z}) = (\mathrm{e}^{2\omega \mathrm{i}} z, \mathrm{e}^{-2\omega \mathrm{i}} \bar{z})$ maps $z^{n+k} \bar{z}^{n-k}$ onto $\mathrm{e}^{4k\omega \mathrm{i}} z^{n+k} \bar{z}^{n-k}$. ■

Now we are ready to give precise statements about the splitting of the separatrix under polynomial perturbations of the form (16). For instance, we shall see that both *nonresonant frequencies* and *high-order resonant frequencies*—that is, $\omega/\pi \notin \mathbb{Q}$ or ω/π is an irreducible fraction with a high denominator—are strong obstructions for the persistence of the separatrix. A homogeneous polynomial $\beta_{2n}(x, y)$ of degree $2n$ is *rotationally invariant* when it has the form

$$\beta_{2n}(x, y) = \tilde{\beta}^{[0]}_{2n} z^n \bar{z}^n = \tilde{\beta}^{[0]}_{2n} |z|^{2n} = \tilde{\beta}^{[0]}_{2n} (x^2 + y^2)^n$$

for some constant $\tilde{\beta}^{[0]}_{2n} \in \mathbb{R}$. These polynomials are the only homogeneous ones that remain invariant under the action of the continuous group of rotations $R_\varphi : \mathbb{R}^2 \to \mathbb{R}^2$.

Theorem 13. *If $\beta(x, y)$ is an even polynomial of degree $2n$, then the perturbation (16) splits the separatrix in any of the following two cases:*

(i) *n odd and $\mathrm{e}^{4k\omega \mathrm{i}} \neq -1$ for $k = 1, \dots, n$; or*

(ii) *n even, $\beta_{2n}(x, y)$ not rotationally invariant, and $\mathrm{e}^{4k\omega \mathrm{i}} \neq 1$ for $k = 1, \dots, n$.*

Proof. From Proposition 11 and Lemma 12, we know that (21) is a necessary condition for the persistence of the separatrix. Let us check that this condition is incompatible with the two listed cases.

(i) If n is odd and $\mathrm{e}^{4k\omega \mathrm{i}} \neq -1$ for $k = 1, \dots, n$, condition (21) implies that $\tilde{\beta}^{[k]}_{2n} = 0$ for all $k = -n, \dots, n$. Therefore, the homogeneous polynomial $\beta_{2n}(x, y)$ is zero, which contradicts the fact that $\beta(x, y)$ has degree $2n$.

(ii) If n is even and $\mathrm{e}^{4k\omega \mathrm{i}} \neq 1$ for $k = 1, \dots, n$, then condition (21) implies that $\tilde{\beta}^{[k]}_{2n} = 0$ for all $k \neq 0$, which contradicts the fact that $\beta(x, y)$ has degree $2n$ and $\beta_{2n}(x, y)$ is not rotationally invariant. ■

It would be interesting to know whether there exist some entire perturbations of the form (16) that preserve the separatrix. Of course, such perturbations cannot verify the sufficient condition for splitting given in Theorem 10. We have not found any perturbation of this

kind, which is not so strange because in similar contexts, related with other McMillan maps, they simply do not exist. An area-preserving example of this situation can be found in [13], and a high-dimensional symplectic one in [14].

7. Bifurcations of the zero-level set in an example. In this section, we shall study the bifurcations in $\omega \in \mathbb{T}$ of the topological shape of the zero-set $Z \subset \mathbb{T} \times \mathbb{R}$ for the perturbation $\kappa(x, y, z) = (0, x, 0)$. We note that, according to Lemma 5 and Remark 4, Z contains (and is symmetric with regard to) the eight symmetric points shown in Figure 3 and it is also $(\pi/2, 0)$-periodic.

Based on detailed numerical computations and several analytical arguments, we conjecture that 0 is a singular value of the Melnikov function if and only if $\omega = \pm\pi/2$, and so the only bifurcations of $Z = M^{-1}(0)$ take place at those values. For instance, we show in Figure 4 the numerically computed shape of $\tilde{Z}$ for several values of the frequency in the range $0 \le \omega \le \pi$.

We give a dynamical interpretation of these Melnikov-like results. If $\omega \neq \pm\pi/2$, the set of primary intersections $P_\epsilon = Z + \mathrm{O}(\epsilon)$ consists of four vertical curves for ϵ small enough; see Theorem 1. Each curve has a symmetric twin, because P_ϵ is invariant under the axial symmetry $S(x, y, z) = (-x, -y, z)$. The fixed sets of the reversors are

$$\operatorname{Fix} R_\epsilon = \{(x, y, z) \in \mathbb{R}^3 : y = \epsilon x/2,\ z = 0\},$$
$$\operatorname{Fix} T_\epsilon = \{(x, y, z) \in \mathbb{R}^3 : x = 0,\ z = 0\}.$$

Hence, two heteroclinic vertical curves cross a curve $\mathrm{O}(\epsilon)$-close to the x-axis, and the other pair of heteroclinic curves cross the y-axis. The rotation number of these vertical curves is equal to ω in the range $-\pi/2 < \omega < \pi/2$, but it jumps to $\omega \mp \pi$ when the bifurcation values $\omega = \pm\pi/2$ are crossed. The shape of the primary set when $\omega = \pm\pi/2$ is not completely clear, because then 0 is a singular value of the Melnikov function and Theorem 1 cannot be applied. This is an open question.

The rest of the section is devoted to presenting some rigorous results supporting the previous conjecture, although we have found a complete proof only for $h \ge h_0 \approx 2.28$. Nevertheless, we have been able to prove the following results. If ω is a *regular frequency* (that is, if 0 is a regular value of the Melnikov function), then Z contains just four vertical curves. Otherwise, we say that ω is a *singular frequency* and Z contains the four vertical curves jointly with the images and preimages of exactly one horizontal straight line, in which case the degenerate zeroes of the Melnikov function are just the points in the intersections between the horizontal and vertical curves. The number of singular frequencies is finite. The frequencies $\omega = 0$ and $\omega = \pi$ are regular. The frequencies $\omega = \pm\pi/2$ are singular and are the only singular ones when the characteristic exponent is big enough: $h \ge h_0 := \log 16 - \log(\sqrt{113} - 9) \approx 2.28$.

In order to lighten the computations, we write the Melnikov function in its simplest form. Let $\chi : \mathbb{R} \to \mathbb{R}$ be the function

$$\chi(t) = \frac{4c^2 s^2}{\cosh((t+h)/2)\cosh^2(t/2)\cosh((t-h)/2)}. \tag{22}$$

Then, using that $\alpha(x) = x$ and $\beta(x, y) = 0$, the Melnikov function (10) becomes

$$M(\theta, t) = a(t)\sin 2\theta + b(t)\cos 2\theta, \tag{23}$$

where $a(t)$ and $b(t)$ are given by the absolutely convergent series

$$a(t) = \sum_{k\in\mathbb{Z}} \cos(2k\omega)\chi(t+kh), \qquad b(t) = \sum_{k\in\mathbb{Z}} \sin(2k\omega)\chi(t+kh).$$

We also introduce the complex-valued function

$$E(t) = E_\omega(t) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{2k\omega\mathrm{i}}\chi(t+kh) = a(t) + b(t)\mathrm{i}, \tag{24}$$

which plays a crucial role in the digression because of the relation

$$\partial_\theta M(\theta,t)/2 + M(\theta,t)\mathrm{i} = E(t)\mathrm{e}^{2\theta\mathrm{i}}. \tag{25}$$

This relation has interesting consequences. For instance, if $(\theta_0, t_0) \in \mathbb{T}\times\mathbb{R}$ is a degenerate zero of the Melnikov function, $E(t_0)$ must be zero. In particular, 0 is a regular value of the Melnikov function when $E(t)$ has no real zeroes. Therefore, we are naturally led to the study of the sets

$$\Omega = \Omega_h = \{\omega\in\mathbb{T} : E_\omega(t) \text{ has some real zero}\}, \qquad h>0. \tag{26}$$

Their main properties are addressed in the following lemma, whose proof is deferred to Appendix A. The proof is based on some nice properties of quasi-elliptic functions that can be deduced from elementary facts of complex variable theory contained in any basic textbook, like, for instance, [34].

Lemma 14. *Given any $h>0$, the set (26) is finite, π-periodic, and symmetric: $\Omega = -\Omega$. If the function (24) has some real zero, all of them are simple, and the set of its real zeroes is either $h\mathbb{Z}$ or $h/2+h\mathbb{Z}$, so Ω is the disjoint union of the sets*

$$\Omega^0 = \{\omega\in\mathbb{T} : E_\omega(0)=0\}, \qquad \Omega^1 = \{\omega\in\mathbb{T} : E_\omega(h/2)=0\}.$$

Further, $\pm\pi/2\in\Omega^1$, $0\notin\Omega$, and $\pi\notin\Omega$. Finally, $\Omega^1 = \{\pm\pi/2\}$ for $h\ge h_1 := 2\log\frac{20}{9}\approx 1.60$ and $\Omega^0=\emptyset$ for $h\ge h_0 := \log 16 - \log(\sqrt{113}-9)\approx 2.28$.

Conjecture 15. *$\Omega = \Omega^1 = \{\pm\pi/2\}$ and $\Omega^0=\emptyset$ for all $h>0$.*

We present in Appendix B strong numerical evidence for this conjecture.

Lemma 16. *Let $E:\mathbb{R}\to\mathbb{C}$ be an analytic function such that $E(-t)=\overline{E(t)}$.*

(i) *If $E(t)$ has no real zeroes, then there exists a unique odd analytic function $\varphi:\mathbb{R}\to\mathbb{R}$ and an integer $n\in\{0,1\}$ such that $E(t) = |E(t)|\,\mathrm{e}^{(\varphi(t)+\pi n)\mathrm{i}}$ for all $t\in\mathbb{R}$.*

(ii) *If $E(t)$ has no real multiple zeroes, then there exists a unique odd analytic function $\varphi:\mathbb{R}\to\mathbb{R}$ and a function $n:\mathbb{R}\to\{0,1\}$ such that $E(t) = |E(t)|\,\mathrm{e}^{(\varphi(t)+\pi n(t))\mathrm{i}}$ for all $t\in\mathbb{R}$. The function $n(t)$ is constant, but at the zeroes of $E(t)$.*

Proof. (i) If a function is analytic and never zero on a convex subset of the complex plane, then it has an analytic argument on that convex subset. This is an elementary result in complex variable theory; see [4, section 2.1]. Let $\varphi(t)$ be an analytic argument of $E(t)/E(0)$; that is, any analytic function $\varphi:\mathbb{R}\to\mathbb{R}$ such that

$$E(t) = E(0)\,|E(t)/E(0)|\,\mathrm{e}^{\varphi(t)\mathrm{i}} = |E(t)|\,\mathrm{e}^{(\varphi(t)+\pi n)\mathrm{i}}.$$

Obviously, $n = 0$ if $E(0) > 0$ and $n = 1$ if $E(0) < 0$. The argument is not unique, but it is determined up to a multiple of 2π; that is, it is determined once we choose the value of $\varphi(0)$ from the set $2\pi\mathbb{Z}$. The condition $E(-t) = \overline{E(t)}$ implies that $\varphi(-t) + \varphi(t) = 2\varphi(0)$ for all real t. If we want an odd argument, $\varphi(0) = 0$ is the only possible choice.

(ii) It suffices to realize that, at any simple zero, the argument undergoes a jump by a multiple of π. When these jumps are stored in the discrete-valued function $n(t)$, the function $\varphi(t)$ remains analytic. ∎

These two lemmas are the basis for the next theorem, in which the shape and bifurcations of the zero-level set $Z = M^{-1}(0) \subset \mathbb{T} \times \mathbb{R}$ are described.

Theorem 17. *Let Z be the zero-level set of the Melnikov function (23). Let $\Omega = \Omega^0 \cup \Omega^1$ be the decomposition of the set (26) given in Lemma 14. Let $\omega \in \mathbb{T}$. There exists a unique odd analytic function $\bar{\theta}_\omega : \mathbb{R} \to \mathbb{R}$ such that the following hold:*

(i) *If $\omega \notin \Omega$, then $Z = \{\theta = \bar{\theta}_\omega(t) \pmod{\pi/2}\}$ and ω is a regular frequency.*

(ii) *If $\omega \in \Omega^j$, then $Z = \{t = jh/2 \pmod{h}\} \cup \{\theta = \bar{\theta}_\omega(t) \pmod{\pi/2}\}$ and ω is a singular frequency. The degenerate zeroes of the Melnikov function are just the points in the intersections between the horizontal lines and the vertical curves.*

(iii) *The function $\bar{\Theta}(t, \omega) := \bar{\theta}_\omega(t)$ is analytic on $\mathbb{R} \times (\mathbb{T} \setminus \Omega)$.*

(iv) *If $\omega = 0 \pmod{\pi/2}$, then $\bar{\theta}_\omega(t) \equiv 0$.*

(v) $\bar{\theta}_\omega(t + h) = \bar{\theta}_\omega(t) + \omega \pmod{\pi/2}$.

(vi) $\bar{\theta}_\omega(h/2 - t) + \bar{\theta}_\omega(h/2 + t) = \omega \pmod{\pi/2}$.

(vii) $\bar{\theta}_{-\omega}(t) = -\bar{\theta}_\omega(t)$ *and* $\bar{\theta}_{\omega+\pi}(t) = \bar{\theta}_\omega(t)$.

Proof. Sometimes, we do not write explicitly the dependence on the frequency. Using that $a(t)$ is even and $b(t)$ is odd, we see that the function (24) verifies the relation $E(-t) = \overline{E(t)}$. This is important, because it was a hypothesis in Lemma 16.

(i) If $\omega \notin \Omega$, then $E(t)$ has no real zeroes, and so ω is a regular frequency. It remains to prove that Z is composed of four vertical curves of the form $\{\theta = \bar{\theta}(t) \pmod{\pi/2}\}$ for some function $\bar{\theta}(t)$. Let $\bar{\theta} : \mathbb{R} \to \mathbb{R}$ be the odd analytic function $\bar{\theta}(t) = -\varphi(t)/2$, where $\varphi(t) = \arg(E(t)/E(0))$ is the argument introduced in Lemma 16. Let n be the integer mentioned in the same lemma. Using that $E(t)$ has no real zeroes jointly with relation (25), we have

$$\begin{aligned} Z = M^{-1}(0) &= \{(\theta, t) \in \mathbb{T} \times \mathbb{R} : E(t)\mathrm{e}^{2\theta \mathrm{i}} \in \mathbb{R}\} \\ &= \{(\theta, t) \in \mathbb{T} \times \mathbb{R} : \sin(\varphi(t) + \pi n + 2\theta) = 0\} \\ &= \{(\theta, t) \in \mathbb{T} \times \mathbb{R} : \theta = \bar{\theta}(t) \pmod{\pi/2}\}. \end{aligned}$$

(ii) We know that $E(t)$ has no real multiple zeroes. Let $\varphi(t)$ and $n(t)$ be the functions given in Lemma 16. Let $\bar{\theta}(t) = -\varphi(t)/2$. Then

$$\begin{aligned} Z &= \{(\theta, t) \in \mathbb{T} \times \mathbb{R} : E(t)\mathrm{e}^{2\theta \mathrm{i}} \in \mathbb{R}\} \\ &= \{(\theta, t) : E(t) = 0\} \cup \{(\theta, t) : \sin(\varphi(t) + \pi n(t) + 2\theta) = 0\} \\ &= \{t = jh/2 \pmod{h}\} \cup \{\theta = \bar{\theta}(t) \pmod{\pi/2}\}. \end{aligned}$$

Therefore, $Z = M^{-1}(0)$ contains four vertical curves that intersect infinitely many horizontal straight lines. Obviously, these intersections are degenerate zeroes of the Melnikov function. Next, we shall prove that the other zeroes are nondegenerate.

Let (θ_0, t_0) be a zero of the Melnikov function not contained in any horizontal line: $M(\theta_0, t_0) = 0$ and $E(t_0) \neq 0$. Then $\partial_\theta M(\theta_0, t_0) = 2E(t_0)\mathrm{e}^{2\theta_0 \mathrm{i}} \neq 0$; see (25).

On the other hand, let (θ, t_1) be a point contained in some horizontal line: $M(\theta, t_1) = 0$, $E(t_1) = 0$, and $E'(t_1) \neq 0$. Again from relation (25), we get $\partial_\theta M(\theta, t_1) = 0$ and $\partial_t M(\theta, t_1) = \Im E'(t_1)\mathrm{e}^{2\theta \mathrm{i}}$, where $\Im$ denotes imaginary part. Hence, the degenerate zeroes on the horizontal line $\{t = t_1\}$ are just those that verify the condition $\Im E'(t_1)\mathrm{e}^{2\theta \mathrm{i}} = 0$. But, since $E'(t_1) \neq 0$, there are exactly four of such angles $\theta \in \mathbb{T}$. These four angles are the ones corresponding to the four intersections of the horizontal line $\{t = t_1\}$ with the four vertical curves $\{\theta = \bar{\theta}(t) \pmod{\pi/2}\}$.

(iii) Level sets associated with regular values of analytic functions vary in an analytic way under analytic perturbations.

(iv) If $\omega = 0 \pmod{\pi/2}$, then $\sin(2k\omega) = 0$ for all k, and $b(t) = \Im E(t) \equiv 0$.

(v) This has to do with the fact that the zero-level set Z is (ω, h)-periodic. Given any $t \in \mathbb{R}$, we consider the slice $Z_t = \{\theta \in \mathbb{T} : (\theta, t) \in Z\}$. If $E(t) \neq 0$, then

$$Z_{t+h} = \{\theta \in \mathbb{T} : \theta = \bar{\theta}(t + h) \pmod{\pi/2}\},$$
$$Z_t + \omega = \{\theta \in \mathbb{T} : \theta = \bar{\theta}(t) + \omega \pmod{\pi/2}\}.$$

But these two sets coincide, due to the (ω, h)-periodicity of Z, and so we obtain that $\bar{\theta}(t+h) = \bar{\theta}(t) + \omega \pmod{\pi/2}$ for any real t such that $E(t) \neq 0$. Indeed, by analytic continuation, this equality holds for any real t.

(vi) It follows directly from the previous item and the odd character of $\bar{\theta}(t)$: $\bar{\theta}(h/2 - t) + \bar{\theta}(h/2 + t) = -\bar{\theta}(t - h/2) + \bar{\theta}(t - h/2) + \omega = \omega \pmod{\pi/2}$.

(vii) First, $\bar{\theta}_{-\omega}(t) = -\frac{1}{2}\arg E_{-\omega}(t) = -\frac{1}{2}\arg \overline{E_\omega(t)} = \frac{1}{2}\arg E_\omega(t) = -\bar{\theta}_\omega(t)$. Second, $\bar{\theta}_{\omega+\pi}(t) = -\frac{1}{2}\arg E_{\omega+\pi}(t) = -\frac{1}{2}\arg E_\omega(t) = \bar{\theta}_\omega(t)$. ∎

Remark 9. The numerical computations show that $\bar{\theta}_\omega(t + h) = \bar{\theta}_\omega(t) + \omega$ holds only in the range $-\pi/2 < \omega < \pi/2$; see Figure 4. This does not contradict item (v) in Theorem 17.

Remark 10. Similar results hold for the perturbation $\kappa(x, y, z) = (0, 0, y^2)$. In that case, it turns out that the Melnikov function has the form

$$M(\theta, t) = \hat{a}(t)\sin 2\theta + \hat{b}(t)\cos 2\theta + \hat{c}(t)$$

for some absolutely convergent series $\hat{a}(t)$, $\hat{b}(t)$, and $\hat{c}(t)$. The analytical study is harder because of the additional third term—compare with (23). We have numerically checked that the only bifurcations take place at the singular frequencies $\omega \in \{0, \pm\pi/2, \pi\}$, whereas the zero-level set still contains just four vertical curves for regular frequencies.

8. Conclusion and open problems.

In this paper, we have obtained several analytical results about the splitting of separatrices under perturbations of some integrable volume-preserving maps using a discrete version of the Melnikov method. The integrable maps have a two-dimensional heteroclinic connection of spherical shape between two fixed points of saddle-focus type. We have bounded the topological complexity of the primary heteroclinic set under some polynomial perturbations. We have also given a sufficient condition for the splitting of the separatrices under some entire perturbations. Finally, we have obtained a complete picture of the bifurcations that take place under a simple perturbation. Despite these results, many unsolved questions remain. We indicate four.

The first question is: How accurate are our first-order Melnikov estimates? Of course, it would be necessary to compute numerically the perturbed invariant manifolds and to compare the real distance between them with the distance predicted by using the Melnikov function. A related numerical experiment was performed in [9] for linear perturbations of a four-dimensional symplectic version of the McMillan map. This is a work in progress.

Second, we conjecture that the separatrix splitting studied in this paper is exponentially small in the characteristic exponent, although the role of the frequency is still unclear. Several examples of the effect that resonant frequencies can have in the dynamics of three-dimensional maps near Hopf-saddle-node bifurcations can be found in [11, 12], although not related to a problem about the splitting of separatrices. One could guess an asymptotic exponentially small formula for the splitting using a multiple-precision arithmetic, like in [31]. Such formulae for the splitting of one-dimensional heteroclinic connections between saddle-focus fixed points of volume-preserving systems have already been found in [2] (for maps) and [8] (for flows), but we do not know any similar formula for the two-dimensional case.

Another question is: What about Šil'nikov-like bifurcations in the discrete setting? The perturbation of a spheromak structure in three-dimensional flows is a classical setup for studying Šil'nikov bifurcations. Some results about the continuous case are contained in [10]. It is natural to consider the discrete version of this problem, although the problem seems qualitatively more complicated.

It would be also interesting to study some questions about transport. As a first step, we should compute the geometric flux through the perturbed separatrices. The $\mathrm{O}(\epsilon)$-term of this flux can be computed by integrating certain Melnikov two-forms over a suitable region; see [24]. Next, we could follow the ideas introduced in [26], although we must take into account that the scenario for maps is richer than the one for flows. For instance, we recall that equatorial and bubble-type heteroclinic curves cannot appear in autonomous flows.

Finally, we stress that only volume-preserving (conservative) perturbations have been studied in this paper, although the study of dissipative perturbations is also possible. See, for instance, the theories developed in [9, 7, 25].

Appendix A. Proof of Lemma 14. Here, we study the existence of real zeroes of the function (24). This function has many properties similar to those of elliptic functions, and so we shall study its zeroes using tools typical in the theory of elliptic functions.

We list in the next lemma some basic properties of the function $E(t)$, including that $E(t)$ is meromorphic, $2\pi\mathrm{i}$-periodic, and h-quasi-periodic, and so quasi-elliptic.

Lemma 18. *The function $E(t) = E_\omega(t) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{2k\omega\mathrm{i}}\chi(t+kh)$ verifies the following:*

(i) *It is meromorphic in the complex plane and its poles are the points in the set $\mathcal{P} = \pi\mathrm{i} + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}$ (double ones).*

(ii) *$E(t+2\pi\mathrm{i}) = E(t)$ and $E(t+h) = \mathrm{e}^{-2\omega\mathrm{i}}E(t)$ for all complex t.*

(iii) *$E(-\overline{t}) = \overline{E(t)}$ for all complex t.*

(iv) *If $\omega = \pi/2 \pmod{\pi}$, the points $t = h/2 \pmod{h}$ are the only real zeroes of $E(t)$, all of them being simple ones.*

(v) *If $\omega = 0 \pmod{\pi}$, then $E(t)$ has no real zeroes.*

(vi) *$E_{-\omega}(t) = \overline{E_\omega(\overline{t})}$ and $E_{\omega+\pi}(t) = E_\omega(t)$.*

Proof. (i) The function $\chi(t)$ is meromorphic in the complex plane, and its poles are the

points in $\pi\mathrm{i} + 2\pi\mathrm{i}\mathbb{Z}$ (double ones) and $\pi\mathrm{i} \pm h + 2\pi\mathrm{i}\mathbb{Z}$ (simple ones).

(ii) The function $E(t)$ is $2\pi\mathrm{i}$-periodic because so is $\chi(t)$. On the other hand,

$$E(t+h) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{2k\omega\mathrm{i}}\chi(t+kh+h) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{2(k-1)\omega\mathrm{i}}\chi(t+kh) = \mathrm{e}^{-2\omega\mathrm{i}}E(t).$$

(iii) We recall that $E(t) = a(t) + b(t)\mathrm{i}$, where $a(t)$ and $b(t)$ are real analytic functions such that $a(t)$ is even and $b(t)$ is odd. Hence, $E(-\overline{t}) = a(-\overline{t}) + b(-\overline{t})\mathrm{i} = a(\overline{t}) - b(\overline{t})\mathrm{i} = \overline{a(t)} - \overline{b(t)}\mathrm{i} = \overline{E(t)}$.

(iv) In this case, $\mathrm{e}^{2k\omega\mathrm{i}} = -1$, and so $E(t+h) = -E(t)$. Thus, $E(t)$ becomes an elliptic function with periods $2h$ and $2\pi\mathrm{i}$. Further, $E(t)$ has order four, because it has just four poles (counted with multiplicity) on any cell with periods $2h$ and $2\pi\mathrm{i}$; see (i).

Using that the elliptic function

$$E(t) = \sum_{k\in\mathbb{Z}} (-1)^k \chi(t+kh)$$

is even, h-antiperiodic, and $2\pi\mathrm{i}$-periodic, we get that $E(t)$ vanishes at the four points $h/2$, $3h/2$, $h/2 + \pi\mathrm{i}$, and $3h/2 + \pi\mathrm{i}$. For instance, $E(h/2) = E(-h/2) = -E(h/2)$, so $E(h/2) = 0$. But we know that $E(t)$ has exactly four zeroes (counted with multiplicity) on each cell with periods $2h$ and $2\pi\mathrm{i}$. Since the previous four zeroes belong to the same cell, they are the only ones (modulo periodicities), and they are simple.

(v) $\chi(t) > 0$ in $\mathbb{R}$. Thus, if $\mathrm{e}^{2\omega\mathrm{i}} = 1$, $E(t) = \sum_{k\in\mathbb{Z}} \chi(t+kh) > 0$ for any real t.

(vi) First, $E_{-\omega}(t) = \sum_{k\in\mathbb{Z}} \overline{\mathrm{e}^{2k\omega\mathrm{i}}\chi(t+kh)} = \overline{\sum_{k\in\mathbb{Z}} \mathrm{e}^{2k\omega\mathrm{i}}\chi(\overline{t}+kh)} = \overline{E_\omega(\overline{t})}$, because $\chi(t)$ is real analytic. The second property is trivial. ∎

Next, we gain some insight on the structure of the complex zeroes of the quasi-elliptic function $E(t)$. Roughly speaking, we state in the following lemma that $E(t)$ has order two, and its zeroes look like in Figure 5. The proof is adapted from similar proofs about elliptic functions.

Lemma 19. *The quasi-elliptic function $E(t)$ has order two; that is, it has two zeroes in any cell with periods h and $2\pi\mathrm{i}$. Let t_1 and t_2 be the zeroes in any cell. Then*

$$t_1 + t_2 \in 2\omega\mathrm{i} + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}. \tag{27}$$

Additionally, the set

$$\mathcal{T} = \{t_1, t_2\} + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}$$

formed by the complex zeroes of $E(t)$ is symmetric with regard to the vertical lines $\{\Re t = 0 \pmod h\}$ and $\{\Re t = h/2 \pmod h\}$.

Proof. We recall the following version of the argument principle [34, section 6.3]. Let C be a contour in the complex plane, let $f(t)$ be a function analytic inside and on C, let $g(t)$ be a meromorphic function without zeroes or poles on C, and let $t_1, \dots, t_J$ and $p_1, \dots, p_K$ be the zeroes and poles of $g(t)$ in the interior of C, repeated as many times as their multiplicities and orders, respectively. Then

$$\frac{1}{2\pi\mathrm{i}} \oint_C f(t)\frac{g'(t)}{g(t)}\mathrm{d}t = \sum_j f(t_j) - \sum_k f(p_k). \tag{28}$$

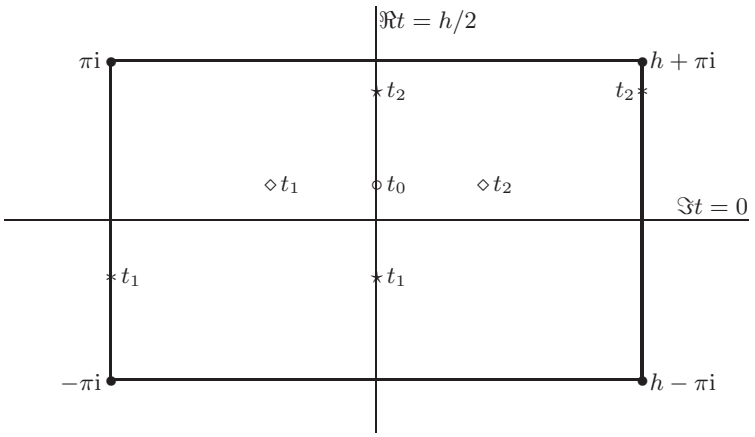


Figure 5. *The three scenarios for the set of zeroes of the quasi-elliptic function $E(t)$. First ($\star$): The zeroes are on the lines $\{\Re t = h/2 \pmod h\}$. Second ($\diamond$): The zeroes have the same imaginary part modulo $2\pi\mathrm{i}$. Third ($*$): The zeroes are on the lines $\{\Re t = 0 \pmod h\}$. The double poles are marked with the symbol $\bullet$. The middle point is $t_0 = h/2 + \omega\mathrm{i}$ in the three cases, because $t_1 + t_2 \in 2\omega\mathrm{i} + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}$.*

This version of the principle assumes that the contour has no self-intersections, and that it is oriented counterclockwise.

If we choose any cell of periods h and $2\pi\mathrm{i}$ as the contour C, and set $f(t) = 1$, $g(t) = E(t)$ in (28), we get that $E(t)$ has $J = K = 2$ zeroes in the cell, because the integrals of the quotient $E'(t)/E(t)$ over opposite sides of the cell cancel out. (We recall that $E(t)$ has exactly one double pole on each cell.)

Let t_1 and t_2 be the zeroes and p be the double pole of $E(t)$ in a cell C. Let the corners of the cell be s, $s + h$, $s + h + 2\pi\mathrm{i}$, $s + 2\pi\mathrm{i}$. Now, if we keep the same contour, but take $f(t) = t$ and $g(t) = E(t)$, we get that

$$\begin{aligned} t_1 + t_2 - 2p &= \frac{1}{2\pi\mathrm{i}} \left(\int_s^{s+h} + \int_{s+h}^{s+h+2\pi\mathrm{i}} + \int_{s+h+2\pi\mathrm{i}}^{s+2\pi\mathrm{i}} + \int_{s+2\pi\mathrm{i}}^{s} \right) \frac{tE'(t)}{E(t)} \mathrm{d}t \\ &= \frac{1}{2\pi\mathrm{i}} \left(h \int_s^{s+2\pi\mathrm{i}} \frac{E'(t)}{E(t)} \mathrm{d}t - 2\pi\mathrm{i} \int_s^{s+h} \frac{E'(t)}{E(t)} \mathrm{d}t \right) \\ &= \frac{h}{2\pi\mathrm{i}} \log E(t)\big|_s^{s+2\pi\mathrm{i}} + \log E(t)\big|_{s+h}^{s} \end{aligned}$$

on making use of the quasi-periodic properties of $E(t)$.

Using again the quasi periodicities $E(s + 2\pi\mathrm{i}) = E(s)$ and $E(s + h) = \mathrm{e}^{-2\omega\mathrm{i}} E(s)$, we see that $\log E(t)\big]_s^{s+2\pi\mathrm{i}} \in 2\pi\mathrm{i}\mathbb{Z}$ and $\log E(t)\big]_{s+h}^{s} \in 2\omega\mathrm{i} + 2\pi\mathrm{i}\mathbb{Z}$. Therefore,

$$t_1 + t_2 \in 2\omega\mathrm{i} + 2p + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}.$$

Now, relation (27) follows because we know that the double pole $p \in \pi\mathrm{i} + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}$.

Finally, the specular symmetries with regard to the vertical lines are a direct consequence of the relations $E(-\overline{t}) = \overline{E(t)}$ and $E(t + h) = \mathrm{e}^{-2\omega\mathrm{i}} E(t)$. ■

Using this lemma, we realize that there are just three possible scenarios for the set of complex zeroes of the quasi-elliptic function $E(t)$, which are listed in the caption of Figure 5.

We have numerically checked that only the first scenario takes place, but we have not found a proof. Therefore, in the following lemma we can only deduce that the set of real zeroes of $E(t)$ is either $h\mathbb{Z}$ or $h/2+h\mathbb{Z}$, although we suspect that the case $h\mathbb{Z}$ is a mirage.

Lemma 20. *If $E(t)$ has some real zeroes, all of them are simple, and the set of its real zeroes is either $h\mathbb{Z}$ or $h/2+h\mathbb{Z}$.*

Proof. We note that $E(t)$ cannot have either a double real zero or two different real zeroes modulo h. On the contrary, we could take $t_1, t_2 \in \mathbb{R}$ in Lemma 19, so that $\mathbb{R} \ni t_1 + t_2 \in 2\omega\mathrm{i} + h\mathbb{Z} + 2\pi\mathrm{i}\mathbb{Z}$. But this would imply that $\omega = 0 \pmod{\pi}$, and then $E(t)$ has no real zeroes; see Lemma 18. Therefore, the set of real zeroes has the form $t_* + h\mathbb{Z}$ for some single real zero $t_* \in [0, h)$. But, since $t_* + h\mathbb{Z}$ must be symmetric with regard to the points 0 and $h/2$, there are only two possibilities: either $t_* = 0$ or $t_* = h/2$. ∎

As a by-product of this lemma, the set defined in (26) can also be defined as

$$\Omega = \Omega_h = \{\omega \in \mathbb{T} : E_\omega(0)E_\omega(h/2) = 0\}.$$

To prove Lemma 14, it suffices to check that (1) Ω is finite; (2) $E_\omega(0) \neq 0$ for all $h \geq h_0$ and all $\omega \in \mathbb{T}$; and (3) $E_\omega(h/2) \neq 0$ for all $h \geq h_1$ and all $\omega \neq \pm\pi/2$.

Lemma 21. *The set $\Omega = \{\omega \in \mathbb{T} : E_\omega(0)E_\omega(h/2) = 0\}$ is finite.*

Proof. The claim follows from the fact that, once any $h > 0$ is fixed, the function

$$\mathbb{T} \ni \omega \mapsto E_\omega(0)E_\omega(h/2) \in \mathbb{C}$$

is analytic, but not identically zero since the series $E_0(t) = \sum_{k\in\mathbb{Z}} \chi(t + kh)$ is positive for all real t. ∎

Lemma 22. *If $h \geq h_0 := \log 16 - \log(\sqrt{113} - 9)$, then $E_\omega(0) > 0$ for all $\omega \in \mathbb{T}$.*

Proof. We introduce the positive quantities

$$\chi_k = \chi_k(h) := \chi(kh) = \frac{4\cosh^2(h/2)\sinh^2(h/2)}{\cosh((k+1)h/2)\cosh^2(kh/2)\cosh((k-1)h/2)},$$

where $\chi(t)$ is the function given in (22). Our goal is to prove that

$$E_\omega(0) = \sum_{k\in\mathbb{Z}} \mathrm{e}^{2k\omega\mathrm{i}}\chi_k = \chi_0 + 2\sum_{k\geq 1} \cos(2k\omega)\chi_k > 0$$

for all $h \geq h_0$ and $\omega \in \mathbb{T}$. Here, we have used the symmetry $\chi_{-k} = \chi_k$. Since $\max_{\omega\in\mathbb{T}} |\cos(2k\omega)| = 1$, in order to prove the lemma it suffices to establish that

$$2\sum_{k\geq 1} \chi_k(h) < \chi_0(h) \qquad \forall h \geq h_0 := \log 16 - \log(\sqrt{113} - 9). \tag{29}$$

The rest of the proof is devoted to obtaining this bound. If we work with the multiplicative variable $x = \mathrm{e}^{-h} \in (0, 1)$, then $\chi_0 = \chi_0(x) = (1-x)^2/x$ and

$$\chi_k = \chi_k(x) = \frac{4(1-x^2)^2x^{2k-2}}{(1+x^{k+1})(1+x^k)^2(1+x^{k-1})} < 4(1-x^2)^2x^{2k-2}$$

for any $k \geq 1$ and $x \in (0,1)$. In particular,

$$\sum_{k\geq 1} \chi_k(x) < 4(1-x^2)^2 \sum_{k\geq 1} x^{2k-2} = 4(1+x)(1-x) \qquad \forall x \in (0,1).$$

Let $x_0 := (\sqrt{113}-9)/16 < 1$ and $h_0 := \log(1/x_0) > 0$. If $h \geq h_0$, then $x = \mathrm{e}^h \in (0, x_0]$ and $8x^2 + 9x - 1 \leq 0$. In particular, we conclude that the bound (29) holds. ■

Lemma 23. *If $h \geq h_1 := 2\log\frac{20}{9}$, then $E_\omega(h/2) \neq 0$ for all $\omega \neq \pm\pi/2$.*

Proof. This proof is similar to the previous one. We introduce the positive quantities

$$\tilde{\chi}_k = \tilde{\chi}_k(h) := \chi(kh + h/2) = \frac{4\cosh^2(h/2)\sinh^2(h/2)}{\cosh\big((\frac{k}{2}+\frac{3}{4})h\big)\cosh^2\big((\frac{k}{2}+\frac{1}{4})h\big)\cosh\big((\frac{k}{2}-\frac{1}{4})h\big)}.$$

Our goal is to prove that if $h \geq h_1$ and $\omega \neq \pm\pi/2$, then

$$\mathrm{e}^{\omega \mathrm{i}} E_\omega(h/2) = 2\sum_{k\geq 0} \cos((2k+1)\omega)\tilde{\chi}_k = 2\left(\tilde{\chi}_0 + \sum_{k\geq 1} a_k \tilde{\chi}_k\right)\cos\omega \neq 0.$$

Here, we have used the symmetry $\tilde{\chi}_{-(k+1)} = \tilde{\chi}_k$ and have introduced the notation

$$a_k = a_k(\omega) := \frac{\cos(2k+1)\omega}{\cos\omega}.$$

That is, $a_k(\omega) = T_{2k+1}(\cos\omega)/\cos\omega$, where $T_n(x)$ denotes the Chebyshev polynomial of first kind and degree n defined by relation $T_n(\cos\omega) = \cos n\omega$. Now, using some standard properties of Chebyshev polynomials contained in [1, entries 22.5.29 and 22.14.1], we realize that $\max_{\omega\in\mathbb{T}} |a_k(\omega)| = 2k+1$. Therefore, in order to prove the lemma it suffices to establish that

$$\sum_{k\geq 1} (2k+1)\tilde{\chi}_k(h) < \tilde{\chi}_0(h) \qquad \forall h \geq h_1 := 2\log(20/9). \tag{30}$$

The rest of the proof is devoted to proving this bound. If we work with the multiplicative variable $x = \mathrm{e}^{-h/2} \in (0,1)$, then $\tilde{\chi}_0 = \tilde{\chi}_0(x) = 4(1-x^4)^2/x(1+x^3)(1+x)^3$ and

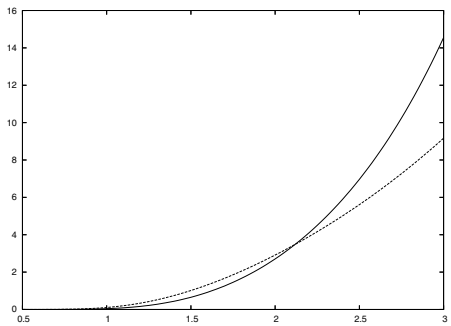
$$\tilde{\chi}_k = \tilde{\chi}_k(x) = \frac{4(1-x^4)^2 x^{4k-2}}{(1+x^{2k+3})(1+x^{2k+1})^2(1+x^{2k-1})} < 4(1-x^4)^2 x^{4k-2}$$

for any $k \geq 1$ and $x \in (0,1)$. In particular,

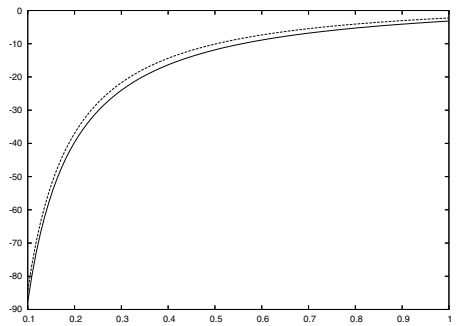
$$\sum_{k\geq 1} (2k+1)\tilde{\chi}_k(x) < 4(1-x^4)^2 \sum_{k\geq 1} (2k+1)x^{4k-2} = 12x^2(1-x^4/3)$$

for all $x \in (0,1)$. Hence, if there exists some $x_1 \in (0,1)$ such that

$$\frac{(1-x^4)^2}{1-x^4/3} =: f(x) \geq g(x) := 3x^3(1+x^3)(1+x)^3 \qquad \forall x \in (0, x_1],$$



(a) $m_0(h)$ and $m_1(h)$ vs $h \in (0.5, 3)$.



(b) $\log m_0(h)$ and $\log m_1(h)$ vs $h \in (0.1, 1)$.

Figure 6. *Graphs of the functions* $m_0(h) = \min\{M_0(\omega, h) : \omega \in \mathbb{T}\}$ *(full curves) and* $m_1(h) = \min\{M_1(\omega, h) : \omega \in \mathbb{T}\}$ *(broken curves) in normal (left) and logarithmic (right) vertical scales.*

then (30) holds for any $h \geq h_1 := 2\log(1/x_1)$. The function $f(x)$ is increasing in the interval $(0, 1)$, whereas $g(x)$ is decreasing in the same interval. On the other hand,

$$f(9/20) = \frac{23543526721}{25250080000} > \frac{465593887647}{512000000000} = g(9/20).$$

Therefore, we can take $x_1 := 9/20$. ∎

Appendix B. Numerical evidence for Conjecture 15. In the proofs of Lemmas 22 and 23, we have shown that the analytic functions $M_j : \mathbb{T} \times \mathbb{R}_+ \to \mathbb{R}$ defined by

$$M_0(\omega, h) = E_\omega(0), \qquad M_1(\omega, h) = \frac{\mathrm{e}^{\omega \mathrm{i}} E_\omega(h/2)}{2\cos\omega}$$

are positive when $h \geq h_0$ and $h \geq h_1$, respectively. We conjecture that, in fact, they are positive everywhere, which is equivalent to Conjecture 15.

Figure 6 provides strong numerical evidence for this conjecture. Concretely, we have numerically checked that the functions

$$m_j : \mathbb{R}_+ \to \mathbb{R}, \qquad m_j(h) = \min\{M_j(\omega, h) : \omega \in \mathbb{T}\}$$

are positive in the range $1/10 \leq h \leq 3$. Greater values of h are already covered by our analytical results. Smaller values of h represent a computational challenge, because the functions $m_j(h)$ are exponentially small in h as $h \to 0^+$; see subfigure 6(b). This is a typical behavior for splitting problems in weakly hyperbolic settings.

The computation of such exponentially small splittings requires the use of a multiple precision arithmetic to mitigate the strong cancellations that take place in such problems. For instance, to compute the functions $m_j(h)$ at $h = 1/10$ it is necessary to work with at least 50 digits.

Acknowledgments. This work was completed while HL was a visitor at the University of Texas at Austin, whose hospitality is gratefully acknowledged. Useful conversations with Jaume Amorós, Amadeu Delshams, Rafael de la Llave, James Meiss, Carles Simó, and Jordi Villanueva are also acknowledged.

REFERENCES

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.

[2] C. Amick, E. S. C. Ching, L. P. Kadanoff, and V. Rom-Kedar, *Beyond all orders: Singular perturbations in a mapping*, J. Nonlinear Sci., 2 (1992), pp. 9–67.

[3] V. I. Arnold and B. A. Khesin, *Topological Methods in Hydrodynamics*, Appl. Math. Sci. 125, Springer-Verlag, New York, 1998.

[4] R. B. Ash, *Complex Variables*, Academic Press, New York, London, 1971.

[5] S. Balasuriya, *Optimal perturbation for enhanced chaotic transport*, Phys. D, 202 (2005), pp. 155–176.

[6] S. Balasuriya, I. Mezić, and C. K. R. T. Jones, *Weak finite-time Melnikov theory and 3D viscous perturbations of Euler flows*, Phys. D, 176 (2003), pp. 82–106.

[7] M. Baldomà and E. Fontich, *Poincaré-Melnikov theory for n-dimensional diffeomorphisms*, Appl. Math. (Warsaw), 25 (1998), pp. 129–152.

[8] I. Baldomà and T. M. Seara, *Breakdown of heteroclinic orbits for some analytic unfoldings of the Hopf-zero singularity*, J. Nonlinear Sci., 16 (2006), pp. 543–582.

[9] T. Bountis, A. Goriely, and M. Kollmann, *A Melnikov vector for N-dimensional mappings*, Phys. Lett. A, 206 (1995), pp. 38–48.

[10] H. W. Broer and G. Vegter, *Subordinate Šil'nikov bifurcations near some singularities of vector fields having low codimension*, Ergodic Theory Dynam. Systems, 4 (1984), pp. 509–525.

[11] H. W. Broer, C. Simó, and R. Vitolo, *The Hopf-saddle-node bifurcation for fixed points of 3D-diffeomorphisms: Analysis of a resonance 'bubble,'* Phys. D, 237 (2008), pp. 1773–1799.

[12] H. W. Broer, C. Simó, and R. Vitolo, *The Hopf-saddle-node bifurcation for fixed points of 3D-diffeomorphisms: The Arnol'd resonance web*, Bull. Belgian Math. Soc. Simon Stevin, in press.

[13] A. Delshams and R. Ramírez-Ros, *Poincaré-Melnikov-Arnold method for analytic planar maps*, Nonlinearity, 9 (1996), pp. 1–26.

[14] A. Delshams and R. Ramírez-Ros, *Melnikov potential for exact symplectic maps*, Comm. Math. Phys., 190 (1997), pp. 213–245.

[15] A. Delshams, Yu. Fedorov, and R. Ramírez-Ros, *Homoclinic billiard orbits inside symmetrically perturbed ellipsoids*, Nonlinearity, 14 (2001), pp. 1141–1195.

[16] R. L. Devaney, *Reversible diffeomorphisms and flows*, Trans. Amer. Math. Soc., 218 (1976), pp. 89–113.

[17] R. W. Easton, *Computing the dependence on a parameter of a family of unstable manifolds: Generalized Melnikov formulas*, Nonlinear Anal., 8 (1984), pp. 1–4.

[18] M. L. Glasser, V. G. Papageorgiou, and T. C. Bountis, *Mel'nikov's function for two-dimensional mappings*, SIAM J. Appl. Math., 49 (1989), pp. 692–703.

[19] P. J. Holmes, *Some remarks on chaotic particle paths in time-periodic, three-dimensional swirling flows*, in Fluids and Plasmas: Geometry and Dynamics, Contemp. Math. 28, AMS, Providence, RI, 1984, pp. 393–404.

[20] Y. T. Lau and J. M. Finn, *Dynamics of a three-dimensional incompressible flow with stagnation points*, Phys. D, 57 (1992), pp. 283–310.

[21] H. E. Lomelí, *Saddle connections and heteroclinic orbits for standard maps*, Nonlinearity, 9 (1996), pp. 649–668.

[22] H. E. Lomelí, *Applications of the Melnikov method to twist maps in higher dimensions using the variational approach*, Ergodic Theory Dynam. Systems, 17 (1997), pp. 445–462.

[23] H. E. Lomelí and J. D. Meiss, *Heteroclinic primary intersections and codimension one Melnikov method for volume-preserving maps*, Chaos, 10 (2000), pp. 109–121.

[24] H. E. Lomelí and J. D. Meiss, *Heteroclinic intersections between invariant circles of volume-preserving maps*, Nonlinearity, 16 (2003), pp. 1573–1595.

[25] H. E. Lomelí, J. D. Meiss, and R. Ramírez-Ros, *Canonical Melnikov theory for diffeomorphisms*, Nonlinearity, 21 (2008), pp. 485–508.

[26] R. S. MacKay, *Transport in 3D volume-preserving flows*, J. Nonlinear Sci., 4 (1994), pp. 329–354.

[27] E. M. McMillan, *A problem in the stability of periodic systems*, in Topics in Modern Physics, E. Brittin and H. Odabasi, eds., Colorado Association University Press, Boulder, CO, 1971, pp. 219–244.

[28] I. Mezić and F. Sotiropoulos, *Ergodic theory and experimental visualization of invariant sets in chaotically advected flows*, Phys. Fluids, 14 (2002), pp. 2235–2243.

[29] I. Mezić and S. Wiggins, *On the integrability and perturbation of three-dimensional fluid flows with symmetry*, J. Nonlinear Sci., 4 (1994), pp. 157–194.
[30] P. Mullowney, K. Julien, and J. D. Meiss, *Blinking rolls: Chaotic advection in a three-dimensional flow with an invariant*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 159–186.
[31] R. Ramírez-Ros, *Exponentially small separatrix splittings and almost invisible homoclinic bifurcations in some billiard tables*, Phys. D, 210 (2005), pp. 149–179.
[32] T. Shinbrot, M. M. Alvarez, J. M. Zalc, and F. J. Muzzio, *Attraction of minute particles to invariant regions of volume preserving flows by transients*, Phys. Rev. Lett., 86 (2001), pp. 1207–1210.
[33] F. Sotiropoulos, Y. Ventikos, and T. C. Lackey, *Chaotic advection in three-dimensional stationary vortex-breakdown bubbles: Šil'nikov's chaos and the devil's staircase*, J. Fluid Mech., 444 (2001), pp. 257–297.
[34] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, Cambridge University Press, Cambridge, UK, 1927.
[35] S. L. Ziglin, *Splitting of separatrices, branching of solutions and nonexistence of an integral in the dynamics of a solid body*, Trans. Moscow Math. Soc., 1 (1982), pp. 283–298.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1558–1581
© 2008 Society for Industrial and Applied Mathematics

Electrical Waves in a One-Dimensional Model of Cardiac Tissue*

Margaret Beck†, Christopher K. R. T. Jones‡, David Schaeffer§, and Martin Wechselberger¶

Abstract. The electrical dynamics in the heart is modeled by a two-component PDE. Using geometric singular perturbation theory, it is shown that a traveling pulse solution, which corresponds to a single heartbeat, exists. One key aspect of the proof involves tracking the solution near a point on the slow manifold that is not normally hyperbolic. This is achieved by desingularizing the vector field using a blow-up technique. This feature is relevant because it distinguishes cardiac impulses from, for example, nerve impulses. Stability of the pulse is also shown, by computing the zeros of the Evans function. Although the spectrum of one of the fast components is only marginally stable, due to essential spectrum that accumulates at the origin, it is shown that the spectrum of the full pulse consists of an isolated eigenvalue at zero and essential spectrum that is bounded away from the imaginary axis. Thus, this model provides an example in a biological application reminiscent of a previously observed mathematical phenomenon: that connecting an unstable—in this case marginally stable—front and back can produce a stable pulse. Finally, remarks are made regarding the existence and stability of spatially periodic pulses, corresponding to successive heartbeats, and their relationship with alternans, irregular action potentials that have been linked with arrhythmia.

Key words. geometric singular perturbation theory, Evans function, blow-up, cardiac model, arrhythmia, alternans

AMS subject classifications. 34E15, 35K57, 37L15, 92C30

DOI. 10.1137/070709980

1. Introduction.

The first model of electrical activity in the cardiac membrane, an adaptation of the classic Hodgkin–Huxley model of neural dynamics, was introduced by Hutter and Noble [HN60] and Noble [Nob62] for the Purkinje fiber. Beeler and Reuter [BR77] introduced the first model for the dynamics of a ventricular myocyte, and more complicated models including additional and new experimental data (see, for example, Luo and Rudy [LR91, LR94]) followed. The main aim of all these models was to describe generic restitution properties of cardiac tissue, as well as to support spiral waves that break up spontaneously. (Restitution refers to the relation between the diastolic interval, the interval between the end of an action

*Received by the editors December 3, 2007; accepted for publication (in revised form) by J. Keener August 11, 2008; published electronically December 10, 2008.
http://www.siam.org/journals/siads/7-4/70998.html

†Department of Mathematics, University of Surrey, Guildford GU2 7XH, UK (M.Beck@surrey.ac.uk). The research of this author was partially supported under NSF grant DMS-0602891.

‡Department of Mathematics, University of North Carolina, Chapel Hill, NC 27599 (ckrtj@email.unc.edu), and Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (C.K.R.T.Jones@warwick.ac.uk). The research of this author was partially supported under NSF grant DMS-0410267.

§Department of Mathematics and Center for Nonlinear and Complex Systems, Box 90320, Duke University, Durham, NC 27708 (dgs@math.duke.edu). The research of this author was partially supported by NSF grant PHY-0549259 and NIH grant 1R01-HL-7283.

¶School of Mathematics & Statistics and Centre for Mathematical Biology, University of Sydney, NSW, 2006, Australia (wm@maths.usyd.edu.au).

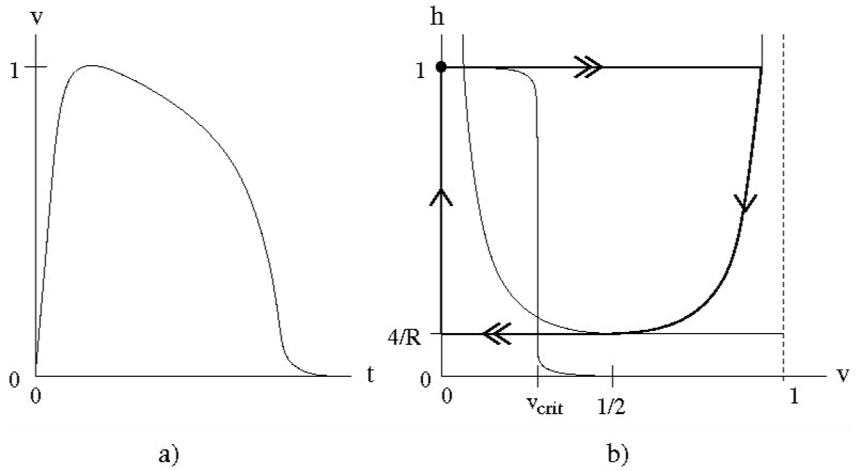


Figure 1. (a) *A sketch of the v component of the pulse.* (b) *A sketch of the nullclines associated with* (1.7) *and the leading order traveling pulse solution (bold) of* (2.1).

potential and the beginning of the next action potential, and the duration of the next action potential pulse.) The length and shape of action potentials in myocardial cells are distinct from those of the Hodgkin–Huxley model. In particular, myocardial cells have substantially prolonged action potentials compared to neural action potentials, which facilitates muscular contraction. See Figure 1(a).

In order to focus on the qualitative aspects of the Hodgkin–Huxley model of membrane activity in the squid giant axon, a simplified version, known as the FitzHugh–Nagumo model, was developed. In the same spirit, Karma [Kar93, Kar94] extracted a model of minimal complexity that reproduces restitution properties of the above-mentioned cardiac cell models. Because these properties are absent in the standard FitzHugh–Nagumo model, it was necessary to introduce a different minimal model.

In this paper, we study a variant of Karma's model that was introduced by Mitchell and Schaeffer [MS03]. In addition, we allow for spatial variation and replace the step functions that appear in their model by smooth functions. The result is

$$\begin{aligned} v_t &= \kappa v_{xx} + I_{in}(v,h) + I_{out}(v) + I_{stim}(t), \\ h_t &= \frac{h_\infty(v) - h}{\tau_h(v)}, \end{aligned} \tag{1.1}$$

where

$$I_{in}(v,h) = \frac{1}{\tau_{in}} h v^2 (1-v), \quad I_{out}(v) = -\frac{1}{\tau_{out}} v,$$

$$h_\infty(v) = \begin{cases} 1 & \text{if } v < v_{crit} - \delta_v, \\ f^h(v) & \text{if } v_{crit} - \delta_v < v < v_{crit} + \delta_v, \\ 0 & \text{if } v_{crit} + \delta_v < v, \end{cases}$$

and

$$\tau_h(v) = \begin{cases} \tau_{open} & \text{if } v < v_{crit} - \delta_v, \\ f^\tau(v) & \text{if } v_{crit} - \delta_v < v < v_{crit} + \delta_v, \\ \tau_{close} & \text{if } v_{crit} + \delta_v < v, \end{cases}$$

where δ_v is sufficiently small, f^h and f^τ are smooth monotonic functions that give continuity of the functions h_∞ and τ_h, and $I_{stim}(t)$ is an external stimulus. In the above equations, $x \in \mathbb{R}$, $t > 0$, $v = v(x,t) \in \mathbb{R}$ corresponds to the membrane potential of myocardial tissue and $h = h(x,t) \in \mathbb{R}$ is a gating variable. Both are dimensionless quantities that can range between zero and one.

The change of membrane potential v is governed by diffusion, through the diffusion coefficient κ, and the ionic currents. The current $I_{in}(v,h)$ denotes a bulk inward current, a combination of all currents that raise the voltage across the membrane—primarily sodium and calcium current. This current is voltage dependent and includes the gating variable h—open when $h = 1$, closed when $h = 0$—that governs the inactivation of the bulk inward current. The current $I_{out}(v)$ denotes a bulk outward current, a combination of all currents that decreases the voltage across the membrane—primarily potassium current. This current is voltage dependent but ungated. The stimulus current I_{stim} is an external current usually applied in brief pulses, either by a pacemaker cell or an experimenter.

The parameters τ_{in} and τ_{out} govern the flow of ions into and out of the cell, respectively, and the parameters τ_{open} and τ_{close} govern the opening and closing rates of the gate h. Based upon physiological information, it is reasonable to assume that

$$\tau_{in},\ \tau_{out} \ll \tau_{open},\ \tau_{close}. \tag{1.2}$$

Therefore, changes in the voltage v occur much faster than changes in the gating variable h, the inactivation of the bulk inward current. We exclude the case where the speed τ_{out} of the bulk outward current is comparable to the speeds of the gating variable h. For a myocardial cell to be able to produce an action potential, it is a necessary to have $\tau_{out}/\tau_{in} =: R$ sufficiently large; i.e., activation must occur before deactivation—the inward current has to be sufficiently faster than the outward current. The exact minimum size for R is dependent on the specific model, and we will see below that for the model we investigate it is $R > 4$. Similar conditions on the gating speeds are not necessary for the existence of action potentials. We remark that, although R is large, it is not asymptotically large with respect to $1/\epsilon$, where ϵ is a small parameter that we define below.

Under these assumptions we will rescale space and time so that

$$\kappa = 1, \quad \tau_{out} = 1, \quad 1/\tau_{in} = R. \tag{1.3}$$

We remark that it is not necessary that $\tau_{open} = \tau_{close}$, but we assume this for convenience. Allowing them to differ, but remain of the same order, would only change the decay/growth rates on the slow manifold and would not qualitatively affect our results. Therefore, we define ϵ via

$$1/\tau_{open} = 1/\tau_{close} = \epsilon. \tag{1.4}$$

We note that similar scalings were used in [MS03]. For typical values of these parameters, see [CS06]. The relevance of different choices of R will be discussed further below. The voltage threshold v_{crit} determines when the gate switches from an opening to a closing state or vice versa, and it is reasonable to assume that $0 < v_{crit} < 1/2$ [MS03, CS06].

In this paper, we will be primarily interested in traveling pulse solutions, corresponding to a single heartbeat, stimulated in the distant past at $x = -\infty$. The reason for this is the following. One particularly interesting biological behavior, found in the ODE version of the model, is a bifurcation in the response to periodic external stimulation as the frequency is increased. For smaller frequencies, the model produces action potentials with constant maximum value in one-to-one correspondence with the stimulus. As the frequency increases, this behavior can destabilize, and alternans, action potentials with beat-to-beat variation in their restitution, can appear. This phenomenon has been linked with ventricular fibrillation and sudden cardiac death [CS06]. In order to eventually understand this phenomenon in spatially dependent models, we first seek to understand the dynamics of a single heartbeat, which is described by a traveling pulse solution to the above model without external stimulus. Therefore, in this paper we will set $I_{stim}(t) \equiv 0$.

Thus, the model we will study throughout is

$$\begin{aligned} v_t &= v_{xx} + Rhv^2(1-v) - v, \\ h_t &= \epsilon g(v,h), \end{aligned} \tag{1.5}$$

where

$$g(v,h) = \begin{cases} 1-h & \text{if } v < v_{crit} - \delta_v, \\ f^h(v) - h & \text{if } v_{crit} - \delta_v < v < v_{crit} + \delta_v, \\ -h & \text{if } v_{crit} + \delta_v < v, \end{cases} \tag{1.6}$$

and we have chosen $f^\tau(v) = 1/\epsilon = \tau_{open} = \tau_{close}$. Notice that $(v,h) \equiv (0,1)$ is a stationary solution of (1.5). We are interested in traveling pulses that are biasymptotic to this stationary solution. In order to study such solutions, we define the moving coordinate $\xi = x + ct$ and analyze the model in the (ξ, t) coordinates:

$$\begin{aligned} v_t &= v_{\xi\xi} - cv_\xi + Rhv^2(1-v) - v, \\ h_t &= -ch_\xi + \epsilon g(v,h). \end{aligned} \tag{1.7}$$

As mentioned above, this model has many similarities with the FitzHugh–Nagumo equation, which was analyzed in [JKL91] and [Jon84] using geometric singular perturbation theory [Fen71, Fen79, Jon94, Kap99]. The above model also contains a small parameter, ϵ, indicating the presence of two separated time-scales on which the dynamics occur, and so we will use similar techniques in our analysis. On the other hand, there are key structural differences manifested predominantly in the inward current term $Rhv^2(1-v)$, which will lead to properties of the associated traveling pulse solution that are distinct from those of the pulse solution to the FitzHugh–Nagumo model. We will point to these differences throughout the paper.

If we consider the ODE associated with spatially independent solutions,

$$\begin{aligned} v_t &= Rhv^2(1-v) - v, \\ h_t &= \epsilon g(v,h), \end{aligned}$$

we see that the v- and h-nullclines are given by the sets $\{v = 0\} \cup \{h = 1/(Rv(1-v))\}$ and $\{g(v,h) = 0\}$, respectively (see Figure 1(b)). This provides intuition for the stationary solutions that one can expect to exist for the PDE (1.7). For the construction of the pulse, the v-nullcline will correspond to the slow manifold. Thus, we expect that the pulse will consist of four pieces: a fast jump in $\{h = 1\}$ from $v = 0$ to $v = v^+$, on the right branch of the v-nullcline; a slow decay on the rightmost branch of the v-nullcline; a second fast jump in $\{h = 4/R\}$ from $v = 1/2$ to $v = 0$; and a slow growth along the leftmost branch of the v-nullcline, back to the point $(v,h) = (0,1)$. We will assume that $R > 4$, which is in accordance with our earlier comment on the size of the ratio τ_{out}/τ_{in} and their relative sizes in (1.3).

The key difference between this pulse and that of the FitzHugh–Nagumo model will be that the second "fast" jump, in the present case, occurs at the knee of the v-nullcline. This point is exactly where the slow manifold of the singularly perturbed system loses normal hyperbolicity. As a direct consequence, the repolarization period of the wave back is much longer than the fast depolarization period of the wave front. (This explains the quotes in *the second "fast" jump*, above.) This is one of the key features of membrane potentials in cardiac tissue.

The outline of the paper is as follows. Section 2 contains the proof of existence of traveling pulses. Section 3 is devoted to the stability analysis of such solutions. Finally, in section 4, we briefly discuss spatially periodic traveling pulses, corresponding to successive heartbeats.

2. Existence of the traveling pulse.

The main result of this section is Theorem 2.1, in which we prove that a traveling pulse, connecting $(v,h) = (0,1)$ at $\xi = \pm\infty$, exists for all $\epsilon \in (0, \epsilon_0)$, where ϵ_0 is sufficiently small.

2.1. Construction of the pulse: $\epsilon = 0$.

The traveling pulse is a stationary solution to the PDE (1.7) and also a solution to the ODE

$$\begin{aligned} v_\xi &= w, \\ w_\xi &= cw - Rhv^2(1-v) + v, \\ h_\xi &= \frac{\epsilon}{c} g(v,h). \end{aligned} \tag{2.1}$$

Notice that v and w are fast variables, while h is a slow variable. By setting $\epsilon = 0$ in the above system, we obtain the reduced fast system:

$$\begin{aligned} v_\xi &= w, \\ w_\xi &= cw - Rhv^2(1-v) + v, \\ h_\xi &= 0. \end{aligned} \tag{2.2}$$

By defining $y = \epsilon\xi$ in (2.1) and setting $\epsilon = 0$, we obtain the reduced slow system:

$$\begin{aligned} 0 &= w, \\ 0 &= cw - Rhv^2(1-v) + v, \\ h_y &= \frac{1}{c} g(v,h). \end{aligned} \tag{2.3}$$

First, consider the reduced slow equation (2.3). The associated leading order slow manifold is given by two pieces:

$$S_l = \{v = w = 0\} \quad \text{and} \quad \tilde{S}_r = \left\{ w = 0,\, h = \frac{1}{Rv(1-v)} \right\}.$$

We will be interested in S_l and the right branch of $\tilde{S}_r$, given by

$$S_r = \left\{ w = 0,\, h = \frac{1}{Rv(1-v)},\, v \geq \frac{1}{2} \right\}. \tag{2.4}$$

The slow dynamics on these manifolds is given by

$$\begin{aligned} h_y &= \frac{1}{c}(1-h) \quad \text{if} \quad (v,w) \in S_l, \\ h_y &= -\frac{1}{c}h \qquad \text{if} \quad (v,w) \in S_r. \end{aligned}$$

Next, consider system (2.2), where, due to the third equation, we may think of h as being a fixed parameter. For $1 \geq h > 4/R$, this equation has three fixed points: $(0,0)$, $(v^+(h),0)$, and $(v^-(h),0)$, where

$$v^{\pm}(h) = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{4}{Rh}}. \tag{2.5}$$

Note that $(0,0) \in S_l$ and $(v^+(h),0) \in S_r$. In addition, this system is equivalent to the traveling wave equation associated with the PDE

$$v_t = v_{\xi\xi} - cv_\xi - Rhv(v - v^+(h))(v - v^-(h)). \tag{2.6}$$

In other words, system (2.2) is the ODE satisfied by stationary solutions of (2.6). Using the change of coordinates $v \to v^+(h)v$, $\xi \to (1/v^+(h)\sqrt{Rh})\xi$, $c \to -(v^+(h)\sqrt{Rh})c$, and $t \to t/[(v^+(h))^2 Rh]$, we see that this equation is just the bistable equation

$$v_t = v_{\xi\xi} + cv_\xi + v(1-v)(v-\mu).$$

In the above equation, $\mu = v^-(h)/v^+(h)$. Traveling wave solutions to the bistable equation are relatively well understood. See, for example, [Xin00]. In particular, for $0 < \mu < 1/2$, there is an exact formula for the unique traveling wave connecting $v = 0$ at $\xi = -\infty$ to $v = 1$ at $\xi = +\infty$. Similarly, for $1/2 < \mu < 1$, there is an exact formula for the unique wave connecting $v = 1$ at $\xi = -\infty$ to $v = 0$ at $\xi = +\infty$. Translating back to the original variables, we obtain the following formulae for the traveling wave solutions of the reduced fast equation:

$$v(\xi) = \frac{v^+(h)}{1 + e^{-\sqrt{\frac{Rh}{2}}v^+(h)\xi}}; \quad c = \sqrt{2Rh}\left(\frac{1}{2}v^+(h) - v^-(h)\right) \quad \text{if} \quad \frac{1}{2}v^+(h) - v^-(h) > 0, \tag{2.7}$$

and

$$v(\xi) = \frac{v^+(h)}{1 + e^{\sqrt{\frac{Rh}{2}}v^+(h)\xi}}; \quad c = -\sqrt{2Rh}\left(\frac{1}{2}v^+(h) - v^-(h)\right) \quad \text{if} \quad \frac{1}{2}v^+(h) - v^-(h) < 0. \tag{2.8}$$

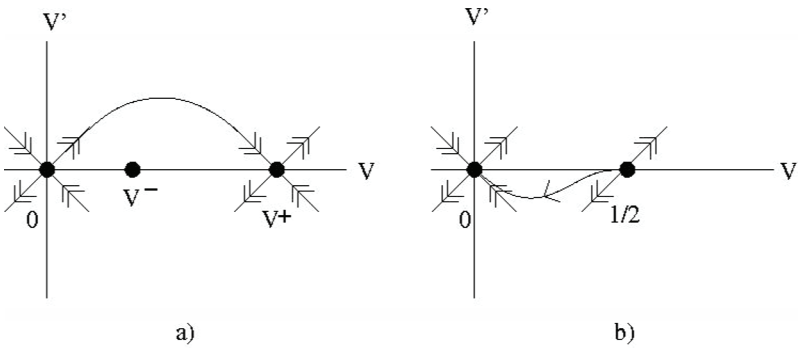


Figure 2. *A sketch of the phase planes for the leading order fast components of the pulse:* (a) *the traveling pulse of* (2.6)*, for* $\frac{1}{2}v^+ - v^- > 0$*;* (b) *the traveling pulse of* (2.9)*, for* $c > \sqrt{2}/2$*, which is asymptotic to the center manifold at* $-\infty$*.*

Notice this implies that, if $\frac{1}{2}v^+(h) - v^-(h) > 0$, then there is a unique heteroclinic connection between $v = 0$ at $-\infty$ and $v = v^+(h)$ at $+\infty$ (see Figure 2(a)). When $\frac{1}{2}v^+(h) - v^-(h) < 0$, then there is a unique connection going in the opposite direction. Using the formulae for $v^\pm(h)$, given in (2.5), one can see that the value of h for which $\frac{1}{2}v^+(h) - v^-(h) = 0$ is given by $\bar{h} = \frac{9}{2R}$.

It may be of interest to note the following. If one considers (2.6) with $c = 0$, then the associated ODE is Hamiltonian with $H(v, v') = (v')^2/2 - v^2/2 + Rhv^3/3 - Rhv^4/4$, where $v' = v_\xi$. For positive wavespeeds $c > 0$, $dH/d\xi = c(v')^2$. As h decreases through $\bar{h}$, the value of $H(v^+(h), 0)$ switches from being positive to negative. As a result, for $h > \bar{h}$, there is no way to have a connection going from $v = v^+(h)$ to $v = 0$, and, for $h < \bar{h}$, there is no way to go in the opposite direction.

The above reduced fast analysis was for $h \in (4/R, 1]$. If $h = 4/R$, however, we have $v^+(h) = v^-(h)$, and the structure of the reduced fast system changes. In that case, (2.2) becomes

$$\begin{aligned} v_\xi &= w, \\ w_\xi &= cw + v(2v-1)^2, \end{aligned}$$

which is equivalent to the traveling wave equation for the PDE

$$v_t = v_{\xi\xi} - cv_\xi - v(2v-1)^2. \tag{2.9}$$

Up to a change of variables, this PDE is exactly the generalized Fisher–KPP (Kolmogorov–Petrovskii–Piskunov) equation of order 2 [Xin00]. For each $c \geq \sqrt{2}/2$, equation (2.9) has a heteroclinic orbit connecting $(1/2, 0)$ at $-\infty$ with $(0, 0)$ at $+\infty$. If $c > \sqrt{2}/2$, then the orbit leaves $(1/2, 0)$ along a center manifold and approaches $(0, 0)$ along its stable manifold (see Figure 2(b)). If $c = \sqrt{2}/2$, known as the critical wavespeed, then the orbit leaves $(1/2, 0)$ along the unstable manifold and approaches $(0, 0)$ along its stable manifold. For more information on critical wavespeeds in this and other related equations, see, for example, [PK06].

Putting the information from the reduced slow and fast dynamics together, we expect that the leading order pulse will consist of four pieces as follows:

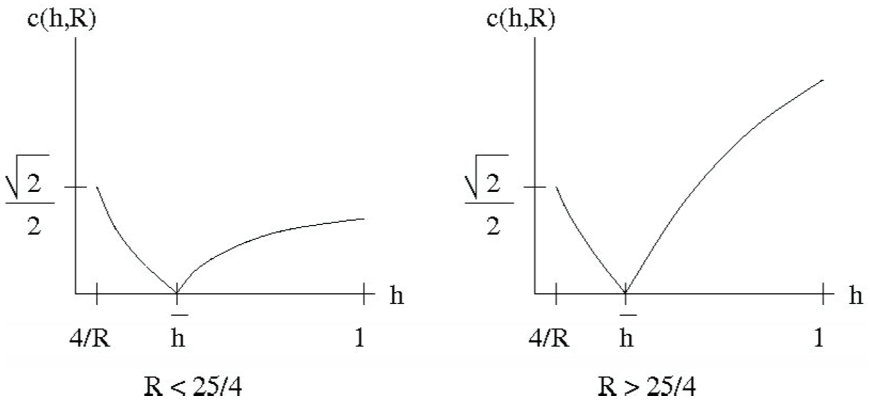


Figure 3. *A sketch of the function $c(R,h)$, as given in (2.7) and (2.8), for fixed R.*

1. a fast jump from $(0,0,1)$ to $(v^+(1),0,1)$, which is given explicitly in (2.7), with $c^* = c(R,1)$;
2. slow decay along S_r to a point $(v^+(h^*),0,h^*)$, where h^* will be determined by the third piece of the pulse;
3. a fast jump from $(v^+(h^*),0,h^*)$ to $(0,0,h^*)$, at the value of $h^* \in [4/R,1)$ such that $c^* = c^*(R,h^*)$; and
4. slow growth along S_l back to $(0,0,1)$.

Notice that the first piece will determine the wavespeed $c^* = c(R,1)$. The value of h^* at which the third piece occurs will be determined by the relation in (2.8) and will be the value of $h \in [4/R,1)$ such that $c(R,1) = c^* = c(R,h^*)$. A sketch of the function $c(R,h)$, for fixed R, is given in Figure 3.

If $4 < R < 25/4$, then there exists an $h^* \in (4/R,\bar{h})$ for which the reduced fast system (2.2) has a heteroclinic orbit of speed c^* connecting $(v+(h^*),0)$ at $-\infty$ with $(0,0)$ at $+\infty$. However, if $R > 25/4$, there is no such $h^* \in (4/R,\bar{h})$.

Recall that, based upon the relative sizes in (1.2), we expect R to be large. As a result, we are interested in the case where $R > 25/4$. Thus, the second fast jump must occur for $h = 4/R$, because for this value of h a connection exists for all $c \geq \sqrt{2}/2$. As a result, the complete, leading order pulse is given by a curve as sketched in Figure 4.

This situation corresponds to the pulse leaving the slow manifold at the knee, which is exactly the point where the manifold ceases to be normally hyperbolic. Therefore, it is not immediately clear what will happen to the pulse for $\epsilon > 0$. Furthermore, because $c(R,1) > \sqrt{2}/2$ (for $R > 25/4$), the second fast jump leaves the knee along the center, rather than the unstable, manifold. This will create a slow transition between the first slow and second fast piece of the pulse, due to the algebraic, rather than exponential, growth away from the slow manifold in the singular limit.

2.2. Persistence of the pulse: $1 \gg \epsilon > 0$. The above analysis tells us what the pulse looks like, to leading order. In order to show that a traveling pulse exists for $0 < \epsilon \ll 1$, we will track the leading order center-unstable manifold of $(0,0,1)$, which we will denote by W^{cu}, forward along the first fast jump, down the slow manifold, and along the second fast jump. In addition, we will track the leading order center-stable manifold of $(0,0,1)$, denoted by W^{cs},

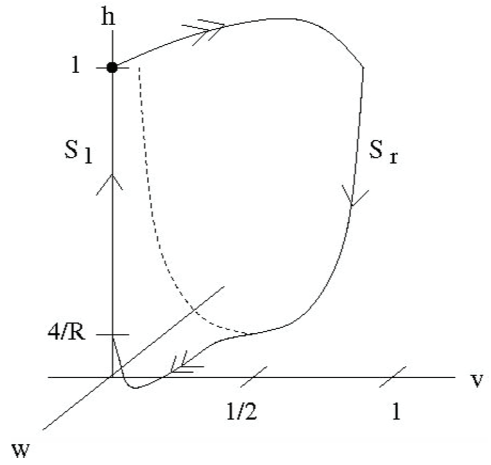


Figure 4. *A sketch of the leading order pulse, consisting of a fast jump from* $(0,0,1)$ *to* $(v^+(1),0,1)$*, slow decay along* S_r*, another fast jump connecting* $(1/2,0,4/R)$ *to* $(0,0,4/R)$ *and leaving at the knee, and slow growth along* S_l*.*

backward down along the leftmost piece of the slow manifold. We will show that, for $\epsilon = 0$, these manifolds intersect transversely, which proves that a traveling pulse must also exist for ϵ positive and sufficiently small.

There are two main difficulties in tracking the center-unstable manifold W^{cu} along the pulse. The first difficulty lies in following W^{cu} along the slow manifold S_r for large times. However, there is a tool, known as the exchange lemma and proved in [JK94], which will allow us to do this. Second, complications can arise due to the lack of normal hyperbolicity at the knee. Analysis in a neighborhood of the knee can be carried out using a geometric desingularization technique known as blow-up [KS01]. Using these techniques, we will prove the following theorem.

Theorem 2.1. *Let* ϕ_0 *denote the formal traveling pulse solution of* (1.5) *for* $\epsilon = 0$*, constructed in section 2.1. More specifically,* ϕ_0 *is defined by the union of the following curves in the* (v, v_ξ, h) *phase space. Define* $S_l = \{v = v_\xi = 0,\ h \in [4/R, 1]\}$*,* $S_r = \{v \in [1/2, v^+(1)],\ v_\xi = 0,\ h = \frac{1}{Rv(1-v)}\}$*,* $F_F = \{(v, v_\xi)$ *given by* (2.7) *for* $\xi \in \mathbb{R},\ h = 1\}$*, and* $F_B = \{(v, v_\xi)$ *the traveling wave solution of* (2.9) *when* $c = c^* > \sqrt{2}/2$ *for* $\xi \in \mathbb{R},\ h = 4/R\}$*. Then*

$$\phi_0 = F_F \cup S_r \cup F_B \cup S_l.$$

If $\epsilon_0 > 0$ *is sufficiently small, then for all* $\epsilon \in (0, \epsilon_0)$ *there exists a traveling pulse solution of* (1.5)*,* $\phi_\epsilon(\xi) = (V(\xi;\epsilon), V_\xi(\xi,\epsilon), H(\xi;\epsilon))$*, satisfying*

$$\lim_{\xi \to \pm\infty} (V(\xi;\epsilon), V_\xi(\xi,\epsilon), H(\xi;\epsilon)) = (0,0,1)$$

and lying as a curve within $\mathcal{O}(\epsilon^{2/3})$ *of the set* ϕ_0*.*

Remark 2.2. *In the statement of the above theorem, the perturbed pulse is parameterized by the spatial variable* ξ*. The leading order pulse,* ϕ_0*, cannot be parameterized in this way. This is because, for* $\epsilon = 0$*, it takes an infinite amount of "time" for the pulse to traverse each piece*

of the leading order orbit in the phase space. This is why the leading order pulse is defined geometrically, in terms of the sets F_F, S_r, F_B, S_l.

Proof. First, we remark that, along various pieces of the leading order pulse, different quantities (i.e., v, w, h, c, ϵ) will determine the key properties of the tracked manifold. Thus, we will consider only the relevant equations for v_ξ, w_ξ, h_ξ, c_ξ, or ϵ_ξ along each of the two fast and two slow components of the pulse. With a slight abuse of notation, we will still refer to the tracked manifold as the center-stable or center-unstable manifold along each piece and hope that it will be clear from the context exactly which variables we are keeping track of at the time.

We begin by tracking W^{cu} along the first fast jump. Consider the reduced fast system, equation (2.2). The plane $\{h = 1\}$ is invariant, and we want to determine how the unstable manifold of $(v, w) = (0, 0)$ intersects the stable manifold of $(v, w) = (v^+(1), 0)$ as we vary the wavespeed c. To do this, we fix $h = 1$ and append to (2.2) the equation $c_\xi = 0$:

$$\begin{aligned} v_\xi &= w, \\ w_\xi &= cw - Rv^2(1 - v) + v, \\ c_\xi &= 0. \end{aligned} \tag{2.10}$$

Based upon the analysis of the bistable equation mentioned above, there is a unique $c^* = c^*(R)$ for which a unique heteroclinic connection between the saddle $(0, 0)$ at $-\infty$ and the saddle $(v^+(1), 0)$ at $+\infty$ exists. We will show that the two-dimensional center-unstable manifold, which is a union of the unstable manifolds of $(v, w) = (0, 0)$ for values of c near c^* and denoted by $W^{cu}(0, 0)$, intersects the two-dimensional center-stable manifold of $(v^+(1), 0)$ (again defined as a union of the stable manifolds for c near c^*), denoted by $W^{cs}(v^+(1), 0)$, and that this intersection is transverse in the c direction. In other words, upon varying c along the fibers within the center-unstable and center-stable manifolds, there is an intersection for a unique value, $c = c^*$, and the manifolds intersect transversely at this point.

One way to track the evolution of two-dimensional manifolds is using two-forms, as in [JKL91]. This essentially allows one to track the evolution of their tangent planes. The vector of one-forms associated with (2.10) is (dv, dw, dc), and its evolution is given by

$$\begin{aligned} dv' &= dw, \\ dw' &= \left(1 - R2v(1 - v) + Rv^2\right) dv + cdw + wdc, \\ dc' &= 0, \end{aligned}$$

where $(\cdot)' = d/d\xi$. The associated two-forms are $P_{vw} = dv \wedge dw$, $P_{vc} = dv \wedge dc$, and $P_{wc} = dw \wedge dc$, with evolution equations

$$\begin{aligned} P'_{vw} &= cP_{vw} + wP_{vc}, \\ P'_{vc} &= P_{wc}, \\ P'_{wc} &= \left(1 - R2v(1 - v) + Rv^2\right) P_{vc} + cP_{wc}. \end{aligned}$$

These equations can be analyzed as in [JKL91], and we restate the details here for convenience. To be precise, we really should think of $\{P_{vw}, P_{vc}, P_{wc}\}$ as the basis for the space of two-forms

in vwc-space and write an arbitrary element as $f_1(v,w,c)P_{vw} + f_2(v,w,c)P_{vc} + f_3(v,w,c)P_{wc}$. It is really the coefficients f_i that we want to determine. The notation in the above system is therefore an abuse of notation: $P_{vw}(\xi)$ denotes the coefficient $f_1(v,w,c)(\xi)$, and so on. As this notation has become somewhat standard (for example, in system (2.1) we are in some sense thinking of $\{v,w,h\}$ as a basis for $\mathbb{R}^3$, particularly when we plot the phase diagram in Figure 4), we use it in the following.

Consider the equation for P_{vw}, and let $P_{vw}^{\pm}$ and $P_{vc}^{\pm}$ be the two forms associated with manifolds $W^{cu}(0,0)$ ($-$, coming from $-\infty$) and $W^{cs}(v^+(1),0)$ ($+$, coming from $+\infty$). We will show that P_{vc}^{+} and P_{vw}^{+} have the same sign, whereas P_{vc}^{-} and P_{vw}^{-} have the opposite sign. This implies that the vectors of two-forms associated with the manifolds, $(P_{vw}^{\pm}, P_{vc}^{\pm}, P_{wc}^{\pm})$, are linearly independent, and hence, that the manifolds intersect transversely.

The manifolds $W^{cu}(0,0)$ and $W^{cs}(v^+(1),0)$ both have the vector field, $(w, cw - Rv^2(1-v) + v, 0)$, as one tangent vector. Denote the other one by $(dv^{\pm}, dw^{\pm}, 1)$, respectively, where we can take $dc = 1$ since $dc' = 0$. This ensures that the two tangent vectors for each manifold are linearly independent. The two-forms for each tangent plane are, up to a positive normalization factor N, given by 2×2 subdeterminants of a 2×3 matrix, whose rows are exactly the above tangent vectors. Thus, we have

$$P_{vc}^{\pm} = N \det \begin{pmatrix} w & 0 \\ dv^{\pm} & 1 \end{pmatrix} = Nw.$$

As a result, the equation for P_{vw} is given by

$$P_{vw}' = cP_{vw} + Nw^2.$$

Since both $W^{cu}(0,0)$ and $W^{cs}(v^+(1),0)$ asymptotically contain a line of fixed points in the c-direction, we know that $P_{vw}^{+} \to 0$ as $\xi \to +\infty$, and $P_{vw}^{-} \to 0$ as $\xi \to -\infty$. Using the above equation, we can then see that $P_{vw}^{-} > 0$ and $P_{vw}^{+} < 0$ along the manifolds. Since $P_{vc}^{\pm} = Nw$ and w is positive along the first fast jump, $P_{vc}^{\pm} > 0$. Thus, the manifolds intersect transversely.

In order to continue tracking W^{cu} along the slow manifold S_r, we will need to use the exchange lemma [JK94]. This lemma tells us how information in the center direction, corresponding to the wavespeed c, at the top of S_r, is exchanged for information in the center direction, corresponding to h, at the bottom of S_r. More specifically, the lemma tells us that, since W^{cu} is transverse to $W^{cs}(v^+(1),0)$, when it leaves a neighborhood of S_r near any point with $1 > h > 4/R$, the tangent plane to W_{cu} will be C^1 $\mathcal{O}(\epsilon)$ close to the plane spanned by the tangent line to S_r in the plane $\{w = 0\}$ and the fast unstable fiber of the point $(v^+(h), 0, h)$. Thus, at a given value of h, the tangent plane to W^{cu} will be spanned, to leading order, by the vectors

$$\left(1, 0, \frac{2v^+(h) - 1}{2(v^+(h))^2(1 - v^+(h))^2}\right), \quad \left(1, \frac{c}{2} + \frac{\sqrt{c^2 + 4(Rhv^+(h) - 2)}}{2}, 0\right). \tag{2.11}$$

The exchange lemma tells us what W^{cu} looks like up to a neighborhood of the knee, when $h = 4/R$. At this point, the slow manifold S_r is not normally hyperbolic, and so we'll need to use blow-up to track the manifold around the knee.

2.2.1. Analysis of the knee. Consider the ODE for the pulse, (2.1), and append to it an equation for ϵ:

$$
\begin{aligned}
v_\xi &= w,\\
w_\xi &= cw - Rhv^2(1-v) + v,\\
h_\xi &= \frac{\epsilon}{c} g(v,h),\\
\epsilon_\xi &= 0.
\end{aligned}
\tag{2.12}
$$

Because the first fast jump selected the wavespeed $c = c(\epsilon)$, where $c(0) = c^*$, we now think of c as being fixed. (For notational convenience we do not explicitly write the epsilon dependence.) The relevant center directions, therefore, are now given by h and also by ϵ.

We are interested in the behavior of this equation near the knee, which corresponds to the fixed point $(1/2, 0, 4/R, 0)$. The Jacobian at this point is given by

$$
J = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & c & -R/8 & 0 \\ 0 & 0 & 0 & -4/(cR) \\ 0 & 0 & 0 & 0 \end{pmatrix}.
$$

This matrix has one positive eigenvalue, $\lambda = c$, with associated eigenvector $(1, c, 0, 0)$. In addition, $\lambda = 0$ is an eigenvalue with algebraic multiplicity three and geometric multiplicity one. The associated eigenvector is $(1, 0, 0, 0)$, and the generalized eigenvectors are $(0, 1, 8c/R, 0)$ and $(0, 0, -8/R, -2c^2)$. In order to do the blow-up, we will need to isolate the nonhyperbolic dynamics, which occur on a three-dimensional center manifold. In a neighborhood of the knee, this manifold can be represented by

$$
\begin{aligned}
w &= F\left((v-1/2), (h-4/R), \epsilon\right)\\
&= \alpha_0(h-4/R) + \alpha_1\epsilon + \beta_0(v-1/2)^2 + \beta_1(h-4/R)^2 + \beta_2\epsilon^2\\
&\quad + \gamma_0(v-1/2)(h-4/R) + \gamma_1(v-1/2)\epsilon + \gamma_2(h-4/R)\epsilon + \mathcal{O}(3),
\end{aligned}
\tag{2.13}
$$

where α_i, β_i, and γ_i are constants. We remark that the center manifold is not unique. However, the analysis that we carry out below is valid up to exponentially small terms, and thus is independent of the choice of the center manifold.

One can explicitly compute the above coefficients. As we will see below, the ones that will be relevant for the blow-up analysis are

$$
\alpha_0 = R/(8c), \quad \beta_0 = -2/c.
$$

Inserting these into (2.12), one sees that the dynamics on the center manifold are given by

$$
\begin{aligned}
v_\xi &= -\frac{2}{c}\left(v - \frac{1}{2}\right)^2 + \frac{R}{8c}\left(h - \frac{4}{R}\right)\\
&\quad + \mathcal{O}\left(\epsilon, \left(h-\frac{4}{R}\right)^2, \epsilon^2, \left(v-\frac{1}{2}\right)\left(h-\frac{4}{R}\right), \epsilon\left(h-\frac{4}{R}\right), \epsilon\left(v-\frac{1}{2}\right)\right),\\
h_\xi &= -\frac{4}{Rc}\epsilon + \mathcal{O}\left(\epsilon\left(h-\frac{4}{R}\right)\right),\\
\epsilon_\xi &= 0.
\end{aligned}
\tag{2.14}
$$

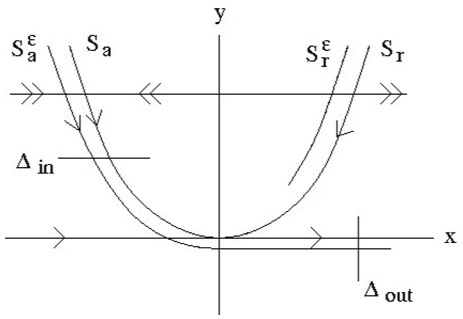


Figure 5. *A sketch of the slow manifolds and sections for the fold, as analyzed in* [KS01].

This system is essentially the normal form for a fold point, given in [KS01]. Their analysis explains why the terms of order $\mathcal{O}\left(\epsilon, (h-\frac{4}{R})^2, \epsilon^2, (v-\frac{1}{2})(h-\frac{4}{R}), \epsilon(h-\frac{4}{R}), \epsilon(v-\frac{1}{2})\right)$ are all indeed higher order. We now collect the results from [KS01] that are relevant for this paper.

In [KS01], the authors analyze systems of the form

$$\begin{aligned} x' &= -y + x^2 + \mathcal{O}(\epsilon, xy, y^2, x^3), \\ y' &= -\epsilon + \mathcal{O}(\epsilon x, \epsilon y, \epsilon^2). \end{aligned} \tag{2.15}$$

This system possesses a slow manifold that, for $\epsilon = 0$, is given by $S = \{(x, y) : y = x^2\}$. It can be divided into the attracting and repelling branches of the parabola, denoted by S_a and S_r, respectively. (To be consistent with the notation in [KS01], we use S_r to denote the repelling branch and hope that it will not be confused with the right branch of the slow manifold, given in (2.4).) Outside a neighborhood of the fold point, $(0, 0)$, these manifolds are normally hyperbolic and, therefore, perturb smoothly to locally invariant manifolds S_a^ϵ and S_r^ϵ, for ϵ positive and sufficiently small (see Figure 5).

The main result of [KS01] describes what happens in a neighborhood of $(0, 0)$ for $0 < \epsilon \ll 1$, and it can be explained as follows. Let $\Delta_{in} = \{(x, \rho^2),\ x \in I\}$ and $\Delta_{out} = \{(\rho, y),\ y \in \mathbb{R}\}$, where $I \subset \mathbb{R}$ is a suitable interval. Let $\pi : \Delta_{in} \to \Delta_{out}$ be the transition map associated with the flow of (2.15) (see Figure 5).

Proposition 2.3 (see [KS01]). *There exists an $\epsilon_0 > 0$ such that, for $\epsilon \in (0, \epsilon_0)$, the following hold:*

1. *The manifold S_a^ϵ passes through Δ_{out} at a point $(\rho, y(\epsilon))$, where $y(\epsilon) = \mathcal{O}(\epsilon^{2/3})$. In particular, $y(\epsilon) = -\Omega_0 \epsilon^{2/3} + o(\epsilon^{2/3})$, where $\Omega_0 > 0$ is known explicitly.*
2. *The transition map π is a contraction with contraction rate $\mathcal{O}(e^{-k/\epsilon})$, where k is a positive constant.*

The first item of this proposition tells us how to track the manifold W^{cu} around the knee. The second tells us that the resulting analysis will be independent of the choice of center manifold. We note the scaling $\epsilon^{2/3}$ is consistent with the higher order asymptotics of the restitution curve, computed in [MS03, CS06].

As mentioned above, for any $h \in (4/R, 1)$, the two-dimensional manifold W^{cu} is spanned by the one-dimensional fast unstable direction of the points $(v, w) = (v^+(h), 0)$ and the tangent line in the plane $\{w = 0\}$ to the one-dimensional slow manifold, given by $h = 1/(Rv(1-v))$.

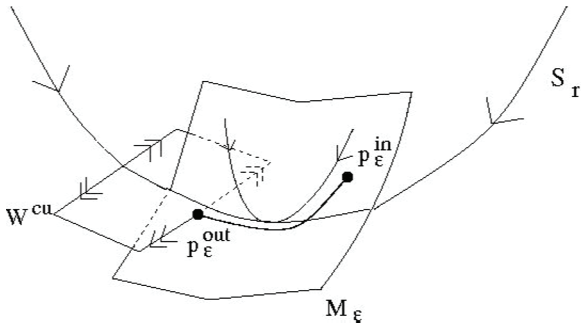


Figure 6. *A schematic diagram, in a neighborhood of the knee, of the tracked manifold W^{cu}. It is guided by the trajectory through the points p_ϵ^{in} and p_ϵ^{out}, which lies on the center manifold M_ϵ.*

As W^{cu} enters a neighborhood of the knee, it will intersect the center manifold of the knee, given in (2.13). We will denote this center manifold by $M_\epsilon = \{w = F(v, h, \epsilon)\}$ and its leading order version by $M_0 = \{w = F(v, h, 0)\}$.

In order to see how the dynamics on the center manifold at the knee will guide W^{cu}, define the following objects (with a slight abuse of notation, we will reuse the quantities $\Delta_{in,out}$ and ρ from the fold analysis):

$$\begin{aligned}
\Delta_{in} &= \left\{(v, w, h) : h = \frac{4}{R} + \rho^2\right\},\\
\Delta_{out} &= \left\{(v, w, h) : v = \frac{1}{2} - \rho\right\},\\
I^{u,in} &= W^{cu} \cap \Delta_{in},\\
p_\epsilon^{in} &= I^{u,in} \cap M_\epsilon,\\
I^{u,out} &= W^{cu} \cap \Delta_{out},\\
p_\epsilon^{out} &= I^{u,out} \cap M_\epsilon.
\end{aligned}$$

Both Δ_{in} and Δ_{out} are two-dimensional objects, as is W^{cu}. $I^{u,in}$ and $I^{u,out}$, however, are one-dimensional, and $p_\epsilon^{in,out}$ are points (at least for fixed ϵ, when M_ϵ is two-dimensional).

Because the point p_ϵ^{in} lies on the center manifold M_ϵ, its evolution will be governed by the dynamics of the fold. The trajectory through p_ϵ^{in} will follow the attracting critical manifold S_a^ϵ around the knee and exit the neighborhood of the knee on the section Δ_{out} at the point p_ϵ^{out}. Since p_ϵ^{in} is also in W^{cu}, the tracked manifold will follow its trajectory around the knee. Upon exiting a neighborhood of the knee, W^{cu} will be spanned by the one-dimensional direction of the flow at p_ϵ^{out} and the one-dimensional fast fibers of M_ϵ at p_ϵ^{out}. See Figure 6.

The important information to take away from the analysis at the knee is the following. When $\epsilon = 0$, the manifold W^{cu} is tangent at the knee to the plane spanned by the vectors $(1, 0, 0)$ and $(1, c, 0)$, which are obtained by setting $v = 1/2$ in (2.11), and has height $h = 4/R$. We need to know how this picture changes for ϵ positive and small. Based on the above analysis, we see that, on leaving a neighborhood of the knee, the tangent plane to W^{cu} is spanned by a vector tangent to the fast unstable fibers of M_ϵ and a vector in the direction of the flow. We know that the unstable fibers of the center manifold M_ϵ perturb smoothly, so their

h component will be $\mathcal{O}(\epsilon)$. The other tangent vector is given by the flow, whose h component is $\mathcal{O}(\epsilon)$. In addition, the height of the tangent plane to W^{cu} is given by $h = 4/R - \mathcal{O}(\epsilon^{2/3})$. Therefore, upon leaving a neighborhood of the knee, the perturbed manifold W^{cu} will be $\mathcal{C}^1$ $\mathcal{O}(\epsilon^{2/3})$ close to the unperturbed one. We remark that the C^1 aspect of the perturbation follows from the fact that the center manifold itself is normally hyperbolic, and therefore its unstable fibers perturb smoothly. It is these fibers that make up W^{cu}. Proposition 2.3 then ensures that the base points of these fibers can change by no more than $\mathcal{O}(\epsilon^{2/3})$.

2.2.2. Completion of the existence proof. We will now follow W^{cu} along the back and W^{cs} backward down the left branch of the slow manifold and show that they intersect transversely. This will complete the existence argument.

We have shown above that when W^{cu} leaves a neighborhood of the knee, its tangent plane is close to a plane that is parallel to the vw plane. One can see directly from (2.1) that, to leading order, any plane parallel to the vw plane is invariant. There is only a finite amount of time between when W^{cu} leaves a neighborhood of the knee and when it enters a neighborhood of the point $(0, 0, 4/R)$. Therefore, by choosing ϵ sufficiently small, we can make the tangent plane to W^{cu}, upon entering this neighborhood, as close to a plane parallel to the vw plane as we like.

The center stable manifold of $(0, 0, 1)$, W^{cs}, consists of the union of the stable manifolds of the saddle points $(0, 0)$ for $h \in [4/R, 1]$. As a result, it will transversely intersect any plane which is parallel to the vw plane. This implies that W^{cs} intersects W^{cu} transversely, which completes the proof. ■

3. Stability of the pulse. The goal of this section will be to prove the following theorem on the linear stability of the traveling pulse.

Theorem 3.1. *The traveling pulse solution, constructed in section 2, is spectrally stable. In other words, the operator obtained by linearizing around the wave (see (3.1)) has no spectrum in* $\{\mathrm{Re}(\lambda) \geq 0\}$ *except for an isolated eigenvalue at the origin of geometric and algebraic multiplicity one.*

We remark that linear stability of the wave follows from a spectral mapping theorem for the strongly continuous semigroup generated by the linear operator in (3.1). In addition, because the zero eigenvalue is isolated, standard arguments, such as invariant manifold theory, can be used to show that the traveling pulse is nonlinearly stable as well [BJ89].

The outline of the proof is as follows. First, we will show that the essential spectrum is bounded to the left of the imaginary axis, although the bound will be dependent on ϵ. We will then construct the Evans function [Eva73, AGJ90] associated with the full problem, for $\epsilon > 0$, followed by the Evans functions associated with the reduced fast pieces along the front and back of the pulse. This will be done in section 3.1. We will then show that eigenvalues of the full Evans function are determined by those of the reduced problems, and use information about the stability of the pulses in the bistable and generalized KPP equations to determine the stability of the pulse. This will be done in section 3.2.

There are two key elements of the argument. First, the spectrum of the reduced fast problem along the back, corresponding to the generalized KPP equation, contains essential spectrum that accumulates at the origin. Thus, one must be careful in analyzing the associated reduced Evans function. We will appeal to the results of [WXY06], in which the Evans function

for algebraically decaying solutions to such equations was analyzed. Second, the loss of normal hyperbolicity at the knee could potentially allow the zeros of the full Evans function to be different than those of the reduced problem. Using techniques similar to those of the existence argument above, we will see that this is not the case.

3.1. Essential spectrum and definition of the Evans function. We will denote the persistent traveling pulse solution to (1.7) by $(V(\xi), H(\xi))$. (Note that this solution is also dependent on the parameter ϵ, although we will suppress this in our notation.) In order to consider the stability of the pulse, assume that solutions to (1.7) have the form $(v,h)(\xi,t) = (V(\xi),H(\xi)) + (p,r)(\xi,t)$. The linearized flow for the new coordinates (p,r), which represent the perturbation of the wave, is then given by

$$\begin{aligned} p_t &= p_{\xi\xi} - cp_\xi + \left(2RHV(1-V) - RHV^2 - 1\right)p + RV^2(1-V)r, \\ r_t &= -cr_\xi + \epsilon g_v(V,H)p - \epsilon r. \end{aligned} \tag{3.1}$$

The associated eigenvalue problem, when written as a first order system, is given by

$$\begin{aligned} p_\xi &= q, \\ q_\xi &= \left(\lambda - 2RHV(1-V) + RHV^2 + 1\right)p + cq - RV^2(1-V)r, \\ r_\xi &= \epsilon\frac{g_v(V,H)}{c}p - \frac{(\lambda+\epsilon)}{c}r. \end{aligned} \tag{3.2}$$

We can write this eigenvalue problem using matrix notation

$$\frac{d}{d\xi}\begin{pmatrix} p \\ q \\ r \end{pmatrix} = A(\xi,\lambda)\begin{pmatrix} p \\ q \\ r \end{pmatrix}, \tag{3.3}$$

where

$$A(\xi,\lambda) = \begin{pmatrix} 0 & 1 & 0 \\ \lambda - 2RHV(1-V) + RHV^2 + 1 & c & -RV^2(1-V) \\ \epsilon\frac{g_v(V,H)}{c} & 0 & -\frac{(\lambda+\epsilon)}{c} \end{pmatrix} \tag{3.4}$$

and the ξ dependence is through the underlying wave, $(V,H) = (V(\xi),H(\xi))$.

3.1.1. Location of the essential spectrum. The essential spectrum is determined by the asymptotic limits of the matrix A, defined in (3.4), which are given by

$$\begin{aligned} A^\infty(\lambda) &= \lim_{\xi\to\pm\infty} A(\xi,\lambda) \\ &= \begin{pmatrix} 0 & 1 & 0 \\ \lambda+1 & c & 0 \\ 0 & 0 & -\frac{(\lambda+\epsilon)}{c} \end{pmatrix}. \end{aligned} \tag{3.5}$$

The boundary of the essential spectrum is given by all values of λ for which this matrix has purely imaginary eigenvalues [Hen81]. This set is given by

$$\{-\epsilon - ick : k \in \mathbb{R}\} \cup \{-k^2 - ick - 1 : k \in \mathbb{R}\}.$$

In addition, by [Hen81], the essential spectrum lies to the left of the above boundary: $\sigma_{ess} \subset \{\lambda \in \mathbb{R} : \mathrm{Re}(\lambda) \leq -\epsilon\}$. In the limit $\epsilon \to 0$, the essential spectrum will approach the imaginary axis. However, this will not affect the definition of the Evans function, as given below. We define $\Omega = \Omega(\epsilon)$ to be the open region in the complex plane that lies to the right of the essential spectrum, containing the right half plane.

3.1.2. Definition of the Evans function. The eigenvalues of the asymptotic matrix $A^{\infty}(\lambda)$, defined in (3.5), are given by

$$\nu_0 = -\frac{(\lambda+\epsilon)}{c}, \quad \nu^{\pm} = \frac{c}{2} \pm \frac{1}{2}\sqrt{c^2 + 4(\lambda+1)}.$$

For $\lambda \in \Omega$, $\nu^+(\lambda)$ is the unique eigenvalue with positive real part. One can check that there exists a b, independent of ϵ, such that, for all ϵ sufficiently small and all $\lambda \in \tilde{\Omega} := \{\lambda : \mathrm{Re}(\lambda) > -b\}$, $\nu^+(\lambda)$ remains the unique eigenvalue with largest real part. (Note that $\tilde{\Omega}$ is slightly larger than, but contains, Ω, for ϵ sufficiently small.) This remains true even when the real part of ν_0 changes sign. The eigenvector associated with ν^+ is given by $X^+ = (1, \nu^+, 0)^t$. As a result, there exists a unique solution to (3.3), $\zeta(\xi, \lambda)$, such that

$$\lim_{\xi \to -\infty} \zeta(\xi, \lambda) e^{-\nu^+(\lambda)\xi} = X^+(\lambda). \tag{3.6}$$

Consider now the associated adjoint problem,

$$\frac{d}{d\xi} Z = -\bar{A}^T(\lambda, \xi) Z. \tag{3.7}$$

Similarly, for $\lambda \in \tilde{\Omega}$, there is a unique eigenvalue of the associated asymptotic matrix with smallest real part. This eigenvalue is given by $\mu^-(\lambda) = -\bar{\nu}^+(\lambda)$, and its associated eigenvector is $Y^- = (\mu^- - c, 1, 0)^t$. For $\lambda \in \tilde{\Omega}$, we have that $\mathrm{Re}(\mu^-) < 0$. There exists a unique solution to (3.7), $\eta(\xi, \lambda)$, such that

$$\lim_{\xi \to +\infty} \eta(\xi, \lambda) e^{-\mu^-(\lambda)\xi} = Y^-(\lambda). \tag{3.8}$$

The Evans function [AGJ90] is then defined by

$$D(\lambda) = \zeta(\xi, \lambda) \cdot \eta(\xi, \lambda). \tag{3.9}$$

As in [Jon84], $D(\lambda)$ can be shown to be analytic on $\tilde{\Omega}$. This is because ν^+ and μ^- are the unique eigenvalues with largest and smallest real part, respectively, in $\tilde{\Omega}$, uniformly in ϵ, even as λ crosses into the essential spectrum. (For additional information on the extension of the Evans function into the essential spectrum, see, for example, [KS98, GZ98].) For $\lambda \in \Omega$, the zeros of $D(\lambda)$, along with their multiplicities, correspond to the eigenfunctions of the linear operator in (3.1). For $\lambda \in \tilde{\Omega}$, however, this relationship does not necessarily hold. Technically, the actual Evans function, $D(\lambda)$, is defined only on Ω, and the Evans function we consider is an analytic extension of it (sometimes denoted by $\tilde{D}(\lambda)$) into $\tilde{\Omega}$. We will not emphasize this distinction here.

3.2. Locating zeros of $D(\lambda)$ in the right half plane. The goal of this section is to show that the only zero of $D(\lambda)$ with $\mathrm{Re}(\lambda) \geq 0$ is $\lambda = 0$, and that its geometric and algebraic multiplicity is one. The argument will be similar to that in [Jon84], although we will have to do a bit of extra work to account for the loss of hyperbolicity in the knee of the existence construction.

The outline of the argument is as follows. First, we will construct the reduced Evans functions associated with the fast flow along the front and back of the pulse and show that the only zero in the closed right half plane is at the origin and associated with the front. The back does not contribute a zero there because of its algebraic decay at $-\infty$. This is a key difference between this problem and the stability of the pulse in the FitzHugh–Nagumo system. Next, it will be shown that any zeros of the full Evans function must be close to zeros of the reduced Evans functions. Thus, because there exists a unique zero of the reduced Evans functions in the right half plane, and we know it remains at zero for the full system due to translation invariance of the underlying wave, the wave must be spectrally stable. Note that it is not necessary to compute the derivative of the Evans function at the origin, as it was for the FitzHugh–Nagumo model, since there is no second zero to locate.

3.2.1. The reduced Evans functions. In this section, we consider the reduced Evans functions for the fast equation along the front and back of the pulse. We will show that both reduced Evans functions have no zeros with $\{\mathrm{Re}(\lambda) \geq 0\} \setminus \{0\}$, and that there is only one zero at $\lambda = 0$, associated with the front.

Recall that, along the front, to leading order we have $h \equiv 1$. As a result, the reduced, fast PDE that governs the dynamics of the front is given by

$$v_t = v_{\xi\xi} - cv_\xi + Rv^2(1-v) - v. \tag{3.10}$$

As mentioned above, this is just the bistable equation, and the front is the heteroclinic connection between 0 at $-\infty$ and $v^+(1)$ at $+\infty$. Up to rescaling, this is also the equation that governs the dynamics of the front of the pulse for the FitzHugh–Nagumo equation. Its Evans function was analyzed in detail in [Jon84], and we summarize those results in the following proposition.

Proposition 3.2 (see [Jon84]). *Let $D_F(\lambda)$ denote the reduced Evans function that one obtains from the stability analysis of the heteroclinic front of (3.10). Then D_F is analytic in $\tilde{\Omega}$ and*

1. $D_F(0) = 0$,
2. $D_F(\lambda) \neq 0$ *for all* $\lambda \in \tilde{\Omega} \setminus \{0\}$.

In [Jon84], it was also shown that $\frac{d}{d\lambda} D_F(\lambda)|_{\lambda=0} > 0$, although we will not need that fact here. However, we do need that the derivative of the Evans function at 0 is nonzero, for simplicity of the eigenvalue. We remark that the b in the definition of $\tilde{\Omega}$ may need to be chosen slightly smaller than above in order for this proposition to hold.

Next, consider the reduced PDE for the back,

$$v_t = v_{\xi\xi} - cv_\xi - v(2v-1)^2, \tag{3.11}$$

where c is the wavespeed that was selected in the analysis of the front (see (2.7) for $h = 1$). As mentioned above, this is the generalized Fisher–KPP equation of order 2. The back is a

heteroclinic connection between $1/2$ at $-\infty$ and 0 at $+\infty$. It is asymptotic to a stable manifold at $+\infty$, but a center manifold at $-\infty$, where it decays only algebraically. In addition, the essential spectrum of the associated linearized operator is contained in a parabolic region of the left half plane that touches the imaginary axis at the origin. As a result, the stability analysis of the back and construction of the associated reduced Evans function is a bit more subtle. However, this analysis has been carried out in [WXY06], and we collect the relevant results.

Proposition 3.3 (see [WXY06]). *Let $D_B(\lambda)$ denote the reduced Evans function that one obtains from the stability analysis of the heteroclinic solution of* (3.11). *Then*

1. *D_B is analytic in $\tilde{\Omega}$,*
2. *$D_B(\lambda) \neq 0$ for all $\lambda \in \tilde{\Omega}$.*

Again, it may be necessary to take b in the definition of $\tilde{\Omega}$ to be slightly smaller than above.

It may be surprising that the derivative of the heteroclinic solution does not lead to a zero of the reduced Evans function at the origin. This is because the Evans function in [WXY06] is constructed using the strong unstable eigenvalue at $-\infty$. Since the wave decays only algebraically there, it is asymptotic to the weak unstable direction and, therefore, does not contribute a zero. See Theorem 2.1 and Lemma 4.1 of [WXY06] for more details.

3.2.2. Approximation of eigenvalues. We now show that any zero of the Evans function $D(\lambda)$ in $\tilde{\Omega}$ must be near the unique zero of $D_F(\lambda)$ and $D_B(\lambda)$ in that region, $\lambda = 0$. Let B_δ denote the ball of radius δ at 0, where δ is chosen sufficiently small so that $B_\delta \subset \tilde{\Omega}$, and let $G = \tilde{\Omega} \setminus B_\delta$.

Lemma 3.4. *$D(\lambda) \neq 0$ for all $\lambda \in G$.*

Proof. First, we note that we need only consider a bounded region within G, say $\hat{G} = \{\lambda \in G : |\lambda| < M\}$ for some fixed M that is independent of ϵ. This can be shown using a scaling argument [San02].

Fix $\lambda \in \hat{G}$ and consider $\zeta(\xi, \lambda)$ and $\eta(\xi, \lambda)$, defined in (3.6) and (3.8). We will track ζ around the pulse until ξ is sufficiently large and then show that it cannot be orthogonal to η.

The main idea is to show that, in the absence of an eigenfunction for the reduced fast flows, the strong unstable direction is always an attractor for the evolution of ζ. As a result, if $\lambda \in \hat{G}$, then the unstable direction completely determines the evolution of ζ, and we can use it to track the evolution around the pulse and show it cannot be orthogonal to η. This argument will be similar to that of [Jon84], and so some details will be omitted.

In order to follow $\zeta(\xi, \lambda)$ around the pulse, we must track its evolution according to (3.2). To make this more precise, one must couple the traveling wave system (2.1) to (3.2) and track the combined solution $((V(\xi), W(\xi), H(\xi)), (p(\xi), q(\xi), r(\xi)))$. However, for our purposes it is sufficient to refer only to system (3.2). In addition, in [Jon84] the wave was parameterized by $\theta \in S^1$, but we will not discuss this further here.

First, we track ζ along the fast front. If $v_F(\xi)$ and $w_F(\xi)$ denote the leading order fast pulse along the front, then the eigenvalue problem associated with (3.10) is

$$\begin{aligned} v_\xi &= w, \\ w_\xi &= [\lambda - 2Rv_F(1 - v_F) + Rv_F^2 + 1]v + cw. \end{aligned} \tag{3.12}$$

The reduced Evans function along the front is defined by $D_F(\lambda) = \zeta_F(\xi, \lambda) \cdot \eta_F(\xi, \lambda)$, where ζ and η are defined analogously to (3.6) and (3.8).

To leading order, $\zeta(\xi, \lambda) = (\zeta_F(\xi, \lambda), 0)$ and $\eta(\xi, \lambda) = (\eta_F(\xi, \lambda), 0)$. In other words, the components of the full Evans function are just the inclusions in $\mathbb{R}^3$ of their reduced counterparts. Therefore, to leading order, the reduced equation can be used to track the evolution of ζ along the front of the wave.

We are really only interested in the direction of the vector $\zeta(\xi, \lambda)$, which can be studied using the projectivized version of (3.2). To that end, define $(a, b) := \pi(p, q, r) = (q/p, r/p)$. Thus, $\pi : \{(p, q, r) \in \mathbb{C}^3 : p \neq 0\} \to \mathbb{CP}^2$. The evolution of (a, b) is governed, to leading order, by

$$\begin{aligned} a' &= [\lambda - 2RHV(1 - V) + RHV^2 + 1] + ca - RV^2(1 - V)b - a^2, \\ b' &= -\frac{\lambda}{c}b - ab. \end{aligned} \tag{3.13}$$

The eigenvector associated with the unique largest eigenvalue of an ODE always corresponds to a stable fixed point of the corresponding projectivized system. If we fix λ and consider the "frozen" version of (3.13), where ξ is fixed on the right-hand side, then the projectivized version of that eigenvector is an attractor for the system. Taking a union over all ξ in some interval, for example, we would obtain an attractor for the evolution of (3.13) in that interval. (Note: this fact relies on the compactness of $\hat{G}$.) If we then let $\hat{\zeta}$ denote the projectivized version of ζ, that attractor would govern the evolution of $\hat{\zeta}$ along that interval.

We could similarly construct the projectivized version of the reduced equation (3.12). The unstable direction is an attractor for $\hat{\zeta}_F$ as it follows the front. For $\lambda \in \hat{G}$, we know that the reduced system does not have an eigenvalue, and so $\hat{\zeta}_F$ will approach the unstable direction as $\xi \to +\infty$. Thus, when the pulse enters a neighborhood of the invariant slow manifold, $\hat{\zeta}(\xi, \lambda)$ will be equal, to leading order, to the direction of the unstable fast fiber of the manifold.

Next, we need to track $\hat{\zeta}$ down along the slow manifold until the pulse enters a neighborhood of the knee. Using the projectivized equations associated with the fast flow for any fixed $h \in (4/R, 1]$, one sees that the union of the unstable eigenvectors associated with the fixed points $(v^+(h), 0)$ of (2.2) is an attractor. As a result, as $\hat{\zeta}$ follows the first slow piece of the pulse, it will remain close to the direction of the unstable fibers of the slow manifold, until it enters a neighborhood of the knee.

Now we must track $\hat{\zeta}$ around the knee. For $\lambda \in \hat{G}$, the presence of the knee does not pose any additional complications, and the analysis follows as in [Jon84]. This is because, when $(V, H) \sim (1/2, 4/R)$ at the knee, (3.2) is hyperbolic for $\lambda \neq 0$. There is still a unique largest eigenvalue, and $\hat{\zeta}$ will be attracted to the direction of its corresponding eigenvector.

The analysis of the evolution of $\hat{\zeta}$ along the back is similar to that of the front, above. The key fact is that, when $\hat{\zeta}$ emerges from a neighborhood of the knee, it will be close to the strong unstable direction. Since this direction corresponds to the unstable fiber of the (normally hyperbolic) center manifold, it will perturb smoothly. In other words, it is a tangent vector living in the tangent space of the traveling wave (which is $\mathcal{O}(\epsilon^{2/3})$ close to the leading order wave), and its direction is C^1 $\mathcal{O}(\epsilon)$ close to the leading order strong unstable direction. Since it is this direction that is used to construct the Evans function in [WXY06], we can

then use those results to conclude that any zero of the full Evans must be near a zero of the reduced Evans function.

Because the reduced equation, the linearization of the generalized KPP equation, does not have a eigenvalue in $\hat{G}$, $\hat{\zeta}$ must be $\mathcal{O}(\epsilon)$ close to the direction of the unstable fibers as the pulse enters a neighborhood of the slow invariant manifold. Thus, we can follow it up along the slow manifold S_l and conclude that it is not orthogonal to $\hat{\eta}(\xi,\lambda)$ when the pulse enters a neighborhood of the fixed point $(0,0,1)$. Because the $\hat{\zeta}$ and $\hat{\eta}$ determine the directions of the vectors ζ and η, this proves that ζ and η are not orthogonal, as well. ■

3.2.3. Winding number calculation. We know from the previous section that any potential unstable eigenvalues must lie in B_δ, the ball of radius δ at the origin. Let $K = \partial B_\delta$ denote the boundary of that ball, and choose δ sufficiently small that zero is the only eigenvalue of either reduced fast system that is contained in B_δ. We will compute the winding number of $D(\lambda)$ along K and show that it is one. This will show that there is exactly one zero of the Evans function in $\tilde{\Omega}$, which implies that there is exactly one eigenvalue of geometric and algebraic multiplicity one in $\tilde{\Omega}$ [AGJ90]. Since we know there must exist an eigenvalue at $\lambda = 0$, it is the only one. This will complete the proof of Theorem 3.1. We remark that, again, this argument is similar to that of [Jon84], and so we do not include all of the details here. As above, we must check that the presence of the knee does not affect the winding number.

Note that an analytic extension of the Evans function is defined for all $\lambda \in B_\delta \cup K$, uniformly in ϵ, and that it is nonzero on K. Any zero inside B_δ necessarily corresponds to an eigenvalue only if it is in $\Omega \cup \{0\}$.

In the previous section, it was shown that the projectivized equations can be used to track $\hat{\zeta}$ around the pulse. If we take an element $(q/p, r/p) \in \mathbb{CP}^2$, then we can associate it with an element in $\mathbb{C}^3$ using $\pi^{-1}(q/p, r/p) = (1, q/p, r/p)$, which is just a normalized version of the vector (p,q,r). When computing the winding number we will need not only the direction of the vector, but its amplitude as well. One can directly check that, for any ξ such that $p(\xi,\lambda) \neq 0$,

$$\zeta(\xi,\lambda) = p(\xi,\lambda)[\pi^{-1}(\hat{\zeta})](\xi,\lambda).$$

The Evans function is independent of ξ, and so we can evaluate the expression on the right-hand side of (3.9) at any value of ξ we choose. It is convenient to pick some sufficiently large value of ξ, denoted by T_4. One can then show that $W(D(K)) = W(p(T_4, K))$. The proof of this fact follows closely that in [Jon84], and so we do not repeat it here.

Let T_0 be the value of ξ for which the underlying wave exits a neighborhood of $(0,1)$, T_1 be the value at which it enters a neighborhood of $(v^+(1),1)$, T_2 be the value at which it exits a neighborhood of $(1/2, 4/R)$, and T_3 be the value at which it enters a neighborhood of $(0, 4/R)$. Similar to the previous section, we will track the evolution of $p(\xi,\lambda)$ around the underlying pulse and evaluate the winding numbers $W(p(T_i, K))$ for $i = 0, \ldots, 4$, showing

1. $W(p(T_0, K)) = 0$,
2. $W(p(T_1, K)) = 1$,
3. $W(p(T_2, K)) = 1$,
4. $W(p(T_3, K)) = 1$,
5. $W(p(T_4, K)) = 1$.

Note that, in [Jon84], it was shown $p(T_i, K) \neq 0$ for $i = 0, 4$. In other words, as p moves along the pulse, the corresponding winding number increases by one only as it moves along the front, due to the eigenvalue of the reduced system there. Along the rest of the wave, it remains constant.

The proof essentially follows from the results in [Jon84]. The only thing one needs to check, due to the presence of the knee, is that $W(p(T_2, K)) = 1$. This would follow if $p(\xi, \lambda) \neq 0$ for all $\xi \in [T_1, T_2]$, uniformly for λ near zero. This is because one could then construct a homotopy between $p(T_1, K)$ and $p(T_2, K)$ to show that their winding numbers are equal.

Consider the projectivized system (3.13) near the knee, i.e., for $(V, H) \sim (1/2, 4/R)$,

$$a' = \lambda + ca - \frac{R}{8}b - a^2,$$
$$b' = -\frac{\lambda}{c}b - ab.$$

The three fixed points of this system are $(a^+(\lambda), 0)$, $(a^-(\lambda), 0)$, and $(-\lambda/c, -8\lambda^2/Rc)$, where $a^\pm = (c \pm \sqrt{c^2 + 4\lambda})/2$. These correspond to the unstable, stable, and center directions of the slow manifold, respectively. As $\lambda \to 0$, $(a^-, 0)$ and $(-\lambda/c, -8\lambda^2/Rc)$ coincide, but the remaining fixed point remains separate, uniformly in λ. It is this direction that defines the attractor that $\hat{\zeta}$ follows as it moves along the wave. Since these fast unstable directions always point along vectors with nonzero p component, this shows that $p(\xi, \lambda) \neq 0$ for all $\xi \in [T_1, T_2]$, uniformly for λ near zero, just as in [Jon84].

In other words, the evolution of $\hat{\zeta}$, and hence the winding number, is determined primarily by the behavior of the strong unstable direction to which it is attracted. Since we have shown that this direction remains unique despite the nonhyperbolicity at the knee, the winding number calculation in essentially the same as in [Jon84].

By continuing to follow ζ along the pulse, we arrive at $W(p(T_4, K)) = W(D(K)) = 1$. This proves that there is only one zero of the Evans function in $\tilde{\Omega}$, and since we know there exists a zero at the origin, it must be the only one. This concludes the proof of the spectral stability of the wave.

4. Spatially periodic waves. We now briefly remark on the existence and biological relevance of spatially periodic waves, which will be the subject of future work.

Recall from (2.7) and (2.8) that the fast jumps exist for explicitly computable values of the wavespeed. If we can find a c_{per} and an $h_{per} \in (4/R, 1)$ such that a fast jump exists at h_{per} from $v = 0$ at $\xi = -\infty$ to $v = v^+(h_{per})$ at $\xi = +\infty$, and $c_{per} = c(h_{per}, R) > \sqrt{2}/2$, then we would be able to construct a leading order periodic solution consisting of the following four pieces:

1. a fast jump from $(0, 0, h_{per})$ to $(v^+(h_{per}), 0, v^+(h_{per}))$, which is given explicitly in (2.7), with $c_{per} = c(R, h_{per})$,
2. slow decay along S_r,
3. a fast jump from $(1/2, 0, 4/R)$ to $(0, 0, 4/R)$, with $c_{per} = c(R, h_{per})$,
4. slow growth along S_l back to $(0, 0, h_{per})$.

Note we require that $c_{per} > \sqrt{2}/2$, so that the second fast jump must occur at the knee along the center manifold. If $c_{per} \leq \sqrt{2}/2$, then the second fast jump would leave the slow manifold

along an unstable manifold, which is not what happens in cardiac dynamics, as discussed in section 1.

In Figure 3, we see that such an orbit is possible only if $R > 25/4$. Otherwise, the first fast jump will necessarily occur for a value of the wavespeed less than $\sqrt{2}/2$. If $R > 25/4$, then a first fast jump with sufficiently large wavespeed exists for $h > 25/4R$. As a result, there exists at leading order a family of periodic orbits, one for each $h_{per} \in (25/4R, 1)$.

Regarding the investigation of the persistence of this family, for $\epsilon > 0$, we expect that a combination of the techniques used to prove the existence of a family of spatially periodic solutions to the FitzHugh–Nagumo equation [Car77] and the blow-up, used in this paper, will be applicable. To investigate stability, it is possible that the theory developed in [Gar97, SS01] will apply. In those papers, the authors investigated the linear stability of families of periodic waves of reaction diffusion equations that are close to a homoclinic orbit—at least for sufficiently large period—and, in particular, applied their results to the family of periodic solutions of the FitzHugh–Nagumo equation.

Periodic solutions are of biological interest because they represent a beating heart. In cardiac models, the stability of these periodic waves cannot persist indefinitely as the period is shortened. Even the ODE model suffers a period-doubling bifurcation as the pacing period is shortened. The bifurcation in the PDE context is more complicated than simple period doubling—see, for example, [EK02, EK06]. We plan to study this bifurcation in a subsequent publication.

Acknowledgments. M. B. wishes to thank Björn Sandstede for helpful discussions on the extension of the Evans function into the essential spectrum, and Steve Schecter for bringing [WXY06] to her attention. M. B., C. J., and M. W. would like to thank the Mathematical Sciences Research Institute in Berkeley, CA, for their hospitality during the spring of 2007, while part of this work was completed.

REFERENCES

[AGJ90] J. Alexander, R. Gardner, and C. Jones, *A topological invariant arising in the stability analysis of traveling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[BJ89] P. W. Bates and C. K. R. T. Jones, *Invariant manifolds for semilinear partial differential equations*, Dynamics Reported, 2 (1989), pp. 1–38.

[BR77] G. W. Beeler and H. Reuter, *Reconstruction of the action potential of ventricular myocardial fibres*, J. Physiol., 268 (1977), pp. 177–210.

[Car77] G. A. Carpenter, *A geometric approach to singular perturbation problems with applications to nerve impulse equations*, J. Differential Equations, 23 (1977), pp. 335–367.

[CS06] J. W. Cain and D. G. Schaeffer, *Two-term asymptotic approximation of a cardiac restitution curve*, SIAM Rev., 48 (2006), pp. 537–546.

[EK02] B. Echebarria and A. Karma, *Instability and spatiotemporal dynamics of alternans in paced cardiac tissue*, Phys. Rev. Lett., 88 (2002), paper 208101.

[EK06] B. Echebarria and A. Karma, *Amplitude equation approach to spatiotemporal dynamics of cardiac alternans*, Phys. Rev. E, 76 (2007), paper 051911.

[Eva73] J. W. Evans, *Nerve axon equations.* III. *Stability of the nerve impulse*, Indiana Univ. Math. J., 22 (1972/73), pp. 577–593.

[Fen71] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–226.

[Fen79] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[Gar97] R. A. Gardner, *Spectral analysis of long wavelength periodic waves and applications*, J. Reine Angew. Math., 491 (1997), pp. 149–181.

[GZ98] R. A. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.

[Hen81] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1981.

[HN60] O. F. Hutter and D. Noble, *Rectifying properties of heart muscle*, Nature, 188 (1960), pp. 495–497.

[JK94] C. K. R. T. Jones and N. Kopell, *Tracking invariant manifolds with differential forms in singularly perturbed systems*, J. Differential Equations, 108 (1994), pp. 64–88.

[JKL91] C. K. R. T. Jones, N. Kopell, and R. Langer, *Construction of the FitzHugh-Nagumo pulse using differential forms*, in Patterns and Dynamics in Reactive Media (Minneapolis, MN, 1989), IMA Vol. Math. Appl. 37, Springer, New York, 1991, pp. 101–115.

[Jon84] C. K. R. T. Jones, *Stability of the traveling wave solution of the FitzHugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[Jon94] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems, Lecture Notes in Math. 1609, R. Johnson, ed., Springer-Verlag, Berlin, 1994, pp. 44–118.

[Kap99] T. J. Kaper, *An introduction to geometric methods and dynamical systems theory for singular perturbation problems*, in Proc. Sympos. Appl. Math. 56, AMS, Providence, RI, 1999, pp. 85–131.

[Kar93] A. Karma, *Spiral breakup in model equations of action potential propagation in cardiac tissue*, Phys. Rev. Lett., 71 (1993), pp. 1103–1106.

[Kar94] A. Karma, *Electrical alternans and spiral wave breakup in cardiac tissue*, Chaos, 4 (1994), pp. 461–472.

[KS98] T. Kapitula and B. Sandstede, *Stability of bright solitary-wave solutions to perturbed nonlinear Schrodinger equations*, Phys. D, 124 (1998), pp. 58–103.

[KS01] M. Krupa and P. Szmolyan, *Extending geometric singular perturbation theory to nonhyperbolic points—Fold and canard points in two dimensions*, SIAM J. Math. Anal., 33 (2001), pp. 286–314.

[LR91] C. H. Luo and Y. Rudy, *A model of the ventricular cardiac action potential. Depolarization, repolarization, and their interaction*, Circ. Res., 68 (1991), pp. 1501–1526.

[LR94] C. H. Luo and Y. Rudy, *A dynamic model of the cardiac ventricular action potential.* I. *Simulations of ionic currents and concentration changes*, Circ. Res., 74 (1994), pp. 1071–1096.

[MS03] C. C. Mitchell and D. G. Schaeffer, *A two-current model for the dynamics of cardiac membrane*, Bull. Math. Biol., 65 (2003), pp. 767–793.

[Nob62] D. Noble, *A modification of the Hodgkin-Huxley equations applicable to Purkinje fibre action and pace-maker potentials*, J. Physiol., 160 (1962), pp. 317–352.

[PK06] N. Popović and T. J. Kaper, *Rigorous asymptotic expansions for critical wave speeds in a family of scalar reaction-diffusion equations*, J. Dynam. Differential Equations, 18 (2006), pp. 103–139.

[San02] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems, Vol. 2, North-Holland, Amsterdam, 2002, pp. 983–1055.

[SS01] B. Sandstede and A. Scheel, *On the stability of periodic traveling waves with large spatial period*, J. Differential Equations, 172 (2001), pp. 134–188.

[WXY06] Y. Wu, X. Xing, and Q. Ye, *Stability of travelling waves with algebraic decay for n-degree Fisher-type equations*, Discrete Contin. Dynam. Systems, 16 (2006), pp. 47–66.

[Xin00] J. Xin, *Front propagation in heterogeneous media*, SIAM Rev., 42 (2000), pp. 161–230.

SIAM J. Applied Dynamical Systems
Vol. 7, No. 4, pp. 1582–1611
© 2008 Society for Industrial and Applied Mathematics

Canard Induced Mixed-Mode Oscillations in a Medial Entorhinal Cortex Layer II Stellate Cell Model*

Horacio G. Rotstein†, Martin Wechselberger‡, and Nancy Kopell§

Abstract. Stellate cells (SCs) of the medial entorhinal cortex (layer II) display mixed-mode oscillatory activity, subthreshold oscillations (small-amplitude) interspersed with spikes (large amplitude), at theta frequencies (8–12 Hz). In this paper we study the mechanism of generation of such patterns in an SC biophysical (conductance-based) model. In particular, we show that the mechanism is based on the three-dimensional canard phenomenon and that the subthreshold oscillatory phenomenon is intrinsically nonlinear, involving the participation of both components (fast and slow) of a hyperpolarization-activated current in addition to the voltage and a persistent sodium current. We discuss some consequences of this mechanism for the SC intrinsic dynamics as well as for the interaction between SCs and external inhibitory inputs.

Key words. theto rhythm, reduction-of-dimensions, mixed-mode oscillations, canards, folded node

AMS subject classifications. 34D15, 34C26, 92C20

DOI. 10.1137/070699093

1. Introduction.

The entorhinal cortex (EC) is the interface between the neocortex and the hippocampus [1], and it plays a very important role in orchestrating the flow of information between these two areas of the brain. Neocortical information flows to the hippocampus, to be processed, through the superficial layers (II and III) of the EC. The spiny stellate cells (SCs) are the most abundant principal cell type in layer II of the medial EC [1, 2]. These cells give rise to the main afferent fiber system to the hippocampus. In addition, in layer II of the EC grid cells are putative SCs [3] (and see references therein). Grid cells are principal neurons that exhibit multiple phase fields arranged in hexagonal patterns [4, 5, 6]. Their recent discovery implies that the EC contains a neural map of the spatial environment which is then transmitted to the hippocampus.

In vitro electrophysiological investigations have shown that, when depolarized, SCs develop small-amplitude rhythmic subthreshold membrane potential oscillations (STOs) at theta frequencies (8–12 Hz). If the membrane potential is depolarized further, then SCs fire action potentials at the peak of the STOs but not necessarily at every STO's cycle [7]. The amplitudes of STOs and spikes differ roughly in an order of magnitude. We refer to the resulting temporal patterns (combination of STOs and spikes) as mixed-mode oscillations (MMOs).

*Received by the editors August 2, 2007; accepted for publication (in revised form) by B. Ermentrout July 6, 2008; published electronically December 17, 2008.
http://www.siam.org/journals/siads/7-4/69909.html
†Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102 (horacio@oak.njit.edu). This author's work was partially supported by NSF grant DMS-0817241.
‡School of Mathematics and Statistics and Centre for Mathematical Biology, University of Sydney, NSW, 2006 Australia (wm@maths.usyd.edu.au).
§Department of Mathematics and Center for Biodynamics, Boston University, Boston, MA 02215 (nk@math.bu.edu). This author's work was partially supported by NSF grant DMS-0717670.

These are a distinctive property of SCs in layer II of the medial EC [8, 9, 10], and they can also be found in in vivo electrophysiological studies [11].

Theta frequency STOs and MMOs in SCs are intrinsic single-cell phenomena [8] and have been shown to result from the interaction between two currents: a persistent sodium current (I_p) and a hyperpolarization activated current (I_h) [9]. Mathematical (conductance-based) models, incorporating I_p and I_h in addition to the spiking currents (transient sodium, delayed rectifier potassium, and leak), have been used to reproduce, via simulations, several aspects of the SC dynamics [12, 13, 14]. However, the mechanistic aspects of the generation of STOs and MMOs are still not fully understood.

The goal of this paper is to uncover the mechanism of generation of STOs and MMOs in the biophysical SC model proposed in [12], and to identify the key parameters controlling the transition among the various types of MMO patterns. We use analytical and computational techniques to show that, as hypothesized in [10], the generation of STOs and the onset of spikes in this model is governed by the three-dimensional canard phenomenon [15, 16, 17]. Qualitatively different mechanisms have been proposed to explain the generation of MMOs in other models. These are break-up of an invariant torus [18], break-up (loss) of stability of a Shilnikov homoclinic orbit [19, 20], and subcritical Hopf-homoclinic bifurcation [21, 22]. See also [17, 23] and other articles in the focus issue on MMOs introduced in [24] for a detailed discussion.

In section 2 we provide some biophysical and mathematical background related to the generation of MMOs in SCs. We briefly describe the key experimental findings and the biophysical SC model that we use in this paper. This model is a three-dimensional (3D) reduction [10] of the 7D model presented in [12]. The former, which we will refer to as the SC model, is a good approximation to the latter in the subthreshold regime where STOs and the onset of spikes are observed [10], thus allowing the investigation of the mechanisms of generation of STOs and MMOs. In addition, we explain some basic aspects on the canard phenomenon.

In section 3 we describe the (dimensional) SC model and nondimensionalize it to uncover the multiple time-scale nature of the model. In particular, we show that the membrane potential evolves on a much faster time scale than the h-current gating variables (r_f and r_s). Although the former is faster than the latter, there is no significant time scale separation between the two gates compared with the time scale separation introduced by the fast voltage dynamics. Therefore both gating dynamics of the h-current are considered as slow. For the remainder of this paper we use this dimensionless SC model. However, the results will be presented in terms of both the dimensional and dimensionless values of the relevant parameters. We also present the result of our simulations using this SC model showing MMO patterns and their corresponding phase-space diagrams.

In section 4 we analyze the mechanism of generation of MMOs in the SC model using numerical and analytical techniques. We show that for the relevant (biophysically plausible) parameters the SC model can be put into the analytic and geometric 3D canard framework described in [16] for the generation of small-amplitude oscillations (see also [17]). We describe this framework using notation tailored to the model. In addition, we describe the return mechanism necessary to bring trajectories back to the subthreshold regime after they escape it towards the spiking one.

Our approach provides a geometric framework to qualitatively understand and predict the dynamic properties of the resulting MMO patterns. In particular, it allows us to study the dependence of these patterns on the relevant parameters: the I_h and I_p maximal conductances, the applied DC (constant) current, and the initial conditions of the participating variables in the subthreshold regime. These initial conditions reflect the reset properties of I_h after a spike has occurred. In addition, following the "canard approach," we explain how inhibitory pulses applied at different times after a spike has occurred may suppress some of the STOs of the unperturbed cell and advance the timing of the next spike. This type of calculation is the first step in the computation of spike-time response curves [12, 25, 26], which are used in the study of synchronization properties of small neural networks. We discuss our results and their implications for the understanding of SC dynamics in section 5.

2. Background.

2.1. Biophysics of subthreshold and mixed-mode oscillations in stellate cells. Voltage changes in single (isolated) neurons are the result of the flow of ionic currents into and out of the cells. Typically, three currents are involved in the generation of spikes: a transient sodium current (I_{Na}), a delayed-rectifier potassium current (I_K), and a leak current (I_L) [27]. We refer to them as the standard spiking currents. Spikes are usually initiated by the activation of I_{Na} and terminated by its inactivation followed by the activation of I_K. Additional (nonstandard or nonspiking) currents may be present and play various different roles in neural dynamics. Two nonstandard currents have been implicated in the pacemaking of single-cell rhythmicity at theta frequencies: a persistent sodium current (I_p) and an h-current (I_h) [7, 8, 9, 13, 28, 29, 30, 31, 32] (see also references therein). The former constitutes a depolarization-activated fast inward current that precisely tracks voltage changes and provides the main drive for the depolarizing phase of the STOs. The latter is a hyperpolarization-activated (noninactivating) current with slow activation kinetics, and it provides a delayed feedback effect that promotes resonance [33].

2.2. MMOs in a biophysically plausible SC model. In recent work, Rotstein et al. [10] showed that the biophysical (conductance-based) SC model presented in [12] displays STOs and MMOs, and they initiated a mechanistic study of these phenomena using computational tools and dynamical systems ideas. In [10], reduction-of-dimension techniques were used to reveal a three-dimensional reduced model that is a good approximation to the "full" seven-dimensional SC model in the subthreshold regime where STOs and the onset of spikes occur. This reduced model describes the evolution of the membrane potential V (mV) and the two (fast and slow) h-current gating variables r_f and r_s (dimensionless). The latter describe the opening/closing of the h-current ion channels. In [10] it was found that both I_{Na} and I_K can be neglected in the subthreshold regime where STOs and MMOs are generated, and that the persistent sodium gating variable p has fast dynamics, so the adiabatic approximation can be made; i.e., p can be well approximated by its corresponding voltage-dependent activation curve (see section 3). The resulting equations are presented in section 3, in formulas (1)–(3). As mentioned above, these equations describe the generation of STOs and the onset of spikes, which occurs in the subthreshold regime, but they do not describe the spike dynamics and the early recovery from spiking, which belong to a different regime (where I_{Na} and I_K are the

main active currents) [10]. If one is not interested in the spike details, the dynamics of the SC can be approximately described by (1)–(3) supplemented with an "artificial spike," operating in a much shorter time scale and reaching a peak of about 50 mV. This model has been called the nonlinear artificially spiking (NAS) SC model [10], a class of models that includes the generalized integrate-and-fire models (see [10, 34, 35] for details). For simplicity, in the remainder of this paper we will refer to it as the SC model. When working with this (NAS) SC model, one has to give appropriate threshold (V_{th}) and reset (V_{rst}) values. The former indicates that the trajectory has reached the spiking regime. The latter is the voltage value after a spike has occurred and represents the initial condition in the subthreshold regime. Note that, differently from other types of NAS models, V_{th} is not part of the mechanism of generation of action potentials which, as shown here and in [10], result from the dynamics of the so-called canard structure.

2.3. The canard phenomenon. Canards [36] were first studied in two-dimensional (2D) relaxation oscillators [37, 38, 39, 40], particularly in the van der Pol oscillator. There, the nature of the classical canard phenomenon is the transition from a small-amplitude oscillatory state created in a Hopf bifurcation to a large-amplitude relaxation oscillatory state within an exponentially small range of a control parameter. This transition, also called canard explosion, occurs through a sequence of canard cycles which can be asymptotically stable and is hard to observe in an experiment because of sensitivity to the control parameter. This is well known in the chemical literature, where a canard explosion is classified as a hard transition [41, 42]. Therefore, 2D slow-fast systems display either STOs or large-amplitude oscillations but no MMOs. However, MMOs are possible by the addition of noise [43, 44].

Deterministic 3D slow-fast systems with one fast and two slow variables can produce MMOs [10, 16, 17, 23, 45, 46, 47, 48, 49] (see also the articles in the focus issue on MMOs introduced in [24]). One way to explain these patterns is based on a generalized canard phenomenon, because a special class of canards in three dimensions called canards of folded node (or folded saddle-node) type can be responsible for small-amplitude oscillations [16, 46]. A good intuition for MMOs is that a system moves dynamically from a small-amplitude oscillatory state to a relaxation oscillatory state, and the feature of the large relaxation oscillation is to bring the system back to the basin of attraction of the small-amplitude oscillatory state. A detailed explanation of this generalized canard phenomenon is given in section 4.

3. The model.

3.1. Dimensional formulation. The dimensional equations are

$$C\frac{dV}{dt} = I_{app} - G_L\,(V - E_L) - G_p\,p_\infty(V)\,(V - E_{Na}) - G_h\,(c_f\,r_f + c_s\,r_s)\,(V - E_h), \tag{1}$$

$$\frac{dr_f}{dt} = \frac{r_{f,\infty}(V) - r_f}{\tau_{r_f}(V)}, \tag{2}$$

$$\frac{dr_s}{dt} = \frac{r_{s,\infty}(V) - r_s}{\tau_{r_s}(V)}, \tag{3}$$

where V is the membrane potential (mV), C is the membrane capacitance (μF/cm^2), I_{app} is the applied bias (DC) current (μA/cm^2), $I_L = G_L\,(V - E_L)$, $I_p = G_p\,p_\infty(V)\,(V - E_{Na})$, and

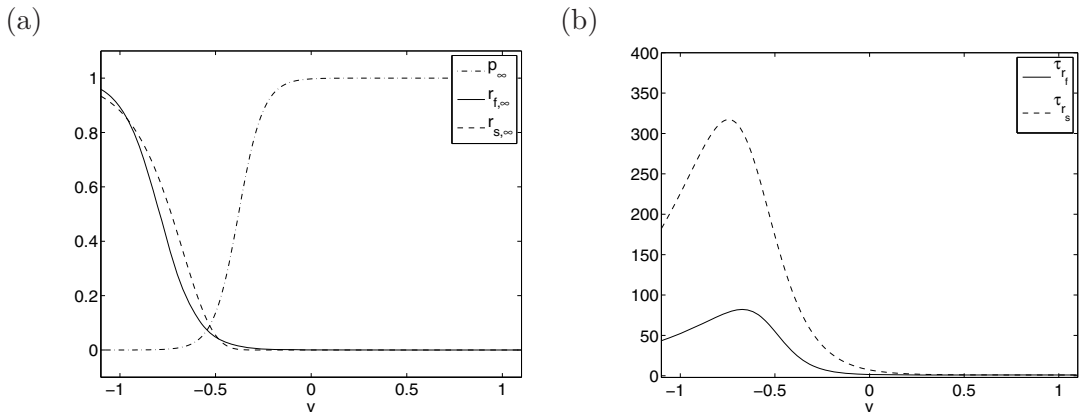


Figure 1. *SC model (1)–(3). (a) Activation/inactivation curves: $p_\infty(v)$, $r_{f,\infty}(v)$, and $r_{s,\infty}(v)$. (b) Voltage-dependent time scales: $\tau_{r_f}(v)$ and $\tau_{r_s}(v)$. The dimensionless voltage v is the result of the rescaling $v = V/K_v$ with $K_v = 100\,mV$.*

$I_h = G_h\,(\,c_f\,r_f + c_s\,r_s\,)\,(\,V - E_h\,)$ [10]. The parameters G_X and E_X $(X = L, p, Na, h)$ are the maximal conductances ($\mathrm{mS/cm^2}$) and reversal potentials (mV), respectively. The units of time are ms. The variables r_f and r_s are the h-current fast and slow gating variables, and the parameters c_f and c_s represent the fraction of the total h-current corresponding to its fast and slow components, respectively. Unless stated otherwise, we will use the following values for the parameters [10, 12]: $E_{Na} = 55$, $E_L = -65$, $E_h = -20$, $G_L = 0.5$, $G_p = 0.5$, $C = 1$, $c_f = 0.65$, and $c_s = 0.35$. The functions $r_{f,\infty}(V)$, $r_{s,\infty}(V)$, and $p_\infty(V)$ are the voltage-dependent activation/inactivation curves, and the functions $\tau_{r_f}(V)$ and $\tau_{r_s}(V)$ are the voltage-dependent time scales. They are given by $r_{f,\infty}(V) = 1/(1 + e^{(V+79.2)/9.78})$, $r_{s,\infty}(V) = 1/(1 + e^{(V+2.83)/15.9})^{58}$, $p_\infty(V) = 1/(1 + e^{-(V+38)/6.5})$, $\tau_{r_f}(V) = 0.51/(e^{(V-1.7)/10} + e^{-(V+340)/52}) + 1$, and $\tau_{r_s}(V) = 5.6/(e^{(V-1.7)/14} + e^{-(V+260)/43}) + 1$. The graphs of these functions are shown in Figure 1.

3.2. Initial and threshold conditions in the subthreshold regime. The initial conditions in the subthreshold regime are given by the reset values of the participating variables after a spike has occurred. For r_f and r_s these reset values can be derived from the 7D SC model [10]. More specifically, during a spike, V increases above zero to a value $V \sim 50\,\mathrm{mV}$. For these values of V, $r_{f,\infty}(V) \sim 0$ and $r_{s,\infty}(V) \sim 0$ (see Figure 1(a)). In addition, for these high values of V, both $\tau_{r_f}(V)$ and $\tau_{r_s}(V)$ are very small (see Figure 1(b)). Therefore, both r_f and r_s quickly decrease to values close to $r_f \sim r_s \sim 0$. The reset value of $V \sim -80\,\mathrm{mV}$ is estimated from numerical simulations of the 7D SC model [10]. Unless stated otherwise, we take $(V, r_f, r_s) = (-80, 0, 0)$ as the initial conditions of system (1)–(3), and we reset the trajectory to these values after each spike has occurred.

Since action potentials in this model are initiated at $V \sim -50\,\mathrm{mV}$ (see, e.g., Figure 2) we may set the voltage threshold value V_{th} for this event to any value $V > -50\,\mathrm{mV}$. Here we choose a value $V_{th} = -40\,\mathrm{mV}$, which is well above the initiation value. We emphasize that the spike results from the dynamics of the SC model, and, consequently, V_{th} only indicates that a spike has occurred and is not a component of the mechanism of spike generation [10].

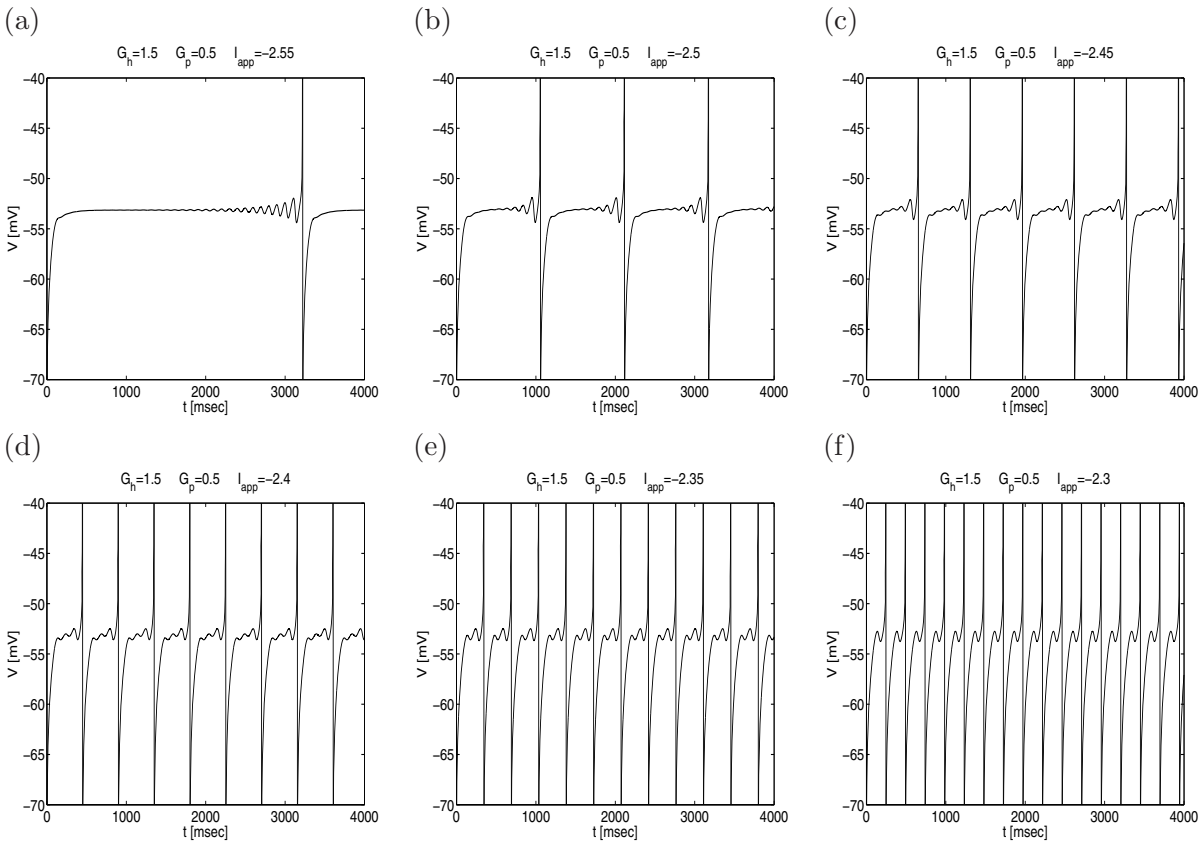


Figure 2. *MMO patterns for the dimensional SC model (1)–(3) with $V_{th} = -40\,mV$ and $V_{rst} = -80\,mV$. The number of subthreshold oscillations per spike decreases with increasing values of I_{app}. Note that V_{th} only indicates that a spike occurs and is not part of the mechanism of spike generation.*

3.3. MMOs in the dimensional model. In Figure 2 we illustrate various MMO patterns generated by the SC model. The voltage traces correspond to $G_h = 1.5$, $G_p = 0.5$, and a sequence of increasing values of I_{app}. We observe that the ratio of subthreshold oscillations to spikes decreases for increasing values of I_{app}. For values of I_{app} below and above these corresponding to Figures 2(a) and (e), the SC becomes silent and fully spiking, respectively (no MMOs). We will use the notation 1^s to indicate that an MMO pattern has a number s of STOs per spike.

Figure 3 shows the 3D phase space corresponding to the voltage traces presented in Figure 2(d) and (b). The V-nullsurface of the SC model is shown as well as the corresponding trajectories of MMO patterns. The trajectories move rapidly from their initial points towards the lower branch of the V-nullsurface and then along it towards the fold-curve (curve of knees of the V-nullsurface). Once the trajectories reach the vicinity of the fold-curve, they start to move almost parallel to the fold-curve and rotate, generating STOs. Finally, the trajectories move rapidly in the direction of increasing values of V, eventually initiating a spike by activating I_{Na}. (The spiking dynamics belongs in a different regime and is not described by this reduced SC model.)

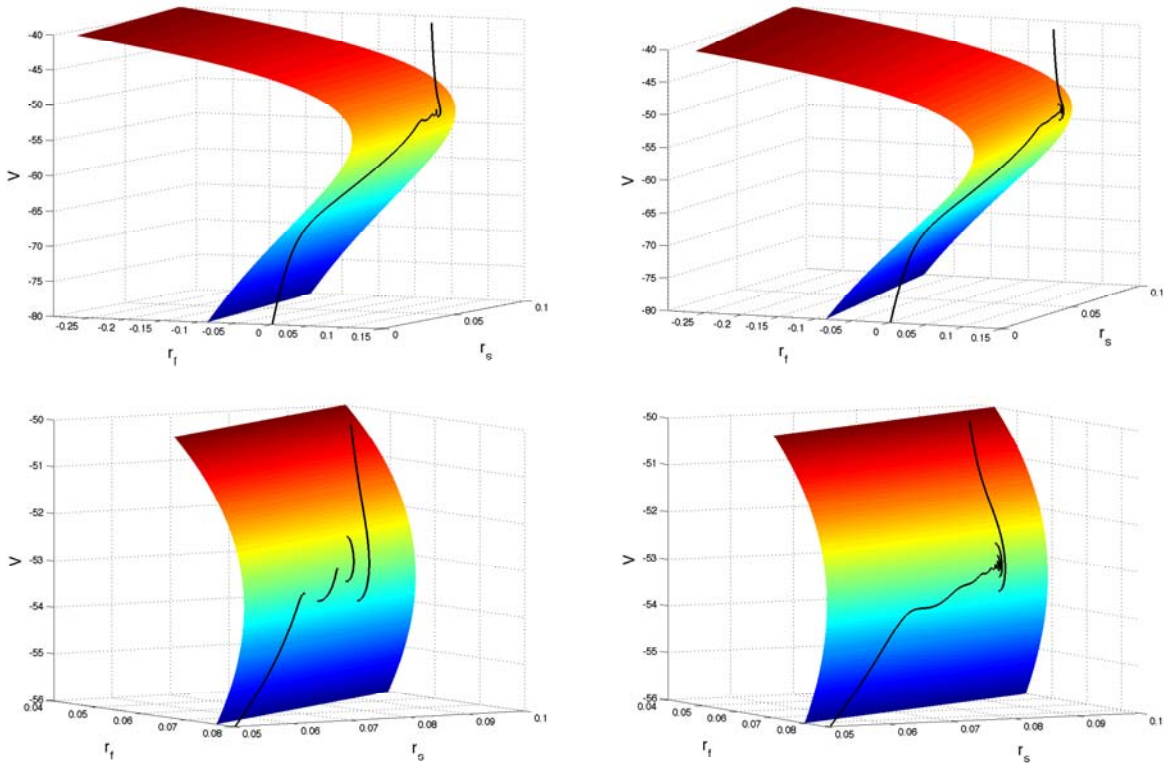


Figure 3. *Phase-space diagrams for the dimensional SC model (1)–(3); $I_{app} = -2.4$ (left) and $I_{app} = -2.5$ (right). The folded V-nullsurface is shown as well as the trajectories corresponding to the time traces shown in Figure 2(d) and (b). Note that the trajectories evolve approximately along the V-nullsurface towards the fold-curve, where they start to create subthreshold oscillations before they escape this regime to fire an action potential.*

3.4. Dimensionless formulation. Here we bring system (1)–(3) to a dimensionless form in order to uncover the different time scales in which the system operates. We first choose appropriate voltage and time scales, K_V and K_t, respectively, and define

$$v = \frac{V}{K_V}, \qquad \bar{t} = \frac{t}{K_t}. \tag{4}$$

From a dimensional analysis point of view one would choose K_V as a combination of the model parameters. A more standard choice would be $K_V = |E_K| = 90\,\mathrm{mV}$, which is the maximum, in absolute value, reversal potential for the full SC model [10], and an upper bound for V. Here we choose $K_V = 100\,\mathrm{mV}$, which is a typical voltage scale for neuronal models, for easier comparison with the original full model as well as with the dimensional reduced model. The dimensionless voltage threshold and reset values are then given by $v_{rst} = -80/K_V = -0.8$ and $v_{th} = -40/K_V = -0.4$. The relevant voltage range for our model in terms of the dimensionless variable v is therefore $[-0.8 : -0.4]$. We define

$$T_f = \min_{v\in[-0.8:-0.4]} \tau_{r_f}(K_V\, v), \qquad T_s = \min_{v\in[-0.8:-0.4]} \tau_{r_s}(K_V\, v), \tag{5}$$

and we choose $K_T = T_f \sim 30\,\text{ms}$ as a typical (slow) time scale (see Figure 1(b)).

We also define a reference maximal conductance $K_G = 1.5\,\text{mS/cm}^2$, which is at the top of the physiologically plausible scale for maximal conductances. This is the value of G_h that we used in the simulations presented in Figures 2 and 3. We define the following dimensionless variables, parameters, and functions:

$$\bar{E}_L = \frac{E_L}{K_V}, \qquad \bar{E}_{Na} = \frac{E_{Na}}{K_V}, \qquad \bar{E}_h = \frac{E_h}{K_V}, \tag{6}$$

$$\bar{G}_p = \frac{G_p}{K_G}, \qquad \bar{G}_h = \frac{G_h}{K_G}, \qquad \bar{G}_L = \frac{G_L}{K_G}, \qquad \bar{I}_{app} = \frac{I_{app}}{K_G\,K_V}, \tag{7}$$

$$\epsilon = \frac{C}{K_T\,K_G} = \frac{C}{T_f\,K_G} \sim 0.023 \ll 1, \qquad \eta = \frac{K_T}{T_s} = \frac{T_f}{T_s} \sim 0.286, \tag{8}$$

$$\bar{r}_{f,\infty}(v) = r_{f,\infty}(K_V\,v), \qquad \bar{r}_{s,\infty}(v) = r_{s,\infty}(K_V\,v), \qquad \bar{p}_\infty(v) = p_\infty(K_V\,v), \tag{9}$$

$$\bar{\tau}_{r_f}(v) = \frac{\tau_{r_f}(K_V\,v)}{T_f}, \qquad \bar{\tau}_{r_s}(v) = \frac{\tau_{r_s}(K_V\,v)}{T_s}. \tag{10}$$

Substituting (4)–(10) into (1)–(3) and deleting the "bar" sign, one gets

$$\epsilon\,\frac{dv}{dt} = I_{app} - G_L\,(v - E_L) - G_p\,p_\infty(v)\,(v - E_{Na}) - G_h\,(c_f\,r_f + c_s\,r_s)\,(v - E_h), \tag{11}$$

$$\frac{dr_f}{dt} = \frac{r_{f,\infty}(v) - r_f}{\tau_{r_f}(v)}, \tag{12}$$

$$\frac{dr_s}{dt} = \eta\,\frac{r_{s,\infty}(v) - r_s}{\tau_{r_s}(v)}. \tag{13}$$

System (11)–(13) is a slow-fast system with v evolving on the fast time scale and both r_f and r_s evolving on a slow scale. These two variables evolve on a similar slow time scale, as becomes apparent by comparing the values of ϵ and η ($\epsilon \ll \eta$).

4. The mechanism of generation of MMOs. MMOs consist of STOs interspersed with spikes (large-amplitude oscillations occurring on a faster time scale). In this section we show that the generation of MMOs in the SC model (11)–(13) is governed by the canard phenomenon. In our explanation we will follow [16, 17]. We use notation tailored to the model. For simplicity we call

$$f(v, r_f, r_s) = I_{app} - G_L\,(v - E_L) - G_p\,p_\infty(v)\,(v - E_{Na}) - G_h\,(c_f\,r_f + c_s\,r_s)\,(v - E_h), \tag{14}$$

$$g(v, r_f) = \frac{r_{f,\infty}(v) - r_f}{\tau_{r_f}(v)}, \tag{15}$$

$$h(v, r_s) = \eta\,\frac{r_{s,\infty}(v) - r_s}{\tau_{r_s}(v)}. \tag{16}$$

As we show below, the existence of MMOs for the SC model (11)–(13) is guaranteed by Theorem 4.1 or Theorem 4.2 in [17]. These theorems require that the v-nullsurface (which we will refer to as S) be folded (parabolic cylinder shape). STOs occur in the vicinity of the fold-curve L (the curve of knees of S), defined as the set of points $\{p \in S : f_v(p) = 0,\ f_{vv}(p) < 0\}$. The lower ($S_a$) and upper ($S_r$) branches of the folded manifold S are attracting ($f_v < 0$) and repelling ($f_v > 0$), respectively. After a finite number of STOs the trajectory moves away from S and escapes the subthreshold regime towards the spiking one, as we explain in section 4.2. For MMOs to occur, the trajectory should be able to come back to the subthreshold regime; i.e., a suitable return mechanism should bring the trajectory back to a region of S where it can evolve towards its curve of knees L (setting the initial conditions in the subthreshold regime). Different models may have different return mechanisms. The one corresponding to this model was described in section 3.2. In the following sections we describe the mechanism of generation of STOs and the onset of spikes for system (11)–(13), and we show their dependence on some of the parameters of the model.

We are mainly interested in understanding the contribution of I_h to the observed mixed-mode oscillatory patterns, since I_h is known to change with development and neuromodulators [50]. Changes in the amounts of I_h are reflected in changes in the maximal conductances G_h. Other effects include changes to the activation curves ($r_{f,\infty}(V)$ and $r_{s,\infty}(V)$) and the voltage-dependent time scales through the values of ϵ and η or the reset properties of I_h (initial conditions in the subthreshold regime). Changes in these values may affect the relative number of STOs, as well as the oscillatory frequency and the amplitude of the STOs.

4.1. A geometric singular perturbation theory approach. The SC model (11)–(13) is a singularly perturbed system with one fast (v) and two slow (r_f, r_s) variables. This system evolves on a slow time scale $t = \varepsilon\tau$. The limiting problem $\varepsilon \to 0$ on this slow time scale t is called the *reduced problem* and describes the evolution of the slow variables (r_f, r_s). The phase space of the reduced problem is the critical manifold S defined by $S := \{(v, r_f, r_s) \in \mathbb{R}^3 : f(v, r_f, r_s) = 0\}$; i.e., S is the v-nullsurface. We represent it by

$$r_f = \phi(v, r_s) = \frac{I_{app} - G_L\,(v - E_L) - G_p\,p_\infty(v)\,(v - E_{Na})}{c_f\ G_h\,(v - E_h)} - \frac{c_s}{c_f}\,r_s. \tag{17}$$

Figure 4 illustrates S for two values of I_{app} and other physiologically plausible parameters.

The second limiting problem is called the layer problem, and it is obtained by rescaling time ($\tau = t/\epsilon$) in system (11)–(13) and setting $\varepsilon \to 0$. The layer problem describes the evolution of v on the fast time scale for fixed values of the gating variables (r_f, r_s); i.e., the slow variables are considered as parameters in this singular limit. Note that the manifold S is a manifold of equilibria for the layer problem.

These two limiting problems, the reduced problem (2D) and layer problem (1D), are lower-dimensional than the full problem (3D) and are therefore more amenable to analysis. Geometric singular perturbation theory [17, 51, 52] provides a way to piece together the information obtained from these lower-dimensional problems in order to provide a unified global description of the observed MMOs in the full 3D system.

4.2. Layer problem. By rescaling time ($\tau = t/\epsilon$) in (11)–(13) and setting $\epsilon = 0$, one obtains the layer problem which describes the fast dynamics away from the critical manifold

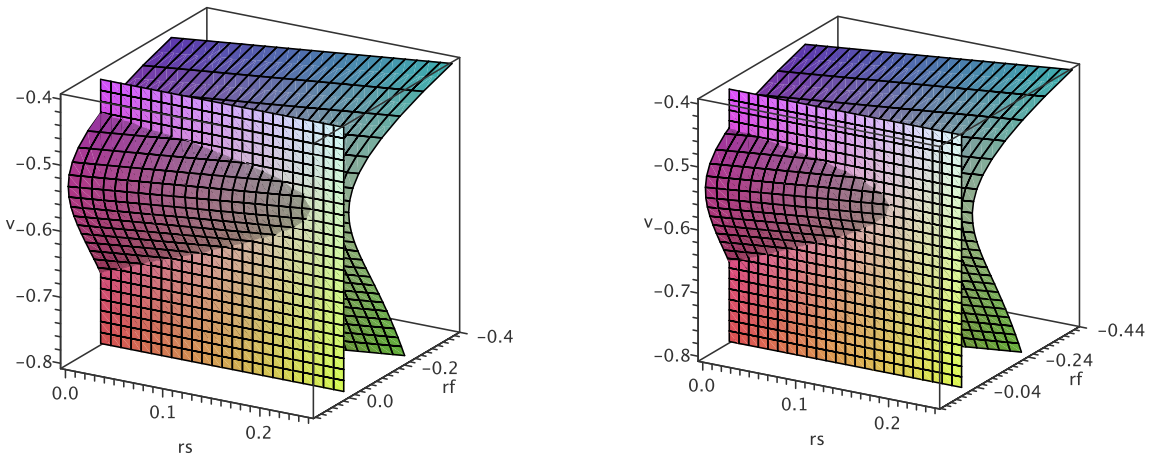


Figure 4. *Domains of possible initial conditions on the critical manifolds S for trajectories starting at the SC reset values. The left and right panels correspond to $I_{app} = -2.64$ (-0.0176) and $I_{app} = -1.86$ (-0.0124), respectively. The dimensionless values of the parameters are given in parentheses. All panels correspond to $G_h = 1.5$ (1.0). This figure illustrates the facts that I_{app} does not change the geometry of the critical manifold significantly and that I_{app} has also no significant influence on the domain of (biologically relevant) initial conditions.*

S, represented by (17):

$$\frac{dv}{d\tau} = f(v, r_f, r_s), \tag{18}$$

$$\frac{dr_f}{d\tau} = 0, \tag{19}$$

$$\frac{dr_s}{d\tau} = 0. \tag{20}$$

Trajectories of the layer problem starting at an initial point $(v_0, r_{s,0}, r_{f,0})$ evolve along 1D sets $(v, r_{s,0}, r_{f,0})$, called fast fibers, near the critical manifold. The critical manifold S is a manifold of equilibria for the layer problem; i.e., the intersection points between S and vertical lines containing the fast fibers define the fixed point corresponding to each trajectory. By linearizing the layer problem at S, one obtains information about the transient behavior of the solutions along the fast fibers. As we illustrate in Figure 4, the critical manifold S is folded. This remains true for parameter variations in the physiologically plausible regime (data not shown). The lower (S_a) and upper (S_r) branches of the folded manifold S are attracting $(f_v < 0)$ and repelling $(f_v > 0)$, respectively. Figure 4 also illustrates that changes in the key parameters of the model (e.g., I_{app}) do not affect the shape of the folded slow manifold S significantly.

4.3. Initial conditions of the reduced problem on the critical manifold. The relevant trajectory in the subthreshold regime starts at $(v, r_f, r_s) = (-0.8, 0, 0)$ (see sections 1 and 3.4).

This trajectory is projected on S_a along the corresponding fast fiber to the equilibrium point $(v_0, r_{f,0} = 0, r_{s,0} = 0)$ of the layer problem. This point on the critical manifold is then used as the initial condition of the reduced flow corresponding to the initial condition $(v, r_f, r_s) = (-0.8, 0, 0)$. Note that we use the same notation for initial conditions on the critical manifold S as for the initial conditions for problem (11)–(13). Figure 4(b) shows the critical manifold intersected with the plane $r_s = 0$. For $r_f = r_s = 0$ as initial conditions (reset values) this figure shows that $V \sim -70\,\mathrm{mV}$ ($v \sim -0.7$) corresponds to the intersection of the fast fiber through $r_f = r_s = 0$ with the critical manifold. Therefore we will take $v_0 \sim -0.7$ as initial condition for the reduced problem. The exact initial condition on the critical manifold depends on the parameters of the model. We make the appropriate calculations for each parameter set.

4.4. The reduced flow. The reduced flow is obtained by setting $\epsilon = 0$ in (11). System (11)–(13) becomes

$$0 = f(v, r_f, r_s), \tag{21}$$

$$\frac{dr_f}{dt} = g(v, r_f), \tag{22}$$

$$\frac{dr_s}{dt} = h(v, r_s). \tag{23}$$

These equations describe the evolution of r_f and r_s on the critical manifold S defined by (17).

Trajectories evolve on the slow manifold S (actually on S_a), from their initial conditions $(v_0, r_{f,0}, r_{s,0})$ towards the fold-curve L. Since S is given as a graph $\phi(v, r_s)$ we project the reduced system (21)–(23) onto the (v, r_s)-plane. By implicitly differentiating the function $f(v, r_f, r_s) = 0$, we obtain the reduced system

$$\begin{pmatrix} -f_v\, v' \\ r_s' \end{pmatrix} = \begin{pmatrix} f_{r_f}\, g + f_{r_s}\, h \\ h \end{pmatrix}_{r_f = \phi(v, r_s)}. \tag{24}$$

This system is singular along the fold-curve L ($f_v = 0$). Therefore we rescale time by a factor $-f_v$ to obtain the desingularized system

$$\begin{pmatrix} \dot{v} \\ \dot{r}_s \end{pmatrix} = \begin{pmatrix} f_{r_f}\, g + f_{r_s}\, h \\ -f_v\, h \end{pmatrix}_{r_f = \phi(v, r_s)}, \tag{25}$$

where the overdot represents differentiation with respect to this new time. Note that system (25) has the same phase portrait as system (24), but the orientation of the flow on S_r (unstable slow manifold) has to be reversed due to the rescaling of time.

There are two types of singularities in system (25): regular and folded singularities. Regular singularities are given by $h = 0$ and $g = 0$ and are therefore also equilibria of the reduced flow (24). These singularities will (generically) persist under small perturbations in system (11)–(13) with $\epsilon \ll 1$. On the other hand, folded singularities are given by $f_v = 0$, which defines the fold-curve L, and $f_{r_f}\, g + f_{r_s}\, h = 0$ (for points $\bar{p}$ on L). Each folded singularity is classified as a folded saddle, folded node, or folded saddle-node based on its classification as a saddle, node, or saddle-node as an equilibrium of (25). Folded singularities are not equilibria

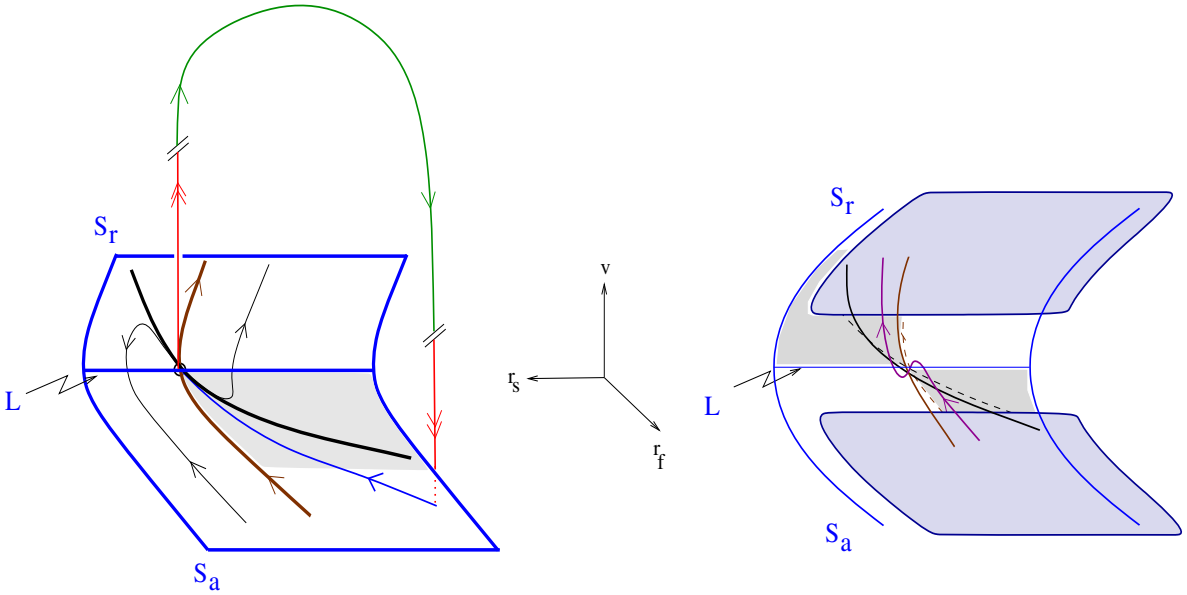


Figure 5. *Schematic representation of the canard mechanism generating MMOs in the SC and related models: A folded node singularity, located on the fold-curve L, forms a singular funnel. A singular periodic orbit consists of a segment on the attracting manifold S_a (blue) within the funnel with the folded node singularity as an endpoint. Then it follows a fast fiber of the layer problem (red), and a global return mechanism (green) projects the singular orbit back into the singular funnel. The return mechanism for the SC model is based on the reset properties of the h-current after a spike has occurred. The right panel shows a schematic representation of a trajectory rotating around the weak canard within the singular funnel.*

of the reduced system (24). However, as shown in [15, 17], their presence gives the opportunity for the reduced flow to cross from S_a to S_r through L in finite time. If there are no folded singularities, then trajectories of the reduced flow which arrive at the fold-curve subsequently jump along the fast fibers and escape the subthreshold regime without generating any STO.

As we show in section 4.6, the singularities found in the SC model, for the relevant biophysically plausible parameters, are folded nodes (or folded saddle-nodes). For more details about folded singularities we refer to [15, 16, 17, 45]. Associated with a folded node there exists a whole sector of trajectories, called singular canards, that are able to pass from S_a to S_r through the folded node. This sector is called the singular funnel. Two singular canards are related to the eigendirections of the folded node. They are the weak and strong canards. They correspond to the smallest and largest (in absolute value) eigenvalues, respectively. The singular funnel is bounded by the fold-curve L and the strong canard. The latter is the strong stable invariant manifold of the folded node. The canards and funnel existing from $\epsilon > 0$ (sufficiently small) arise as perturbation of their singular counterparts. Only trajectories entering the funnel are able to rotate around the weak canard (see Figure 5). For a geometric description of this phenomenon, we refer the reader to [16, 17].

4.5. Existence of MMOs. Based on the singular limit behavior of solutions in both the reduced and the layer problem, Brøns, Krupa, and Wechselberger [17] provided a theorem that guarantees the existence of MMOs for system (11)–(13) with a sufficiently small $0 < \epsilon \ll 1$.

This theorem is based on the following assumptions.

Assumption 1. The singularly perturbed system is (locally) a folded surface, as in system (11)–(13) for parameter sets in the physiologically plausible range. This was shown in section 4.2.

Assumption 2. The problem possesses a folded node singularity. In section 4.6 we will determine parameter ranges (in the physiologically plausible regime) for which folded nodes exist.

Assumption 3. There exists a singular periodic orbit (see Figure 5) which consists of a segment on S_a (blue) within the singular funnel (shadowed region) with the folded node singularity (black circle) as an endpoint, fast fibers (red) of the layer problem, and a global return mechanism (green). The global return mechanism for the SC model was described in sections 3.2 and 4.3.

If Assumptions 1–3 are fulfilled, then Theorem 4.1 in [17] predicts maximal 1^s MMO patterns (for sufficiently small ϵ). There are two limiting cases of the theory related to Assumptions 2 and 3. In Assumption 3, if the global return mechanism is on the border of the singular funnel (the brown trajectory in Figure 5), then Theorem 4.2 in [17] predicts submaximal MMO patterns. In the folded saddle-node limit, Assumption 2 is violated, but we still expect the existence of MMOs (with a large number of STOs). The theory for this limiting case has still to be developed, but we can use the folded node theory to make qualitative predictions of MMOs. We will discuss both limiting cases in section 4.8.

4.6. The folded node singularity and the canard phenomenon. In order to look for parameter ranges in which system (11)–(13) possesses a folded node singularity, we numerically calculated the desingularized reduced flow corresponding to (25) for various values of G_h and I_{app}. We used XPPAUT [53] for these computations and a Runge–Kutta method of order II [54] for the numerical simulations of the 3D system (11)–(13). The results are given in terms of the dimensional values of the parameters. Their corresponding dimensionless values are given in parentheses. We will first consider the case $G_h = 1.5$ (1.0) for which we found the existence of the following:

- a regular node singularity on S_a, a folded saddle singularity, and a regular node singularity on S_r for $I_{app} < -2.64$ (-0.0176);
- a folded saddle-node singularity and a regular node singularity on S_r for $I_{app} \approx -2.64$;
- a regular saddle singularity on S_r, a regular node singularity on S_r, and a folded node singularity for $-2.64 < I_{app} < -1.92$ (-0.0128);
- a saddle-node singularity on S_r and a folded node singularity for $I_{app} \approx -1.92$ (-0.0128);
- a folded node singularity for $-1.92 < I_{app} < -1.86$ (-0.0124);
- a folded focus singularity for $I_{app} > -1.86$.

Therefore a folded node singularity exists for values of $I_{app} \in (-2.64, -1.86)$. To each folded node corresponds a singular funnel which is bounded by the strong canard and the fold-curve. The strong canard is an invariant manifold of the folded node; it is a separatrix of the flow on the reduced phase space S and therefore a borderline for qualitatively different behaviors. Note that the strong canard is related to the strong eigendirection of the folded node and is therefore unique.

Figure 6 shows the folded node, the singular strong canard, the fold-curve, and the singular funnel for three values of I_{app} and the same values of both G_h and the initial conditions on S_a. In Figure 6(a), an initial condition within the singular funnel will be "funneled" into the folded singularity and gives the possibility of STOs [16] for $\epsilon > 0$. The corresponding voltage trace showing STOs is given in Figure 7(a). If an initial condition is outside the funnel (Figure 6(c)), then the reduced flow will reach the fold-curve at an (ordinary) *jump point* where solutions jump off the fold and follow a fast fiber, which leads to the spiking regime without displaying STOs. The corresponding voltage trace showing only spiking activity is given in Figure 7(c). Figures 6(b) and 7(b) correspond to a trajectory within the funnel but very close to its boundary. The voltage trace shows only one STO per spike. From Figure 6 and the voltage traces showed in Figure 7 we observe that as we increase I_{app} the folded node moves to the left and, accordingly, the strong canards (separatrices) intersect the v-axis at higher values. As this occur, trajectories starting at approximately the same initial values on the critical manifold, $v \in (-0.69, -0.68)$, evolve closer to the strong canard, decreasing their number of STOs (per spike), and, for higher values of I_{app}, they are left outside the funnel and generate only spiking activity.

Figure 8 shows the folded nodes and corresponding strong canards for various values of I_{app}. The range of values of v_0 whose corresponding trajectories will be attracted to the funnel shrinks as I_{app} increases. For example, the singular limit analysis predicts that trajectories on the slow manifold with initial conditions $(r_f(0), r_s(0)) = (0, 0)$ and the corresponding values of v_0 will not display STOs for $I_{app} > -2.25$ and for $I_{app} < -2.64$ (corresponding to a lower bound where a folded node singularity exists). Our simulations of the SC model (11)–(13) show the existence of MMOs for $-2.57 < I_{app} < -2.27$; i.e., both the onset of MMOs and the change from MMOs to relaxation oscillations occur for slightly higher values than theoretically predicted. However, in both cases these values are within the order of the singular perturbation parameter ϵ and therefore justify the singular prediction made above.

The observed MMO parameter window $-2.57 < I_{app} < -2.27$ for the SC model is also in good agreement with the MMO parameter window $-2.7 < I_{app} < -2.4$ of the full 7D model. The (small) shift to more depolarized values in the SC model (11)–(13) is basically explained by the lack of the depolarized current I_{Na} given by the full 7D model. A more detailed comparison of the two models can be found in [55].

4.7. Transition between subthreshold oscillatory regimes as a consequence of changes in the h-current reset properties. Changes in the h-current reset properties are reflected in changes in the initial values of its gating variables r_f and r_s in the subthreshold regime, and particularly on the slow manifold. Here we show how this affects the subthreshold oscillatory properties of the MMO 1^s patterns, particularly the number of subthreshold oscillations per spike. This number s depends on the ratio $\mu = \lambda_1 / \lambda_2 < 1$ between the eigenvalues corresponding to the weak and strong eigendirections of the folded node in system (25). In the case $\mu < 1/3$, it was shown in [16] that singularly perturbed systems like (11)–(13) possess $[(1-\mu)/(2\mu)]$ *secondary canards* besides the two primary (weak and strong) canards, where $[(1-\mu)/(2\mu)]$ denotes the greatest integer less than or equal to $(1-\mu)/(2\mu)$.

These secondary canards divide the singular funnel into subsectors (see Figure 9). Each of them is associated with a specific value of s in the MMO 1^s patterns; i.e., trajectories entering

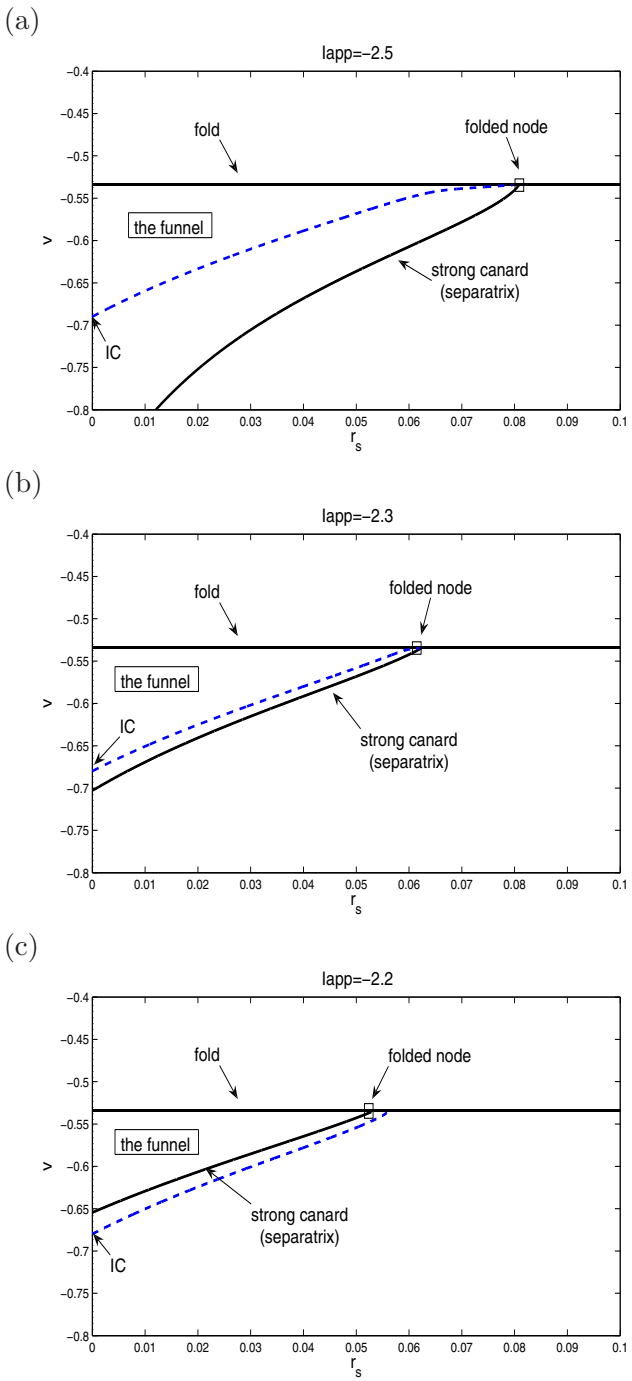


Figure 6. *Schematic representations of the singular funnel for various values of I_{app} and for $G_h = 1.5$ (1.0) and $G_p = 0.5$ (0.3333). The dimensionless values of the parameters are given in parentheses. The singular funnel is bounded by the fold (horizontal) line and the strong canard. The funnel is located on the attracting part of the slow manifold S_a (below the fold). Trajectories (dashed lines) start at their initial conditions (IC) on the slow manifold and evolve towards the folded node.*

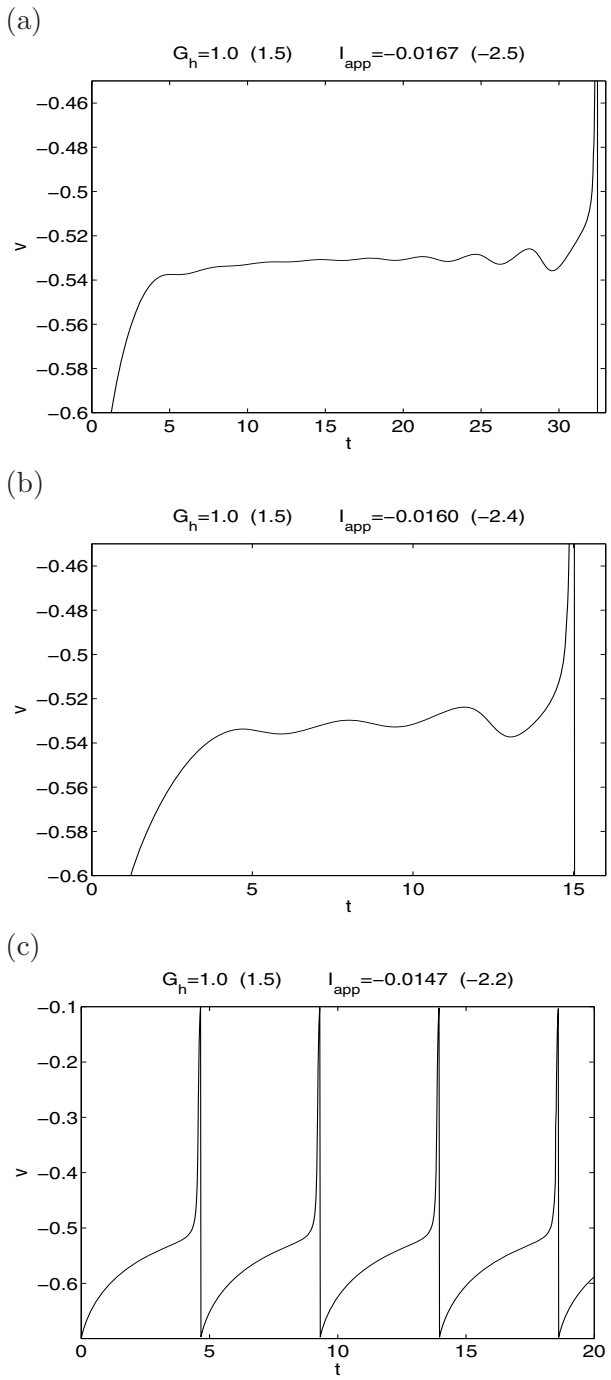


Figure 7. *Voltage traces for the (dimensionless) SC model (11)–(13). $G_h = 0.5$ (0.3333). The dimensional values of G_h and I_{app} are given in parentheses. The initial conditions are located on the slow manifold S. The number of subthreshold oscillations per spike increases with decreasing values of I_{app}.*

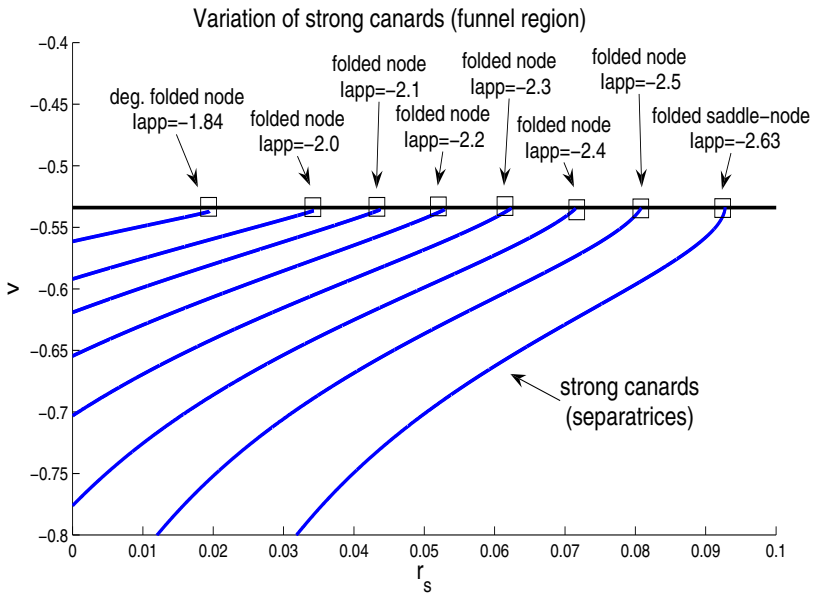


Figure 8. *Schematic representations of the singular funnel for various values of I_{app} and for $G_h = 1.5$ (1.0) and $G_p = 0.5$ (0.3333). The dimensionless values of the parameters are given in parentheses. Each singular funnel is bounded by the fold (horizontal) line and the corresponding strong canard. The funnels are located on the attracting part of the slow manifold S_a (below the fold).*

Table 1

Singular limit predictions of number of secondary canards and maximum number s^* of STOs for $G_h = 1.5$ (1.0) under the variation of I_{app}.

I_{app}	μ	$(1-\mu)/(2\mu)$	s^*
-2.6	0.0097	51	52
-2.5	0.0480	9	10
-2.4	0.0917	4	5
-2.3	0.1430	2	3
-2.25	0.1725	2	3
-2.1	0.2842	1	2
-2.0	0.3940	0	1

different sectors display a different number of STOs per spike. This number increases from the sector bounded by the strong canard to the sector bounded by the fold-curve. The maximum number of STOs is defined by $s^* := (1+\mu)/(2\mu)$, which is a singular limit prediction. Table 1 shows the number of secondary canards and maximum number s^* STOs for $G_h = 1.5$ (1.0) under the variation of I_{app}.

As a consequence of an increase in the initial values of r_f and r_s, the initial voltage on S also changes, and trajectories may enter a sector corresponding to a higher or lower number of STOs. We illustrate this in Figures 9 to 11. We first vary $r_s(0)$ and keep $r_f(0) = 0$ unchanged for $I_{app} = -2.4$ (dimensionless value: $I_{app} = -0.0160$). We illustrate this schematically in Figure 9(a). The voltage traces are presented in Figure 10. The initial conditions on the slow manifold S approximate a line passing through the points $(v_0, r_{f,0}, r_{s,0}) = (-0.69, 0, 0)$

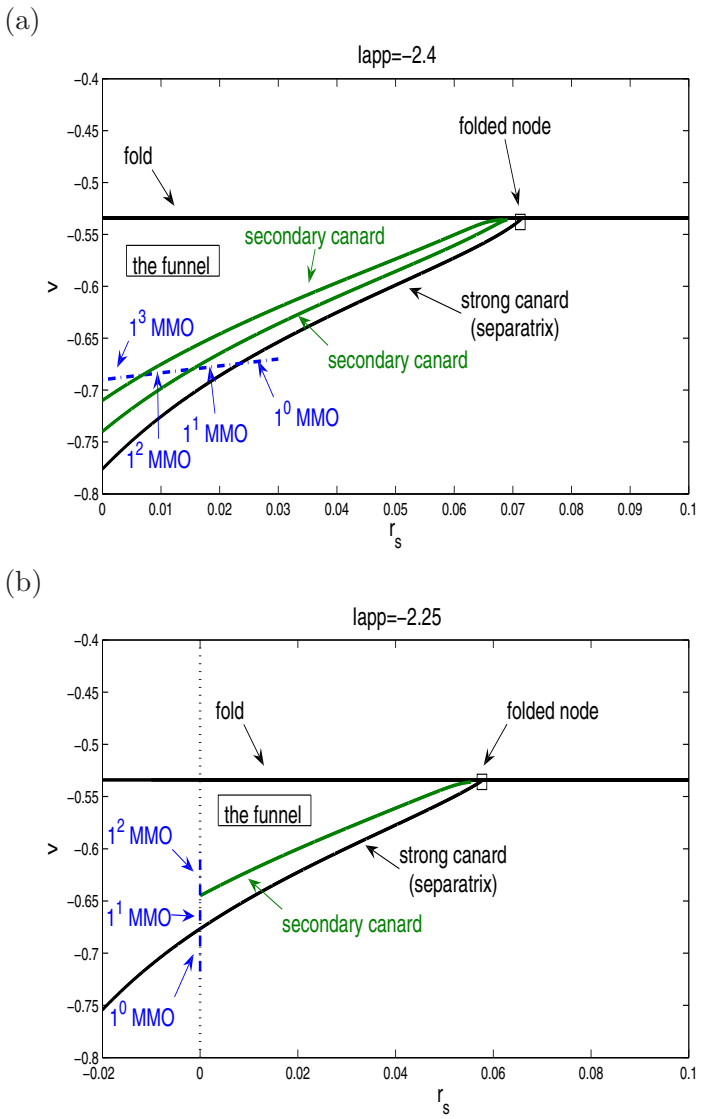


Figure 9. *Schematic illustration of the effect of changes in the initial values of the h-current gating variables r_s and r_f on the mixed-mode oscillatory patterns for $G_h = 1.5$ (1.0), $G_p = 0.5$ (0.3333), and* (a) $I_{app} = -2.4$ (-0.0160) *or* (b) $I_{app} = -2.25$ (-0.0150). *Each singular funnel is bounded by the "fold" line and the strong canard. The funnel is located on the attracting part of the slow manifold S_a (the repelling manifolds S_r are not shown). The secondary canards divide the funnel into sectors. Each sector corresponds to a 1^s mixed-mode oscillatory pattern. Trajectories starting in a given sector display s subthreshold oscillations per spike. Panel* (a) *corresponds to changes in $r_s(0)$ for fixed values of $v(0)$ and $r_f(0)$. The number of subthreshold oscillations per spike, s, decreases from left to right. The corresponding voltage traces are shown in Figure 10. Panel* (b) *corresponds to changes in $r_f(0)$ for fixed values of $v(0)$ and $r_s(0)$. The number of subthreshold oscillations per spike, s, decreases from top to bottom. The corresponding voltage traces are shown in Figure 11.*

and $(v_0, r_{f,0}, r_{s,0}) = (-0.67, 0, 0.03)$ (dashed segment in Figure 9(a)). We observe that the transition from MMO to only spiking patterns occurs for a value of $r_s(0)$ slightly higher than 0.02, and the number s of STOs per spike decreases as we increase $r_s(0)$. The results of

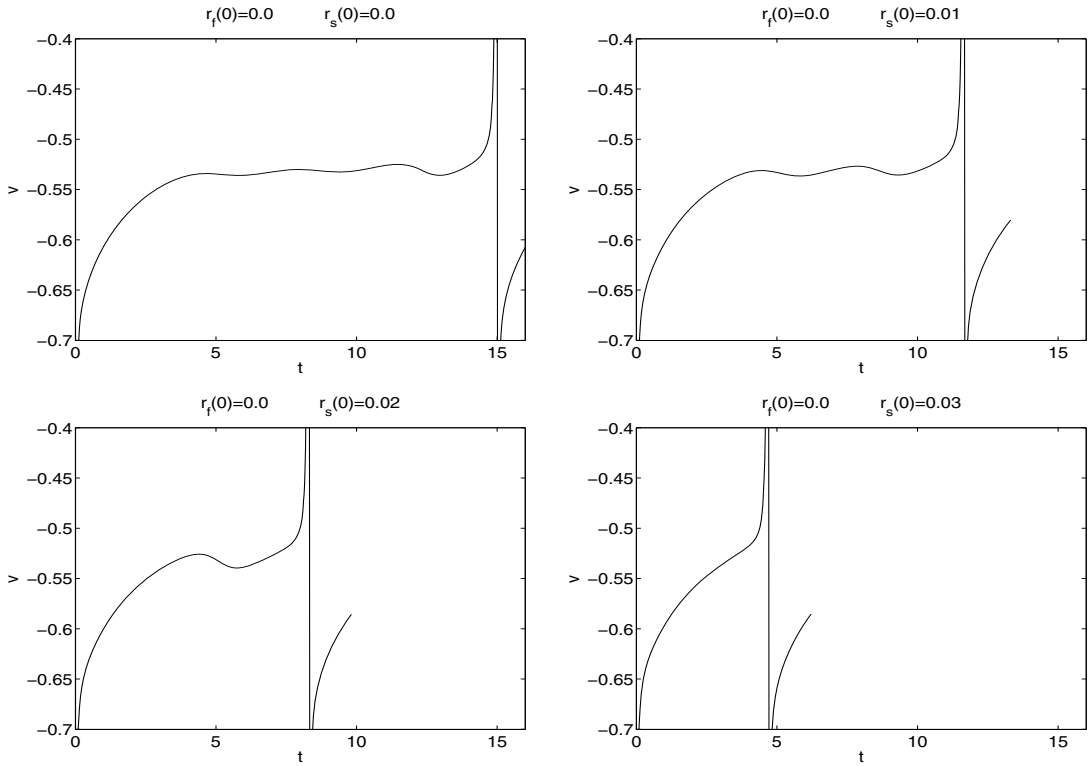


Figure 10. *Voltage traces showing the effect of changes in the initial values of r_s on the mixed-mode oscillatory patterns corresponding to trajectories starting in a vicinity of the slow manifold for fixed initial values of r_f. The dimensionless values of the parameters are given in parentheses: $G_h = 1.5$ (1.0), $G_p = 0.5$ (0.3333), $I_{app} = -2.4$ (−0.0160). The schematic representation of the corresponding singular funnel is given in Figure 9(a).*

our simulations (Figure 10) are in agreement with this prediction. In Figures 9(b) and 11 we vary $r_f(0)$ and keep $r_s(0) = 0$ unchanged for $I_{app} = -2.25$ (dimensionless value: $I_{app} = -0.0150$). The initial conditions on S are located on the v-axis at the points $(-0.69, 0, 0)$, $(-0.66, 0.025, 0)$, and $(-0.63, 0.05, 0)$ (dashed vertical segment in Figure 9(b)). This dashed vertical segment crosses the strong canard at $v \sim -0.675$. According to this, the transition from 1^1 to 1^0 (spikes only) MMO patterns occurs for r_s lower than 0.025. The results of our simulations (Figure 11) are in agreement with this prediction.

4.8. MMO theory and the dependence on the singular perturbation parameter. The existence results on MMOs presented here are based on singular perturbation theory and hold for sufficiently small singular perturbation parameter $0 < \epsilon \ll 1$ [16, 17]. In particular, Theorem 4.1 in [17] states that if Assumptions 1–3 are fulfilled, then maximal 1^{s^*} MMOs are expected for sufficiently small ϵ. Sufficiently small ϵ means here that $\mu \gg \sqrt{\epsilon}$ for the eigenvalue ratio of the folded node, as well as that $\delta \gg \sqrt{\epsilon}$, where δ defines the distance of the initial condition on S_a from the strong canard (relative position of the global return within the funnel). The $\sqrt{\epsilon}$-dependence follows from the canard theory (see, e.g., [16, 17, 55]). If one of these estimates is violated, then we still expect to observe an MMO pattern, but we

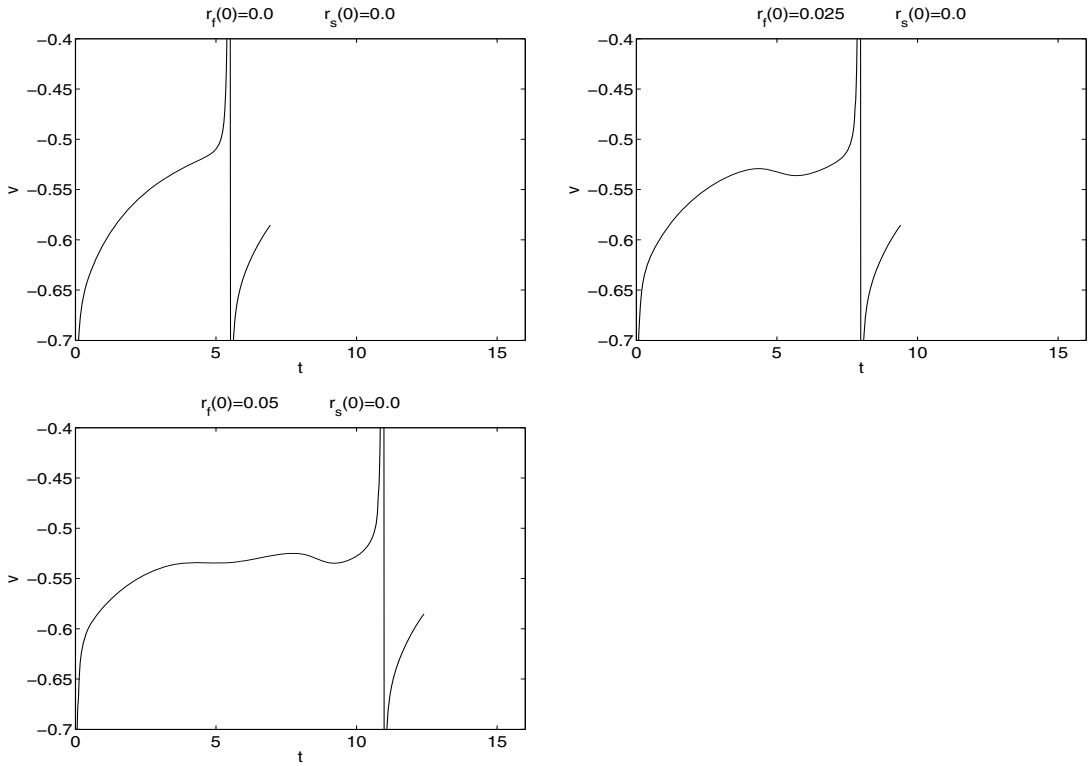


Figure 11. *Voltage traces showing the effect of changes in the initial values of r_f on the mixed-mode oscillatory patterns corresponding to trajectories starting in a vicinity of the slow manifold for fixed initial values of r_s. The dimensionless values of the parameters are given in parentheses: $G_h = 1.5$ (1.0), $G_p = 0.5$ (0.3333), $I_{app} = -2.25$ (−0.0150). The schematic representation of the corresponding singular funnel is given in Figure 9*(b).

cannot predict the exact MMO pattern. Theorem 4.2 in [17] covers the case where δ violates this condition.

For example, Table 1 predicts for $I_{app} = -2.4$ four secondary canards and therefore a maximal number $s^* = 5$ of STOs. If we compare the prediction with Figure 2(d), then we see that a 1^3 MMO pattern is realized with that particular choice of initial conditions. Therefore, Assumption 2 and/or 3 is violated; i.e., μ and/or δ are of order $O(\sqrt{\epsilon})$, where $\sqrt{\epsilon} \sim 0.15$. For $I_{app} = -2.4$, Table 1 shows that $\mu \sim 0.1$, and we estimate from Figure 9(a) that $\delta \sim 0.02$. Hence, both parameters are within the order $O(\sqrt{\epsilon})$, which explains why we do not find the maximal MMO pattern predicted by the singular perturbation theory. Nonetheless, we can explain certain trends. For instance, if we increase δ, i.e., if we decrease the initial conditions on S_a (to physiologically irrelevant negative values), then we observe an increase in STOs until we reach a maximum value of STOs. In the case $I_{app} = -2.4$, 1^8 MMO is the maximum pattern which is observed for initial conditions $r_s < -0.12$ (data not shown). Clearly, the maximum number of STOs, although larger than predicted, is still finite and therefore reflects the characteristics of a folded node induced MMO pattern. The perturbation ϵ is simply too large to give precise estimates on STOs. On the other hand, if we sufficiently decrease ϵ, then

we should observe the maximal 1^5 pattern as predicted. In this case, any $\epsilon < 10^{-5}$ gives this predicted result (data not shown).

Note that in the folded saddle-node limit $\mu \to 0$ an unbounded growth of STOs is expected. Folded saddle-nodes are related to the transition from an excitable to an oscillatory state, and we observe a large number of STOs in our simulations, e.g., for $I_{app} = 2.55$ in Figure 2(a). The MMO theory for this limiting case still has to be developed. So far, MMOs related to folded saddle-node singularities have been studied in [47] for a 3D autocatalator problem and in [23] for a dopaminergic neuron model.

4.9. Dependence of the subthreshold oscillatory frequency on the amount of the h-current. The dynamic picture described in section 4.6 is qualitatively affected by changes in the amount of the h-current, measured by the parameter G_h. We now consider values of G_h lower than the one considered in section 4.6. Table 2 shows that the ranges of values of I_{app} for which the system (11)–(13) has a folded node shrink with increasing values of G_h and $G_p = 0.5$ (dimensionless value: $G_p = 0.3333$). Note that the decrease in the amount of I_h results in an increase of the amount of I_{app}.

Table 2
Folded node regimes for different G_h values.

Dimensional			Dimensionless		
G_h	Folded nodes for	Interval size	G_h	Folded nodes for	Interval size
1.5	$-2.64 < I_{app} < -1.86$	0.78	1.0	$-0.0176 < I_{app} < -0.0124$	0.0052
1.4	$-2.43 < I_{app} < -1.72$	0.71	0.933	$-0.0162 < I_{app} < -0.0115$	0.0047
1.3	$-2.21 < I_{app} < -1.58$	0.63	0.866	$-0.0147 < I_{app} < -0.0105$	0.0042
1.2	$-1.98 < I_{app} < -1.43$	0.55	0.8	$-0.0132 < I_{app} < -0.0095$	0.0037
1.0	$-1.51 < I_{app} < -1.12$	0.39	0.666	$-0.0101 < I_{app} < -0.0075$	0.0026
0.5	$-0.24 < I_{app} < -0.11$	0.13	0.333	$-0.0016 < I_{app} < -0.0007$	0.0009

Figure 12 illustrates the effect that "balanced" changes in the values of G_h and I_{app} have on the MMO patterns for a constant value of G_p. To compensate for the decrease in the amount of I_h (decrease in G_h) we increased I_{app} so that the number of STOs per spike is kept constant, and for a fixed value of G_h, a lower value of I_{app} would produce one less STO per spike. We observe that as we decrease G_h, the STO frequency slightly decreases.

A decrease in the amount of I_h can be also compensated by an increase in the amount of I_p. In Figures 13(a) and (b) the values of I_{app} are kept constant and the values of G_h and G_p were chosen following the principle described in the previous paragraph. Similarly to the previous case, as we decrease G_h, the STO frequency slightly decreases.

4.10. Inhibitory pulses decrease the number of STOs per spike and increase the SC firing frequency. Inhibitory pulses typically delay firing in the postsynaptic cell. However, this is partially reversed when the postsynaptic cell has an h-current [26, 56]. Heuristically, we expect a pulse of inhibition applied at different times t_{pert} to have a differential effect on the postsynaptic spike-time. When the pulse of inhibition is applied to an SC whose trajectory is still far away from the "rotational region," it will have little effect, since the trajectory will have enough time to recover before starting to rotate. In contrast, a pulse of inhibition of the same magnitude applied to an SC that is already rotating may cause it to move from one

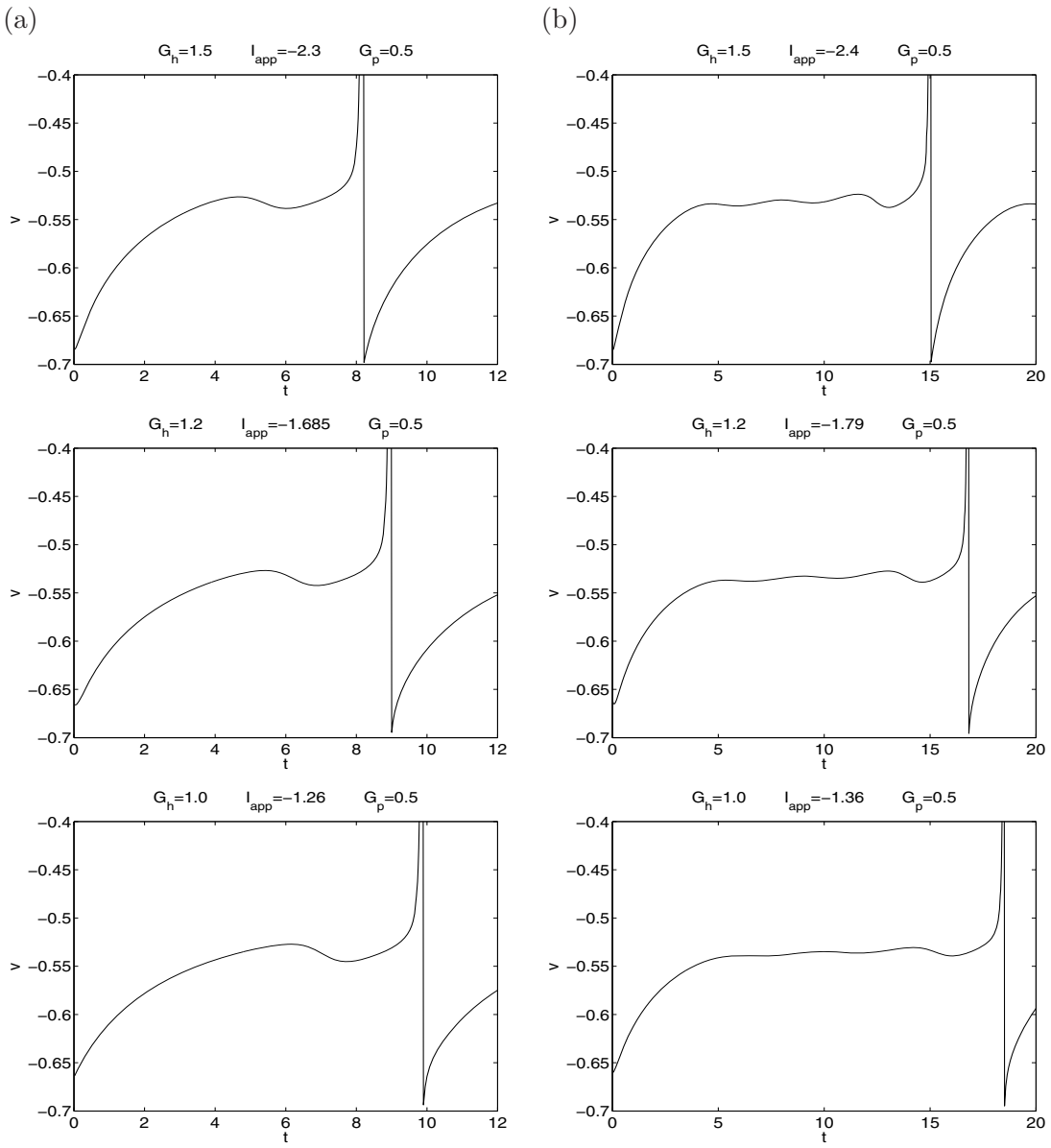


Figure 12. *Voltage traces showing the effect of changes in the values of G_h and I_{app} on the mixed-mode oscillatory patterns corresponding to trajectories starting in a vicinity of the slow manifold. The dimensionless values of the parameters are given in parentheses. $G_p = 0.5$ (0.3333). Top: $G_h = 1.5$ (1.0) and* (a) *$I_{app} = -2.3$ (-0.0153),* (b) *$I_{app} = -2.4$ (-0.0160). Middle: $G_h = 1.2$ (0.8) and* (a) *$I_{app} = -1.685$ (-0.0112),* (b) *$I_{app} = -1.79$ (-0.0119). Bottom: $G_h = 1.0$ (0.6667) and* (a) *$I_{app} = -1.26$ (-0.0084),* (b) *$I_{app} = -1.36$ (-0.0091).*

rotational sector to another with a lower number of STOs or no STOs at all. As a consequence, the spike-time of the perturbed SC will be considerably advanced with respect to that of the unperturbed cell. We illustrate this in Figures 14 and 15 for two different values of I_{app}. In Figure 14, the unperturbed SC displays one STO per spike. The pulses of inhibition applied

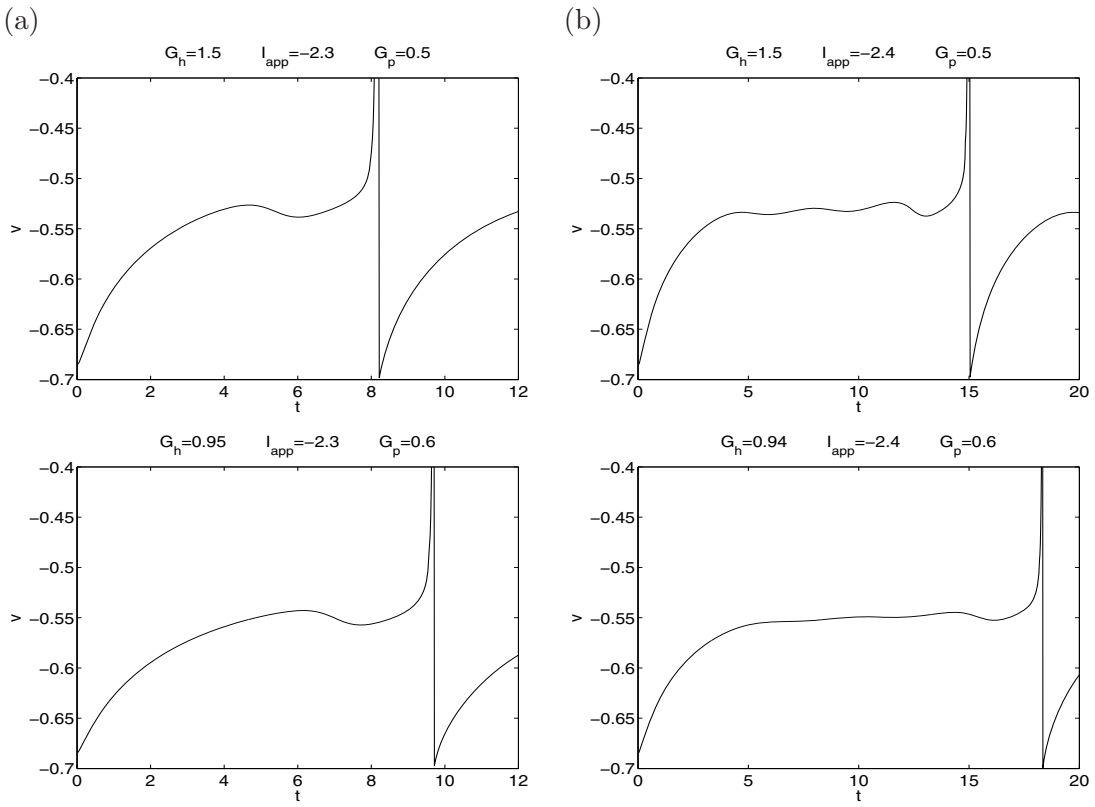


Figure 13. *Voltage traces showing the effects of changes in the value of G_h and G_p on the mixed-mode oscillatory patterns corresponding to trajectories starting in a vicinity of the slow manifold. The dimensionless values of the parameters are given in parentheses.* (a) $I_{app} = -2.3$ *(*-0.0153*) and* (b) $I_{app} = -2.4$ *(*-0.0160*). Top:* $G_h = 1.5$ *(*1.0*) and* $G_p = 0.5$ *(*0.3333*); bottom:* $G_p = 0.6$ *(*0.4*) and* (a) $G_h = 0.95$ *(*0.6333*),* (b) $G_h = 0.94$ *(*0.6267*).*

at $t_{pert} = 2.66$ and $t_{pert} = 2.83$ move the trajectory outside the funnel and cause the SC to spike with no STOs. In Figure 15, the unperturbed SC has three STOs per spike. Pulses of inhibition applied at $t_{pert} = 3.33$ and $t_{pert} = 5.0$ move the trajectory to sectors corresponding to two and one STO per spike, respectively, and a pulse of inhibition applied at $t_{pert} = 5.83$ moves the trajectory outside the funnel and causes the SC to spike with no STOs.

5. Discussion. Stellate cells (SCs) of the medial entorhinal cortex (MEC) display subthreshold oscillations (STOs) and mixed-mode oscillations (MMOs) at theta frequencies [8, 9]. In these MMO patterns, spikes occur at the peak of STOs, though not at every cycle. STOs found in single SCs are the result of the interaction between I_h and I_p [7, 8, 9]. It was recently found that their frequency varies along the dorsal-to-ventral axis of the MEC, scaling with the MEC grid-cells [3] that are believed to contain a neural map of the spatial environment [4, 5, 6].

Previous theoretical work has focused on various aspects of SC dynamics using biophysical (conductance-based) models: reproducing, via simulations, STOs, MMOs, and fully spiking patterns [10, 12, 13, 14, 30], the computational study of resonant properties of SCs [14, 57] and

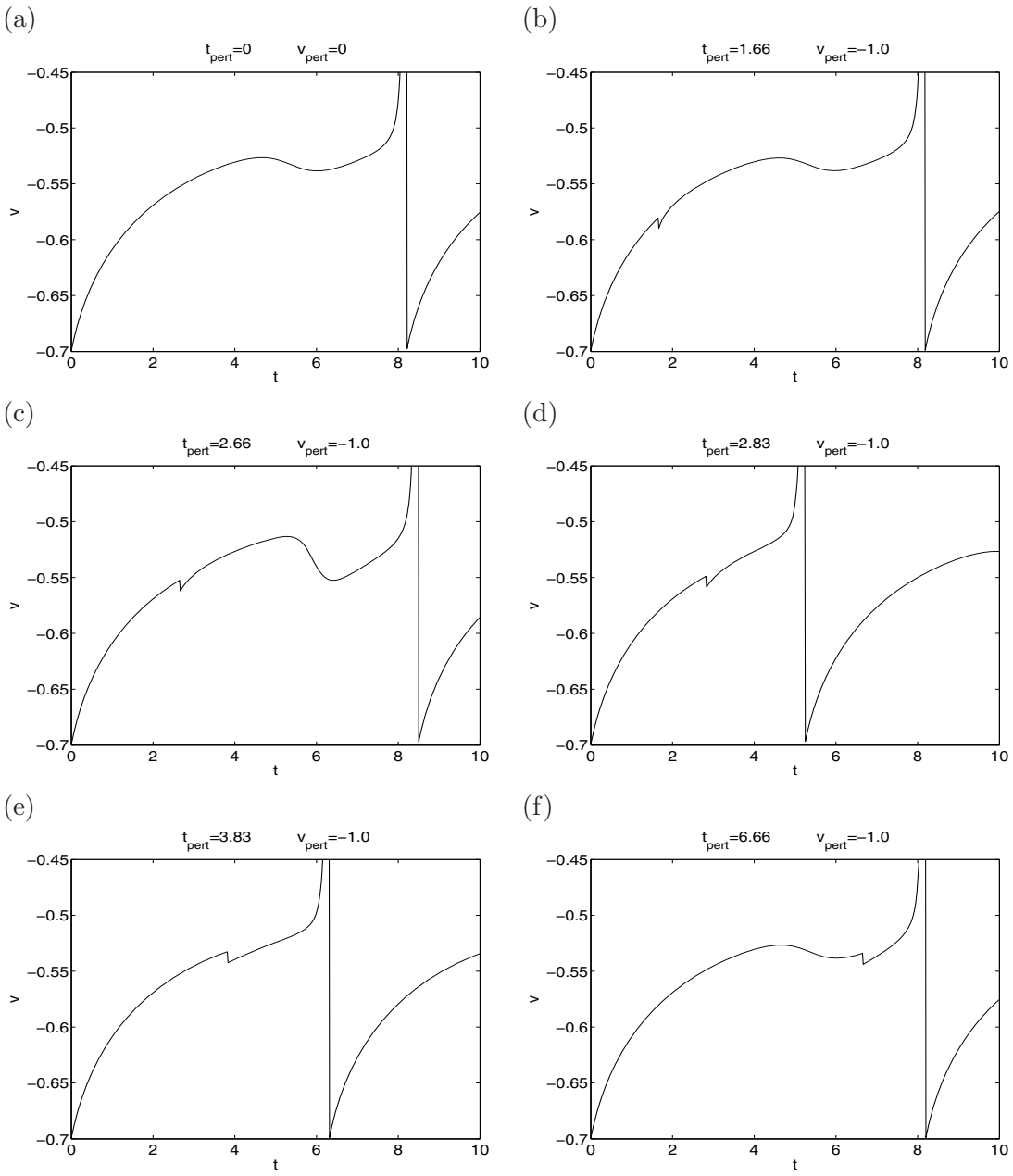


Figure 14. *Voltage traces showing the effect of inhibitory perturbations at various times t_{pert} on the mixed-mode oscillatory patterns corresponding to trajectories starting in a vicinity of the slow manifold for $G_h = 1.5$ (1.0), $I_{app} = -2.3$ (-0.0153). The dimensionless values of G_h and I_{app} are given in parentheses.*

the computational study of synchronization properties of networks including SCs [12, 10, 26].

In [10] we initiated a more detailed mechanistic study of the generation of STOs and MMOs in the 7D SC model proposed in [12]. We considered there a rather restricted set of parameters. Using reduction-of-dimensions techniques, we argued that this model can

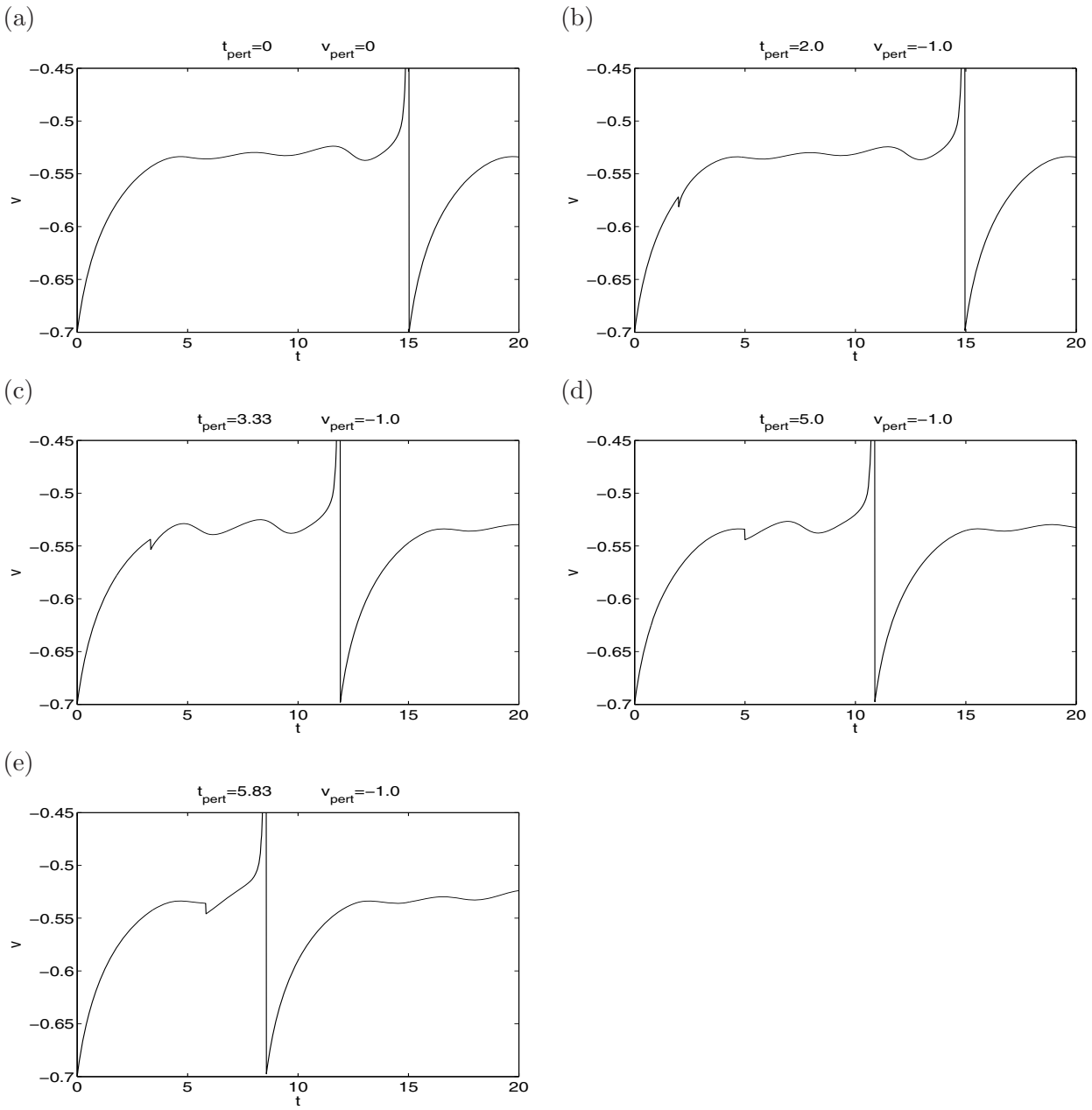


Figure 15. *Voltage traces showing the effect of inhibitory perturbations at various times t_{pert} on the mixed-mode oscillatory patterns corresponding to trajectories starting in a vicinity of the slow manifold for $G_h = 1.5$ (1.0), $I_{app} = -2.4$ (-0.0160). The dimensionless values of G_h and I_{app} are given in parentheses.*

be approximated in the subthreshold regime by the 3D slow-fast system (1)–(3) studied in this paper (the dimensional SC model), which accounts for the interactions between I_p and I_h. Using computational techniques and dynamical systems ideas, we hypothesized that this interaction is intrinsically nonlinear and that the generation of STOs and MMOs is governed

by the 3D canard phenomenon.

In this paper we used analytical and computational tools to show that the 3D canard phenomenon is the main player in the mechanism of generation of STOs in the SC model (1)–(3) for a broad range of biophysically plausible parameter values that include those considered in [10]. The underlying dynamic structure also describes the onset of spikes but not the spike dynamics. Once the oscillating trajectory escapes the subthreshold regime to the spiking one, the reset properties of the h-current provide the additional mechanism needed for the trajectory to return to the subthreshold regime and produce MMOs.

From the mathematical point of view, our results show that the SC model (1)–(3) satisfies the conditions required by the theorems proposed in [17] (see section 4.4) to guarantee the existence of MMOs. Thus, the SC model becomes a biologically plausible example of the theory developed in [17]. Care has to be taken in the limiting case of a folded saddle-node ($\mu \to 0$), since the corresponding theory has still to be developed. Nonetheless, trends can be deduced from the folded node analysis and by sufficiently decreasing the singular perturbation parameter as shown in section 4.8. Examples related to other phenomena include those studied in [16, 45].

From the mechanistic point of view, our results provide an analytical and geometric framework to study various dynamic aspects of SCs: the role that the participating currents and their interaction play in shaping the observed MMO patterns, some of the consequences of these interactions when SCs receive external inputs (sinusoidal or synaptic-like), and the prediction of the effect that different amounts of the participating currents may have on the observed MMO patterns. These differences could be due to a heterogeneous distribution of channels in the EC, to the effect of neuromodulators, or to changes due to development.

A very important component of the framework we refer to above is the singular funnel. By calculating the strong canard it is possible to make qualitative predictions about whether STOs are expected within the spiking pattern or not, i.e., which initial conditions in the subthreshold regime lead to MMOs and which lead to fully spiking patterns. One can be more explicit about the precise number of STOs by further calculating the secondary canards. Changes in the parameters of the model are reflected in the singular funnel via changes in the strong canard, changes in the secondary canards, or changes in the initial conditions in the subthreshold regime due to variations in the trajectory reset values. So, for example, the trajectory corresponding to a specific initial condition and a certain set of parameters may enter the funnel while the trajectory corresponding to the same initial conditions and a different set of parameters may be left outside. The tools required to make more accurate quantitative predictions call for further research.

The results of this paper support our previous hypothesis [10] of the existence of a canard geometric/dynamic structure as the main player in the mechanism of generation of STOs in SCs. This canard structure is generated by the (nonlinear) nullsurfaces and the time scale separation between the participating variables, and it has the potential of producing the canard phenomenon. More specifically, STOs in SCs have been proposed to be sustained by a "push-pull" interplay between I_p and I_h [7, 8]. In particular, I_h has been proposed to play a major pacemaker role in the generation of STOs, providing a delayed feedback mechanism to the voltage changes led by I_p. This mechanism depends on the dynamic (kinetic) properties of I_h. There are various players in this interaction: the amount of I_h and I_p (expressed by the

maximal conductances G_h and G_p), the relative speeds of the voltage and gating variables, and the nonlinearities associated with the dynamic equations. Our results show that these are encoded in the canard structure. Different sets of parameters lead to similar canard structures which produce similar voltage traces: In our model, a decrease in G_h (reflecting a decrease in the amount of I_h) leads to an increase in the "height" of the V-nullsurface. Since the other nullsurfaces do not depend on the parameter of the model, for fixed values of I_{app} and low enough values of G_h, folded nodes no longer exist and trajectories will be attracted to a stable fixed point (see Table 2). However, this can be reversed by increasing the value of I_{app}, which lowers the V-nullsurface. The fact that MMOs with the same number of STOs per spike with roughly the same amplitude are obtained for different pairs of (G_h, I_{app}) and (G_h, G_p) (see Figures 12 and 13) shows that it is not a specific property of I_h and I_p that creates the STOs but rather a property of the various appropriate balances these currents can create, which are reflected in the generation of similar or "equivalent" canard structures. This raises the question of whether internal homeostatic mechanisms exist that keep the number of STOs per spike unchanged. On the other hand, as we mentioned earlier, the fact that the STO frequency changes with I_h shows a way in which the canard mechanism can account for the difference in frequencies experimentally observed in SCs along the ventral-to-dorsal axis of the EC [3].

Spike-time response curve (STRC) techniques have been used to study the synchronization properties of small networks including SCs [12, 25, 26], particularly to investigate how synchronization depends on key ionic currents known to be important to the theta rhythm. STRCs are functions that measure the effect of a spike of a presynaptic cell on the timing of the next spike of the postsynaptic cell. (STRCs are essentially the same as phase response curves [58, 59].) Our results also support the hypothesis that the canard structure is an important component in the mechanism of synchronization in networks including SCs and other excitatory and inhibitory neurons.

There has been some controversy in the literature about whether the observed subthreshold oscillations in SCs are intrinsically linear or nonlinear phenomena [14, 57]. As predicted in [10], our results show that the latter is the most plausible case. Our results also show that the interaction between I_p and I_h responsible for the generation of STOs involves both components of I_h, not only the fast one; i.e., the slow component of I_h does not simply play a modulatory role in the generation of STOs, but rather a dynamic one.

As occurs for many conductance-based models, the concept of a spiking threshold is not well defined in the SC model we study here (see [60] for a discussion on the topic). There are two ways in which spikes are generated in the SC model, leading to patterns whose interspike intervals differ by roughly an order of magnitude. The first type corresponds to initial conditions such that the trajectory is "captured" by the slow manifold S. The spiking period between two such spikes is determined by the time it takes the trajectory to move along the slow manifold. These trajectories may or may not enter the funnel. In either case, the onset of spikes occurs when the trajectory moves away from the slow manifold along a fast fiber towards the spiking regime. Once this occurs, spiking is unavoidable; i.e., the trajectory does not have to cross any voltage threshold value to spike. In the nonlinear artificially spiking model, the value of V_{th} only indicates that a spike has occurred so that it can be (artificially) added to the voltage trace. The second type of spike corresponds to initial conditions such

that the trajectory is never "captured" by the slow manifold S. These initial conditions are above the fold-curve. The period between two such spikes is determined by the time it takes the trajectory to move along a fast fiber (vertical direction); i.e., its order of magnitude is ϵ. As for the first type, once the trajectory enters the subthreshold regime a spike will occur without the need of a voltage threshold value. Spontaneous spikes of the second type are expected to be rare, since reset voltage values are typically lower than that corresponding to the fold-curve. However, trajectories may be forced to change from one spiking regime to the other by external inputs. This may have consequences for network dynamics in the EC.

There are several aspects of the SC dynamics not studied in this paper. One of them is how noise affects the MMO patterns. One important source of noise is the set of persistent sodium channels [13, 61]. In [10] we showed that STOs are less regular and more robust when persistent sodium channel noise is added to the SC model studied here. A second aspect is spike clustering, where two or more spikes occur without interspersed STOs. Noisy MMOs and clustering have been reproduced via simulations in the biophysical (conductance-based) model presented in [13]. Whether or not this model has an underlying canard structure qualitatively similar to the one we uncovered in this paper is still an open question.

Acknowledgments. We thank Tasso Kaper for his useful comments on an earlier draft of this manuscript, and Amit Bose, Morten Brøns, Martin Krupa, and John White for useful discussions.

REFERENCES

[1] D. Johnston and D. G. Amaral, *Hippocampus*, in The Synaptic Organization of the Brain, G. M. Sheperd, ed., Oxford University Press, Lodnon, 2004, pp. 455–498.

[2] H. E. Scharfman, M. P. Witter, and R. Schwarcz, *The Parahippocampal Region: Implications for Neurological and Psychiatric Diseases*, Ann. New York Acad. Sci. 911, New York Academy of Sciences, New York, 2000.

[3] L. M. Giacomo, E. A. Zilli, E. Fransén, and M. E. Hasselmo, *Temporal frequency of subthreshold oscillations scales with entorhinal grid cell field spacing*, Science, 315 (2007), pp. 1719–1722.

[4] M. Fyhn, S. Molden, M. P. Witter, E. I. Moser, and M. B. Moser, *Spatial representation in the entorhinal cortex*, Science, 305 (2004), pp. 1258–1264.

[5] T. Hafting, M. Fyhn, S. Molden, M. B. Moser, and E. I. Moser, *Microstructure of a spatial map in the entorhinal cortex*, Nature, 436 (2005), pp. 801–806.

[6] F. Sargolini, M. Fyhn, T. Hafting, B. McNaughton, M. P. Witter, E. I. Moser, and M. B. Moser, *Conjunctive representation of position, direction, and velocity in the entorhinal cortex*, Science, 312 (2006), pp. 758–762.

[7] C. T. Dickson, J. Magistretti, M. H. Shalinsky, E. Fransén, M. Hasselmo, and A. A. Alonso, *Properties and role of I_h in the pacing of subthreshold oscillation in entorhinal cortex layer* II *neurons*, J. Neurophysiol., 83 (2000), pp. 2562–2579.

[8] A. A. Alonso and R. R. Llinás, *Subthreshold Na^+-dependent theta like rhythmicity in stellate cells of entorhinal cortex layer* II, Nature, 342 (1989), pp. 175–177.

[9] C. T. Dickson, J. Magistretti, M. Shalinsky, B. Hamam, and A. A. Alonso, *Oscillatory activity in entorhinal neurons and circuits*, Ann. New York Acad. Sci., 911 (2000), pp. 127–150.

[10] H. G. Rotstein, T. Oppermann, J. A. White, and N. Kopell, *A reduced model for medial entorhinal cortex stellate cells: Subthreshold oscillations, spiking and synchronization*, J. Comput. Neurosci., 21 (2006), pp. 271–292.

[11] A. A. Alonso and E. García Austt, *Neuronal sources of theta rhythm in the entorhinal cortex of the rat.* II. *Phase relations between unit discharges and theta field potentials*, Exp. Brain Res., 67 (1987), pp. 493–501.

[12] C. D. Acker, N. Kopell, and J. A. White, *Synchronization of strongly coupled excitatory neurons: Relating network behavior to biophysics*, J. Comput. Neurosci., 15 (2003), pp. 71–90.
[13] E. Fransén, A. A. Alonso, C. T. Dickson, M. E. Magistretti, and J. Hasselmo, *Ionic mechanisms in the generation of subthreshold oscillations and action potential clustering in entorhinal layer* II *stellate neurons*, Hippocampus, 14 (2004), pp. 368–384.
[14] S. Schreiber, I. Erchova, U. Heinemann, and A. V. Herz, *Subthreshold resonance explains the frequency-dependent integration of periodic as well as random stimuli in the entorhinal cortex*, J. Neurophysiol., 92 (2004), pp. 408–415.
[15] P. Szmolyan and M. Wechselberger, *Canards in R^3*, J. Differential Equations, 177 (2001), pp. 419–453.
[16] M. Wechselberger, *Existence and bifurcation of canards in R^3 in the case of a folded node*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 101–139.
[17] M. Brøns, M. Krupa, and M. Wechselberger, *Mixed mode oscillations due to the generalized canard phenomenon*, in Bifurcation Theory and Spatio-temporal Pattern Formation, W. Nagata and N. Sri Namachivaya, eds., Fields Institute Communications 49, AMS, New York, 2006, pp. 39–63.
[18] R. Larter and C. G. Steinmetz, *Chaos via mixed mode oscillations*, Philos. Trans. R. Soc. London A, 337 (1991), pp. 291–298.
[19] A. Arneado, F. Argoul, J. Elezgaray, and P. Richetti, *Homoclinic chaos in chemical systems*, Phys. D, 62 (1993), pp. 134–168.
[20] N. Kopell, *Toward a theory of modelling central pattern generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. H. Cohen, S. Rossignol, and S. Grillner, eds., Wiley, New York, 1988, pp. 369–413.
[21] J. Guckenheimer, R. Harris-Warrick, J. Peck, and A. Willms, *Bifurcation, bursting and spike frequency adaptation*, J. Comput. Neurosci., 4 (1997), pp. 257–277.
[22] J. Guckenheimer and A. R. Willms, *Asymptotic analysis of subcritical Hopf-homoclinic bifurcation*, Phys. D, 139 (2000), pp. 159–216.
[23] M. Krupa, N. Popovic, N. Kopell, and H. G. Rotstein, *Mixed-mode oscillations in a three time-scale model for the dopaminergic neuron*, Chaos, 18 (2008), 015106.
[24] M. Brøns, T. J. Kaper, and H. G. Rotstein, *Introduction to focus issue: Mixed mode oscillations: Experiment, computation, and analysis*, Chaos, 18 (2008), 015101.
[25] T. I. Netoff, M. I. Banks, A. D. Dorval, C. D. Acker, J. S. Haas, N. Kopell, and J. A. White, *Synchronization in hybrid neuronal networks of the hippocampal formation*, J. Neurophysiol., 93 (2005), pp. 1197–1208.
[26] D. D Pervouchine, T. I. Netoff, H. G. Rotstein, J. A. White, M. O. Cunningham, M. A. Whittington, and N. Kopell, *Low-dimensional maps encoding dynamics in entorhinal cortex and hippocampus*, Neural Computation, 18 (2006), pp. 2617–2650.
[27] D. Johnston and S. M.-S. Wu, *Foundations of Cellular Neurophysiology*, The MIT Press, Cambridge, MA, 1995.
[28] R. M. Klink and A. A. Alonso, *Ionic mechanisms of muscarinic depolarization in entorhinal cortex layer* II *neurons*, J. Neurophysiol., 77 (1997), pp. 1829–1843.
[29] J. Magistretti and A. A. Alonso, *Biophysical properties and slow voltage-dependent inactivation of a sustained sodium current in entorhinal cortex layer*-II *principal neurons. A whole-cell and single-channel study*, J. Gen. Physiol., 114 (1999), pp. 491–509.
[30] E. Fransén, C. T. Dickson, J. Magistretti, A. A. Alonso, and M. E. Hasselmo, *Modeling the generation of subthreshold membrane potential oscillations of entorhinal cortex layer* II *stellate cells*, Soc. Neurosci. Abstr., 24 (1998), 814.815.
[31] E. Fransén, G. V. Wallestein, A. A. Alonso, C. T. Dickson, and M. E. Hasselmo, *A biophysical simulation of intrinsic and network properties of entorhinal cortex*, Neurocomputing, 26–27 (1999), pp. 375–380.
[32] R. B. Robinson and S. A. Siegelbaum, *Hyperpolarization-activated cation currents: From molecules to physiological function*, Annu. Rev. Physiol., 65 (2003), pp. 453–480.
[33] B. Hutcheon and Y. Yarom, *Resonance oscillations and the intrinsic frequency preferences in neurons*, Trends in Pharmacological Sciences, 23 (2000), pp. 216–222.
[34] M. J. E. Richardson, N. Brunel, and V. Hakim, *From subthreshold to firing-rate resonance*, J. Neurophysiol., 89 (2003), pp. 2538–2554.

[35] E. M. Izhikevich, *Resonate-and-fire neurons*, Neural Networks, 14 (2001), pp. 883–894.
[36] M. Wechselberger, *Canards*, Scholarpedia, 2 (2007), p. 1356; online at http://www.scholarpedia.org/article/Canards.
[37] E. Benoit, J. L. Callot, F. Diener, and M. Diener, *Chasse au Canard*, Collect. Math., 31 (1981), pp. 1–3, 37–119.
[38] W. Eckhaus, *Relaxation oscillations including a standard chase on French ducks*, in Lecture Notes in Math. 985, Springer-Verlag, Berlin, 1983, pp. 449–497.
[39] F. Dumortier and R. Roussarie, *Canard cycles and center manifolds*, Mem. Amer. Math. Soc., 121 (1996), pp. 1–100.
[40] M. Krupa and P. Szmolyan, *Relaxation oscillation and canard explosion*, J. Differential Equations, 174 (2001), pp. 312–368.
[41] M. Brøns and K. Bar-Eli, *Canard explosion and excitation in a model of the Belousov-Zhabotinsky reaction*, J. Phys. Chem., 95 (1991), pp. 8706–8713.
[42] B. Peng, V. Gaspar, and K. Showalter, *False bifurcations in chemical systems: Canards*, Philos. Trans. R. Soc. London A, 337 (1991), pp. 275–289.
[43] V. A. Makarov, V. I. Nekorkin, and M. G. Velarde, *Spiking behavior in a noise-driven system combining oscillatory and excitatory properties*, Phys. Rev. Lett., 15 (2001), pp. 3031–3034.
[44] C. B. Muratov and E. Vanden-Eijnden, *Noise-induced mixed-mode oscillations in a relaxation oscillator near the onset of a limit cycle*, Chaos, 18 (2008), 015111.
[45] J. Rubin and M. Wechselberger, *Giant squid-hidden canard: The 3d geometry of the Hodgkin-Huxley model*, Biol. Cybern., 97 (2007), pp. 5–32.
[46] A. Milik and P. Szmolyan, *Multiple time scales and canards in a chemical oscillator*, in Multiple Time-Scale Dynamical systems, IMA Volume 122, C. K. R. T. Jones and A. Khibnik, eds., Springer, New York, 2000, pp. 117–140.
[47] A. Milik, P. Szmolyan, H. Löffelmann, and E. Groller, *Geometry of mixed-mode oscillations in the 3d-autocatalator*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 8 (1998), pp. 505–519.
[48] J. Drover, J. Rubin, J. Su, and B. Ermentrout, *Analysis of a canard mechanism by which excitatory synaptic coupling can synchronize neurons at low firing frequencies*, SIAM J. Appl. Math., 65 (2004), pp. 69–92.
[49] L. P. Shilnikov, *A case of the existence of a denumerable set of periodic motions*, Sov. Math. Dokl., 6 (1965), pp. 163–166.
[50] H. Richter, R. Klee, U. Heinemann, and C. Eder, *Developmental changes of inward rectifier currents in neurons of the rat entorhinal cortex*, Neurosci. Lett., 228 (1997), pp. 139–141.
[51] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–225.
[52] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Lecture Notes in Math. 1609, Springer, New York, 1994, pp. 44–118.
[53] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems. A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.
[54] R. L. Burden and J. D. Faires, *Numerical Analysis*, PWS Publishing, Boston, 1980.
[55] M. Wechselberger and W. Weckesser, *Bifurcations of Mixed-Mode Oscillations in a Stellate Cell Model*, preprint, School of Mathematics and Statistics, University of Sydney, 2008.
[56] H. G. Rotstein, N. Kopell, A. M. Zhabotinsky, and I. R. Epstein, *A canard mechanism for localization in systems of globally coupled oscillators*, SIAM J. Appl. Math., 63 (2003), pp. 1998–2019.
[57] J. S. Haas and J. A. White, *Frequency selectivity of layer* II *stellate cells in the medial entorhinal cortex*, J. Neurophysiol., 88 (2002), pp. 2422–2429.
[58] J. Winson, *Loss of hippocampal theta rhythm results in spatial memory deficit in the rat*, Science, 201 (1978), pp. 160–163.
[59] G. B. Ermentrout, M. Pascal, and B. Gutkin, *The effects of spike frequency adaptation and negative feedback on the synchronization of neural oscillators*, Neural Computat., 13 (2001), pp. 1285–1310.
[60] E. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press, Cambridge, MA, 2007.
[61] J. A. White, R. Klink, A. Alonso, and A. R. A. Kay, *Noise from voltage-gated ion channels may influence neuronal dynamics in the entorhinal cortex*, J. Neurophysiol., 80 (1998), pp. 262–269.