SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 2, No. 1, pp. 1–35

The Forced van der Pol Equation I: The Slow Flow and Its Bifurcations*

John Guckenheimer†, Kathleen Hoffman‡, and Warren Weckesser§

Abstract. The forced van der Pol oscillator has been the focus of scientific scrutiny for almost a century, yet its global bifurcation structure is still poorly understood. In this paper, we present a hybrid system consisting of the dynamics of the trajectories on the slow manifold coupled with "jumps" at the folds in the critical manifold to approximate the fast subsystem. The global bifurcations of the fixed points and periodic points of this hybrid system lead to an understanding of the bifurcations in the periodic orbits (without canards) of the forced van der Pol system.

Key words. van der Pol oscillator, hybrid dynamical system, bifurcations, chaotic attractor, periodic solutions

AMS subject classifications. 37C10, 37C27, 37G15

PII. S1111111102404738

1. Introduction.

During the first half of the twentieth century, Balthazar van der Pol pioneered the fields of radio and telecommunications [6, 7, 29, 32, 33, 34]. In an era when these areas were much less advanced than they are today, vacuum tubes were used to control the flow of electricity in the circuitry of transmitters and receivers. Contemporary with Lorenz, Thompson, and Appleton, van der Pol experimented with oscillations in a vacuum tube triode circuit and concluded that all initial conditions converged to the same periodic orbit of finite amplitude. Since this behavior is different from the behavior of solutions of linear equations, van der Pol proposed a nonlinear differential equation

$$x'' + \mu(x^2 - 1)x' + x = 0, \tag{1.1}$$

commonly referred to as the (unforced) van der Pol equation [32], as a model for the behavior observed in the experiment. In studying the case $\mu \gg 1$, van der Pol discovered the importance of what has become known as *relaxation oscillations* [33]. These oscillations have become the cornerstone of geometric singular perturbation theory and play a significant role in the analysis presented here. Van der Pol went on to propose a version of (1.1) that includes a periodic forcing term:

$$x'' + \mu(x^2 - 1)x' + x = a\sin(2\pi\nu\tau). \tag{1.2}$$

In a similar equation, he and van der Mark first noted the existence of two stable periodic solutions with different periods for a particular value of the parameters and observed noisy

*Received by the editors April 1, 2002; accepted for publication (in revised form) by M. Golubitsky September 24, 2002; published electronically February 19, 2003.
http://www.siam.org/journals/siads/2-1/40473.html

†Mathematics Department, Cornell University, Ithaca, NY 14853 (gucken@com.cornell.edu).
‡Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250 (khoffman@math.umbc.edu).
§Mathematics Department, Colgate University, Hamilton, NY 13346 (wweckesser@mail.colgate.edu).

behavior in an electrical circuit modeled with (1.2) [35]. Van der Pol further speculated that (1.2) also had this property.

Van der Pol's work on nonlinear oscillations and circuit theory provided motivation for the seminal work of Cartwright and Littlewood [22]. In 1938, just prior to World War II, the British Radio Research Board issued a request for mathematicians to consider the differential equations that arise in radio engineering. Responding to this request, Cartwright and Littlewood began studying the forced van der Pol equation and showed that it does indeed have bistable parameter regimes. In addition, they showed that there does not exist a smooth boundary between the basins of attraction of the stable periodic orbits. They discovered what is now called chaotic dynamics by detailed investigation of this system [7, 8, 9, 20, 21].

Since its introduction in the 1920's, the van der Pol equation has been a prototype for systems with self-excited limit cycle oscillations. The equation has been studied over wide parameter regimes, from perturbations of harmonic motion to relaxation oscillations. It has been used by scientists to model a variety of physical and biological phenomena. For instance, in biology, the van der Pol equation has been used as the basis of a model of coupled neurons in the gastric mill circuit of the stomatogastric ganglion [15, 26]. The Fitzhugh–Nagumo equation [12] is a planar vector field that extends the van der Pol equation as a model for action potentials of neurons [18]. In seismology, the van der Pol equation has been used in the development a model of the interaction of two plates in a geological fault [5].

Despite the continuing work of many scientists and mathematicians (see [13, 16, 17, 19, 26, 28, 30, 31], for example), bifurcations of this system have been studied little. This paper is the first in a series that seeks to give a thorough analysis of the dynamics inherent in the forced van der Pol equation in the relaxation regime. Here we focus upon a two-dimensional reduced system derived from the forced van der Pol equation, classifying the bifurcations of the simplest periodic orbits in this reduced system.

2. Fast subsystems and the slow flow. We begin by setting notation and defining the fast subsystems and slow flow of the forced van der Pol equation. First, we change variables to express (1.2) in a more convenient form by rescaling time $t = \tau/\mu$, so $x' \to \dot{x}/\mu$ and $x'' \to \ddot{x}/\mu^2$, where $\dot{x} \equiv dx/dt$. By defining $y = \dot{x}/\mu^2 + x^3/3 - x$, we transform (1.2) into the system

$$\begin{aligned} \frac{1}{\mu^2}\dot{x} &= y - \frac{x^3}{3} + x, \\ \dot{y} &= -x + a\sin(2\pi\nu\mu t). \end{aligned}$$

Defining new parameters $\varepsilon = 1/\mu^2$ and $\omega = \nu\mu$ and converting these equations into an autonomous system by defining $\theta = \omega t$, we obtain the system

$$\begin{aligned} \varepsilon\dot{x} &= y + x - \frac{x^3}{3}, \\ \dot{y} &= -x + a\sin(2\pi\theta), \\ \dot{\theta} &= \omega \end{aligned} \tag{2.1}$$

as a vector field on $\mathbb{R}^2 \times S^1$. We regard $S^1 = \mathbb{R}/\mathbb{Z}$ and use coordinates $[0, 1]$ for S^1, understanding that the endpoints are identified. We also use the system in the form

$$\begin{aligned} \frac{dx}{ds} &= y + x - \frac{x^3}{3}, \\ \frac{dy}{ds} &= \varepsilon(-x + a\sin(2\pi\theta)), \\ \frac{d\theta}{ds} &= \varepsilon\omega, \end{aligned} \tag{2.2}$$

obtained by rescaling time by $s = \varepsilon t$. In the relaxation regime, namely, $\varepsilon \ll 1$, these equations can be analyzed using singular perturbation theory [14]. The variable x is commonly referred to as the *fast variable*, and the variables y and θ are the *slow variables*. Exploiting the two time scales in (2.1) and (2.2), trajectories are decomposed into fast and slow segments. This decomposition is achieved by considering the dynamics of the singular limit $\varepsilon = 0$ in (2.1) and (2.2).

The limit $\varepsilon = 0$ in (2.2) results in the family of *fast subsystems*. For each y and θ, the fast subsystem is a one-dimensional differential equation for x, with y and θ acting as parameters. If $|y| > 2/3$, there is a single stable equilibrium, and if $|y| < 2/3$, there are two stable and one unstable equilibria of the fast subsystem. The forward limit of each trajectory is one of the stable equilibria. The singular limit $\varepsilon = 0$ of (2.1) results in a differential algebraic equation. The first equation in (2.1) becomes $y = x^3/3 - x$, which defines a two-dimensional manifold called the *critical manifold*. The critical manifold is the union of equilibria of the fast subsystems. The remaining two equations of (2.1) define implicitly a vector field on the critical manifold at regular points of its projection onto the (y, θ) coordinate plane. This vector field on the critical manifold is called the *slow flow* of the system. We call the pieces of the critical manifold composed of stable equilibria of the fast subsystems the *stable sheets* (there are two disjoint stable sheets—one where $x > 1$ and one where $x < 1$), and we call the piece composed of unstable equilibria the *unstable sheet* (where $|x| < 1$).

The relationship between the full system ($\varepsilon \neq 0$, (2.1)) and the slow flow was studied by Tikhonov [2, 24] and later by Fenichel [11]. Fenichel proved that there exists a *slow manifold*, that is, an invariant manifold, within distance $O(\varepsilon)$ of the critical manifold on compact regions of regular points. Moreover, the slow flow on the critical manifold gives $O(\varepsilon)$ approximations to trajectories of the forced van der Pol equation.

For $\varepsilon \neq 0$, systems (2.1) and (2.2) are equivalent, but the limit systems as $\varepsilon \to 0$ are quite different. The trajectories of (2.1) consist of segments for which the fast subsystem is a good approximation to the dynamics and segments for which the slow subsystem is a good approximation to the dynamics. The transitions between these segments occur at folds. Folds are the singular points of the projection of the critical manifold onto the (y, θ) plane. For the forced van der Pol equation, these folds occur at $x = \pm 1, y = \mp 2/3$. As $\varepsilon \to 0$, many solutions of the full system approach curves that are concatenations of trajectories of the slow flow and trajectories of the fast subsystem, joined at fold points of the critical manifold.

Figure 2.1 shows a trajectory of the full system ((2.1), $\varepsilon \neq 0$) along with the critical manifold and the slow flow on that manifold. This figure illustrates the roles of the fast subsystems and slow flow in understanding solutions to the full system. The yellow curve is a

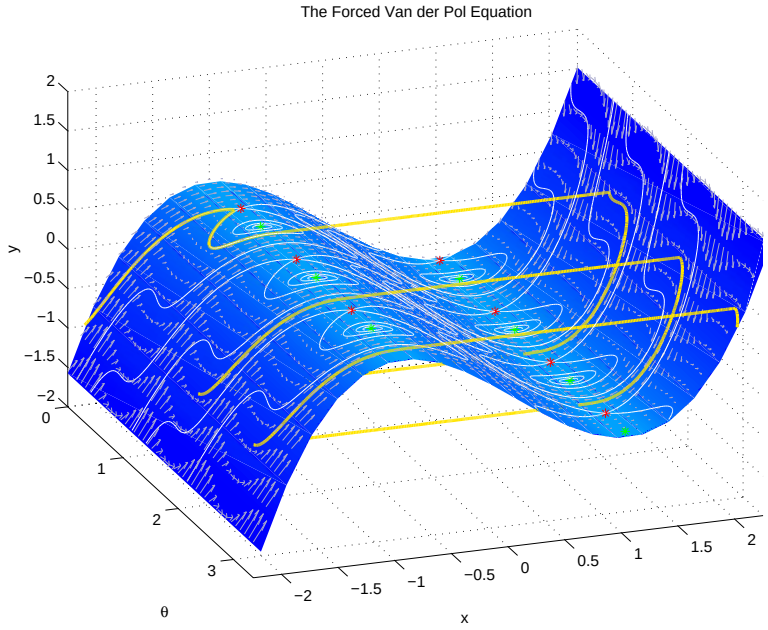
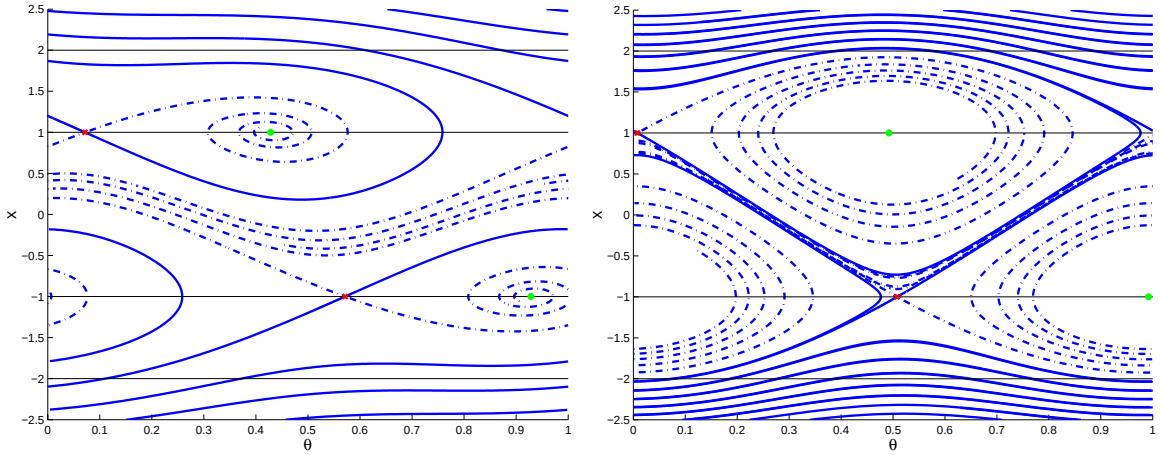


Figure 2.1. *The two-dimensional surface illustrates the critical manifold of the forced van der Pol equations* (2.1)*. In this example,* $a = 2.3$ *and* $\omega = 1$*. The white arrows are the vector field of the slow flow on the critical manifold, and the white curves are corresponding trajectories of that vector field. The green and red asterisks located at the folds of the critical manifold are the folded singularities, foci and saddles, respectively. The yellow curve represents a solution to* (2.1) *for* $\varepsilon = 10^{-4}$*.*



(a) (b)

Figure 2.2. *These are examples of the slow flow. In* (a)*, the parameters are* $a = 2.3$ *and* $\omega = 1$*, the same as in Figure* 2.1*. In* (b)*,* $a = 20$ *and* $\omega = 5$*, which are the same values shown in Figures* 4.1(a) *and* 4.2*. The plots show the stable and unstable manifolds of the folded saddles, along with the circles* $x = \pm 1$ *and* $x = \pm 2$*.*

solution to the full system (2.1), with $\varepsilon = 10^{-4}$, $a = 2.3$, and $\omega = 1$. The blue two-dimensional surface is the critical manifold, and the white curves and arrows show the trajectories of the slow flow and its vector field on the critical manifold, respectively. Notice that the solution to the full system appears to follow the slow flow until it reaches a fold in the critical manifold, where it then appears to follow the fast subsystem until again reaching the critical manifold. The red and green asterisks mark *folded singularities*, points on the fold curves of the critical manifold that play a central role in our analysis of bifurcations. Figure 2.2 gives examples of the slow flow for two sets of parameter values. The folds of the critical manifold are the lines $x = \pm 1$, and the "landing points" of the jumps from the folds (that is, the projections of the fold lines along the fast variable onto the critical manifold) are the lines $x = \pm 2$.

The next section defines a reduced hybrid system whose solutions come from piecewise smooth curves that approximate solutions of the system (2.1) and consist of segments that are solutions to the fast subsystems and slow flow.

3. The reduced system. The main idea of this paper is to study the global bifurcations of the forced van der Pol equation by studying a *reduced system* Ψ_h, a hybrid dynamical system that combines solutions to the slow flow of (2.1) with discrete time transformations along trajectories of the fast subsystems. Specifically, the reduced system is defined to follow the slow flow on the stable sheets $|x| \geq 1$ of the critical manifold and to have discrete time jumps from $(\theta, \pm 1)$ to $(\theta, \mp 2)$ on the boundaries of the two stable sheets. The boundaries of the two stable sheets are *folds* of the projection of the critical manifold (see Figure 2.1) onto the two-dimensional space of slow variables. Trajectories of the system Ψ_h give approximations to many trajectories of the full three-dimensional flow. We explain the nature of this approximation in section 4.

The projection $\pi(x, y, \theta) = (y, \theta)$ of the critical manifold onto the (θ, y) space of slow variables is singular on the *fold* curves $x = \pm 1, y = \mp 2/3$. Due to these singularities, we use (θ, x) as coordinates for the slow flow, with (θ, x) lying in one of the two closed half-cylinders $x \geq 1$ or $x \leq 1$. We reparametrize the slow flow to obtain the vector field

$$\begin{aligned} \theta' &= \omega(x^2 - 1), \\ x' &= -x + a\sin(2\pi\theta). \end{aligned} \tag{3.1}$$

The vector field (3.1) is obtained from the forced van der Pol equation (2.1) by differentiating the algebraic equation $y = x^3/3 - x$ to obtain $\dot{y} = (x^2 - 1)\dot{x}$, substituting the result into the van der Pol equation (2.1), and rescaling the time by $(x^2 - 1)$. We shall henceforth call the system (3.1) the slow flow since it has the same trajectories as the original slow flow on the stable sheets of the critical manifold. However, there are several ways in which this rescaled system (3.1) differs from the slow flow in its original coordinates. First, it is defined on the fold curves, while the original equations are not due to the singularity of the projection π. Second, it reverses time on the unstable sheet, represented by the cylinder $|x| < 1$ in the (θ, x) coordinates. Therefore, the direction of the slow flow on the unstable sheet of the critical manifold is opposite that defined by (3.1). Note that the system (3.1) has equilibrium points on the circles $x = \pm 1$ if $a \geq 1$. These equilibria are called *folded equilibria*. They approximate points where the flow of the van der Pol equation (2.1) is tangent to the fold curves. The van der Pol equation (2.1) has no equilibria.

We establish a few basic properties of the slow flow. If $x > a > 0$, then $x' < 0$, while if $x < -a < 0$, then $x' > 0$. Consequently, the region $|x| < a$ is forward invariant for system (3.1), and all trajectories flow into this region. The vector field is symmetric with respect to the symmetry given by $T(\theta, x) = (\theta + \frac{1}{2}, -x)$. We also observe that the divergence of the vector field is constant, namely, -1. Therefore, its flow can have at most a single periodic orbit. Periodic orbits must be T-symmetric (otherwise, there would be a symmetric partner) and consequently cross $x = 0$. The equilibrium points of the slow flow lie on the circles $x = \pm 1$. We conclude that all trajectories with initial conditions satisfying $|x| > 1$ reach the circles $x = \pm 1$ unless they lie in the stable manifold of an equilibrium point, in which case they approach $x = \pm 1$ as $t \to \infty$.

The equilibrium points of the slow flow lie at the points $(\theta, x) = (\pm \sin^{-1}(1/a)/2\pi, \pm 1)$. Here $\sin^{-1}$ is regarded to be a double valued "function" on $(-1, 1)$. If $a = 1$, there are two equilibrium points; if $a > 1$, there are four equilibrium points. The linear stability of the equilibria is easily computed [28] using the Jacobian of the slow flow equations:

$$\begin{pmatrix} -1 & 2\pi a \cos(2\pi\theta) \\ 2\omega x & 0 \end{pmatrix}.$$

At $a = 1$, the equilibria are saddle-nodes. For $a > 1$, two of the equilibria are saddles. In the parameter interval $1 < a < \sqrt{1 + 1/(16\pi\omega)^2}$, the two remaining equilibria are stable nodes. When $a = \sqrt{1 + 1/(16\pi\omega)^2}$, these equilibria are resonant with a single negative eigenvalue, and when $a > \sqrt{1 + 1/(16\pi\omega)^2}$, they are stable foci. We see below that some global bifurcations of Ψ_h depend upon the type of stable equilibrium points found in the slow flow.

We label the coordinates of the saddle point on $x = 1$ as $(\theta, x) = (\theta_{1s}, 1)$ and the coordinates of the node as $(\theta_{1n}, 1)$. The unstable and stable manifolds of the saddle will be denoted by the standard notation W_u and W_s, respectively. We further define the point $p_{1u} = (\theta_{1u}, 1)$ to be the first intersection of the unstable manifold W_u of $(\theta_{1s}, 1)$ with $x = 1$ and set θ_{2si} to be the θ coordinate of the ith intersection of the stable manifold W_s of the saddle $(\theta_{1s}, 1)$ with $x = 2$ as W_s is traversed backward from $x = 1$ into the region $x > 1$. This notation will be used in our description of the properties of the hybrid system Ψ_h and its bifurcations. Figure 4.1 displays phase portraits of the slow flow for two sets of parameter values.

4. Return maps. When $\varepsilon > 0$ is small, trajectories of the van der Pol equation (2.1) with initial conditions near the stable sheets of the critical manifold that do not pass close to the folded singularities are approximated by trajectories of the reduced system Ψ_h [19]. Consequently, bifurcations of the reduced system identify the location of parameters at which bifurcations of the van der Pol equation are expected. Our primary goal in this paper is to identify bifurcations of the reduced system Ψ_h. (Later papers in this series will investigate the relationship between bifurcations of Ψ_h and those of the van der Pol equation (2.1).) In this section, we develop and describe a return map from the circle S_2 (defined by the equation $x = 2$) to itself as a means of describing periodic solutions to the hybrid system Ψ_h.

4.1. Derivation of the half-return map H. Using the observations in the last section, we define two maps $P_\pm$ along trajectories from the circles $S_{\pm 2}$ defined by $x = \pm 2$ to the circles $S_{\pm 1}$ defined by $x = \pm 1$. At points that are not in the stable manifolds of the equilibria, the maps $P_\pm$ are well defined and smooth. At transverse intersections of the stable manifolds of

the saddles with $S_{\pm 2}$, the maps $P_\pm$ are discontinuous. At the circles $x = \pm 1$ on the critical manifold, trajectories of the forced van der Pol equation jump from a fold to the circles $x = \mp 2$ on the critical manifold. For the two-dimensional reduced system, we define the operators $J_+(\theta, 1) = (\theta, -2)$ and $J_-(\theta, -1) = (\theta, 2)$ that describe the discrete jumps in Ψ_h. The return map for Ψ_h to the circle S_2 is then given by the composition $J_-P_-J_+P_+$ since

$$\begin{aligned} P_+ &: S_2 \to S_1, \\ J_+ &: S_1 \to S_{-2}, \\ P_- &: S_{-2} \to S_{-1}, \\ J_- &: S_{-1} \to S_2. \end{aligned}$$

Actually, the map $J_-P_-J_+P_+$ is the perfect square of a half-return map H. To define H, we use the symmetry operator $T(\theta, x) = (\theta + \frac{1}{2}, -x)$. We note that the slow flow and the reduced system Ψ_h are symmetric with respect to the transformation T, T^2 is the identity on $S^1 \times \mathbb{R}$, and the following relations hold: $TP_+ = P_-T$ and $TJ_+ = J_-T$. Then the return map $J_-P_-J_+P_+ = J_-P_-TTJ_+P_+ = (TJ_+P_+)(TJ_+P_+)$ is the square of the map $H = (TJ_+P_+)$ on the circle S_2. Consequently, the periodic orbits of Ψ_h can be divided into those that are fixed by the half-return map H and those that are not. Because T phase shifts θ by $\frac{1}{2}$, the fixed points of H all yield T-symmetric periodic orbits that make exactly two jumps—one from $x = 1$ to $x = -2$ and one from $x = -1$ to $x = 2$. In this paper, we study the bifurcations of periodic orbits of the reduced system Ψ_h with a focus upon fixed points and period 2 points of the half-return map H. These bifurcations correspond to bifurcations of periodic orbits in the forced van der Pol oscillator that are T-symmetric and have just two jumps. The stable periodic orbits studied by Cartwright and Littlewood [9, 20, 21] are in this class: they are T-symmetric with two jumps and yield fixed points of H.

4.2. Properties of H. The half-return map H depends on the two parameters a and ω. The lines $a = 1$ and $a = 2$ divide the a-ω plane into regions in which H has fundamentally different properties. The map P_+ is a diffeomorphism of the circle S_2 to the circle S_1 for $0 < a < 1$. In this regime, x decreases along all trajectories in the strip $1 < x < 2$, implying that H is a circle diffeomorphism. Its rotation number depends upon ω, increasing with ω. All rotation numbers in $[\frac{1}{2}, \infty)$ are realized as ω varies in $(0, \infty)$.

When $1 < a < 2$, the map P_+ no longer maps the circle S_2 onto the circle S_1. Its image I_1 excludes the portion of S_1 that lies below the unstable manifold W_u defined in the previous section. The discontinuities in the domain of P_+ occur at points in $W_s \cap S_2$. There is a single point of discontinuity since the circle S_2 is a cross-section for the flow and W_s crosses S_2 only once. It also follows that the map P_+ remains increasing in this parameter regime. Thus H is a family of increasing maps of the circle into itself with a single point of jump discontinuity in this parameter regime. This implies that H still has a well-defined rotation number, and the period of all of its periodic orbits is the denominator of the rotation number. Quasi-periodic trajectories are still possible, but the set of parameter values yielding quasi-periodic trajectories is likely to have measure zero [16].

When $2 < a$, the map P_+ is no longer monotone. There are two points $p_{2l} = (\theta_{2l}, 2) = (\frac{1}{2\pi}\sin^{-1}(\frac{2}{a}), 2)$ and $p_{2r} = (\theta_{2r}, 2) = (\frac{1}{2} - \frac{1}{2\pi}\sin^{-1}(\frac{2}{a}), 2)$ at which P_+ has a local maximum

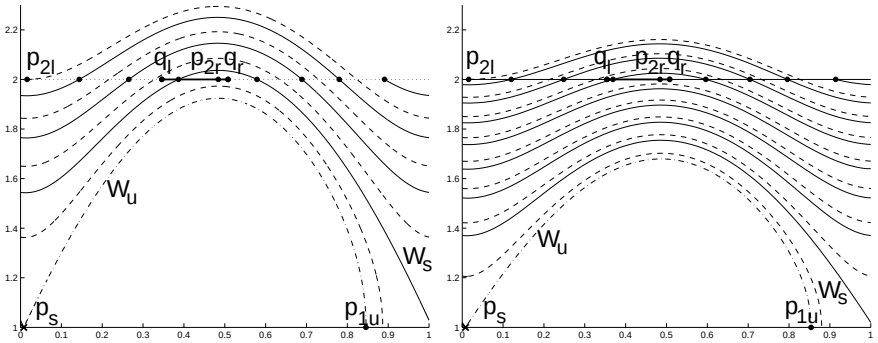


Figure 4.1. *The structure of the reduced system in the half-cylinder* $1 < x$ *for* (a) $(\omega, a) = (5, 20)$ *and* (b) $(\omega, a) = (10, 20)$. *Unstable manifolds* W_u *are drawn with dot-dash curves, stable manifolds* W_s *are drawn solid, the trajectories originating at the points* p_{2l} *are drawn dashed, and the circles* $x = 2$ *are drawn as dotted lines.*

and minimum, respectively. On the interval $D = (\theta_{2l}, \theta_{2r})$, P_+ has negative slope, while on $S^1 - \bar{D}$, it has positive slope. There are two crucial additional aspects to the structure of H as a piecewise continuous and piecewise monotone mapping of the circle. First, there are discontinuities of P_+ at intersections of D with W_s. (There may be only one such intersection point.) At the points of discontinuity in $W_s \cap S_2$, there is a jump with limit values $\frac{1}{2} + \theta_{1s} = \theta_r$ and $\frac{1}{2} + \theta_{1u} = \theta_l$. We denote by q_l and q_r the points $(\theta_l, 2)$ and $(\theta_r, 2)$ in S_2. Second, we observe that the maximum height of W_u is a decreasing function of ω and is unbounded as $\omega \to 0$. Therefore, if $\omega > 0$ is small enough, W_u intersects the circle S_2. When this happens, it divides S_2 into two intervals. The points in S_2 above W_u have their images in $I_H = [q_l, q_r]$, while the points in S_2 below W_u have their images to the left of q_l. (If $0 < \theta_{1s} < \frac{1}{2} < \theta_{1u} < 1$, then $I_H \subset [0, 1]$. Otherwise, if $0 < \theta_{1u} < \frac{1}{2}$, the circular arc I_H contains 0, and it is convenient to choose a fundamental domain for the universal cover of the circle S_2 that contains $[q_l, q_r]$.) Note that W_s lies above W_u.

Figure 4.1 shows the structure of the flow in the strip $1 < x < 2.25$ for $a = 20$, with $\omega = 5$ on the left and $\omega = 10$ on the right. (Note that $\theta = 0$ and $\theta = 1$ are identified, so the flow is actually on a cylinder: $(\theta, x) \in S^1 \times \mathbb{R}$.) The folded saddles p_s are located by the symbol $\times$. Their stable separatrices are drawn as solid curves, and their unstable manifolds are drawn as dot-dashed curves. The circles S_2 are drawn dotted, and the points p_{2l} and p_{2r} are labeled. The dashed trajectories have initial condition p_{2l}. The intervals $I_H = [q_l, q_r]$ that are the images of most branches of H are drawn as thick lines. The points $p_{1u} \in W_u \cap S_1$ are labeled, and the points in $W_s \cap S_2$ are marked by large dots. The graph of the half-return map H for $(\omega, a) = (5, 20)$ is shown in Figure 4.2. The map H is discontinuous at the points of $W_s \cap S_2$ and has a local maximum at p_{2l} and a local minimum at p_{2r}.

The topological theory of one-dimensional maps is based upon partitioning the domain of a map into intervals on which it is continuous and monotone. Here the graph of H can contain the following types of intervals on which it is continuous and monotone:

- a decreasing branch with domain $[p_{2l}, p_{2r}]$ (this occurs if $a > 2$ but W_s intersects S_2 in a single point),

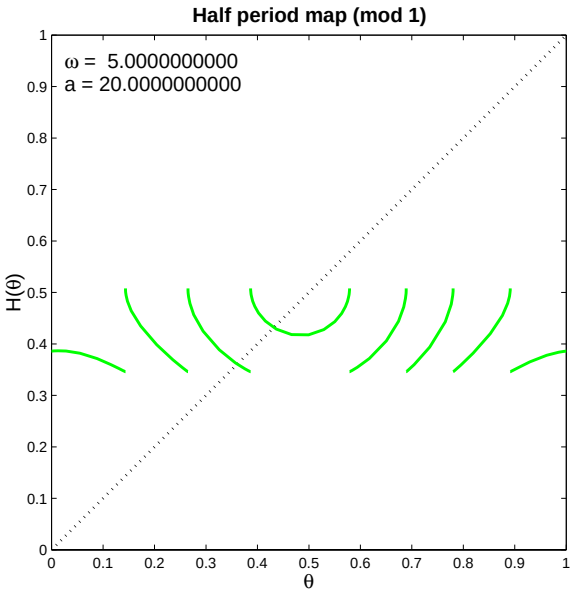


Figure 4.2. *The graph of H when $a = 20$ and $\omega = 5$. These are the same parameter values as in Figure 4.1(a). Note that the discontinuities of H occur at the intersections of W_s with S_2, and H has a local maximum and a local minimum at the θ coordinates of p_{2l} and p_{2r}, respectively.*

- a branch containing p_{2r} with a local minimum,
- a branch containing p_{2l} with a local maximum,
- monotone decreasing branches in (p_{2l}, p_{2r}),
- monotone increasing branches in the complement of $[p_{2l}, p_{2r}]$.

We assume for the moment that all intersections of W_s with S_2 are transverse. Then W_s must have an odd number of intersections with S_2, and every intersection in $[\theta_{2l}, \theta_{2r}]$ is preceded by an intersection in the complement of this interval. Therefore, the number of monotone increasing branches is one larger than the number of monotone decreasing branches. Moreover, the image of all branches is contained in I_H with the possible exception of the branch with a local minimum. The branch structure of H will play an important role in defining the different types of bifurcations in section 5. In particular, saddle-node bifurcations will occur on the branch with a local minimum, and homoclinic points occur at the endpoints of certain branches.

The local maximum and minimum of H occur at points where the vector field is tangent to the circle $x = 2$. The circles $x = \pm 2$ are the images of the fold curves $x = \mp 1$ in the reduced system. Differentiating the equation $x' = -x + a\sin(2\pi\theta)$, we obtain $x'' = -x' + \frac{a}{2\pi}\cos(2\pi\theta)\theta' = \frac{3a\omega}{2\pi}\cos(2\pi\theta) \neq 0$ since $x' = 0$ and $|\sin(2\pi\theta)| < 1$. Therefore, the curvature of the trajectories is not zero at the tangencies with $x = 2$, and there are local extrema for H at these points.

Next we analyze the points of discontinuity for the map H. The end of a branch behaves quite differently depending upon whether the trajectories of points near the end of the branch of H lie to the left or right of the stable manifold W_s of the saddle and upon whether there is

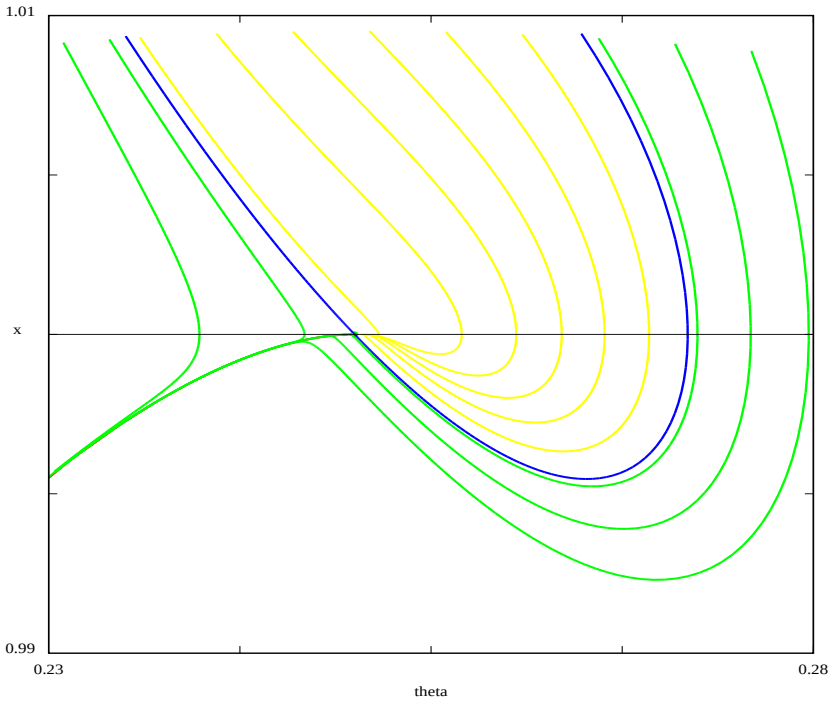


Figure 4.3. *The flow of Ψ_h near a saddle-node. Trajectories in the strong stable manifold are drawn in blue, trajectories in the interior of the stable manifold are drawn in yellow, and trajectories in the hyperbolic region are drawn in green.*

a folded node or a folded focus. In the case in which the trajectories lie to the left of W_s, they cross the circle $x = 1$ immediately to the left of the saddle. We give an asymptotic analysis of the slope of H at the branch end by approximating the flow near the saddle by a linear flow. Introducing coordinates for which the linear unstable and stable manifolds of the saddle are the u and v axes respectively, the linear approximation of the slow flow is

$$\begin{aligned} u' &= \alpha u, \\ v' &= -\beta v \end{aligned} \tag{4.1}$$

with $0 < \alpha$ and $\beta = \alpha+1$ since the trace of the Jacobian at the saddle is -1. The function $u^\beta v^\alpha$ is constant along trajectories of system (4.1). In these coordinates, the circle $x = 1$ becomes a line of the form $v = cu$. The flow from any cross-section to the stable manifold along the v axis to this line will have a derivative that becomes infinite. For example, the intersection (u_1, v_1) of the trajectory through $(u_0, 1)$ with this line will satisfy $|u_0|^\beta = |u_1|^\beta |v_1|^\alpha$ and $|u_1| = |c|^{-\alpha/(\alpha+\beta)} |u_0|^{\beta/(\alpha+\beta)}$. Since $\beta/(\alpha+\beta) < 1$, the slope of this function tends to infinity as $u_0 \to 0$. This argument implies that the slope of H is unbounded for points to the left of W_s.

In the case in which trajectories at the end of the branch pass to the right of W_s, they proceed along the unstable manifold W_u before they cross the circle $x = 1$. If there is a folded

Table 5.1
A summary of the types of bifurcations and their defining equations.

Type of bifurcation	Defining equation
Saddle-node	$H(x) = x,\ H'(x) = 1$
Saddle left homoclinic	$\theta_{2s1} - \theta_{1s} - 0.5 = 0 \pmod 1$
Saddle right homoclinic 1	$\theta_{2s1} - \theta_{1u} - 0.5 = 0 \pmod 1$
Saddle right homoclinic 2	$\theta_{2s2} - \theta_{1u} - 0.5 = 0 \pmod 1$
Saddle right homoclinic 3	$\theta_{2s3} - \theta_{1u} - 0.5 = 0 \pmod 1$
Nodal homoclinic	$\theta_{2ss1} - \theta_{1n} - 0.5 = 0 \pmod 1$
Heteroclinic	$\theta_{1n} - \theta_{2s1} - 0.5 = 0 \pmod 1$

focus, they cross $x = 1$ to the right of W_u. If there is a folded node, these trajectories tend to the node. In the case of the focus, it is well known that the mapping from a cross-section to W_s to a cross-section to W_u will behave asymptotically like $u^{\beta/\alpha}$ at its endpoint. Since $\beta/\alpha > 1$, the derivative approaches 0 as u tends to 0. We conclude that at points of discontinuity for H, the slope is unbounded on one side of the discontinuity (the side of trajectories to the left of W_s), and the slope approaches zero on the opposite side (the side of trajectories to the right of W_s). In the case of a folded node, the half-return map H is undefined on the interval of points between W_s and the strong stable manifold of the node. We choose to extend H to this interval, giving it the constant value that is its limit as points approach the strong stable manifold of the node from the right. See Figure 4.3.

Our analysis of bifurcations will make substantial use of one additional concept that we call the *circuit number* of a trajectory for the reduced system and for H. We lift the slow flow to the universal cover of $\mathbb{R} \times S^1$, and consider trajectories that flow from $(\theta_0, 2)$ to $(\theta_1, 1)$ on the universal cover. The integer part of $\theta_1 - \theta_0$ will be called the circuit number of the trajectory. As ω increases, θ' increases, and the circuit number of the trajectories increase. This leads to a repetitive structure in the bifurcation diagram for the reduced system, in which the same structures reappear in the parameter space, once for each circuit number.

5. Codimension one bifurcations of fixed points.

5.1. Bifurcation types and defining equations. This section characterizes the codimension one bifurcations of fixed points that we find for the half-return map H. We have found at most three fixed points for any parameter value. These fixed points are confined to two branches of H: the branch containing a local minimum and the branch immediately to the left of the branch containing a local minimum. We have not proved that these are the only possible locations of fixed points, but this is consistent with the bifurcation diagram described in section 5.2. Codimension one bifurcations of fixed points for H fall into three classes:

- saddle-node bifurcations,
- homoclinic bifurcations, and
- heteroclinic bifurcations.

For the first two types of bifurcations, we make further distinctions, described in sections 5.1.1 and 5.1.2. Table 5.1 contains a summary of the bifurcations and their defining equations that will be described in the next sections.

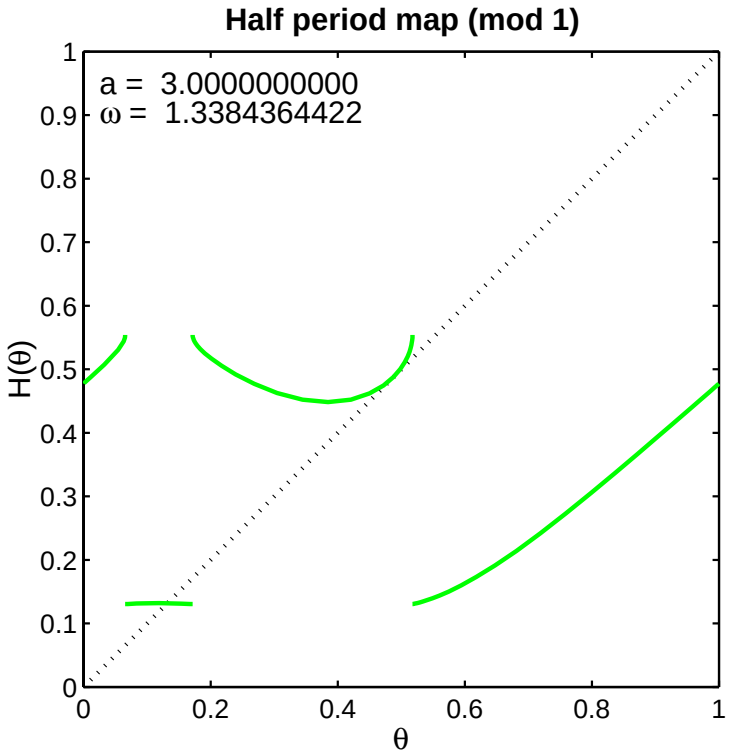


Figure 5.1. *The graph of the half-return map for the given values of a and ω shows a min saddle-node bifurcation, as can be seen by the tangency of the third branch of H with the dotted line $H(\theta) = \theta$ at a minimum of $H - I$.*

5.1.1. Saddle-node bifurcations. The saddle-nodes are distinguished by whether $H - I$ has a local minimum or maximum at the saddle-node point. We call these min and max saddle-nodes, respectively. Figure 5.1 is an example of a min saddle-node, and Figure 5.2 is an example of a max saddle-node. The defining equations for saddle-node bifurcations are $H(x) = x$ together with $H'(x) = 1$.

5.1.2. Homoclinic bifurcations. We distinguish two types of homoclinic orbits—those with a homoclinic connection to the folded saddle, called *saddle homoclinic orbits* and those with a homoclinic connection to the folded node along its strong stable manifold, called *nodal homoclinic orbits*. We further classify the saddle homoclinic orbits as *left homoclinic* or *right homoclinic*, depending on whether the orbit is the limit of trajectories lying to the left or to the right of the stable manifold of the saddle.

The defining equation for a left homoclinic bifurcation is $\theta_{2s1} - \theta_{1s} - 0.5 = 0 \pmod 1$, which guarantees that the first intersection of the stable manifold of the saddle with $x = 2$, namely, θ_{2s1}, is the image of θ_{1s} under TJ_+ (see Figure 5.3). Thus a saddle left homoclinic orbit flows from $(\theta_{2s1}, 2)$ to the saddle (in infinite time) and then is mapped back to its starting point by TJ_+. Fixed points near a left homoclinic saddle bifurcation leave $x = 1$ to the left of the folded saddle point $(\theta_{1s}, 1)$ and return to $x = 2$ near the stable manifold of the folded

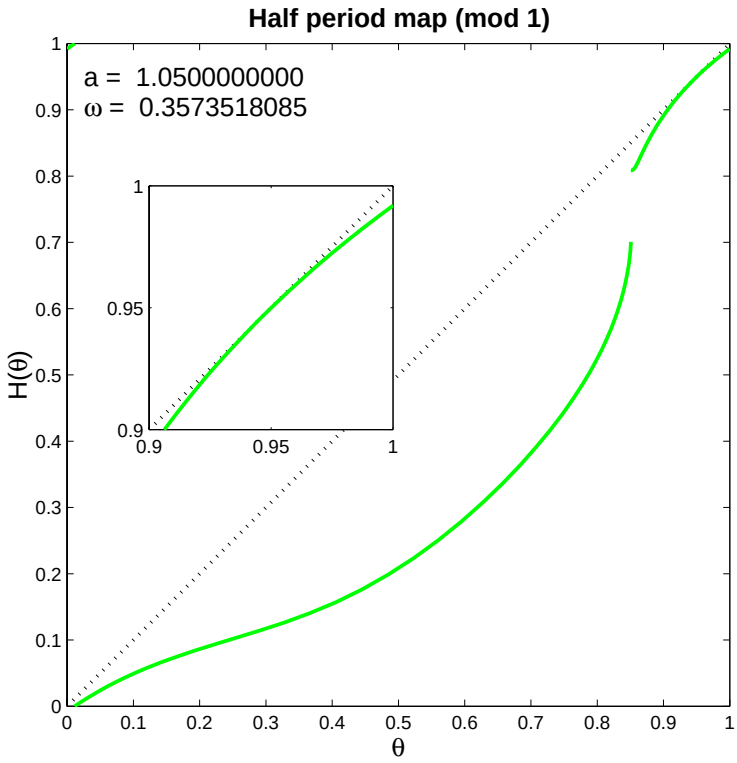
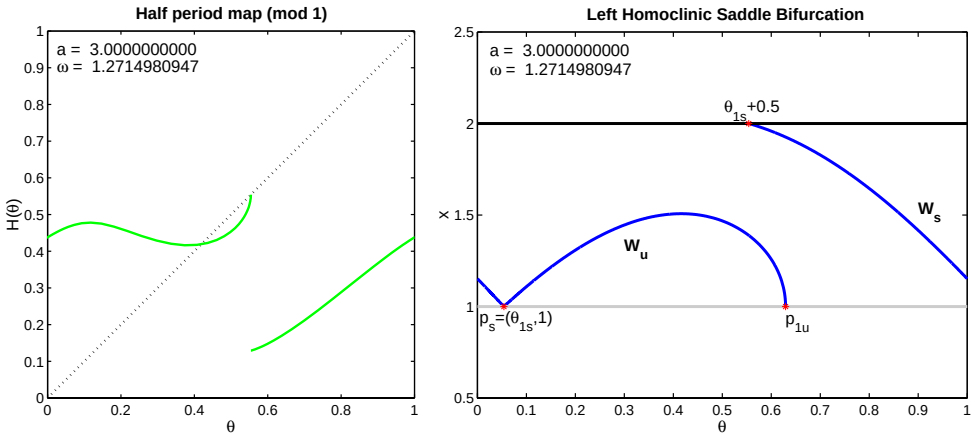
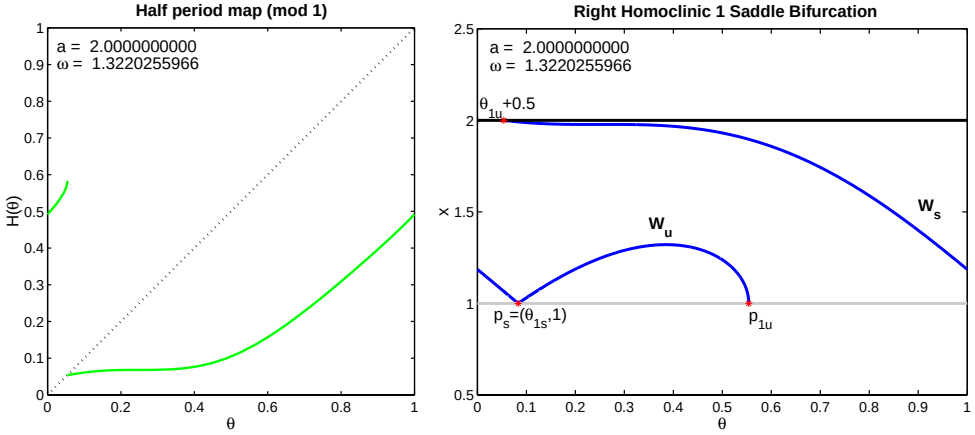


Figure 5.2. *The graph of the half-return map for the given values of a and ω shows a max saddle-node bifurcation, as can be seen by the tangency of the third branch of H with the dotted line $H(\theta) = \theta$ at a maximum of $H - I$.*



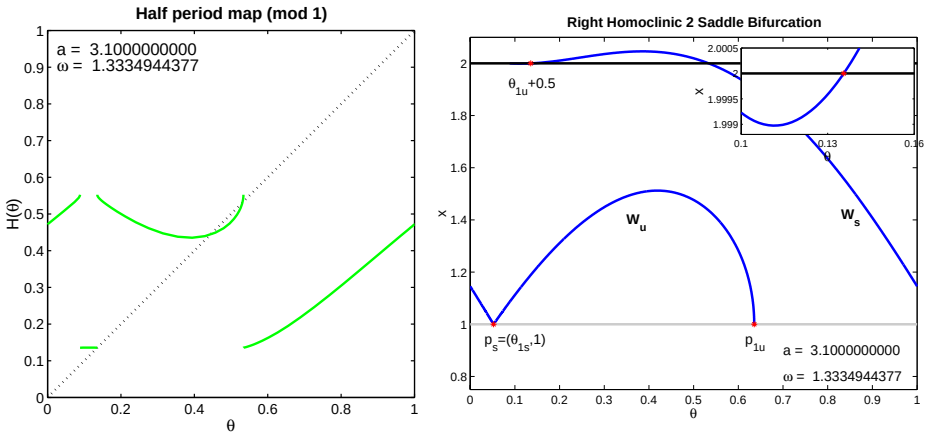
(a) (b)

Figure 5.3. (a) *is a plot of H, and* (b) *illustrates the phase portrait of the orbit at a left homoclinic saddle bifurcation. This bifurcation is apparent in* (a) *since the right end of the first branch of H corresponds to a fixed point.* (b) *shows that the computed orbit satisfies the algebraic condition for a left homoclinic saddle bifurcation $\theta_{2s1} - \theta_{1s} - 0.5 = 0 \pmod 1$.*



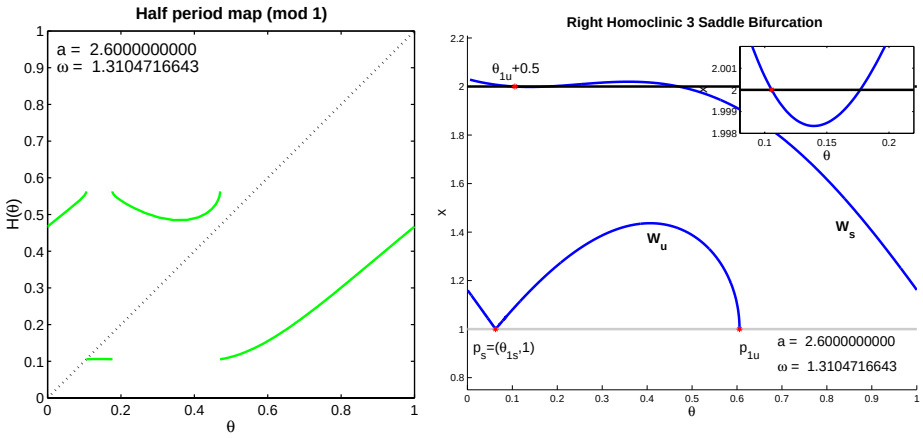
(a) (b)

Figure 5.4. (a) *is a plot of* H*, and* (b) *illustrates the phase portrait of the orbit at a right* 1 *homoclinic saddle bifurcation. This bifurcation is apparent in* (a) *since the left end of the second branch of* H *corresponds to a fixed point.* (b) *shows that the computed orbit satisfies the algebraic condition for a right* 1 *homoclinic saddle bifurcation* $\theta_{2s1} - \theta_{1u} - 0.5 = 0 \pmod 1$.



(a) (b)

Figure 5.5. (a) *is a plot of* H*, and* (b) *illustrates the phase portrait of the orbit at a right* 2 *homoclinic saddle bifurcation. This bifurcation is apparent in* (a) *since the right end of the second branch of* H *corresponds to a fixed point. Figure* (b) *shows that the computed orbit satisfies the algebraic condition for a right* 2 *homoclinic saddle bifurcation* $\theta_{2s2} - \theta_{1u} - 0.5 = 0 \pmod 1$.



(a) (b)

Figure 5.6. (a) *is a plot of* H*, and* (b) *illustrates the phase portrait of the orbit at a right* 3 *homoclinic saddle bifurcation. This bifurcation is apparent in* (a) *since the left end of the second branch of* H *corresponds to a fixed point.* (b) *shows that the computed orbit satisfies the algebraic condition for a right* 3 *homoclinic saddle bifurcation* $\theta_{2s3} - \theta_{1u} - 0.5 = 0 \pmod 1$*, and the inset shows the last two intersections of the stable manifold* W_s *with the circle* S_2*.*

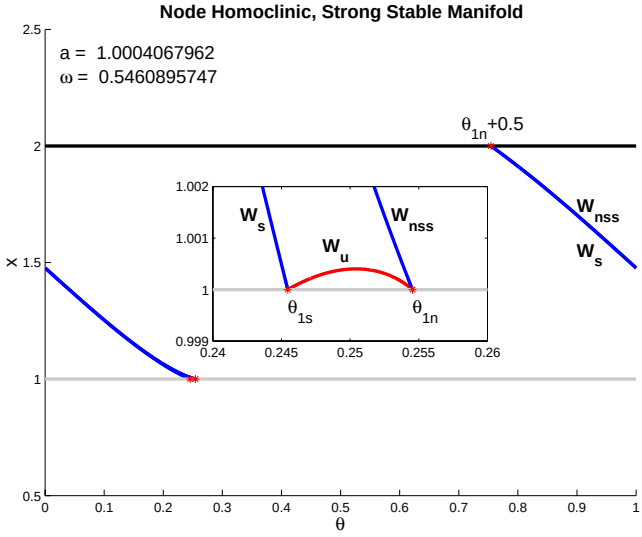


Figure 5.7. *This figure illustrates a nodal homoclinic bifurcation. Note that the computed solution satisfies the algebraic condition* $\theta_{2ss1} - \theta_{1n} - 0.5 = 0 \pmod 1$*. The inset shows the connection between the saddle and the node. We note that it is this same connection that appears in the heteroclinic bifurcations. The plot of the half-return map* H *for these parameter values appears in Figure* 6.3(b).

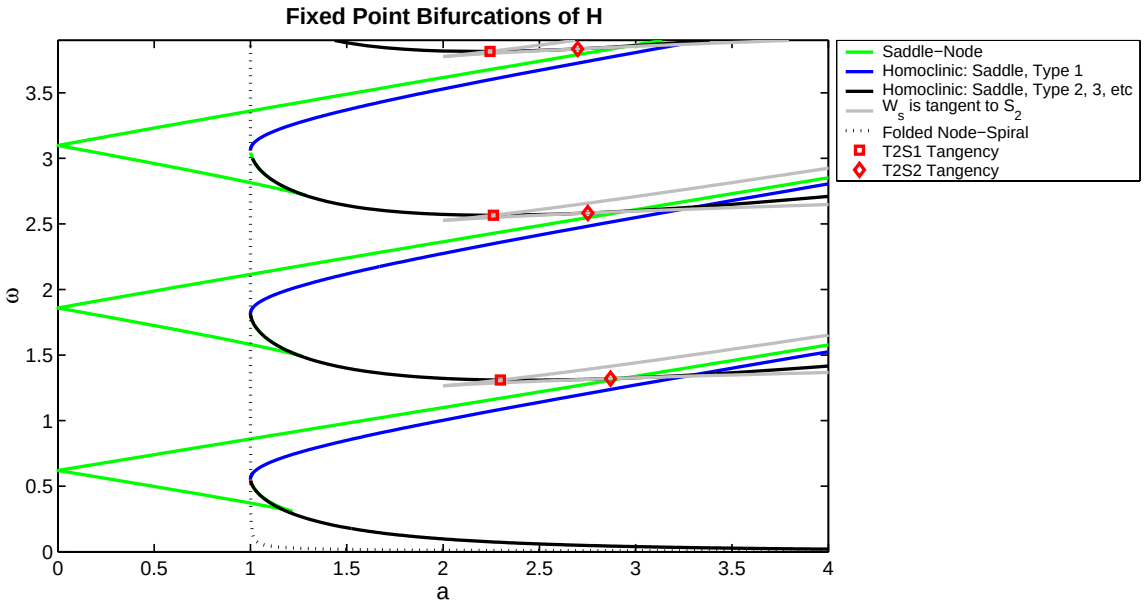
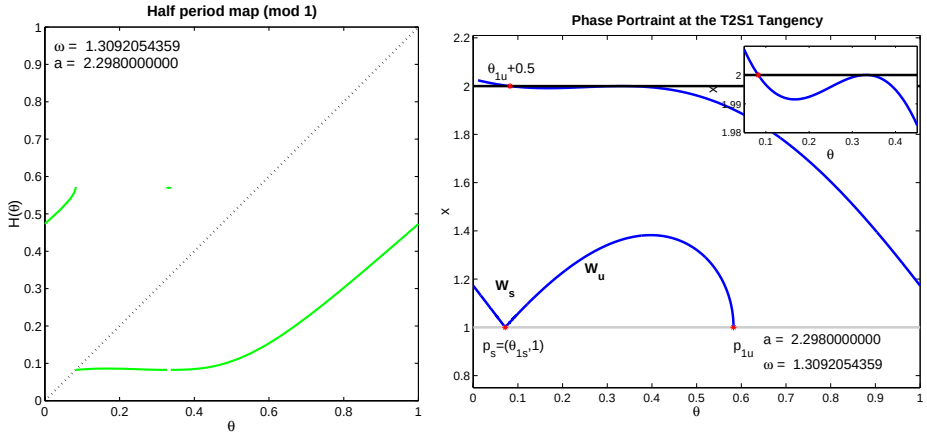
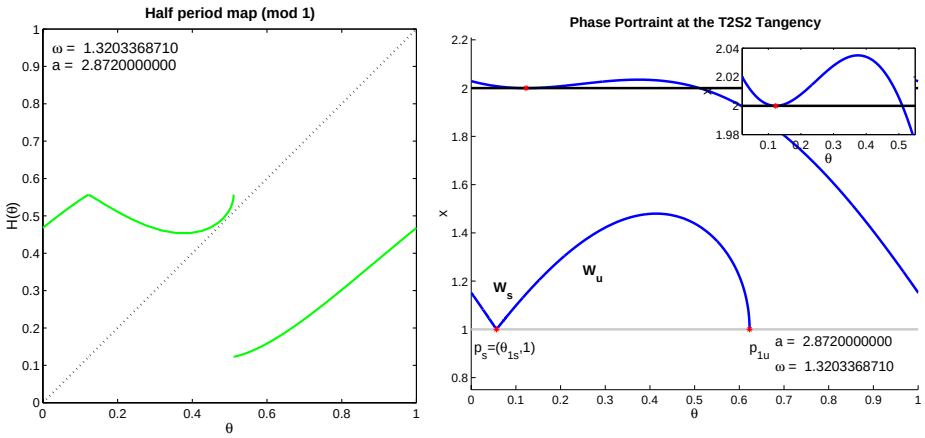


Figure 5.8. *This diagram displays the bifurcation of fixed points of H. The green, blue, and black curves indicate curves of saddle-node points, left homoclinics, and right homoclinics, respectively. The grey curves indicate the parameter values where the stable manifold of the saddle is tangent to the circle S_2. The red squares and red diamonds indicate parameter values where there is a homoclinic orbit and for which the stable manifold of the saddle is also tangent to the circle S_2. Additional features of the diagram that appear on a smaller scale are highlighted in Figures 6.1 and 6.4.*



(a) (b)

Figure 5.9. (a) *is a plot of H near a T2S1 tangency, and* (b) *illustrates the phase portrait of the orbit at the same parameter values. The tangency can be seen in the inset of* (b). *Note that the tangency occurs at the first intersection of W_s with $x = 2$.*



(a) (b)

Figure 5.10. (a) *is a plot of* H *near a* $T2S2$ *tangency, and* (b) *illustrates the phase portrait of the orbit at the same parameter values. The tangency can be seen in the inset of* (b). *Note that the tangency occurs at the second intersection of* W_s *with* $x = 2$.

saddle. Because $(\theta_{2s1}, 2)$ is always a point of S_2 where $x' < 0$, the slope of H is always positive near left homoclinic points. Moreover, the nearby fixed points are unstable since the slope of H is unbounded near the branch ends of Ψ_h orbits that pass to the left of W_s.

Fixed points of H approach a right homoclinic saddle bifurcation if limiting Ψ_h trajectories lie to the right of the stable manifold W_s of the folded saddle (see Figure 5.4). These trajectories then follow the unstable manifold W_u of the saddle and jump to the right of p_{1u}, the intersection of the unstable manifold with $x = 1$. Thus trajectories that approach a right homoclinic saddle bifurcation hit $x = 1$ near θ_{1u} and return to $x = 2$ near the stable manifold of the saddle. The stable manifold W_s may intersect $x = 2$ several times, and the homoclinic point need not jump to the first intersection point. We classify right homoclinic saddle bifurcations by the number of times the stable manifold of the saddle crosses $x = 2$ before returning to the saddle. Thus right homoclinic 1 saddle bifurcations occur when the trajectory returns to $x = 2$ near θ_{2s1}, the first crossing of the stable manifold with $x = 2$. The defining equation for this bifurcation is $\theta_{2s1} - \theta_{1u} - 0.5 = 0 \pmod 1$. Similarly, right homoclinic 2 saddle bifurcations refer to the trajectories that return near the second crossing of the stable manifold with $x = 2$, that is, $\theta_{2s2} - \theta_{1u} - 0.5 = 0 \pmod 1$; an example is shown in Figure 5.5. The only observed right homoclinic saddle bifurcations are of types 1, 2, and 3. Figure 5.6 shows an example of type 3. The analysis of the previous section establishes that the fixed points near a right homoclinic orbit are stable.

Defining equations for the nodal homoclinic bifurcations are such that the intersection of the strong stable manifold of the node with $x = 2$ should have $\theta_{2ss1} = \theta_{1n} - 0.5 \pmod 1$ (see Figure 5.7). Since these bifurcations occur for $a < 2$, the strong stable manifold of the node intersects $x = 2$ only once. Trajectories that lie to the right of the strong stable manifold of the node cross the circle S_1 before approaching the node. Therefore, they make jumps before the trajectories reach the node. Trajectories that lie to the left of the strong stable manifold of the node do not cross the circle S_1. The extension of H to the interval between the

stable manifold of the saddle and the strong stable manifold of the node is constant. For this extended map, the number of fixed points does not change at the nodal bifurcation points, but the fixed points cross from a region where H has positive slope to the extended region, where H is flat.

A qualitative picture of the behavior of H to the right of the strong stable manifold of the folded node can be found by considering the linear flow near a node with an analysis similar to that which we used to determine the slope of H near its branch endpoints. Consider the linear system

$$\begin{aligned}\dot{u} &= -\alpha u,\\ \dot{v} &= -\beta v,\end{aligned}$$

where $\beta > \alpha > 0$. This is a stable node with eigenvectors along the coordinate axes, and the v axis is the strong stable manifold. To obtain a qualitative picture of the behavior of H, we consider a map $u_0 \mapsto u_1$ given by the flow from a line segment $v = v_0$ (and, say, $0 < u_0 < 1$) to the line $v = mu$ (where $m > 0$; this line corresponds to the fold line). The function $u^\beta v^{-\alpha}$ is constant along trajectories; with this we find

$$u_1 = C u_0^{\left(\frac{\beta}{\beta-\alpha}\right)},$$

where $C = \left(\frac{m}{v_0}\right)^{\left(\frac{\alpha}{\beta-\alpha}\right)}$. Thus the map is given by a power law, and since $\beta > \alpha > 0$, the exponent is greater than one. We conclude that the fixed points of H near a nodal homoclinic bifurcation are stable.

5.1.3. Heteroclinic bifurcations. Heteroclinic bifurcations describe curves that are unions of two trajectories that asymptote to both a folded saddle and a folded node. The unstable manifold of the folded saddle lies in the stable manifold of the folded node, giving rise to one segment of the heteroclinic orbit. The second trajectory lies in the stable manifold of the saddle. The defining condition is that the node jumps to the intersection of the stable manifold with S_2. The defining equation is that $\theta_{1n} - \theta_{2s1} - 0.5 = 0 \pmod 1$. Thus the heteroclinic cycle is composed of trajectories lying in the stable and unstable manifolds of the saddle. The primary difference between these bifurcations and the right homoclinics of the saddle is that the unstable manifold W_u approaches the node here, while in the right homoclinic it reaches $x = 1$ and then jumps.

5.2. Description of the bifurcation diagram. Figure 5.8 shows the (numerically computed) bifurcation diagram for fixed points of H in the region $(a, \omega) \in [0, 4] \times [0, 4]$. In this diagram, saddle-node curves are drawn in green, left homoclinic curves in blue, and right homoclinic curves in black. The dotted line separates the region to the right of $a = 1$ with folded nodes from the region with folded foci. The region of the bifurcation diagram close to $a = 1$ will be described in detail in section 6.2 since the bifurcations at $a = 1$ have codimension two.

For $0 < a < 1$, the fixed point bifurcations consist of max and min saddle-node curves, drawn in green. For each circuit number, the min saddle-nodes are the upper branch that appear to extend to infinity. For a sufficiently large, these branches appear to be approximately linear. The lower branch of saddle-nodes in this region are max saddle-nodes. These curves

extend slightly past $a = 1$ but do not extend to infinity as with the curve of min saddle-nodes. We find that these curves end in a cusp that connects to short curves of min saddle-node bifurcation which follow the black curve of right homoclinics back to $a = 1$. This region will be described in more detail in section 6.1.

In the region $a > 1$ and for each circuit number, there are three bifurcation curves that appear to extend to infinity: the min saddle-node bifurcations, the left homoclinic bifurcations, and the right homoclinic bifurcations. Along each right homoclinic curve, two points are marked where the homoclinic orbit is tangent with the line $x = 2$. Changes in the number of intersections of the stable manifold of the saddle with $x = 2$ occur at these locations. The red square separates right saddle homoclinics 1 to the left of the red square from right saddle homoclinics 3 to the right of the red square. See Figure 5.9. The number of crossings switches from 1 to 3 because the stable manifold W_s has a point of tangency with the circle S_2 at the parameter values marked with the red square. To the right of this point along the right saddle homoclinic curve, W_s has 3 crossings of S_2, the third being the homoclinic point. Similarly, the red diamond separates the right saddle homoclinics 3 to its left from the right saddle homoclinics 2 to the right of the red diamond. See Figure 5.10. Here the stable manifold W_s is tangent to the circle S_2 at a local minimum. See Figure 5.10. To the left of the red diamond along the homoclinic curve, the homoclinic points lie to the left of a local minimum in W_s at its second crossing with S_2. To the right of the red diamond along the homoclinic curve, the homoclinic points lie to the right of a local minimum in W_s at its second crossing with S_2.

The types of bifurcations in the bifurcation diagram can be related to properties of the graph of H. The first two crossings of W_s with S_2 bound an interval containing the point $p_{2r} = \frac{1}{2} - \frac{1}{2\pi}\sin^{-1}(\frac{2}{a}, 2)$, where the trajectories of the slow flow have a tangency with S_2 from below. Thus this interval is the branch of H with a local minimum. We call the branch with the local minimum the *central branch* of H. Left homoclinics always occur on the central branch, at its right endpoint. Right homoclinics 2 and right homoclinics 3 occur on the branch immediately to the left of the central branch. In the case of right homoclinics 3, this branch contains the local maximum $\theta_{2l} = \frac{1}{2\pi}\sin^{-1}(\frac{2}{a})$ of H, H has only two branches, and the homoclinic point is the left endpoint of the branch. In the case of right homoclinics 2, the branch is monotonically decreasing and the homoclinic point is the right endpoint of the branch. We also note that, with decreasing ω, new central branches form above the diagonal and then grow in length. Fixed points first appear on the central branch at a min saddle-node bifurcation where the graph of H on the central branch becomes tangent to the diagonal. As ω decreases further, this is quickly followed by a left homoclinic point at which the right endpoint of the central branch crosses the diagonal.

For values of a at which there are folded nodes, there is another curve in the bifurcation diagram corresponding to nodal homoclinic orbits that lie in the strong stable manifold of the folded node. This curve is drawn in magenta on the bifurcation diagram but is sufficiently short that it is difficult to see. Figure 6.4 gives a blown up picture of the region containing the strong nodal homoclinic bifurcation curve on the lowest curve with circuit number 0.

6. Codimension two bifurcations. There are several different points in the bifurcation diagram Figure 5.8 at which bifurcation curves meet or cross. These are codimension two bifurcations of Ψ_h. The previous section discussed tangencies of the stable manifold W_s with

S_2 at right homoclinics where there is no singularity of the bifurcation curve, but its type changes. This section analyzes the remaining codimension two bifurcations of fixed points. We describe each separately and indicate how these bifurcations fit into the diagram of fixed point bifurcations described in the previous section.

6.1. Cusps. Cusps are codimension two bifurcations occurring along the curve of max saddle-node bifurcations. The lower green curve in Figure 6.1 corresponds to the curve of max saddle-nodes that were discussed in the section 5. These curves start at $a = 0$ at the point where the min saddle-node curve meets the max saddle-node curve, and they continue past $a = 1$ to a cusp point just beyond the black curve of right saddle homoclinic orbits (see Figure 6.1). At the cusp, the branch meets a min saddle-node curve that follows the black right homoclinic curve back to $a = 1$. As with the green and blue curves discussed in section 5.2, the location of this curve of saddle-nodes so close to a black curve of homoclinics indicates that these two features are close together in the graph of H. The proximity of these two curves is investigated further in section 6.2. Figure 6.2 shows the half-return map H at a min saddle-node bifurcation on the short branch. Notice that, for nearby parameter values, a homoclinic bifurcation will appear, as can be seen in Figure 6.2.

From the theoretical perspective of singularity theory, we should establish that the fixed point of H has nonzero third derivative at the cusp parameter values. We have not attempted to do this, but the numerical results suggest that this is true.

6.2. Bifurcations at $a = 1$. When $a = 1$, the slow flow has a folded saddle-node equilibrium. At discrete values of ω (one for each circuit number), there is a homoclinic connection along the strong stable manifold of the folded saddle-node. These codimension two points mark the beginning (as a increases) of the more complicated curves of homoclinic points and saddle-node points of H.

In the region of parameter space between $a = 1$ and $a = \sqrt{1 + 1/(16\pi\omega)^2}$, the folded equilibria are saddles and stable nodes. The basin of attraction of the node includes an interval in S_2, and, on this interval, the flow map P_+ is defined to be θ_{1n}, the θ coordinate of the node. Thus the graph of H on this interval is a horizontal line. The left and right endpoints of this interval are the intersections of the stable manifold of the saddle and the strong stable manifold of the node, respectively, with S_2. The value of H on this interval is $\theta_{1n} - \frac{1}{2}$. At the left end of the interval, H has a discontinuity: $\lim_{\theta\to\theta_{2s1}^-} H(\theta) = \theta_{1s} - \frac{1}{2}$, but $\lim_{\theta\to\theta_{2s1}^+} H(\theta) = \theta_{1n} - \frac{1}{2}$.

For each circuit number, there is a region in parameter space between $a = 1$ and $a = \sqrt{1 + 1/(16\pi\omega^2)}$, where the horizontal segment of the graph of H includes a fixed point. If the right end of the interval is a fixed point, there is homoclinic connection along the strong stable manifold of the node. (See Figure 6.3(b).) For circuit number 1, these codimension one points are plotted as a magenta line in Figure 6.4. In this figure, the magenta line ends at the curve $a = \sqrt{1 + 1/(16\pi\omega)^2}$ (the dotted line). Increasing a beyond this point changes the homoclinic point to a fixed point of H (i.e., a periodic orbit of Ψ_h).

If the left end of the interval is a fixed point, there is a heteroclinic connection containing the stable and unstable manifolds of the saddle. (See Figure 6.3(a).) These codimension one points are plotted as a red line in Figure 6.4. At the transition from folded node to

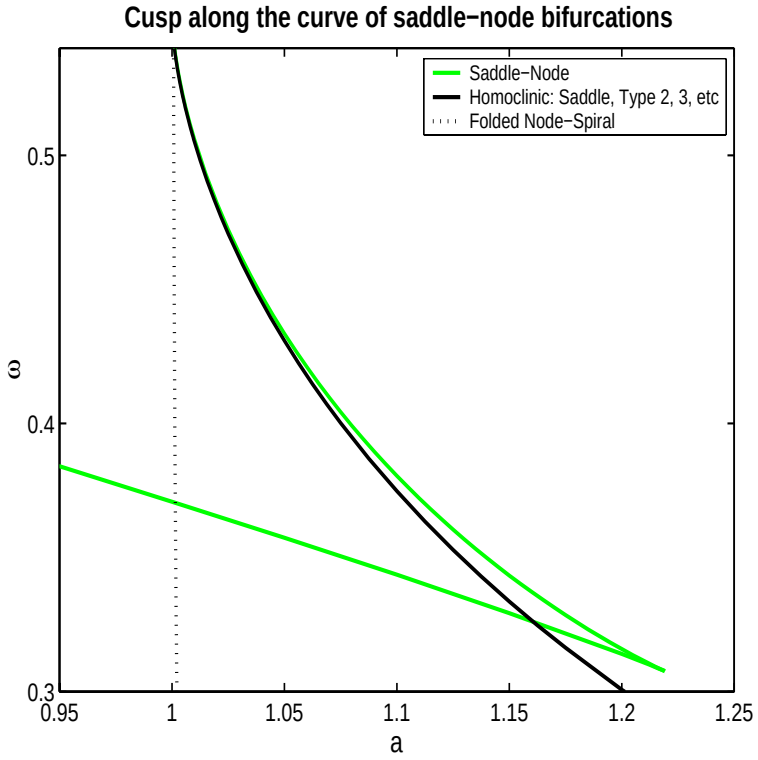


Figure 6.1. *The branches of min saddle-node bifurcations that begin at $a = 0$ continue past $a = 1$ to the cusp shown in this diagram.*

folded spiral, heteroclinic points become right homoclinic points; thus the red curve connects continuously to the black curve in Figure 6.4.

To summarize, the region in parameter space for which the horizontal segment of the graph of H contains a fixed point is the region bounded by the red curve (the heteroclinic points), the magenta curve (the homoclinic points for the strong stable manifold of the node), and the dotted curve (the transition of the stable folded equilibrium from folded node to folded spiral). These curves are shown in Figure 6.4. There are four curves that start at the codimension two point $a = 1$ and $\omega \approx 0.5546$ (where the folded saddle-node has a homoclinic connection). The one upper curve is the blue curve of left saddle homoclinics. There are three lower curves. The lowest curve is the red curve of heteroclinic connections. The middle curve is the magenta curve, where there is a homoclinic connection along the strong stable manifold of the node. Slightly above this curve is the green curve of saddle-node bifurcations. It can be seen in Figure 6.3(b) that only a very small change in the parameters is necessary to perturb the system from the situation of having a stable nodal homoclinic bifurcation to having a saddle-node bifurcation in H. Note that, for values of a smaller than 1, there are no folded equilibria, and the only bifurcations of H are saddle-nodes.

At $a = 1$, the stability of the homoclinic cycles is more subtle to analyze than the stability of the right and left homoclinic cycles because the center manifold of the equilibrium point

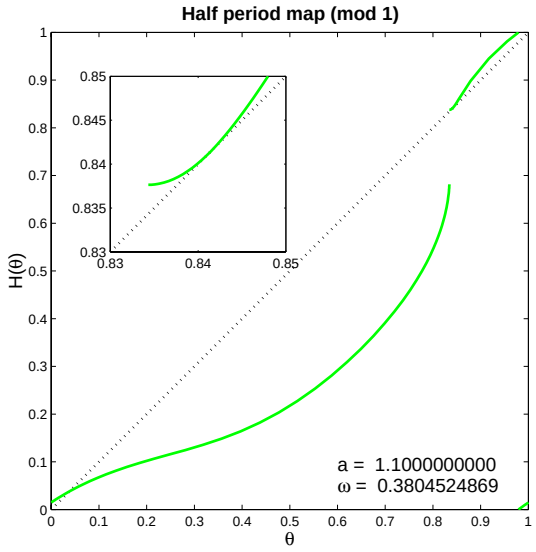


Figure 6.2. *The graph of H at a saddle-node bifurcation point. This point is on the curve of min saddle-node bifurcation shown in Figure 6.1. The inset in the figure shows the tangency of the graph at the point where $H(\theta) = \theta$. It can also be seen that a small change in the parameters can move this to a right homoclinic point, in which the left end of the branch of the graph shown in the inset becomes a fixed point.*

at $(0.25, 1)$ is tangent to the circle S_1. The map from the circle S_2 to the circle S_1 along trajectories adjacent to the strong stable manifold of the equilibrium needs to be determined. We begin with an analysis in normal form of coordinates of a saddle-node that gives most of the information that we need to determine the stability of the map from S_2 to S_1. Consider the (truncated) normal form for a saddle-node equilibrium point of a two-dimensional flow:

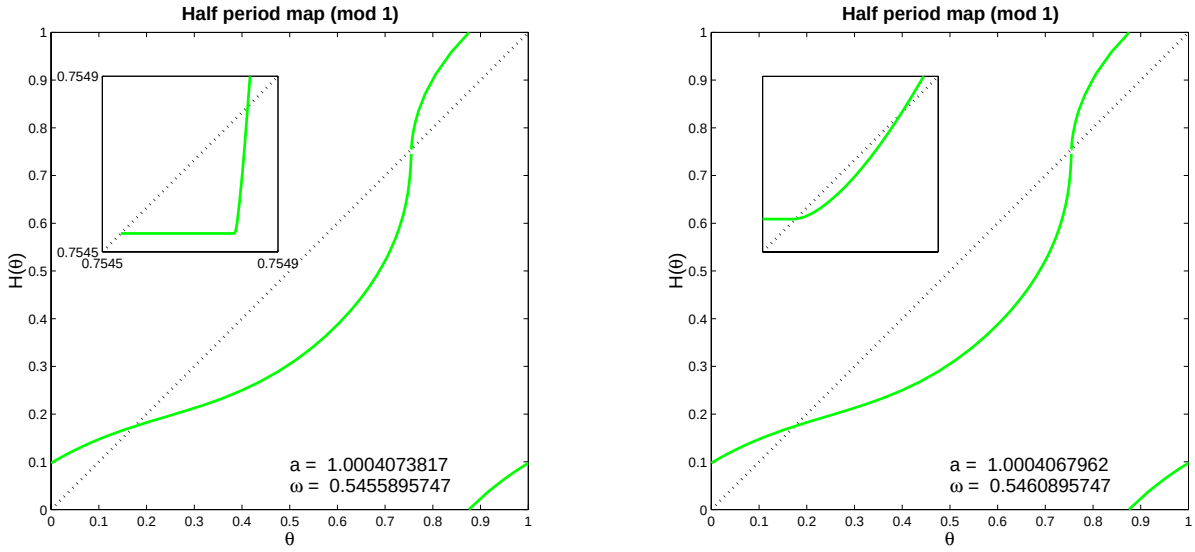
$$\begin{aligned} \dot{u} &= u^2, \\ \dot{v} &= -\alpha v. \end{aligned}$$

The solutions of this system are

$$\begin{aligned} u(t) &= \frac{u_0}{1 - tu_0}, \\ v(t) &= v_0 e^{-\alpha t}. \end{aligned} \tag{6.1}$$

We want to compute the map along trajectories from a cross-section of the strong stable manifold to a curve tangent to the center manifold. Take the cross-section to the strong stable manifold to be $v = 1$ and the curve to be the graph of the function $v = h(u)$ with $h(0) = h'(0) = 0$. If the trajectory with initial condition $(u_0, 1)$ flows to (u_1, v_1) with $v_1 = h(u_1)$, then we eliminate t from (6.1) to obtain the implicit equation

$$\exp\left(-\alpha\left(\frac{1}{u_0} - \frac{1}{u_1}\right)\right) = h(u_1),$$



(a) (b)

Figure 6.3. (a) *shows an example of the graph of H when there is a heteroclinic connection. There is a fixed point at the left end of the horizontal segment of the graph of H.* (b) *shows an example when there is a homoclinic connection along the strong stable manifold of the node. The square inset in* (b) *has a width of 5×10^{-7}; the lower intersection of the graph of H with the diagonal dotted line occurs at $\theta \approx 0.754539$. It is clear from this picture that only a very small change in the parameters is required to produce a saddle-node bifurcation in H.*

which can be solved for the inverse of the map along trajectories

$$u_0 = \frac{\alpha u_1}{\alpha - u_1 \ln(h(u_1))}.$$

If $\lim u \ln(h(u)) = 0$ as $u \to 0$, then $\lim \frac{du_0}{du_1} = 1$ as $u_1 \to 0$. This criterion and the tangency conditions are satisfied if $h(u)$ is asymptotic to u^β for some $\beta > 1$.

Now, the center manifold of (3.1) at its saddle-node equilibrium point has quadratic tangency with the curve defined by $x = 1 - 2\pi^2(\theta - 1/4)^2$. This is verified by differentiating this equation and using (3.1) to see that, along the curve,

$$\begin{aligned}(x - 1 + 2\pi^2(\theta - 1/4)^2)' &= -x + \sin(2\pi\theta) + 4\pi^2\omega(\theta - 1/4)(x^2 - 1) \\ &= -(1 - 2\pi^2(\theta - 1/4)^2) + \cos(2\pi(\theta - 1/4) \\ &\qquad + 4\pi^2\omega(\theta - 1/4)(x^2 - 1) \\ &= o((\theta - 1/4)^2)\end{aligned}$$

since $(x^2 - 1) = O((\theta - 1/4)^2)$. We conclude that the slope of the half-return map H approaches a finite slope as θ approaches the strong stable manifold of the saddle-node equilibrium from the right.

We have computed H in a small neighborhood of this strong stable manifold and found that the limit slope is large but finite. At the codimension two point $(\omega, a) = (0.554586, 1)$

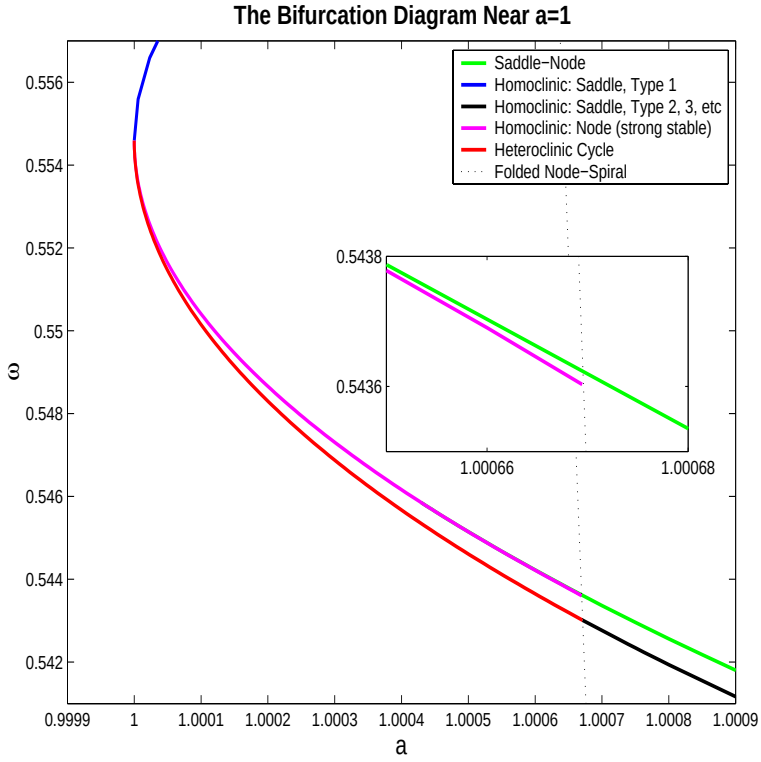


Figure 6.4. *A detail from the bifurcation diagram. This plot shows the curves that emanate from the codimension two point, where there is a homoclinic connection to the folded saddle-node. For circuit number* 0*, this occurs at* $a = 1$ *and* $\omega \approx 0.5546$.

with circuit number 0, the slope is approximately 200. This calculation helps us understand the structure of the bifurcation diagram near $a = 1$. It implies that, for values of a slightly larger than 1, $H' \to 0$ as θ approaches the strong stable manifold of the node from the right but that the slope will increase rapidly and quickly become much larger than one. Thus there will be a local minimum of $H - I$ near the endpoint. (See, for example, Figure 6.3.) When this local minimum of $H - I$ is a fixed point of H, we have a min saddle-node bifurcation. In this regime, there are two intervals in which H has slope smaller than 1 and two intervals in which H has slope larger than 1. As a increases, the intervals which are not adjacent to the discontinuity of H shrink in size, disappearing when there is an inflection point with slope 1. Cusps occur when these inflection points of slope 1 are also fixed points.

6.3. Transversal crossings and the bistable regimes. The simplest codimension two bifurcations for the reduced system are points at which two codimension one bifurcations occur at different places in the phase space. At such points, two codimension one bifurcation curves intersect in the bifurcation diagram. We expect, and find in our numerical computations, that these intersection points are transverse. The crossings play a significant role in the phenomenon of bistability that has been an important part of the history of the forced van der Pol system.

For each value of a larger than those for which cusps occur, the values of ω for which H has a stable fixed point with a given circuit number is an interval bounded below by a right homoclinic curve and bounded above by a point on the curve of min saddle-node bifurcations that originates at $a = 0$. For values of a close to 3, we observe that the min saddle-node curve of circuit number k and the right homoclinic of circuit number $k + 1$ intersect one another transversally. To the right of these intersection points, there are sectors in which there are simultaneously stable fixed points of H with circuit numbers k and $k + 1$. The parameters for which Littlewood [20, 21] deduced the existence of chaotic invariant sets for the flow lie in these regions. To the left of the intersections of min saddle-nodes and right homoclinics, there are regions with no fixed points for H. To the right of these intersection points, H has fixed points for all values of ω. In addition to the two stable fixed points of H, there are also unstable fixed points of H for parameter values just below the min saddle-node curves of circuit number k. These disappear along left homoclinic curves with decreasing ω. We observe that the left homoclinic curve of circuit number k also intersects the right homoclinic curve of circuit number $k + 1$ transversally. In section 7, we show that there is a rich dynamical structure near these codimension two bifurcations, including new families of chaotic invariant sets.

The one additional transversal crossing of bifurcation curves for fixed points that we observe is in the region near $a = 1$, where the max saddle-node curves cross the right homoclinic curves with the same circuit numbers.

7. Period 2 points of H.

Thus far, our discussion of the bifurcation diagram of the forced van der Pol equation has focused on the bifurcations of fixed points of H, the half-return map. In the parameter region $0 \le a < 1$, the theory of families of circle diffeomorphisms gives a qualitative description of bifurcations of all periodic orbits. Following a brief review of this theory, this section presents a numerical study of periodic orbits of period 2 with circuit number 1. Period 2 orbits are solutions to $H^2(x) = x$. We display graphs of the second iterate H^2 of the half-return map that give insight into the bifurcations of the period 2 orbits. There are many common features between the bifurcations of fixed points and periodic orbits of period 2 for H, but there are significant differences as well. We find new types of codimension two bifurcations involving homoclinic orbits and identify small parameter regimes in which there are new classes of chaotic invariant sets.

7.1. The parameter region $0 \le a < 1$.

For $0 < a < 1$, the map H is a diffeomorphism of the circle S_2 that varies smoothly with both a and ω. Moreover, the slow flow has a *rotational* property with respect to ω. As ω increases, the direction of the slow flow rotates toward the x axis at a nonzero rate. Since the slow flow is in the sector with decreasing x and increasing ω, two slow flows with the same parameter a and different parameters ω are transverse. These two facts imply that the image of H is strictly increasing with ω. We also conjecture, based upon our numerical computations, that the function $H - I$ appears to have a single local maximum and a single local minimum. This conjecture is sufficient to determine the qualitative properties of the bifurcation diagram in the region $0 < a < 1$ based on the theory of circle diffeomorphisms [1]. There is an *Arnold tongue* of rotation number $\frac{p}{q}$ for each

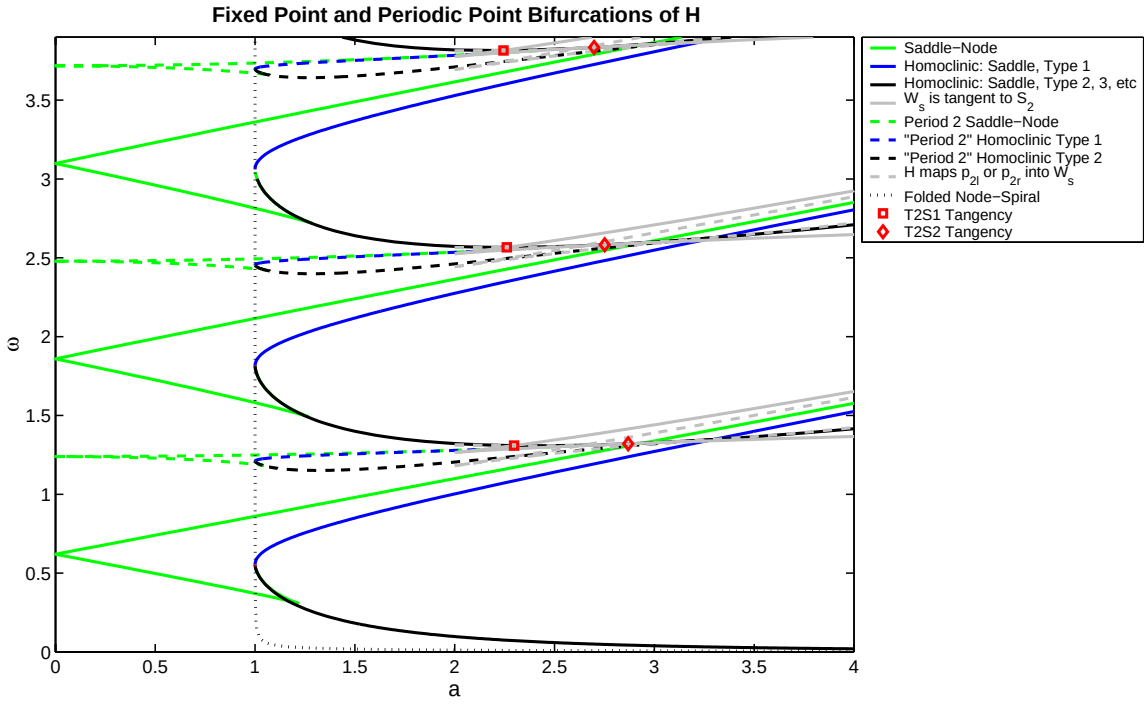


Figure 7.1. *This bifurcation diagram expands the diagram displayed in Figure 5.8 to include the bifurcations associated with period 2 orbits. Note that the same line color is used to distinguish the same types of bifurcation, but the line style indicates whether it is a bifurcation of fixed points or period 2 points.*

rational number $\frac{p}{q}$. This is a strip that begins at[1] $\omega = \frac{2p+q}{(3-2\ln 2)q}$ and intersects each line $0 < a = \text{const} < 1$ in an interval. In the interior of the Arnold tongue, there are two periodic orbits of period q. On the lower boundary of the Arnold tongue, there is a max saddle-node of period q, while on the upper boundary of the Arnold tongue, there is a min saddle-node of period q. For each irrational rotation number ρ and parameter a, there is a single value of ω for which H has rotation number ρ. The set of parameter values with irrational rotation numbers forms a set of positive Lebesgue measure on each line $0 < a = \text{const} < 1$ in the parameter plane.

7.2. An augmented bifurcation diagram. Figure 7.1 is similar to the bifurcation diagram displayed in Figure 5.8 but adds new information about bifurcations of period 2 orbits of H. The bifurcation curves of period 2 orbits are dashed: saddle-node curves are green, left homoclinic curves are blue, and right homoclinic curves are black. Also drawn as gray solid curves are parameters where the stable manifold W_s is tangent to S_2, and gray dashed curves represent parameters where H maps one of the points p_{2l} or p_{2r} into W_s. Figure 7.2 shows the region with period 2 orbits of circuit number 1 in more detail.

[1]When $a = 0$, the slow flow is easily integrated, and we find $H(\theta) = \theta + \omega\left(\frac{3}{2} - \ln 2\right) - \frac{1}{2}$. This is actually the function lifted to the universal cover of S^1; i.e., for the moment, we do not compute H modulo 1. By solving $H^q(\theta) = p$ for relatively prime integers p (the circuit number) and q (the period), we obtain the given formula for ω. The saddle-node curves, in particular, begin at $\omega = \frac{2p+1}{3-2\ln 2}$.

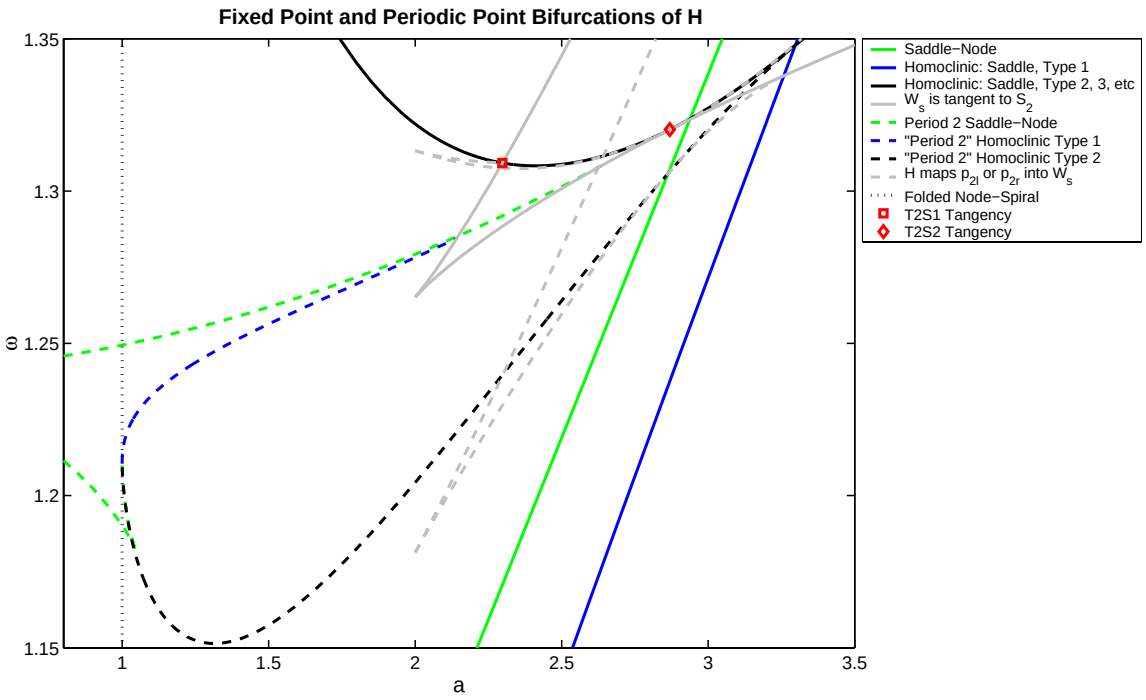


Figure 7.2. *A closer look at part of the bifurcation diagram shown in Figure 7.1.*

Several features of these figures are notable. The gray curves originate at $a = 2$. From each endpoint, two curves emerge that correspond to parameters at which p_{2l} lies in or maps to W_s (lower curves) and parameters at which p_{2r} lies in or maps to W_s (upper curves). "Inside" the solid gray curve, the half-return map H has three points of discontinuity, while outside the cusp it has a single discontinuity. Similarly, the number of discontinuities of H^2 changes as parameters cross the dashed gray curves. The curves where W_s is tangent to S_2 intersect the right homoclinic curves of fixed points at the codimension two points where the type of the right homoclinic curve changes. The intersection at p_{2l}, where the type changes from right homoclinic 3 to right homoclinic 2, is tangential: the tangency curve remains below the right homoclinic curve. There are similar points of tangential intersection along the curves of period 2 right homoclinic bifurcations.

The curve of period 2 min saddle-nodes that begins at $a = 0$ meets the curve of period 2 left homoclinics at a codimension two bifurcation discussed in section 7.3. The curve of period 2 left homoclinics crosses the curve of fixed point right homoclinics. The curves of period 2 left and right homoclinics end at the codimension two bifurcation, where the circuit number 1 right and circuit number 0 left homoclinic curves cross. The coordinates of this point are approximately $(a, \omega) = (3.29725, 1.34783)$. The sequences of bifurcations for the period 2 orbits appear to undergo similar bifurcations to those of the fixed points in the region close to $a = 1$, but we have not explored the bifurcations here thoroughly. Instead, we focus upon the homoclinic bifurcations.

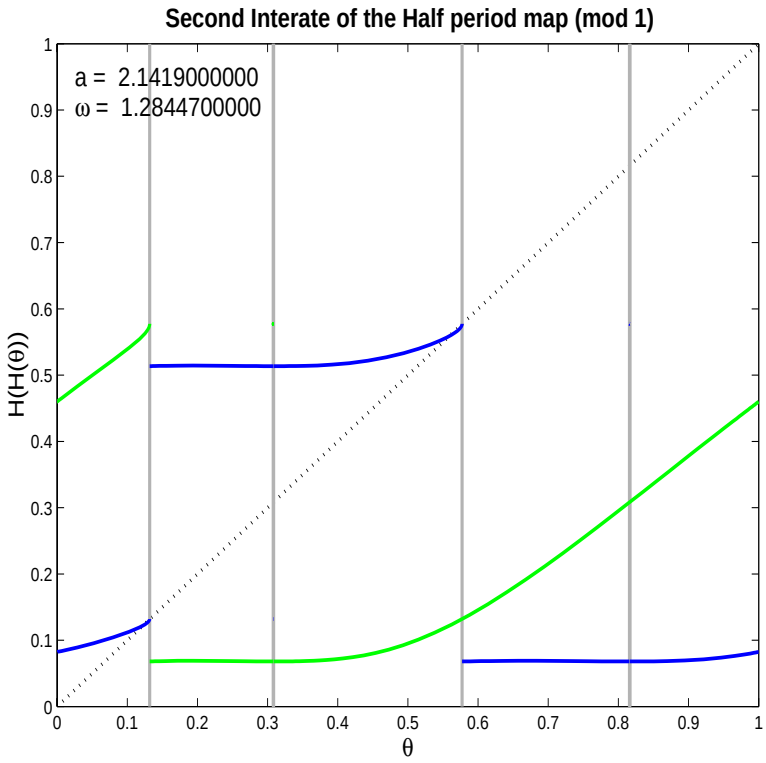


Figure 7.3. *Graphs of H (green) and H^2 (blue) near a point of codimension two bifurcation.*

7.3. Period 2 homoclinic bifurcations. The "morphology" of period 2 homoclinics is substantially more intricate than that of the fixed point homoclinic orbits. In particular, the number of discontinuities of H^2 changes several times along the period 2 homoclinic curves, and there are codimension two bifurcations at which the branches containing the period 2 orbit change. Here we traverse these period 2 homoclinic curves, analyzing the transitions that occur and displaying the graphs of H^2 in different parameter regions.

The upper left homoclinic curve encounters the min tangency curve where the stable manifold of the saddle passes through the point p_{2r} of tangency of the vector field with $x = 2$. The coordinates of the intersection point are approximately $(a, \omega) = (2.14190, 1.28447)$. See Figure 7.3. The map H has three branches for a to the right of the min tangency curve, and the homoclinic point is the third intersection of the stable manifold with $x = 2$. The max tangency curve where the stable manifold of the saddle passes through the point p_{2l} of tangency of the vector field with $x = 2$ crosses the parameter curve where H^2 has a min tangency for $2.62 < a < 2.63$ (i.e., the point p_{2r} maps to the intersection of W^s with $x = 2$.) This has no immediate impact upon the bifurcation curves but creates additional discontinuities of H^2. For example, when $(a, \omega) = (2.63, 1.3084)$, H^2 has eight discontinuities, and its graph has eight branches (see Figure 7.4). For values of a to the right of the max tangency curve, H once again has a single discontinuity.

For (a, ω) near $(2.66384, 1.30982)$, the min saddle-node curve appears to end at parame-

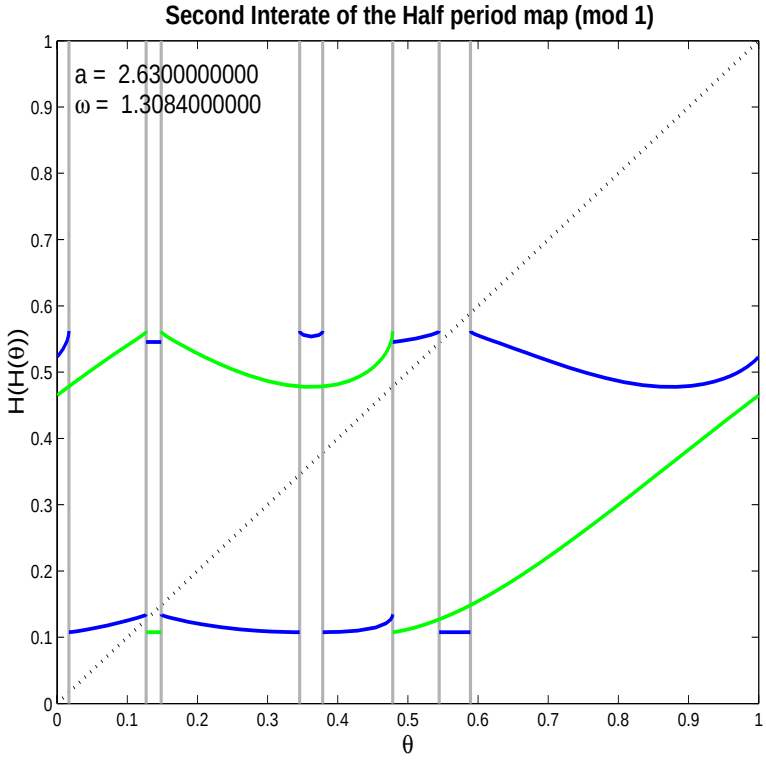


Figure 7.4. *Graphs of H (green) and H^2 (blue) at a point where H^2 has eight branches.*

ters where the left homoclinic curve and the max tangency curve have a point of tangential intersection (see Figure 7.5). Defining equations for this codimension two bifurcation are based upon the conditions that the stable manifold of the saddle passes through p_{2l} and that $H^2(\theta_{2l}) = \theta_{2l}$. The unfolding of the bifurcation depends upon the singularities of H and H^2 at the tangency point. These maps can be modeled as the composition of a quadratic folding map with the power law $|u_1| = |u_0|^{\beta/(\alpha+\beta)}$ as we determined in section 4.2. Thus the composition is the power law $|u_1| = |u_0|^{2\beta/(\alpha+\beta)}$. Since $\beta = \alpha+1$, the exponent is $1+1/(2\alpha+1) > 1$, and the maps H and H^2 have zero derivative at the singularity, which is a local maximum. For values of a smaller than the codimension two point, the homoclinic points lie at the right endpoint of the branch of H^2 to the left of θ_{2l}. This branch has positive slope. For values of a larger than the codimension two point, the homoclinic points lie at the left endpoint of the branch of H^2 to the right of θ_{2l}. This branch has negative slope, so there can no longer be a saddle-node bifurcation of these periodic orbits. Instead, there will be a period-doubling bifurcation curve that lies below (i.e., smaller values of ω) the homoclinic curve. The period-doubling curve affects the stability of the period 2 orbits and signals the presence of period 4 orbits, but new period 2 orbits are not born at the period-doubling bifurcation.

The next codimension two bifurcation along the period 2 left homoclinic curve is its crossing of the period 1 right homoclinic curve at a parameter value $2.79 < a < 2.80$. Here there is simultaneously a fixed point at the left end of the branch of H containing θ_{2l} and a

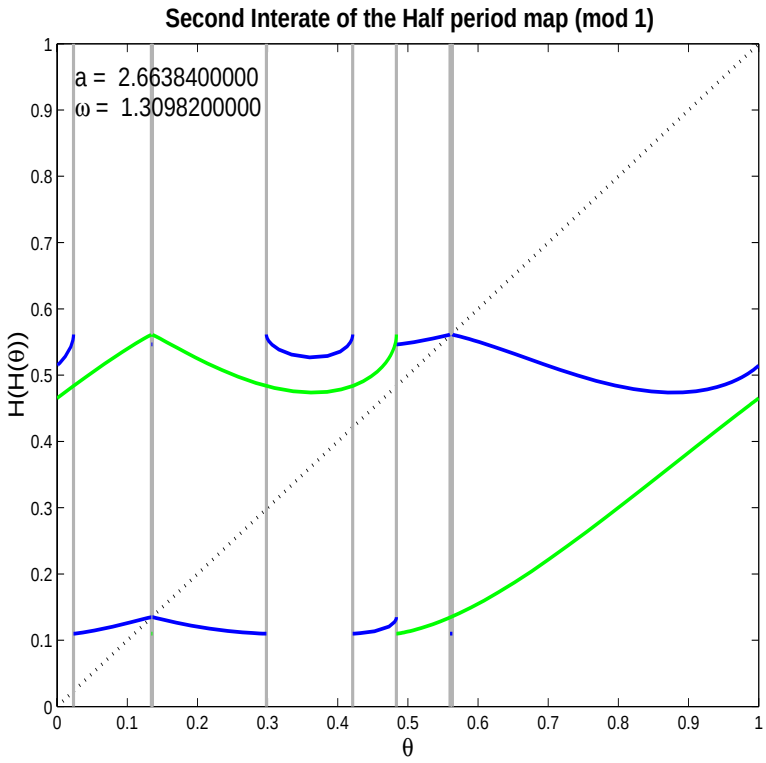


Figure 7.5. *Graphs of H (green) and H^2 (blue) near the end of the min saddle-node curve.*

homoclinic point of H^2 on the branch immediately to the right of the one containing θ_{2l}. For larger values of a, there is a region of bistability with both a fixed point and a period 2 orbit.

The endpoint of the period 2 left homoclinic curve lies at the codimension two point near $(a, \omega) = (3.29725, 1.34782)$, where the right homoclinic curve of fixed points with circuit number 1 crosses the left homoclinic curve of fixed points with circuit number 0 (see Figure 7.6). The period 2 right homoclinic curve also ends at this codimension two point. Apparently, this codimension two point, defined as a transversal crossing of two codimension one bifurcations, is much more complicated than it seems at first glance. Indeed, we find that there are perturbations from this parameter value at which the map H has a chaotic invariant set.

7.4. Chaos without canards. The half-return map H appears to have chaotic trajectories for parameter values $(a, \omega) \approx (3.2, 1.34008531)$. Figure 7.7 shows the graphs of H and H^2, and Figure 7.8 shows detail that illustrates the chaotic invariant set. The parameters $(3.2, 1.34008531)$ are above the left homoclinic curve of fixed points with circuit number 0 and below the right homoclinic curve of fixed points with circuit number 1, in the vicinity of the intersection of these two curves. The branch of H that contains a local minimum has two fixed points—one stable point and an unstable fixed point θ_u close to the right end of the branch. The end of the branch occurs at θ_{2s1}, the first intersection of the stable manifold of the saddle point p_{1l} with $x = 2$. The map H is increasing on $J = [\theta_u, \theta_{2s1})$ with $H(J) = [\theta_u, \theta_r)$. Note that $H(J)$ is bounded on the right by the supremum of H, and on J, H has a single point of

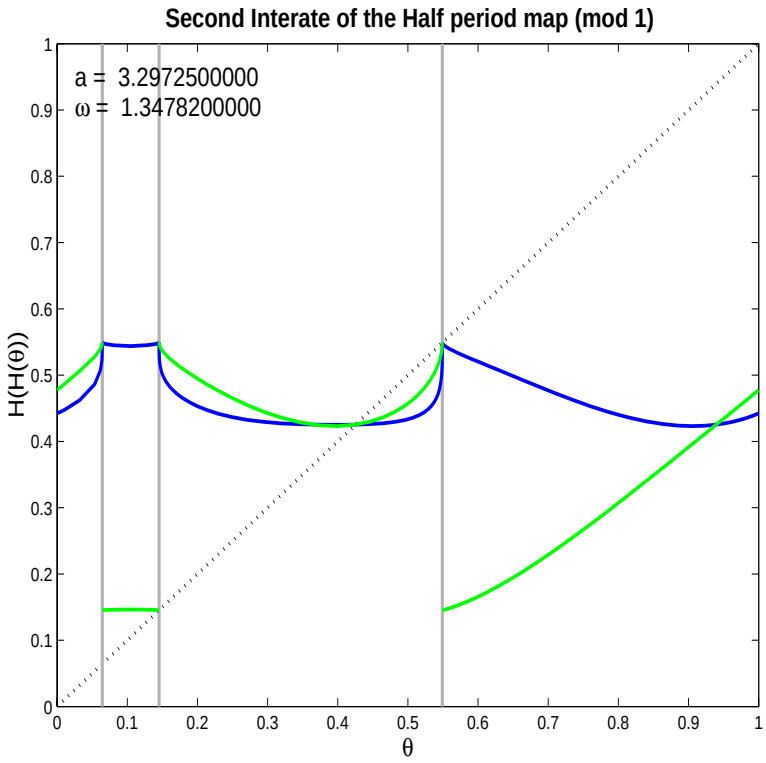


Figure 7.6. *The graphs of H and H^2 at the codimension two point of transversal crossing of right and left homoclinic bifurcations of fixed points.*

discontinuity at θ_{2s1}. We follow the interval $H(J) - J$ for two iterates. First, $H(H(J) - J)$ is an interval whose left endpoint is θ_l, the infimum of H. For the parameter values that lie on a right homoclinic curve of fixed points, θ_l is a homoclinic point of H. For parameter values below this homoclinic curve, $H(\theta_l)$ lies to the right of the discontinuity point θ_{2s2}, inside the branch of H with a local minimum. For parameters near $(a, \omega) = (3.2, 1.34008531)$, the map H is decreasing on $H(H(J) - J)$. We find in numerical calculations that there is a value of $\omega \in (1.34008530, 1.34008531)$ so that $H^2(\theta_r) = \theta_u$, the unstable fixed point. We observe that $H^2(\theta_{2s1}^+) \in H(J) - J$, so $H^2(H(J) - J) \subset H(J)$. We conclude that the union I of the two intervals $H(J)$ and $H(H(J) - J)$ is an invariant set for H. Partitioning I into the three intervals $I_1 = J$, $I_2 = H(J) - J$, and $I_3 = H(H(J) - J)$, we have $H(I_1) = I_1 \cup I_2$, $H(I_2) = I_3$, and $H(I_3) \supset I_1$. This implies that H has a chaotic invariant set that contains a subshift of finite type [27] with transition matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

It is quite possible that H not only has a chaotic invariant set for the parameters described above but that I is an attractor. This is true, for example, if the Schwarzian derivative of H on the interval I is negative. Since the map H can be approximated by functions of the form

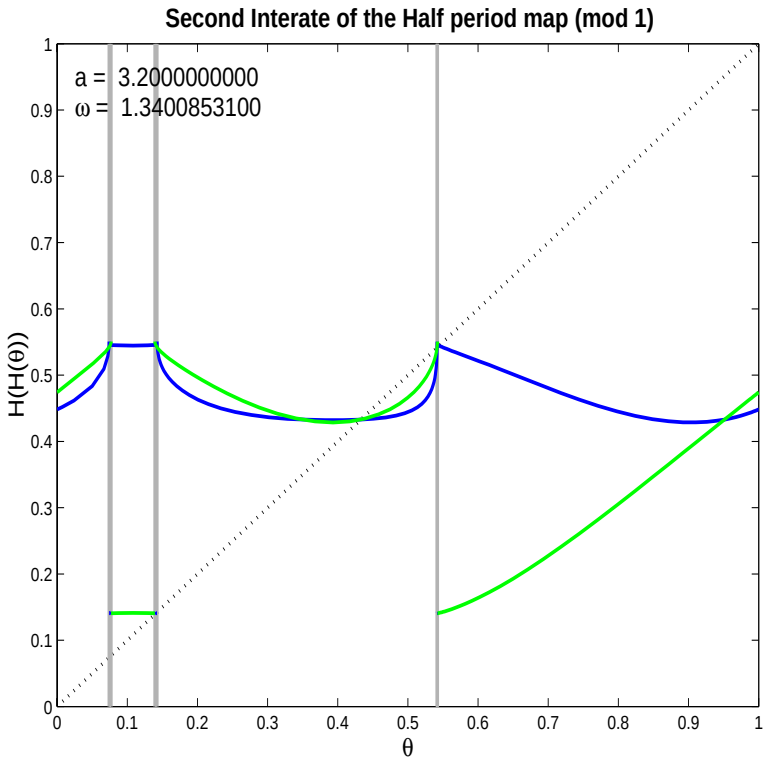


Figure 7.7. *The graphs of H and H^2 at parameters with an apparent chaotic invariant set.*

$f(x) = x^\alpha$ and these functions do have negative Schwarzian derivative, we conjecture that there are parameter values for which H has chaotic attractors. Furthermore, we conjecture that the set of such parameter values has positive measure.

8. Concluding remarks. This section concludes with remarks about

- the methods we used to compute the values of the half-return map H and the bifurcation diagram of the reduced system and
- aspects of the bifurcations of the forced van der Pol equation that are not addressed by the analysis of the slow flow and the reduced system.

Our computations of the half-return map H were performed by numerical integration of the slow flow equations (3.1) from initial conditions on S_2 to their intersection with the circle S_1. We used variable step size Runge–Kutta methods for these integrations, making the error tolerances sufficiently stringent that the apparent errors in the calculations were small enough to produce reliable and robust results. To compute bifurcations of the half-return map H, these numerical integrations were embedded in algorithms that used Newton's method to solve defining equations for each type of bifurcation. For saddle-node bifurcations, the defining equations were obtained by simultaneously solving the fixed point equation $H(\theta) = \theta$ and the equation $H'(\theta) = 1$. For left homoclinic bifurcations, we computed the intersections (θ_{kl}, x_{kl}) of W_s with the line $\theta + 0.5 + k = \theta_{1s}$ in the universal cover of the cylinder $S^1 \times \mathbb{R}$ and then solved the equation $x_{kl} = 2$ using Newton's method. For the right homoclinics, we

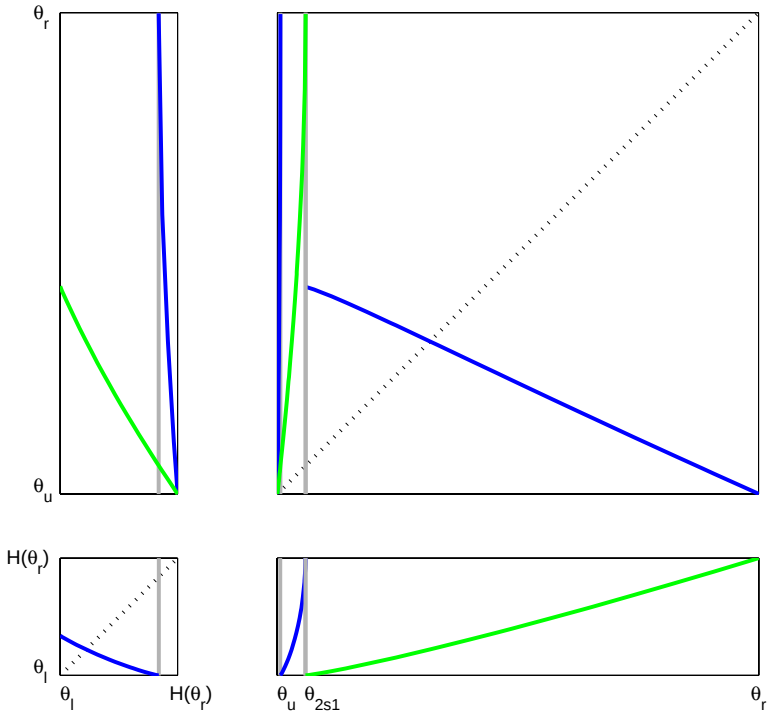


Figure 7.8. *A closer look at H and H^2 for the same parameter values as in Figure 7.7. As in that figure, the green curves are the graph of H, and the blue curves are the graph of H^2. The right interval is $H(J) = [\theta_u, \theta_r]$, and the left interval is $H(H(J) - J) = [\theta_l, H(\theta_r)]$ (where $J = [\theta_u, \theta_{2s1}]$). The union of these two intervals is an invariant set. Its return map consists of the branch of H to the left of the discontinuity and the branch of H^2 to the right of the discontinuity. The coordinate values are $\theta_u \approx 0.541589$, $\theta_{2s1} \approx 0.542117$, $\theta_r \approx 0.550583$, $\theta_l \approx 0.140666$, and $H(\theta_r) \approx 0.142859$.*

computed the intersections (θ_{kr}, x_{kr}) of W_s with the line $\theta + 0.5 + k = \theta_{1u}$ and then solved the equation $x_{kr} = 2$. This procedure of integrating the equation to the desired value of θ and then solving $x = 2$ gave better results than integrating to $x = 2$ because the angles between the vector field and lines of constant θ were much larger than those between the vector field and the lines of constant x at the bifurcation points. This procedure also avoided the necessity of distinguishing the type of right homoclinic to be computed. This was determined after computing the bifurcation point. Similar methods were applied to H^2 to compute the bifurcation curves of period 2 orbits.

The forced van der Pol system is a smooth vector field that does not have discontinuities of its flow. Asymptotic analysis of the behavior of the flow near the folds of its critical manifold is needed to interpret what our results say about bifurcations of the forced van der Pol system. For trajectories that avoid neighborhoods of the folded singularities, classical theory of singularly perturbed systems [2, 19] can be applied for this purpose. However, the analysis of trajectories that pass near the folded singularities is complicated. There even remain gaps in the theory that describes the local geometry of trajectories in the vicinity of the folded nodes [3, 25, 28]. For the global geometry of the flow, there is another level of phenomena

that is not touched upon in this paper. In particular, *canards* are trajectories of the forced van der Pol system that pass near a folded singularity and then continue along the unstable sheet of the critical manifold. These trajectories are not approximated by trajectories of the reduced system as defined here. The canards can be approximated by trajectories of the slow flow that obey different rules about when they jump from one sheet of the slow flow to another, and they give rise to a whole new set of geometric structures in the flow of the forced van der Pol system. The chaotic invariant sets described by Cartwright and Littlewood [8, 9] consist entirely of trajectories that contain canards. The extension of the half-return map to include the canard trajectories, the associated bifurcations, and the existence of horseshoes is addressed in a subsequent paper [4].

REFERENCES

[1] V. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1977.

[2] V. Arnold, V. Afrajmovich, Yu. Ilyashenko, and L. Shil'nikov, *Dynamical Systems* V: *Bifurcation Theory and Catastrophe Theory*, Encyclopaedia Math. Sci., Springer-Verlag, Berlin, 1994.

[3] É. Benoît, *Canards et enlacements*, Inst. Hautes Études Sci. Publ. Math., 72 (1990), pp. 63–91.

[4] 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*, in preparation.

[5] J. Cartwright, V. Eguiluz, E. Hernandez-Garcia, and O. Piro, *Dynamics of elastic excitable media*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 9 (1999), pp. 2197–2202.

[6] M. L. Cartwright, *Balthazar van der Pol*, J. London Math. Soc., 35 (1960), pp. 367–376.

[7] M. L. Cartwright, *van der Pol's equation for relaxation oscillations*, in Contributions to the Theory of Nonlinear Oscillations II, Princeton Ann. Math. Stud. 2, Princeton University Press, Princeton, NJ, 1952, pp. 3–18.

[8] M. Cartwright and J. Littlewood, *On nonlinear differential equations of the second order* I: *The equation* $\ddot{y} - k(1-y^2)\dot{y} + y = bk\cos(\lambda t + a)$, *k large*, J. London Math. Soc., 20 (1945), pp. 180–189.

[9] M. Cartwright and J. Littlewood, *On nonlinear differential equations of the second order* II: *The equation* $\ddot{y} - kf(y,\dot{y})\dot{y} + g(y,k) = p(t) = p_1(t) + kp_2(t)$, $k > 0$, $f(y) \geq 1$, Ann. of Math. (2), 48 (1947), pp. 472–494.

[10] M. Cartwright and J. Littlewood, *Addendum to "On nonlinear differential equations of the second order* II, Ann. of Math. (2), 50 (1949), pp.,504–505.

[11] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–225.

[12] R. FitzHugh, *Impulses and physiological states in models of nerve membrane*, Biophysics J., 1 (1961), pp. 445–466.

[13] J. Flaherty and F. Hoppensteadt, *Frequency entrainment of a forced van der Pol oscillator*, Stud. Appl. Math., 58 (1978), pp. 5–15.

[14] J. Grasman, *Asymptotic Methods for Relaxation Oscillations and Applications*, Springer-Verlag, New York, 1987.

[15] J. Guckenheimer, K. Hoffman, and W. Weckesser, *Numerical computation of canards*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 10 (2000), pp. 2669–2687.

[16] J. Keener, *Chaotic behavior in a piecewise continuous difference equation*, Trans. Amer. Math. Soc., 261 (1980), pp. 589–604.

[17] M. Levi, *Qualitative analysis of the periodically forced relaxation oscillations*, Mem. Amer. Math. Soc., 244 (1981), pp. 1–147.

[18] C. Koch and I. Segev, *Methods in Neuronal Modeling*, MIT Press, Cambridge, MA, 1998.

[19] N. Levinson, *A second order differential equation with singular solutions*, Ann. of Math. (2), 50 (1949), pp. 127–153.

[20] J. Littlewood, *On nonlinear differential equations of the second order* III: *The equation* $\ddot{y} - k(1-y^2)\dot{y} + y = bk\cos(\lambda t + a)$ *for large* k *and its generalizations*, Acta Math., 97 (1957), pp. 267–308 (errata in Acta. Math., 98 (1957), p. 110).

[21] J. Littlewood, *On nonlinear differential equations of the second order* IV: *The equation* $\ddot{y} - kf(y)\dot{y} + g(y) = bkp(\phi), \phi = t + a$ *for large* k *and its generalizations*, Acta Math., 98 (1957), pp. 1–110.

[22] S. McMurran and J. Tattersall, *The mathematical collaboration of M. L. Cartwright and J. E. Littlewood*, Amer. Math. Monthly, 103 (1996), pp. 833–845.

[23] R. Mettin, U. Parlitz, and W. Lauterborn, *Bifurcation structure of the driven van der Pol oscillator*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 3 (1993), pp. 1529–1555.

[24] E. Mischenko and N. Rozov, *Differential Equations with Small Parameters and Relaxation Oscillations*, Plenum Press, New York, 1980.

[25] E. Mischenko, Yu. Kolesov, A. Kolesov, and N. Rozov, *Asymptotic Methods in Singularly Perturbed Systems*, Consultants Bureau, New York, 1994 (translated from the Russian by Irene Aleksanova).

[26] P. F. Rowat and A. I. Selverston, *Modeling the gastric mill central pattern generator of the lobster with a relaxation-oscillator network*, J. Neurophysiology, 70 (1993), pp. 1030–1053.

[27] S. Smale, *Diffeomorphism with many periodic points*, in Differential and Combinatorial Topology, S. Cairns, ed., Princeton University Press, Princeton, NJ, 1963, pp. 63–80.

[28] P. Szmolyan and M. Wechselberger, *Canards in* R^3, J. Differential Equations, 177 (2001), pp. 419–453.

[29] F. L. H. M. Stumpers, *Balth. van der Pol's work on nonlinear circuits*, IRE Trans. Circuit Theory, 7 (1960), pp. 366–367.

[30] F. Takens, *Forced oscillations and bifurcations*, Comm. Math. Inst. Rijkuniv. Utrecht, 3 (1974), pp. 1–59.

[31] F. Takens, *Constrained Equations: A Study of Implicit Differential Equations and Their Discontinuous Solutions*, Report ZW-75-03, Mathematisch Institut, Rijksuniversiteit, Groningen, The Netherlands, 1975.

[32] B. van der Pol, *A theory of the amplitude of free and forced triode vibrations*, Radio Review, 1 (1920), pp. 701–710, 754–762.

[33] B. van der Pol, *On "Relaxation Oscillations"* I, Phil. Mag., 2 (1926), pp. 978–992.

[34] B. van der Pol, *The nonlinear theory of electric oscillations*, Proc. IRE, 22 (1934), pp. 1051–1086.

[35] B. van der Pol and J. van der Mark, *Frequency demultiplication*, Nature, 120 (1927), pp. 363–364.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 1, pp. 36–52

Stability by KAM Confinement of Certain Wild, Nongeneric Relative Equilibria of Underwater Vehicles with Coincident Centers of Mass and Buoyancy*

George W. Patrick†

Abstract. Purely rotational relative equilibria of an ellipsoidal underwater vehicle occur at nongeneric momentum where the symplectic reduced spaces change dimension. The stability of these relative equilibria under momentum changing perturbations is not accessible by Lyapunov functions obtained from energy and momentum. A blow-up construction transforms the stability problem to the analysis of symmetry-breaking perturbations of Hamiltonian relative equilibria. As such, the stability follows by KAM theory rather than energy-momentum confinement.

Key words. relative equilibria, stability, blow-up, KAM

AMS subject classifications. 37J25, 70H14, 37J40, 70H09, 37J15, 70H33

PII. S1111111102404751

Introduction. The phase space $TSE(3)$ with Lagrangian

$$L(A, a, \Omega, v) \equiv \frac{1}{2}\Omega^t \mathbf{I}\Omega + \frac{1}{2}v^t \mathbf{M}v \tag{0.1}$$

approximately models the motion of a neutrally buoyant vehicle submerged in an inviscid irrotational fluid (see [1] and the references therein) in the case of coincident centers of mass and buoyancy. Here tangent vectors of $SE(3)$ are represented by left translation, and elements of $SE(3)$ parameterize the configurations of the vehicle by embedding a reference vehicle into the fluid. $\mathbf{I}$ and $\mathbf{M}$ are constant, positive definite, 3×3 matrices that can be calculated from the shape and mass distribution of the vehicle. This system admits the $SE(3)$-symmetry of the left action of $SE(3)$ on itself.

If the vehicle is an ellipsoid with principal axes of inertia along the axes of the ellipsoid, then $\mathbf{I} = \mathrm{diag}(I_1, I_2, I_3)$ and $\mathbf{M} = \mathrm{diag}(M_1, M_2, M_3)$ (i.e., $\mathbf{I}$ and $\mathbf{M}$ are diagonal). If $M_1 = M_2$ and $I_1 = I_2$ (or, similarly, if $M_1 = M_3$ and $I_1 = I_3$, etc.), then there is a further material symmetry of the system: $SO(2) = \{\exp(\mathbf{k}^\wedge\theta)\}$ acts as a subgroup of $SE(3)$ by inverse multiplication on the right. If $\mathbf{I}$ and $\mathbf{M}$ are both constants of the identity, then the material symmetry is $SO(3)$.

Lie–Poisson reduction yields the Poisson phase space $se(3)^* = \{(\pi, p)\}$, where $\pi = \mathbf{I}\Omega$ and $p = \mathbf{M}v$. The equations of motion are

$$\frac{d\pi}{dt} = \pi \times \Omega + p \times v, \qquad \frac{dp}{dt} = p \times \Omega, \tag{0.2}$$

*Received by the editors April 2, 2002; accepted for publication (in revised form) by J. Marsden October 3, 2002; published electronically February 19, 2003. This work was supported by an NSERC individual research grant and an EPSRC visiting fellowship.
http://www.siam.org/journals/siads/2-1/40475.html
†Department of Mathematics and Statistics, University of Saskatchewan Saskatoon, Saskatoon, Saskatchewan, S7N 5E6, Canada (patrick@math.usask.ca).

and by direct substitution, for each $\alpha_e \in \mathbb{R}$,

$$p_e^{\alpha_e}: \quad \pi = \alpha_e \mathbf{k}, \qquad p = 0,$$

is an equilibrium of the Poisson reduced systems and hence a relative equilibrium of the original system. The generator is

$$\Omega_e^{\alpha_e} \equiv \frac{\alpha_e}{I_3}\mathbf{k}, \quad v_e^{\alpha_e} \equiv 0,$$

so the relative equilibrium corresponds to a stationary vehicle rotating about a principal axis which is aligned with the vertical. This article is concerned with the stability of these relative equilibria in the case in which I_3 is not an intermediate principal axis, i.e., assuming $I_1 < I_2 < I_3$ or $I_3 < I_2 < I_1$.

The symplectic leaves of $se(3)^*$ are as follows. On the complement of $p = 0$ lie the generic symplectic leaves, all diffeomorphic to TS^2, and which are the level sets of the two Casimirs $|p|$ and $\pi \cdot p$. Nongeneric leaves occur within the set $p = 0$ and are the level sets of the subcasimir $|\pi|$. Thus the relative equilibria $p_e^{\alpha_e}$ correspond to Lyapunov stable equilibria on the (nongeneric) symplectic leaves of $se(3)^*$ since the energy has a definite critical point when restricted to those leaves. Were the symmetry group compact, this *leafwise stability* would imply stability of the equilibrium modulo the isotropy group of the momentum. $SE(3)$, of course, is not compact. The question is whether or not $p_e^{\alpha_e}$ are stable under perturbations from nongeneric leaves into nearby generic leaves.

Leonard and Marsden [2] have identified this question as particularly delicate, and the theory of [5], the sharpest possible for the problem of establishing the stability of relative equilibria by energy-momentum confinement in the case of noncompact symmetry, corroborates that opinion. Patrick, Roberts, and Wulff separate generators of relative equilibria into two complementary classes, *tame* and *wild*. The generator of $p_e^{\alpha_e}$ is tame if and only if $\alpha_e = 0$, corresponding to a stationary, nonrotating vehicle, in which case $SE(3)$-stability follows directly since the energy has zero derivative at $p_e^{\alpha_e}$ and has positive definite Hessian there. However, if $\alpha_e \neq 0$, the generator is wild and the theory does not imply stability.

So it is an open question whether the relative equilibria $p_e^{\alpha_e}$, $\alpha_e \neq 0$, are $SE(3)$-stable or not, and the problem appears inaccessible by energy-momentum confinement. This is due to the presence of a noncompact symmetry group and wild generators.

1. The blow-up construction. The stability issue is one of perturbation from a nongeneric symplectic leaf to nearby, higher dimensional generic leaves. In order to bridge to Hamiltonian perturbation theory, which is usually cast in a setting of a fixed canonical phase space, it is natural to begin by normalizing the generic leaves. The leaf corresponding to

$$|p| = a, \quad \pi \cdot p = b$$

for $a > 0$ is diffeomorphic to

$$TS^2 = \left\{(w, \dot{w}) \in \mathbb{R}^3 \times \mathbb{R}^3 : |w| = 1, w \cdot \dot{w} = 0\right\}$$

by the map

$$w = \frac{p}{|p|}, \quad \dot{w} = \pi - \frac{\pi \cdot p}{|p|^2} p,$$

the inverse map being given by, for fixed $a > 0$ and $b \in \mathbb{R}$,

$$p = aw, \quad \pi = \dot{w} + \frac{b}{a} w. \tag{1.1}$$

Having normalized the symplectic leaves to the constant manifold TS^2, one seeks to extend this to the nongeneric leaves within $p = 0$, which means extending it to $a = 0$ since $p = aw$. As it stands, (1.1) is poorly defined for $a = 0$, but for fixed ratios of b/a it is well defined even for arbitrarily small a, suggesting that the proper way to approach the nongeneric leaves from generic ones is through constant $\pi \cdot p/|p|$. *Setting $\gamma \equiv b/a$ and using the parameters a and γ instead of a and b code the generic leaves so they fit smoothly into the nongeneric ones, thus allowing the possibility of an effective perturbation approach.* The map $p = aw$, $\pi = \dot{w} + \gamma w$ for $a = 0$ is many-to-one, and so the three dimensional set of nongeneric leaves $p = 0$ is "blown up" by this map to the five dimensional set of $TS^2 \times \mathbb{R} = \{((w, \dot{w}), \gamma)\}$. Thus one is led to define the *blown-up space* of $se(3)^*$ as

$$\hat{P} \equiv TS^2 \times \mathbb{R}_{\geq 0} \times \mathbb{R} \equiv \{(w, \dot{w}, a, \gamma) : |w| = 1, w \cdot \dot{w} = 0, a \geq 0\}$$

with *blow-down map*

$$p = aw, \quad \pi = \dot{w} + \gamma w$$

and corresponding *blow-up map*, defined on the generic ($p \neq 0$) leaves only,

$$w = \frac{p}{|p|}, \quad \dot{w} = \pi - \frac{\pi \cdot p}{|p|^2} p, \quad a = |p|, \quad \gamma = \frac{b}{a}.$$

The blow-up map is a diffeomorphism from the (open) set of generic leaves to the (open) set $a > 0$ in the blown-up space (the *generic sector*) such that each generic leaf is sent to the constant manifold TS^2. The evolution of the generic leaves is transformed to an evolution on TS^2 parameterized by the Casimir values a and γ. The blow-down map takes the set $a = 0$ in the blown-up space (the *nongeneric sector*) to the set of nongeneric leaves and is many-to-one on that sector. Increasing the parameter a from zero corresponds to leaving the nongeneric leaves and moving to the generic ones, while γ parameterizes the possible avenues of departure.

The utility of the blow-up to support perturbation arguments depends on whether or not the dynamics of the generic sector can be continued smoothly to the nongeneric sector. On

the generic sector, the vector field that generates the dynamics is

$$
\begin{aligned}
\frac{dw}{dt} &= \frac{1}{a}\frac{dp}{dt} = \frac{1}{a}p \times \mathbf{I}^{-1}\pi = w \times \mathbf{I}^{-1}(\dot{w} + \gamma w),\\
\frac{d\dot{w}}{dt} &= \frac{d\pi}{dt} - \gamma\frac{dw}{dt}\\
&= \pi \times \mathbf{I}^{-1}\pi + p \times \mathbf{M}^{-1}p - \gamma\frac{dw}{dt}\\
&= (\dot{w} + \gamma w) \times \mathbf{I}^{-1}(\dot{w} + \gamma w) + aw \times \mathbf{M}^{-1}aw - \gamma w \times \mathbf{I}^{-1}(\dot{w} + \gamma w)\\
&= \dot{w} \times \mathbf{I}^{-1}(\dot{w} + \gamma w) + a^2 w \times \mathbf{M}^{-1}w.
\end{aligned}
\tag{1.2}
$$

This is smooth in a for all $a \geq 0$, as required. Dynamics on $a = 0$ robust enough to continue through perturbation to small positive a will have implications for the original system. By continuity in a, the blown-up vector field is a lift by the smooth blow-down map of the vector field for the original system, even through the nongeneric sector. Thus the flow on the nongeneric sector corresponds through the blow-down map to the flow of the original system on the union of the nongeneric leaves.

The blow-up has a very transparent reformulation since $\hat{P}$ is diffeomorphic to $S^2 \times \mathbb{R}^3 \times \mathbb{R}_{\geq 0} = \{(w, \pi, a)\}$ by the map $\pi = \gamma w + \dot{w}$. Through this diffeomorphism, the blow-down map is simply $p = aw$, which is to say that w by itself is enough to desingularize the foliation by symplectic leaves but not enough to normalize the leaves. The blow-up map is a proper map since the map $(a, w) \mapsto aw$ is proper.

Some exploration of the nongeneric sector of the blown-up space may be helpful for visualization purposes. For fixed π_0, the equation $\gamma w + \dot{w} = \pi_0$ has a solution $w \in S^2$ and $\gamma = \pi_0 \cdot w$. Consequently, the blow-up of the point $(\pi, p) = (\pi, 0)$ is in all cases a two sphere. This sphere intersects fixed γ such that $|\gamma| < |\pi_0|$ in a circle, $\gamma = \pm|\pi_0|$ in a point, and $|\gamma| > |\pi_0|$ not at all. Thus departure from the point $p = 0$, $\pi = \pi_0$ along $\gamma > |\pi_0|$ is impossible. The nongeneric symplectic leaf $|\pi| = r > 0$ blows up to $\gamma^2 + |\dot{w}|^2 = r^2$, which is diffeomorphic to $S^2 \times S^2$ and which for fixed $|\gamma| < r$ is a circle bundle and for $|\gamma| = r$ is a sphere.

The Hamiltonian pulls back through the smooth blow-down map to

$$
\begin{aligned}
\hat{H} &\equiv \frac{1}{2}\pi^t \mathbf{I}^{-1}\pi + \frac{1}{2}p^t \mathbf{M}^{-1}p = \hat{H}^0 + a^2\hat{H}^1,\\
\hat{H}^0 &\equiv \frac{1}{2}(\dot{w} + \gamma w)^t \mathbf{I}^{-1}(\dot{w} + \gamma w), \quad \hat{H}^1 \equiv \frac{1}{2}w^t \mathbf{M}^{-1}w;
\end{aligned}
$$

$\hat{H}$ is written this way in anticipation of perturbation arguments from $a = 0$ to small nonzero a. The symplectic form $\hat{\omega}$ on the nongeneric sector can be calculated from the formula for the coadjoint orbit symplectic forms of $SE(3)$ in [3], with the result that

$$
\begin{aligned}
&\hat{\omega}(w, \dot{w})\big((\delta w_1, \delta\dot{w}_1), (\delta w_2, \delta\dot{w}_2)\big)\\
&\qquad = -w \cdot (\delta w_1 \times \delta\dot{w}_2 - \delta w_2 \times \delta\dot{w}_1) - \gamma w \cdot (\delta w_1 \times \delta w_2).
\end{aligned}
$$

By continuity, the relation $i_{X_{\hat{H}}}\hat{\omega} = d\hat{H}$ persists from $a = 0$ to $a > 0$, so the vector field (1.2) is Hamiltonian at $a = 0$ with symplectic form $\hat{\omega}$ and Hamiltonian $\hat{H}^0$. Thus the evolution

on the nongeneric sector is Hamiltonian in a way that smoothly continues the Hamiltonian structure of the generic sector.

The dynamics on the invariant submanifold $p = 0$ in the original space P admits the subcasimir $|\pi|$. This conserved quantity (*conserved on $p = 0$ only*) pulls back to a conserved quantity $|w + \gamma\dot{w}|$ for the *nongeneric sector* of the blow-up space $\hat{P}$. Since $w \cdot \dot{w} = 0$ and $|w| = 1$, this gives the conserved quantity $|\dot{w}|^2$ and hence the conserved quantity $f(|\dot{w}|)$, where f is any function. The Hamiltonian vector field of $f(|\dot{w}|^2)$ is

$$\frac{dw}{dt} = -\frac{f'(|\dot{w}|)}{|\dot{w}|}\dot{w} \times w, \quad \frac{d\dot{w}}{dt} = -\gamma\frac{f'(|\dot{w}|)}{|\dot{w}|}w \times \dot{w}.$$

Note that $\tilde{m} \equiv \dot{w} + \gamma w$ is conserved by these equations so that

$$\frac{dw}{dt} = -\frac{f'(|\dot{w}|)}{|\dot{w}|}\tilde{m} \times w, \quad \frac{d\dot{w}}{dt} = -\gamma\frac{f'(|\dot{w}|)}{|\dot{w}|}\tilde{m} \times \dot{w},$$

the solution of which is rotations about $\tilde{m}$. To normalize the period at 2π and the right-hand sense about $\tilde{m}$, choose

$$\frac{f'(|\dot{w}|)}{|\dot{w}|}|\tilde{m}| = \frac{f'(|\dot{w}|)}{|\dot{w}|}\sqrt{\gamma^2 + |\dot{w}|^2} = -1,$$

which gives $f(|\dot{w}|) = -\sqrt{\gamma^2 + |\dot{w}|^2}$. *Thus the nongeneric sector has an additional $SO(2)$-symmetry*, which acts by

$$\theta \cdot (w, \dot{w}) \equiv (\exp(m^\wedge\theta)w, \exp(m^\wedge\theta)\dot{w}), \quad m \equiv \frac{\dot{w} + \gamma w}{\sqrt{\gamma^2 + |\dot{w}|^2}},$$

and has momentum

$$\hat{J} \equiv -\sqrt{\gamma^2 + |\dot{w}|^2}.$$

This extra $SO(2)$-symmetry arises from a subcasimir of the original system. The action and the corresponding momentum are defined on the nongeneric sector where γ and $\dot{w}$ are not both zero. The set where $a = \gamma = 0$ and $\dot{w} = 0$ exactly corresponds through the blow-down/up to the set where $p = 0$ and $\pi = 0$, so the relative equilibria $p_e^{\alpha_e}$ are in the domain of definition of the action and its momentum.

Here are some aspects of the $SO(2)$-action and its relation to the blow-down/up map.

1. The action is free except on the set $\dot{w} = 0$, which is a two sphere of fixed points. This two sphere is also the level $-|\gamma|$ of the momentum $\hat{J}$, is a symplectic submanifold of $\hat{P}$, and as such is equal to its own singular reduction.
2. The orbit relation of the action together with the parameter γ exactly absorb the additional phase space from blowing up the nongeneric leaves. Indeed, for fixed γ, the blow-down map is a quotient map for the action, and the orbit space is therefore smooth, irrespective of the fact that the action is not free.
3. The blow-down map restricts to a quotient map for the (singular or nonsingular) symplectic reduced space associated to the $\hat{J} = \hat{\mu}$ level set. As such, this reduced space is symplectomorphic to the nongeneric leaf $p = 0$, $|\pi| = -\hat{\mu}$.

Only the verification of the third item in the nonsingular ($\hat{\mu} < -|\gamma|$) case is troublesome. For that, it is easily verified that the map $\pi = \gamma w + \dot{w}$ is a quotient map for the $SO(2)$-action on $\hat{J}^{-1}(\hat{\mu})$ which has image the sphere $TS^2_{-\hat{\mu}} = \{\pi : |\pi| = -\hat{\mu}\}$. To pull down the symplectic form $\hat{\omega}$ by that map, first let $(\pi, \delta\pi_i) \in TS^2_{-\hat{\mu}}$, $i = 1, 2$, and seek $(w, \dot{w}, \delta w_i, \delta\dot{w}_i)$ such that

$$|w| = 1, \quad w \cdot \dot{w} = 0, \quad w \cdot \delta w_i = 0, \quad \delta w_i \cdot \dot{w} + w \cdot \delta\dot{w}_i = 0,$$
$$-\sqrt{\gamma^2 + |\dot{w}|^2} = \hat{\mu}, \quad \delta\dot{w}_i \cdot \dot{w} = 0, \quad \pi = \gamma w + \dot{w}, \quad \delta\pi_i = \gamma\delta w_i + \delta\dot{w}_i.$$

To solve these equations, choose a w such that $w \cdot \dot{\pi} = \gamma$, and set $\dot{w} = \pi - \gamma w$. Expanding δw_i and $\delta\dot{w}_i$ in the basis w, $\dot{w}$, $w \times \dot{w}$ gives

$$\delta w_i = -\frac{w \cdot \delta\pi_i}{\hat{\mu}^2 - \gamma^2}(\pi - \gamma w), \quad \delta\dot{w}_i = (w \cdot \delta\pi_i)w + \frac{(w \times \pi) \cdot \delta\pi_i}{\hat{\mu}^2 - \gamma^2} w \times \pi.$$

Substitution into $\hat{\omega}$ then gives

$$\begin{aligned}
&\hat{\omega}(w, \dot{w})\big((\delta w_1, \delta\dot{w}_1), (\delta w_2, \delta\dot{w}_2)\big) \\
&\quad = \frac{1}{\hat{\mu}^2 - \gamma^2}\big((\delta\pi_1 \cdot w)((\pi \times \delta\pi_2) \cdot w) - (\delta\pi_2 \cdot w)((\pi \times \delta\pi_1) \cdot w)\big) \\
&\quad = \frac{1}{\hat{\mu}^2 - \gamma^2}(w \times (w \times \pi)) \cdot (\delta\pi_1 \times \delta\pi_2) \\
&\quad = \frac{1}{\hat{\mu}^2 - \gamma^2}((w \cdot \pi)w - \pi) \cdot (\delta\pi_1 \times \delta\pi_2) \\
&\quad = \frac{1}{\hat{\mu}^2 - \gamma^2}\left((w \cdot \pi)\frac{w \cdot \pi}{|\pi|^2}\pi - \pi\right) \cdot (\delta\pi_1 \times \delta\pi_2) \\
&\quad = \frac{1}{\hat{\mu}^2 - \gamma^2}\left(\frac{\gamma^2}{\hat{\mu}^2} - 1\right)\pi \cdot (\delta\pi_1 \times \delta\pi_2) \\
&\quad = -\frac{1}{|\pi|^2}\,\pi \cdot (\delta\pi_1 \times \delta\pi_2).
\end{aligned}$$

This last expression is the symplectic form on the nongeneric leaf $|\pi| = -\hat{\mu}$, as required.

In short, the $SO(2)$-symmetry arises and exactly absorbs the additional dimensions resulting from the blow-up construction. The symplectic reductions of the nongeneric sector by this symmetry exactly coincide with the original system restricted to the nongeneric symplectic leaves of the phase space P.

The pull-back of the relative equilibria $p_e^{\alpha_e}$ by the blow-down map is the set of $(w, \dot{w}, a, \gamma)$ such that

$$p = aw = 0, \quad \pi = \alpha_e \mathbf{k} = \dot{w} + \gamma w.$$

Since $|w| = 1$, the first equation is equivalent to $a = 0$ (the relative equilibria are, of course, in the nongeneric sector), and dotting the second with w shows that it is equivalent to $\gamma = \alpha_e \mathbf{k} \cdot w$ and $\dot{w} = \alpha_e \mathbf{k} - \gamma w$. Since $\gamma^2 + |\dot{w}|^2 = {\alpha_e}^2$ and $\alpha_e \neq 0$, all of these solutions are within the

open set where the $SO(2)$-action is defined (i.e., where γ and $\dot{w}$ are not both zero). Thus the relative equilibria $p_e^{\alpha_e}$ blow up to

$$\hat{p}_e^{\alpha_e}: \quad |w| = 1, \quad \dot{w} = \alpha_e \mathbf{k} - \alpha_e(\mathbf{k}\cdot w)w, \quad a = 0, \quad \gamma = \alpha_e \mathbf{k}\cdot w. \tag{1.3}$$

Since w is unconstrained in (1.3) except for the first equation, $\hat{p}_e^{\alpha_e}$ is diffeomorphic to S^2. Since $m = \mathbf{k}$ on $\hat{p}_e^{\alpha_e}$, the $SO(2)$-action on $\hat{p}_e^{\alpha_e}$ is by rotation of the pair $(w, \dot{w})$ about $\mathbf{k}$. By substitution of $\hat{p}_e^{\alpha_e}$ into (1.2), each point of $\hat{p}_e^{\alpha_e}$ is a relative equilibrium for the $SO(2)$-symmetry, except for the two points $w = \pm\mathbf{k}, \dot{w} = 0, a = 0, \gamma = \pm\alpha_e$, which are equilibria that reside at singular points of the action. Each relative equilibrium in $\hat{p}_e^{\alpha_e}$ has the same generator, namely,

$$\hat{\xi}_e^{\alpha_e} \equiv -\frac{\alpha_e}{I_3}, \tag{1.4}$$

and, by substitution into $\hat{J}$, the same momentum, namely,

$$\hat{\mu}_e^{\alpha_e} \equiv -\alpha_e.$$

Fixing γ, which means fixing a parameter, $\alpha_e = \gamma/\cos\phi$, where ϕ is the angle between $\mathbf{k}$ and w. Thus, for fixed γ, there are two-dimensional submanifolds of relative equilibria, as α_e is varied, as expected for an $SO(2)$-symmetric Hamiltonian system. Along those submanifolds there is the *momentum-generator* relation

$$\hat{\xi}_e^{\alpha_e} = \frac{1}{I_3}\hat{\mu}_e^{\alpha_e}, \tag{1.5}$$

which will give a crucial component in the KAM twist condition to follow.

2. Normal forms in the blown-up system. The stability of the relative equilibrium $p_e^{\alpha_e}$ follows from the stability of its blow-up $\hat{p}_e^{\alpha_e}$.

Proposition 2.1. *Suppose that, for some fixed α_e, $\hat{p}_e^{\alpha_e}$ is stable for the flow $\hat{F}_t$ on $\hat{P}$ in the sense that, for all neighborhoods $\hat{U}$ of $\hat{p}_e^{\alpha_e}$, there is a neighborhood $\hat{V}$ of $\hat{p}_e^{\alpha_e}$ such that $\hat{F}_t(\hat{p}) \in \hat{U}$ for all $\hat{p} \in \hat{V}$. Then $p_e^{\alpha_e}$ is a stable relative equilibrium.*

Proof. Suppose U is a neighborhood of $p_e^{\alpha_e}$. U pulls back by the blow-down map to an open neighborhood $\hat{U}$ of $\hat{p}_e^{\alpha_e}$. Let $\hat{V}$ be a neighborhood as in the statement of the proposition. Then it suffices to show that $\hat{V}$ pushes forward by the blow-down map to a neighborhood of $p_e^{\alpha_e}$. However, this follows since the blow-down map is proper. ∎

In particular, if all of the relative equilibria and both equilibria in $\hat{p}_e^{\alpha_e}$ are stable under perturbation both within the phase space TS^2 and in the parameters a and γ, then the original relative equilibrium $p_e^{\alpha_e}$ is stable. When a is perturbed away from 0, this is an $SO(2)$-symmetry breaking perturbation. As $\dim TS^2 = 4$, the blown-up system is integrable when $a = 0$, and hence the stability issue is one of the stability of periodic orbits of a nearly integrable Hamiltonian system.

Assume, without loss of generality, that $\alpha_e > 0$. Since the $SO(2)$-symmetry on $\hat{p}_e^{\alpha_e}$ is by rotation about $\mathbf{k}$, it suffices to consider the stability of orbits in $\hat{p}_e^{\alpha_e}$ emanating from points $\hat{p}_e^{\alpha_e,\theta}$ obtained by substituting $w = \sin\theta\mathbf{i} + \cos\theta\mathbf{k}$ into (1.3) for $\theta \in [0, \pi]$.

2.1. Normal form for the relative equilibria. Consider first the *relative equilibria* in $\hat{p}_e^{\alpha_e,\theta}$; i.e., exclude the *equilibria* corresponding to $\theta = 0$ and $\theta = \pi$. The argument proceeds by adapting and incrementally refining, to the order required for the stability analysis, the normal form for relative equilibria developed in [6].

Below $O(x;y)^k$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ will denote the set of smooth y-dependent functions such that $O(x;y)/|x|^k$ is bounded near 0. The product $O(x;y)^k O(x';y')^{k'}$ denotes the set of finite sums of products of elements of $O(x;y)^k$ and $O(x';y')^{k'}$.

2.1.1. Initial normal form. This is constructed from the linearization of the relative equilibrium, which means the linearization at $\hat{p}_e^{\alpha_e,\theta}$ of Hamiltonian vector field $X_{\hat{H}_{\xi_e^{\alpha_e}}}$, where

$$\hat{H}^0_{\xi_e^{\alpha_e}} \equiv \hat{H}^0 - \hat{\xi}_e \hat{J}.$$

The characteristic polynomial of the linearization is $x \mapsto x^2(x^2 + {\omega_e}^2)$, where

$$\omega_e \equiv \pm\alpha_e \sqrt{\left(\frac{1}{I_3} - \frac{1}{I_1}\right)\left(\frac{1}{I_3} - \frac{1}{I_2}\right)}. \tag{2.1}$$

For later convenience, define ω_e to be positive if $I_3 > I_1$ and $I_3 > I_2$ and negative if $I_3 < I_1$ and $I_3 < I_2$. The linearization has a 0 and $\pm i\omega_e$ generalized eigenspaces, both of dimension 2. Introducing the parameter

$$D \equiv \frac{I_2(I_3 - I_1)}{I_1(I_3 - I_2)},$$

the vectors

$$\begin{aligned}
v_1 &\equiv \frac{D^{\frac{1}{4}}}{\sqrt{\alpha_e}} \begin{bmatrix} 0 & \cos\theta & 0 & 0 & \alpha_e \sin^2\theta & 0 \end{bmatrix}, \\
v_2 &\equiv \frac{D^{-\frac{1}{4}}}{\sqrt{\alpha_e}} \begin{bmatrix} \cos\theta & 0 & -\sin\theta & \alpha_e \sin^2\theta & 0 & \alpha_e \sin\theta\cos\theta \end{bmatrix}, \\
v_3 &\equiv \sin\theta \begin{bmatrix} 0 & 1 & 0 & 0 & -\alpha_e\cos\theta & 0 \end{bmatrix}, \\
v_4 &\equiv \frac{1}{\alpha_e \sin\theta} \begin{bmatrix} -\cos^2\theta & 0 & \cos\theta\sin\theta & \alpha_e\cos^3\theta & 0 & -\alpha_e\sin\theta(1+\cos^2\theta) \end{bmatrix}
\end{aligned}$$

form a basis of $T_{\hat{p}_e^{\alpha_e,\theta}} S^2$ which satisfies the following:

1. the basis is symplectically canonical so that the symplectic form with respect to it is

$$\omega(\hat{p}_e^{\alpha_e,\theta}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix};$$

2. with respect to the basis, the derivative of the momentum is

$$d\hat{J}(\hat{p}_e^{\alpha_e,\theta}) = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix};$$

3. the third basis vector v_3 is the infinitesimal generator action corresponding to $1 \in so(2)$;
4. the first two vectors v_1, v_2 span the ω_e generalized eigenspace, and the last two v_3, v_4 span the 0 generalized eigenspace of the linearization;
5. the linearization of the relative equilibrium is

$$dX_{\hat{H}^0_{\xi_e^{\alpha_e}}}(\hat{p}_e^{\alpha_e,\theta}) = \begin{bmatrix} 0 & \omega_e & 0 & 0 \\ -\omega_e & 0 & 0 & 0 \\ 0 & 0 & 0 & \kappa_e \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.2}$$

where

$$\kappa_e \equiv 1/I_3. \tag{2.3}$$

Consequently, the basis effects a Witt–Moncrief decomposition

$$T_{\hat{p}_e^{\alpha_e,\theta}}\hat{P} = N_1 \oplus so(2) \oplus so(2)^* \equiv \operatorname{span}(v_1, v_2) \oplus \mathbb{R}v_3 \oplus \mathbb{R}v_4.$$

Here N_1, the *symplectic normal*, may be identified with the tangent space to the symplectic reduced space at $\hat{p}_e^{\alpha_e,\theta}$ for the $SO(2)$-action. The appearance of the nilpotent part of the linearization is the foundational element of [6]. The value of κ_e coincides with the derivative $d\hat{\xi}_e^{\alpha_e}/d\hat{\mu}_e^{\alpha_e}$ from (1.5), as predicted by the general theory.

The initial normal form can now be transcribed from the data above and is

$$\begin{aligned} \hat{H} &= \frac{\omega_e}{2}(q^2+p^2) + \hat{\xi}_e^{\alpha_e}\nu + \frac{1}{2}\kappa_e\nu^2 + R(q,p,\nu) + \frac{a^2}{2}\hat{H}^1(q,p,\varphi,\nu), \\ R &= O(q,p,\nu)^3 \end{aligned} \tag{2.4}$$

on the product of $\mathbb{R}^2 \times T^*SO(2) = \{(q,p),(\varphi,\nu)\}$ with the product symplectic form $dq \wedge dp + d\varphi \wedge d\nu$, with $SO(2)$ acting by lifts of its left action on itself, and with the momentum map $\nu - \alpha_e$. The transcription is that there is an $SO(2)$ intertwining symplectic diffeomorphism from a neighborhood of the $SO(2)$-orbit of $\hat{p}_e^{\alpha_e}$ to a neighborhood of 0 times the zero section of $T^*SO(2)$ which

1. sends the relative equilibrium $\hat{p}_e^{\alpha_e,\theta}$ to $p = q = \nu = \varphi = 0$;
2. has derivative at $\hat{p}_e^{\alpha_e}$, the identity map with respect to the basis v_i and the standard basis of $\mathbb{R}^2 \times T^*SO(2)$;
3. intertwines the momentum maps $\hat{J}$ and $\nu - \alpha_e$.

Thus the transcription is *structure-preserving* in that it is symplectic and it preserves the $SO(2)$-symmetry and momentum, so the blown-up system near the group orbit of the relative equilibrium $\hat{p}_e^{\alpha_e,\theta}$ can be replaced by the entirely equivalent system (2.4) near $q = p = \nu = 0$.

2.1.2. Elimination of $qO(\nu)^2$, $pO(\nu)^2$, and $(q^2-p^2)\nu$. The remainder term of (2.4) can be expanded as

$$\begin{aligned} R =& c_1(\nu)q + c_2(\nu)p + c_3\nu(q^2-p^2) + c_4\nu(q^2+p^2) \\ &+ O(q,p)^3 + O(\nu)^3 + O(q,p)^2O(q,p,\nu)^2, \end{aligned}$$

where $c_1(\nu) = O(\nu)^2$, $c_2(\nu) = O(\nu)^2$, and c_3, c_4 are constants. The transformation

$$\tilde{q} = q + \frac{c_1}{\omega_e}, \quad \tilde{\varphi} = \varphi + \frac{p\nu}{\omega_e}\frac{dc_1}{d\nu},$$

suggested by completing the square in $\frac{1}{2}\omega_e q^2 + c_1 q\nu^2$, is structure-preserving and changes the Hamiltonian to the same form but without terms of the form $qO(\nu)^2$. Similarly, one eliminates $pO(\nu)^2$. The transformation

$$\tilde{q} = \frac{q}{f(\nu)}, \quad \tilde{p} = f(\nu)p, \quad f(\nu) = \left(\frac{1 - \frac{2c_3}{\omega_e}\nu}{1 + \frac{2c_3}{\omega_e}\nu}\right)^{\frac{1}{4}}, \tag{2.5}$$

takes the fragment $\frac{\omega_e}{2}(q^2 + p^2) + c_3(q^2 - p^2)\nu$ to

$$\left(\frac{\omega_e}{2} + c_3\right) q^2 + \left(\frac{\omega_e}{2} - c_3\right) p^2 = \frac{\omega_e}{2}(\tilde{q}^2 + \tilde{p}^2) + \tilde{q}^2 O(\nu)^2 + \tilde{p}^2 O(\nu)^2,$$

while the symplectic form becomes

$$\begin{aligned} dq \wedge dp + d\varphi \wedge d\nu &= d\tilde{q} \wedge dp + d\varphi \wedge d\nu + \frac{f'}{f}(qdp + pdq) \wedge d\nu \\ &= d\tilde{q} \wedge dp + d\left(\varphi + \frac{f'}{f}qp\right) \wedge d\nu. \end{aligned}$$

Adjoining $\tilde{\varphi} = \varphi + (f'/f)qp$ to (2.5) gives a structure-preserving symplectic transformation that eliminates the term $c_3(q^2 - p^2)$. Thus, without loss of generality,

$$R = c_4(q^2 + p^2) + O(q,p)^3 + O(\nu)^3 + O(q,p)^2 O(q,p,\nu)^2. \tag{2.6}$$

2.1.3. Normal form for the rigid body. We will require the first two terms of the normal form corresponding to the equilibrium $\pi = \alpha_e \mathbf{k}$ of the blown-up system reduced by its $SO(2)$-symmetry, i.e., the symplectic reduced spaces of the rigid body $\frac{1}{2}\pi \mathbf{I}^{-1}\pi$. The map

$$\pi = \left(\left(\alpha_e - \frac{1}{4}(Q^2 + P^2)\right)^{\frac{1}{2}} P, \left(\alpha_e - \frac{1}{4}(Q^2 + P^2)\right)^{\frac{1}{2}} Q, \alpha_e - \frac{1}{2}(Q^2 + P^2)\right)$$

is a symplectic chart on the reduced space $|\pi| = \alpha_e$, and in these coordinates, the Hamiltonian becomes, up to a constant,

$$\frac{1}{2}\pi^t \mathbf{I}^{-1}\pi = \frac{1}{2}\left(\alpha_e - \frac{1}{4}(Q^2 + P^2)\right)\left(\left(\frac{1}{I_1} - \frac{1}{I_3}\right)P^2 + \left(\frac{1}{I_2} - \frac{1}{I_3}\right)Q^2\right).$$

Action-angle variables for the linearized flow are

$$Q = \sqrt{2I}D^{\frac{1}{4}}\sin\psi, \qquad P = \sqrt{2I}D^{-\frac{1}{4}}\cos\psi,$$

and the Hamiltonian is then

$$\frac{1}{2}\pi^t \mathbf{I}^{-1}\pi = \omega_e I - \frac{\omega_e}{2\alpha_e}(D^{\frac{1}{2}}\sin^2\psi + D^{-\frac{1}{2}}\cos^2\psi)I^2.$$

By averaging over ψ,

$$\frac{1}{2}\pi^t \mathbf{I}\pi = \omega_e I + \frac{1}{2}\upsilon_e I^2 + O(Q,P)^5, \tag{2.7}$$

where

$$\upsilon_e = -\frac{\omega_e}{2\alpha_e}(D^{\frac{1}{2}} + D^{-\frac{1}{2}}) = \frac{1}{2}\left(\frac{2}{I_3} - \frac{1}{I_1} - \frac{1}{I_2}\right). \tag{2.8}$$

2.1.4. Matching and normalizing the reduced spaces at $\hat{p}_e^{\alpha_e,\theta}$. For $a = 0$, the symplectic reduced space through $\hat{p}_e^{\alpha_e,\theta}$ of the blown-up system is the $|\pi| = \alpha_e$ symplectic reduced space of the rigid body $\frac{1}{2}\pi^t\mathbf{I}\pi$. For $a = 0$, the symplectic reduced space of the normal form (2.4) through $q = p = \varphi = \nu = 0$ is $\mathbb{R}^2$ with symplectic form $dq \wedge dp$ and Hamiltonian $\hat{H}|_{\nu=0}$. Since the intertwining map between the blown-up system and the normal form is structure-preserving, it descends to symplectomorphisms of reductions of these two systems. Consequently, by symplectomorphism on (q,p) only, the normal-form Hamiltonian (2.4) at $q = p = \nu = 0$ can be equated to the rigid body normal form (2.7), after which the normal form is correct to fourth order in pure q and p and

$$R = O(q,p)^5 + \nu O(q,p)^2 O(q,p,\nu).$$

2.1.5. Refinement by matching the normal forms and generators along the relative equilibria near $\hat{p}_e^{\alpha_e,\theta}$. An advantage may be obtained by comparing the reduced normal along the relative equilibria $\hat{p}_e^{\alpha_e+z,\theta}$ as z varies. These relative equilibria occur (for both systems) at momentum $-(\alpha_e + z)$. For the rigid body, the only z-dependent adjustment is in the α_e dependence of the linearized frequency, which becomes $\omega_e(\alpha_e + z)/\alpha_e$, so the normal form is

$$\left(\omega_e + \frac{\omega_e}{\alpha_e}z\right) I + \frac{1}{2}\upsilon_e I^2 + O(q,p;z)^5. \tag{2.9}$$

For (2.4), it is the normal form of the reduction at $\nu = -z$, so it is the normal form of the Hamiltonian

$$\omega_e I + \frac{1}{2}\upsilon_e I^2 - 2c_4 I z + O(q,p)^2 O(q,p,z)^2,$$

which is

$$(\omega_e - 2c_4 z + O(z)^2) I + O(q,p;z)^3. \tag{2.10}$$

Comparing (2.9) and (2.10) at first order in I gives a crucial fact:

$$c_4 = -\frac{\omega_e}{2\alpha_e}.$$

Also, the $SO(2)$-generator of the blown-up system at $\hat{p}_e^{\alpha_e+z,\theta}$, which is $-(\alpha_e + z)/I_3$, and the $SO(2)$-generator of system (2.4) at the relative equilibrium $q = p = 0$ are the same. Equating these gives

$$\hat{\xi}_e^{\alpha_e} + \kappa_e \nu + \left.\frac{\partial R}{\partial \nu}\right|_{\substack{q=p=0 \\ \nu=-z}} = -\frac{1}{I_3}(\alpha_e + z) = \hat{\xi}_e^{\alpha_e} - \kappa_e z,$$

which means that R has no pure ν terms. Particularly, the $O(\nu)^3$ term in (2.6) is zero.

2.1.6. Symmetry-breaking term. The transcription to the initial normal form is known to first order since the derivative of the transcription is the identity map along the $SO(2)$-orbit of $\hat{p}_e^{\alpha_e,\theta}$. Consequently, $\hat{H}^1$ can be calculated to first order by substitution of

$$w = \exp(\varphi\mathbf{k})P_w(\hat{p}_e^{\alpha_e,\theta} + (qv_1 + pv_2 + \nu v_4))$$

into $w^t\mathbf{M}^{-1}w$, where $P_w(w,\dot{w}) = w$.

2.1.7. Altogether. Putting all the foregoing together, the normal form is

$$\begin{aligned}\hat{H} = \omega_e I + \frac{1}{2}\upsilon_e I^2 + \hat{\xi}_e^{\alpha_e}\nu + \frac{1}{2}\kappa_e\nu^2 - \frac{\omega_e}{\alpha_e}I\nu \\ + \ O(q,p)^5 + \nu O(q,p)^2 O(q,p,\nu) \\ + \ a^2\hat{H}^{1,0}(q,p,\varphi,\nu) + a^2\hat{H}^{1,1}(q,p,\varphi,\nu) + a^2 O(q,p,\nu;\varphi)^2,\end{aligned} \tag{2.11}$$

where

$$\hat{H}^{1,0} \equiv \frac{M_2 - M_1}{2M_1M_2}\sin^2\theta\cos^2\varphi \tag{2.12}$$

and

$$\begin{aligned}\hat{H}^{1,1} \equiv \frac{(M_2 - M_1)\sin 2\theta}{4M_1M_2\sqrt{\alpha_e}}\big(-D^{\frac{1}{4}}q\sin 2\varphi + D^{-\frac{1}{4}}p\cos 2\varphi\big) \\ - \frac{\cos^2\theta}{\alpha_e}\left(\frac{1}{M_1}\cos^2\varphi + \frac{1}{M_2}\sin^2\varphi - \frac{1}{M_3}\right)\nu \\ - \frac{\sin 2\theta}{4D^{\frac{1}{4}}\sqrt{\alpha_e}}\left(\frac{2}{M_3} - \frac{1}{M_1} - \frac{1}{M_2}\right)p.\end{aligned}$$

The details of the symmetry-breaking term $\hat{H}^1$ are not required for the stability analysis and are displayed here for the sake of completeness. The functional form of $\hat{H}^{1,1}$ depends on the choice of the basis v_i, and further normalization or analysis would be required to extract information from it.

2.2. Normal form for the equilibria. The two equilibria $\hat{p}_e^{\alpha_e,0}$ and $\hat{p}_e^{\alpha_e,\pi}$ corresponding to $w = \mathbf{k}$ and $w = -\mathbf{k}$, respectively, remain to be considered. These equilibria are fixed points of the action of $SO(2)$, and the analysis requires a transparent extension of the normal form in [6] to *equilibria* which have $SO(2)$-isotropy.

It suffices to consider $\hat{p}_e^{\alpha_e,0}$; the case of $\hat{p}_e^{\alpha_e,\pi}$ is similar. There is a one-parameter family of possible linearizations of the equilibrium, namely, the linearizations at $\hat{p}_e^{\alpha_e,0}$ of the Hamiltonian vector fields $X_{\hat{H}_\lambda^0}$, where $\hat{H}_\lambda^0 - \lambda\hat{J}$. These linearizations have characteristic polynomials

$$x \mapsto (x^2 + (\alpha_e + \lambda I_3)^2)(x^2 + {\omega_e}^2).$$

Choosing $\lambda = -\alpha_e/I_3$ gives the largest possible null space and therefore the largest number of intrinsically defined higher order terms. The vectors

$$\begin{aligned}v_{1,0} \equiv \frac{D^{\frac{1}{4}}}{\sqrt{\alpha_e}}\begin{bmatrix}0 & 1 & 0 & 0 & 0 & 0\end{bmatrix}, \quad v_{2,0} \equiv \frac{D^{-\frac{1}{4}}}{\sqrt{\alpha_e}}\begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\end{bmatrix}, \\ v_{3,0} \equiv \frac{1}{\sqrt{\alpha_e}}\begin{bmatrix}0 & 1 & 0 & 0 & -\alpha_e & 0\end{bmatrix}, \quad v_{4,0} \equiv \frac{1}{\sqrt{\alpha_e}}\begin{bmatrix}-1 & 0 & 0 & \alpha_e & 0 & 0\end{bmatrix}\end{aligned}$$

form a basis of $T_{\hat{p}_e^{\alpha_e,0}}S^2$ which is symplectically canonical and such that the linearization $dX_{H^0_\lambda}(\hat{p}_e^{\alpha_e,0})$ is the linearization (2.2) with κ_e replaced by zero. The first two vectors span the tangent space to the (singular) reduced space through $p_e^{\alpha_e,0}$, which is the $|\pi| = \alpha_e$ symplectic reduced space of the rigid body $\frac{1}{2}\pi^t\mathbf{I}^{-1}\pi$.

Higher order terms of the Taylor expansion of $X_{H^0_\lambda}$ are intrinsically polynomials on $\ker dX_{H^0_\lambda}(\hat{p}_e^{\alpha_e,0})$. Alternately, one can compute the higher order terms of the Taylor expansion of the Hamiltonian on the null space. Letting (x, y) be the coordinates on $\ker dX_{H^0_\lambda}(\hat{p}_e^{\alpha_e,0})$ indicated by the last of the two basis vectors above, the Hamiltonian on the null space is easily computed to be

$$\frac{1}{8I_3}(x^2+y^2)^2 + O(x,y)^5 = \frac{\kappa_e}{8}(x^2+y^2)^2 + O(x,y)^5.$$

The initial normal form, obtained by the equivariant Darboux theorem, is the Hamiltonian

$$\begin{aligned}\hat{H} = \frac{\omega_e}{2}(q^2+p^2) - \frac{\alpha_e}{2I_3}(x^2+y^2) + \frac{\kappa_e}{8}(x^2+y^2)^2 \\ + \ O(q,p)O(q,p,x,y)^2 + O(x,y)^5 + a^2\hat{H}^1(q,p,x,y)\end{aligned}$$

on the phase space $\mathbb{R}^2\times\mathbb{R}^2 = \{(q,p),(x,y)\}$, with symplectic form $dq\wedge dp + dx\wedge dy$, with $SO(2)$ acting by counterclockwise rotation on (x, y), and with the momentum mapping $-\frac{1}{2}(x^2+y^2)$. The transcription is by local symplectic diffeomorphism with analogous properties to those stated in section 2.1.1.

Manipulations similar to those in section 2.1 are required, as follows:

1. Linear terms in q of the form $qO(x,y)^2$ can be removed as in section 2.1.2, and so can linear terms in p, x, and y. Quartic terms of the form $O(q,p)^2O(x,y)^2$ must by $SO(2)$-invariance be in $(x^2+y^2)O(q,p)$ and so can be written as sums of $(q^2+p^2)(x^2+y^2)$ and $(q^2-p^2)(x^2+y^2)$, and the latter kind can be removed, as in section 2.1.2.
2. Pure q and p terms up to order 4 can be found by matching the reduced system of the initial normal form to rigid body reduced spaces.
3. By items 1 and 2, all terms up to and including order 4 are removed or calculated, except for the coefficient of the term $(q^2+p^2)(x^2+y^2)$. This can be found by matching normal forms along the equilibria $q = p = 0$ (which are fixed points of the action of $SO(2)$), and the resulting term is $-\frac{\omega_e}{4\alpha_e}(q^2+p^2)(x^2+y^2)$.
4. The remainder after all of that, having no terms linear in any variable, and being of degree at least 5, is of the form $O(q,p,x,y)^2O(q,p,x,y)^3$ and is $SO(2)$-invariant.
5. The symmetry-breaking term $\hat{H}^1$ can be calculated as in section 2.1.6 by substituting

$$w = P_w(\hat{p}_e^{\alpha_e,0} + qv_{1,0} + pv_{2,0} + xv_{3,0} + yv_{4,0})$$

into $w^t\mathbf{M}^{-1}w$ and keeping the leading terms, which are of order 2.

Altogether, the normal form is

$$\begin{aligned}\hat{H} =& \omega_e I + \frac{1}{2}\upsilon_e I^2 + \xi_e^{\alpha_e}\nu + \frac{1}{2}\kappa_e\nu^2 - \frac{\omega_e}{\alpha_e}I\nu + O(q,p,x,y)^2O(q,p,x,y)^3 \\ &+ \frac{a^2}{2}\hat{H}^{1,1}(q,p,x,y) + a^2O(q,p,x,y)^3,\end{aligned} \tag{2.13}$$

where $I = \frac{1}{2}(q^2+p^2)$ and $\nu = \frac{1}{2}(x^2+y^2)$, and

$$\hat{H}^{1,1} \equiv \frac{(M_3-M_1)(D^{-\frac{1}{4}}p-y)^2}{\alpha_e M_1 M_3} + \frac{(M_3-M_2)(D^{\frac{1}{4}}q+x)^2}{\alpha_e M_2 M_3}.$$

3. Stability. The rescaling

$$I = a^{2c}\tilde{I}, \quad \nu = a^{2c}\tilde{\nu}$$

is symplectic with multiplier a^{2c}. Substituting into (2.11) and dropping the tildes give

$$\begin{aligned}\frac{1}{a^{2c}}\hat{H} = \omega_e I + \frac{1}{2}\upsilon_e I^2 a^{2c} + \hat{\xi}_e^{\alpha_e}\nu + \frac{1}{2}\kappa_e \nu^2 a^{2c} - \omega_e I \nu a^{2c} + O(a;q,p,\varphi,\nu)^{3c} \\ + \ a^{2-2c}\hat{H}^{1,0} + O(a;q,p,\varphi,\nu)^{2-c}.\end{aligned}$$

Matching the exponents of a in the first nontrivial terms of the integrable part, i.e., $\frac{1}{2}\kappa_e\nu^2 a^{2c}$ and $\frac{1}{2}\upsilon_e I^2 a^{2c}$, with the first term of the nonintegrable part, gives $2c = 2-2c$ or $c = \frac{1}{2}$. After putting $\epsilon = \sqrt{a}$, and disposing of the factor $1/a^{2c}$ of $\hat{H}$, which merely reparameterizes time, one has

$$\hat{H} = \omega_e I + \hat{\xi}_e^{\alpha_e}\nu + \left(\frac{1}{2}\kappa_e\nu^2 - \frac{\omega_e}{\alpha_e}I\nu + \frac{1}{2}\upsilon_e I^2 + \hat{H}^{1,0}\right)\epsilon^2 + O(\epsilon)^3, \tag{3.1}$$

where the dependence of $O(\epsilon)^3$ on all of q, p, φ, and ν has been notationally suppressed. For $\epsilon = 0$, the Hamiltonian (3.1) has a periodic orbit cylinder by varying φ and ν with $I = 0$. The orbit $I = \nu = 0$ corresponds to the relative equilibrium $\hat{p}_e^{\alpha_e,\theta}$.

For determining stability, it suffices to approximate the Poincaré map for the orbit corresponding to $\nu = 0$ in the zero energy level. Solving (3.1) for ν when $H = 0$ gives

$$\nu = -\frac{\omega_e I}{\hat{\xi}_e^{\alpha_e}} + O(\epsilon)^2,$$

and the equations of motion for (3.1) are

$$\begin{aligned}\frac{d\psi}{dt} &= \omega_e + \left(\upsilon_e I - \frac{\omega_e}{\alpha_e}\nu\right)\epsilon^2 + O(\epsilon)^3, \quad \frac{dI}{dt} = O(\epsilon)^3, \\ \frac{d\varphi}{dt} &= \hat{\xi}_e^{\alpha_e} + \left(\kappa_e\nu - \frac{\omega_e}{\alpha_e}I\right)\epsilon^2 + O(\epsilon)^3, \quad \frac{d\nu}{dt} = -\frac{\partial \hat{H}^{1,0}}{\partial\varphi}\epsilon^2 + O(\epsilon)^3.\end{aligned} \tag{3.2}$$

On the zero energy level, the equation for the evolution of φ is

$$\frac{d\varphi}{dt} = \hat{\xi}_e^{\alpha_e} - \frac{\omega_e(\kappa_e\alpha_e + \hat{\xi}_e^{\alpha_e})}{\alpha_e\hat{\xi}_e^{\alpha_e}}I\epsilon^2 + O(\epsilon)^3. \tag{3.3}$$

With an initial condition $I = I_0$, the second equation of (3.2) gives $I = I_0 + O(\epsilon)^3$, so the return time of the Poincaré map is, from (3.3),

$$T \equiv -\frac{2\pi}{\hat{\xi}_e^{\alpha_e}}\left(1 + \frac{(\alpha_e\kappa_e + \hat{\xi}_e^{\alpha_e})\omega_e I_0}{\alpha_e(\hat{\xi}_e^{\alpha_e})^2}\epsilon^2\right) + O(\epsilon)^3.$$

Solving the first two equations of (3.2) over this period, dropping the subscript 0 for the initial conditions, and using the relation $\hat{\xi}_e^{\alpha_e} = -\kappa_e \alpha_e$ to eliminate α_e give

$$
\begin{aligned}
I' &= I + O(\epsilon)^3, \\
\psi' &= \psi - \frac{2\pi\omega_e}{\hat{\xi}_e^{\alpha_e}} - \epsilon^2 \frac{2\pi}{(\hat{\xi}_e^{\alpha_e})^3}(\upsilon_e(\hat{\xi}_e^{\alpha_e})^2 - \kappa_e {\omega_e}^2) I + O(\epsilon)^3.
\end{aligned} \tag{3.4}
$$

This is of the form [4, Theorem 2, p. 231], namely, $(I, \psi) \mapsto (I', \psi')$ by

$$
\begin{aligned}
I' &= I + \epsilon^{r+s} c(I, \psi, \epsilon), \\
\psi' &= \psi + \omega + \epsilon^s h(I) + \epsilon^{s+r} d(I, \psi, \epsilon),
\end{aligned}
$$

with $r = 1$, $s = 2$, and

$$
h(I) \equiv -\frac{2\pi}{(\hat{\xi}_e^{\alpha_e})^3}(\upsilon_e(\hat{\xi}_e^{\alpha_e})^2 - \kappa_e {\omega_e}^2) I.
$$

The twist condition $dh/dI \neq 0$ is

$$
\upsilon_e(\hat{\xi}_e^{\alpha_e})^2 - \kappa_e {\omega_e}^2 \neq 0, \tag{3.5}
$$

which, after substituting (1.4), (2.1), (2.3), and (2.8), is

$$
\begin{aligned}
&-\frac{1}{2}\left(\frac{2}{I_3} - \frac{1}{I_1} - \frac{1}{I_2}\right) + I_3 \left(\frac{1}{I_3} - \frac{1}{I_1}\right)\left(\frac{1}{I_3} - \frac{1}{I_2}\right) \\
&\qquad = \frac{1}{I_1 I_2}\left(I_3 - \frac{1}{2}(I_1 + I_2)\right) \neq 0.
\end{aligned}
$$

This is certainly true if I_3 is not between I_1 and I_2.

As for the equilibria $\hat{p}_e^{\alpha_e,0}$ of section 2.2, Arnold's stability theorem [4, Theorem 1, p. 235] together with the normal form (2.13) imply that the equilibrium is stable when $a = 0$ if

$$
\omega_e I + \frac{1}{2}\upsilon_e I^2 + \hat{\xi}_e^{\alpha_e}\nu + \frac{1}{2}\kappa_e \nu^2 - \frac{\omega_e}{\alpha_e} I\nu \Big|_{\substack{I = \hat{\xi}_e^{\alpha_e} \\ \nu = -\omega_e}} = \frac{1}{2}(\upsilon_e(\hat{\xi}_e^{\alpha_e})^2 - \kappa_e {\omega_e}^2) \neq 0,
$$

which is the same as the twist condition (3.5). Stability follows for sufficiently small nonzero a since a contributes continuously.

Thus $p_e^{\alpha_e}$ are stable as equilibria on the Poisson reduced space $\{(\pi, p)\}$, corresponding to $SE(3)$-stability on the full phase space. As shown by [5], depending on the isotropy group of the momentum, $SE(3)$-stability can be sharpened, irrespective of whether it arises from KAM theory or energy-momentum confinement. The momentum at $p_e^{\alpha_e}$ has zero translational part and rotational part parallel to $\mathbf{k}$, and, in this case, one has $A_{\epsilon_0,\epsilon_1}(\mathbf{k})$-stability for any $\epsilon_0 > 0$ and $\epsilon_1 > 0$, where

$$
A_{\epsilon_0,\epsilon_1}(\mathbf{k}) \equiv \left\{(A, a) \in SE(3) : |\sin\theta_{A,\mathbf{k}}| < \epsilon_1 |a| + \epsilon_0\right\},
$$

$\theta_{A,\mathbf{k}}$ being the angle between $A\mathbf{k}$ and $\mathbf{k}$. $A_{\epsilon_0,\epsilon_1}$ contains (A, a), where A is arbitrary for arbitrarily small ϵ_0, ϵ_1 as long as a is large enough, implying an absence of orientation stability

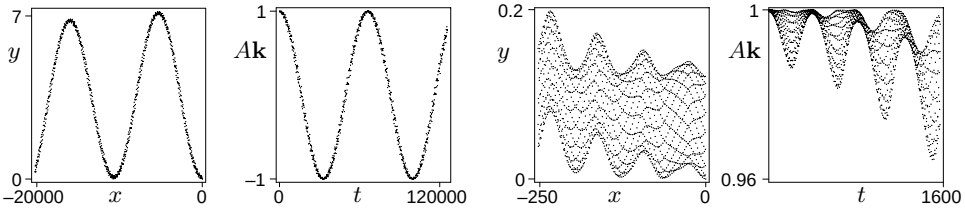


Figure 1. *Absence of orientation stability as shown by simulations of a perturbation of the relative equilibrium corresponding to* $\alpha_e = 1$*. The perturbation is such that the rotational energy and the translational energy are changed by* $-.16\%$ *and* 1.1% *of the energy of the relative equilibrium, respectively. At left is the position of the rigid body projected to a plane orthogonal to* $\mathbf{k}$*. Second from the left is the corresponding plot of* $A\mathbf{k}$ *against time; were the orientation stable, then* $A\mathbf{k}$ *would remain near* 1*, but the rotation axis flips over as evidenced by* $A\mathbf{k}$ *reaching* -1*. The vehicle rotates* $20,000$ *times over the course of the simulation. The two plots at right correspond to the two plots at left but over a shorter simulation corresponding to the first* 250 *rotations of the vehicle. All the plots are time-sampled and show envelopes over which the vehicle oscillates many times.*

when translation is unconfined. This is consistent with simulations of perturbations of $p_e^{\alpha_e}$, one of which is shown in Figure 1, where orientation stability is seen to be ruined by a slow motion of the direction of rotation of the relative equilibrium. There is also a corresponding slow dynamics in translation. The explanation of these slow motions presumably awaits extensions of the drift analysis in [6] and [7] to the context of noncompact symmetry.

The foregoing sort of analysis will always lead to some stability condition, irrespective of possible errors in the derivation, so it is necessary to check (3.4) by comparing it with numerically generated Poincaré maps. Substitution of $\epsilon = 1$ after truncation of $O(\epsilon)^3$ into (3.4) gives the leading behavior of the Poincaré map when a, ν, and I are of comparable order. (They are all of order ϵ^2.) The Poincaré map is determined to leading order in I by the first order twist term $h(I)$ when the zero order term $2\pi\omega_e/\hat{\xi}_e^{\alpha_e}$ has a vanishing effect (i.e., is a multiple of 2π). This happens to occur when $I_1 + I_2 = I_3$, as is easily verified. After substitution of $I_2 = I_3 - I_1$, the first order twist is

$$h(I) = -\frac{\pi {I_3}^2}{I_1(I_3 - I_1)} I,$$

whereupon

$$\frac{I}{h(I)} = -\frac{1}{\pi {I_3}^2} I_1(I_3 - I_1),$$

which is a parabola in I_1/I_3. As can be seen in Figure 2, this compares well with numerical integrations of the original (as opposed to the blown-up) system.

Summary. The following theorem has been proved.

Theorem 3.1. *Within the context of the Lagrangian system* (0.1)*, the motion of an underwater ellipsoid rotating about an axis* $\mathbf{n}$ *is* $A_{\epsilon_0,\epsilon_1}(\mathbf{n})$*-stable if* $\mathbf{n}$ *is a long or short principal axis of inertia.*

This theorem follows from KAM confinement after a blow-up construction and normal-form analysis rather than confinement by Lyapunov functions derived from energy and momentum.

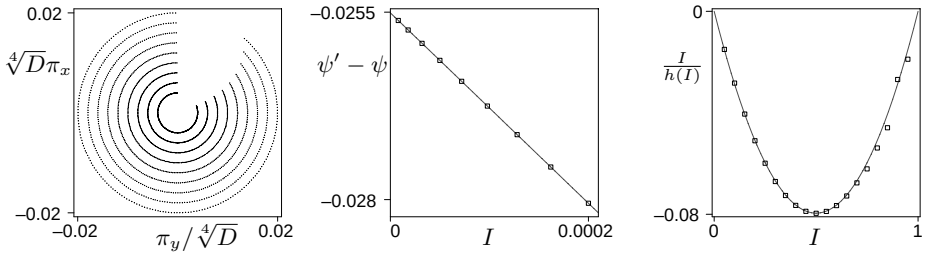


Figure 2. *Numerical verification of the twist predicted by* (3.4). *Leftmost: The curved leading edge of the numerically computed Poincaré map indicates a twist map by visibly showing faster rotation as I increases. The twist overlays a constant rotation (in I) which is caused by high order terms in a and decreases as a decreases. Center: The rotation angle per iteration of the Poincaré map on the right as a function of I. The slope corresponds to the twist predicted by* (3.4). *Right: Reciprocal of the twist for $I_3 = 1$, $I_2 = I_3 - I_1$ as I_1 ranges from* 0 *to* 1 *compared to the parabola predicted by* (3.4).

Acknowledgment. I thank the University of Warwick Mathematics Institute for its hospitality during the sabbatical visit while this paper was written.

REFERENCES

[1] N. E. Leonard, *Stability of a bottom-heavy underwater vehicle*, Automatica J. IFAC, 33 (1997), pp. 331–346.

[2] 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.

[3] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Texts Appl. Math. 17, Springer-Verlag, New York, 1994.

[4] K. R. Meyer and G. R. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Springer-Verlag, New York, 1991.

[5] G. W. Patrick, R. M. Roberts, and C. Wulff, *Stability of Poisson Equilibria and Hamiltonian Relative Equilibria by Energy Methods*, preprint, math.DS/0201239, 2002.

[6] G. W. Patrick, *Relative equilibria of Hamiltonian systems with symmetry: Linearization, smoothness, and drift*, J. Nonlinear Sci., 5 (1995), pp. 373–418.

[7] G. W. Patrick, *Dynamics near relative equilibria: Nongeneric momenta at a* 1:1 *group-reduced resonance*, Math. Z., 232 (1999), pp. 747–788.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 1, pp. 53–96

Semistrong Pulse Interactions in a Class of Coupled Reaction-Diffusion Equations*

Arjen Doelman† and Tasso J. Kaper‡

This article is dedicated in memory of Wiktor Eckhaus.

Abstract. Pulse-pulse interactions play central roles in a variety of pattern formation phenomena, including self-replication. In this article, we develop a theory for the semistrong interaction of pulses in a class of singularly perturbed coupled reaction-diffusion equations that includes the (generalized) Gierer–Meinhardt, Gray–Scott, Schnakenberg, and Thomas models, among others. Geometric conditions are determined on the reaction kinetics for whether the pulses in a two-pulse solution attract or repel, and ODEs are derived for the time-dependent separation distance between their centers and for their wave speeds. In addition, conditions for the existence of stationary two-pulse solutions are identified, and the interactions between stationary and dynamically evolving two-pulse solutions are studied. The theoretical results are illustrated in a series of examples. In two of these, which are related to the classical Gierer–Meinhardt equation, we find that the pulse amplitudes blow up in finite time. Moreover, the blowup of stationary one-pulse solutions and of dynamically varying two-pulse solutions occurs precisely at the parameter values for which the theory we develop predicts that these solutions should cease to exist as bounded solutions. Finally, generalizations to N-pulse solutions are presented.

Key words. reaction-diffusion equations, semistrong pulse interactions, multipulse traveling waves, geometric singular perturbation theory, finite-time blowup, self-replicating pulses

AMS subject classifications. 35K57, 35K45, 35B25, 35B32, 35B40, 34C30, 34C37, 92E20

PII. S1111111102405719

1. Introduction.

In this article, we study the semistrong interaction of pulses in a general class of coupled reaction-diffusion equations in one spatial dimension. The focus is on pattern formation in far-from-equilibrium systems in which the pulses are localized large-amplitude perturbations of a linearly stable, homogeneous, background state. The label semistrong signifies that the concentration of one of the species is far from equilibrium on the domain between the pulse centers, and hence the pulse interaction is governed to leading order through the slow spatial variation of that species.

Semistrong pulse interactions play central roles in a series of pattern formation phenomena, including repelling pulse pairs and self-replication. Repelling pulse pairs in the Gray–Scott model were studied in [3, 4], and this work may be viewed as a continuation of the program

*Received by the editors April 17, 2002; accepted for publication (in revised form) by M. Golubitsky November 5, 2002; published electronically March 31, 2003.
http://www.siam.org/journals/siads/2-1/40571.html
†Korteweg-deVries Instituut, Universiteit van Amsterdam, Plantage Muidergracht 24, 1018TV Amsterdam, The Netherlands (doelman@science.uva.nl). This author was supported by the Organization for Scientific Research (NWO).
‡Department of Mathematics and Center for BioDynamics, Boston University, 111 Cummington Street, Boston, MA 02215 (tasso@math.bu.edu). This author was supported by the National Science Foundation through grant DMS-0072596.

begun there to a broader class of equations. For the Gray–Scott model, ODEs were derived for the time-dependent separation distance between the pulse centers and for their wave speeds. Quantification of the slow spatial variation of the inhibitor concentration in the domain between the pulses was shown to be essential for determining the pulse dynamics.

Self-replication of pulses, discovered and analyzed in the Gray–Scott model [22], is a phenomenon in which a localized pulse splits into two pulses, and as these two pulses move apart from one another, they each split into a pair of pulses so that there are four; see [3, 4, 5, 9, 30, 35, 34, 40, 43, 44]. This process continues, depending on the parameter values and the domain size, until a stationary, spatially periodic pattern is attained. During the entire self-replication process, except near the instants of splitting when the pulse interaction is truly strong, the pulse interactions are semistrong. While the activator concentration is exponentially close to the background state between the pulses due to the singularly perturbed nature of the reaction-diffusion equations, the inhibitor concentration there is not near the value corresponding to the homogeneous steady state and in fact varies on a long spatial-length scale. For instance, in the Gray–Scott model, the pulse interactions are determined largely by the slow spatial variation of the inhibitor concentration on the domain between the pulses, with the local maximum of the inhibitor concentration there increasing slowly in time and in turn causing the pulses to repel each other. We showed in [4] that the time scale of the evolution of this local maximum corresponds to the time scale of pulse-splitting. In addition, recent investigations have shown that self-replication is a rather generic phenomenon that occurs in many reaction-diffusion equations, including the Gierer–Meinhardt equation [21], as has been shown in [10]. Again, the semistrong interaction of pulses plays an essential role.

For a general pair of pulses whose interaction is semistrong, the shapes and amplitudes of the localized pulse-components generally change significantly in time. They are far from being copies of a solitary homoclinic pulse. In fact, there are explicit examples (including the Gray–Scott model and others introduced below) in which pairs of semistrong interacting pulses exist in regions in parameter space where there are no single-pulse homoclinic solutions. In these and other respects, the case of semistrong pulse interaction differs from the weak interaction case (see [12, 14, 42]), where the pulses are to leading order exact, asymptotically stable, single-pulse, homoclinic solutions of the associated reaction-diffusion equation. There, the pulses are assumed to be so far apart that the composition of two pulses is also close to being a solution of the system, and hence the pulses interact through exponentially small tails, they do not change shape to leading order, and they move exponentially slowly away from (or toward) each other in the weak interaction limit. Nevertheless, it is important to note that weak interaction theory applies to a wider range of diffusivities than is considered here; in particular, the diffusivities of both species may be $\mathcal{O}(1)$.

Our study of semistrong pulse interactions is primarily carried out for the following class of coupled reaction-diffusion equations:

$$\begin{cases} \varepsilon^2 U_t &= U_{xx} - \varepsilon^2 \mu U + f(U)V^2, \\ V_t &= \varepsilon^2 V_{xx} - V + g(U)V^2, \end{cases} \tag{1.1}$$

where U and V are positive functions defined for $(x,t) \in \mathbb{R} \times \mathbb{R}^+$, $\mu > 0$ is a parameter, $0 < \varepsilon \ll 1$, and f and g are smooth positive functions on $U > 0$ which may have mild

singularities (poles) at $U = 0$. Many named systems, including the Gierer–Meinhardt, Gray–Scott, Schnakenberg, and Thomas equations (see, for example, [16, 32, 47]) are brought into this specific form (1.1) by scalings that express the asymptotic character of the pulse solutions of singularly perturbed reaction-diffusion equations. See the appendix and Remark 1.1 below. Although ε appears at three places in (1.1), there is, in general, only one small parameter in the original unscaled equation. (ε corresponds to the ratio of the diffusion constants in the Gierer–Meinhardt model, while ε is in essence a certain ratio between constants describing the linear feed terms in the Gray–Scott model; see the appendix and [3, 4, 7].)

The choice of a relatively general system (1.1) was motivated by the following types of questions. In the Gray–Scott model [3, 4] and in the Gierer–Meinhardt equation (section 5), the pulses in a two-pulse solution repel. What mechanism(s) causes the pulses to repel, and is it possible to find systems for which pulses attract? Next, two repelling pulses might eventually be so far apart that their interaction can be studied by the theory of weak interactions. So, we ask, Will a pair of pulses necessarily enter the domain in which the interactions can be considered as weak? We will see in sections 5 and 6 that this is not the case and that several interesting things may happen depending on parameters and the nonlinearities. In addition, we focus on the edge of the parameter domains in which two-pulse solutions exist. A hierarchy of saddle-node points at which stable N-pulse solutions, $N = 1, 2, \ldots$, disappear has been found for the Gray–Scott and Gierer–Meinhardt equations, and this hierarchy governs the self-replication process; see [3, 4, 5, 10, 34, 35]. Are there new phenomena, in addition to self-replication, that occur at the edge of the parameter domains in which the pulse solutions exist? Here, we will see that the answer is yes, and we will determine their relation to self-replication. Finally, we have formally determined the stability of two-pulse solutions for the Gray–Scott model in [3]. Hence we also briefly ask about the stability of the pulse solutions for the general class of systems and about the related issue of the validity of the asymptotic constructions.

Based on the analysis presented in this article, we find that the essential properties of the reaction terms in (1.1) are

(i) the number and location of the zeros of $g'(U)$ and the sign of $g'(U)$ for values of U that are not zeros; and

(ii) the number and location of the zeros of the function $H(U) \equiv (3f(U)/g^2(U)) - \sqrt{\mu}U$.

Properties (i) and (ii), which emerge naturally from the system geometry, determine the salient features of the pulse dynamics. First, the number of zeros of H in (ii) is the number of different stationary, homoclinic, one-pulse solutions the equation has. Second, the zeros of H also determine the allowable locations (in the $U - V$ plane) of the pulses in multipulse solutions. Specifically, each zero corresponds to an allowable pair of jump off and jump on points, which mark the left and right "edges" of the pulses, on a certain invariant manifold. Third, the information in property (i) determines whether adjacent pulses attract or repel each other, as well as whether asymmetric two- and N-pulse solutions can be constructed in addition to symmetric ones.

Our first set of results concerns slowly varying two-pulse solutions, i.e., two-pulse solutions in which the pulse centers move with slowly varying speeds $c_1(t)$ and $c_2(t)$ and the pulse shapes (amplitudes and widths) vary slowly in time for systems of the form (1.1), in which (i) $g'(U)$ has no zeros on $U > 0$ and (ii) the function $H(U) = (3f(U)/g^2(U)) - \sqrt{\mu}U$ has one positive

simple zero. We derive the ODE for the pulse separation distance, denoted $\Delta\Gamma(t)$, as well as the extremal values of the pulse speeds. We find that $\Delta\Gamma(t)$ decreases (and the pulses attract) for configurations in which $g'(U) > 0$, whereas $\Delta\Gamma(t)$ increases (and the pulses repel each other) in those cases for which $g'(U) < 0$. The different signs of $g'(U)$ also appear to be directly connected to the stability of these solutions.

The second set of results concerns system (1.1) for which (i) $g'(U)$ has one positive simple zero and (ii) the function $H(U)$ again has one positive simple zero. For these systems, both stationary and slowly varying two-pulse solutions can exist. Moreover, for the slowly varying solutions, the pulses can either attract or repel each other, and the rates of approach or repulsion can be increasing or decreasing or can even change dynamically in time from being increasing to decreasing or vice versa. The particulars of whether a stationary two-pulse solution exists or of what dynamics a slowly varying two-pulse solution exhibits are determined by where the zero of $g'(U)$ lies with respect to the U value that corresponds to the maximum of a stationary one-pulse solution and to the local maximum of U for the pulse solutions with extremal speed. Here, we see that a pair of repelling pulses may limit, in time, on a stationary two-pulse solution. Such a pair, therefore, will not reach the weak interaction limit.

We also present extensions of the above two principal sets of results. The first extension is to a more general class of systems,

$$\left\{ \begin{array}{rcl} \varepsilon^{2+\sigma} U_t &=& U_{xx} + \varepsilon^2 F_1(U) + V F_2(U,V), \\ V_t &=& \varepsilon^2 V_{xx} + G(U,V). \end{array} \right. \tag{1.2}$$

Here, $\sigma \geq 0$, $0 < \varepsilon \ll 1$, and the functions F_1, F_2, and G are smooth for $U > 0$ and $V \geq 0$ and admit at most poles at $U = 0$. They must also satisfy additional assumptions, which we state in section 7.

The second extension is to N-pulse solutions of (1.1) for $N \geq 2$, including stationary solutions, classical traveling wave solutions, in which the entire N-pulse solutions moves with constant wave speed, and slowly varying solutions. The slowly varying N-pulse solutions are useful for understanding more about self-replication. At the stage of self-replication in which the data has N pulses, one can use the ODEs for the positions of the N-pulse centers and knowledge of the parameter regimes in which they exist and are stable to determine if the pulses (some or all) will split again or if instead the self-replication process ends and the asymptotic state is an N-pulse solution.

The classical Gierer–Meinhardt model is of the form (1.1) with $f(U) \equiv 1$ and $g(U) = 1/U$, and the generalized Gierer–Meinhardt model [7, 33] is of the form (1.2). The ODEs we derive for the pulse separation distances and for the pulse speeds show that the pulses of a slowly varying two-pulse solution repel each other; see also Figure 1.1. Moreover, these results can be used to further understand the self-replication recently discovered [10] in the Gierer–Meinhardt model in the same way as the ODEs for the pulse separation distances were used for the Gray–Scott model in [4].

In addition to obtaining the above results, we have the overarching goal of determining the boundaries of the existence domains in parameter space for the various two-pulse solutions and N-pulse solutions in the semistrong interaction regime. These boundaries have already been shown to be important for self-replication (see the discussion below). Here, we show that

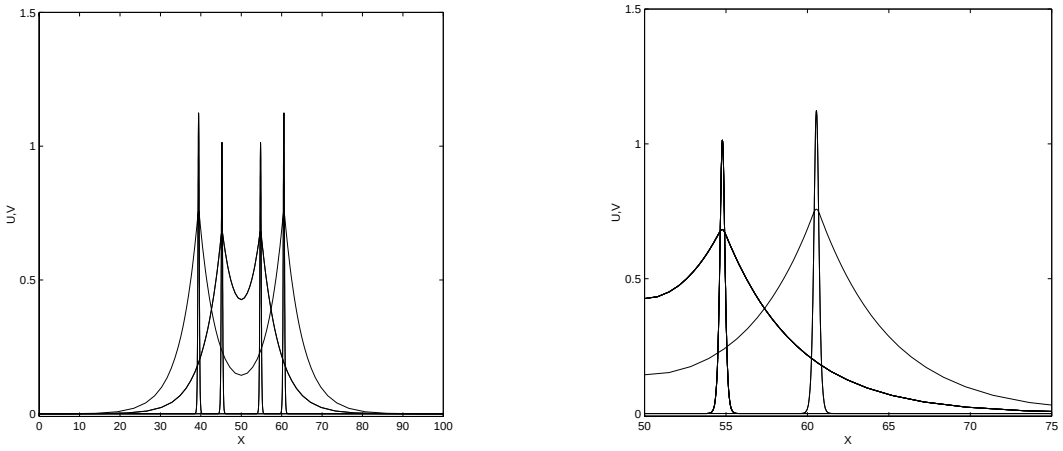


Figure 1.1. *A symmetric, slowly varying, two-pulse solution obtained from direct numerical simulations of the classical Gierer–Meinhardt equation for* $\varepsilon^2 = 0.01$ *and* $\mu = 5$ *shown at two instants of time. The pulses repel each other, as is shown in section 5, and the ODE (5.6) governs the time-dependent separation distance between the pulses. In the left frame, the slowly varying two-pulse solution is shown at two instants of time; in the first instant, the two pulses are the inner pair with smaller maxima, while in the second instant, the pulses are further separated and have larger maxima. The right frame is a magnification of part of the right half of the domain in the left frame, shown so that the pulse structure is more clearly visible. The* V *(activator) component has the narrow needle-like pulses and is vanishingly small on the intervals in between the pulses, while the* U *(inhibitor) component has pulses with lower maxima and varies over a much longer length scale. The inhibitor concentration* U *in between the two pulses is not near zero, the value of the homogeneous steady state, and its local minimum in between the pulses slowly decreases. The numerically observed values of the maxima of* U *and* V *at the pulse peaks agree well with the theory presented here. Homogeneous Neumann boundary conditions and* $N = 201$ *moving grid points were used on the scaled interval* $[0, 1]$*. This and all other numerical simulations in this article have been performed using the code presented in* [2].

the boundaries of the existence domains can also correspond to bifurcation curves associated to the blowup in finite time of localized pulse solutions. On one side of these curves, blowup is observed, but none is observed on the other side. Moreover, we have found an example in which there is a self-replication bifurcation curve that meets a blowup bifurcation curve in a new type of codimension two point, which we label a self-replication/blowup bifurcation point.

The blowup phenomenon was discovered in two examples introduced here to illustrate the theory of semistrong pulse interaction. The first example is (1.1) with $f(U) \equiv 1$ and $g(U) = (1/U)+\alpha$, and the second example is (1.1) with $f(U) \equiv 1$ and $g(U) = (1/U)+(\beta/\sqrt{U})$. The classical Gierer–Meinhardt equation is a special case of both examples, with $\alpha = 0$ and $\beta = 0$, respectively, although the blowup we find in both problems occurs for nonzero values of these parameters.

For the first example, slowly varying two-pulse solutions can exist in parameter regimes in which there is no stationary, homoclinic, one-pulse solution. Since $g'(U) < 0$ for all $U > 0$ and for all $\alpha \geq 0$, the pulses in a slowly varying two-pulse solution repel each other. As a consequence, (to leading order) both pulses necessarily evolve toward a copy of the nonexisting stationary one-pulse solution. It follows from the theory developed in this article that there

is a critical separation distance, $\Delta\Gamma_{bu}$, beyond which the two-pulse solution can no longer be constructed. Numerical simulations show that at $t = t_{bu}$, the time when a solution reaches the critical separation distance, which of course depends on initial conditions, one of the pulses starts to grow extremely rapidly, without bound.

In the second modified Gierer–Meinhardt example, the two interacting pulses can also exist in parameter regimes in which there is no stationary, homoclinic, one-pulse solution, and they also repel (again $g'(U) < 0$ for all $U > 0$ and for all $\beta \geq 0$). Most significantly, they can develop asymptotically large amplitudes as a result of the semistrong interaction. This finite-time blowup can be understood in full analytical detail, with explicit formulae for how both components grow and for the time of the blowup. It differs in some important respects from that observed in the first example.

There are many important characteristics of the Gray–Scott and Gierer–Meinhardt equations (see [4, 10]) that are shared by the two new examples introduced here. As just stated, one of the more interesting of these characteristics is that the slowly varying two-pulse solutions not only exist in the parameter regions where the stationary, homoclinic, one-pulse solutions exist, but they can also exist outside of those regions. Then, for parameter values outside of those regions, the individual pulses of a two-pulse solution tend to copies of the one-pulse, leaving one with the question of what should happen to them since the one-pulse solutions do not exist there. This question was addressed in the context of the self-replication process in the Gray–Scott model in [4] by introducing the notion of an effective small parameter whose size depends on the slowly changing magnitude of the U- (inhibitor) component at the local maximum in between the pulses. For two-pulse solutions in which the pulses are still close together, that magnitude is sufficiently small, and the effective small parameter is such that one is still back in the regime where the one-pulse exists (and is stable; see [5, 6]). However, as that magnitude increases slowly in time, the effective small parameter eventually crosses the self-replication threshold so that the pulses, which could not continue to exist as copies of the nonexistent one-pulse solutions, split just as the initial one-pulse data did. Correspondingly, there is a hierarchy of disappearance (or saddle-node) bifurcations that governs the boundary of the self-replication domain in parameter space, as reported in [5, 4, 34, 35], and a similar analysis can be carried out for splitting in the Gierer–Meinhardt model; see [10].

The analysis presented here shows that this same question arises for the two modified Gierer–Meinhardt examples and has a similar type of answer; only here, the slowly varying two-pulse solutions and the stationary, homoclinic, one-pulse solutions can exhibit finite-time blowup, as well as splitting, depending on parameters.

This article is organized as follows. Section 2 contains the essential geometric information about the invariant manifolds. The basic construction of slowly varying two-pulse solutions is presented in section 3. Section 4 contains the study of symmetric, slowly varying, two-pulse solutions and the presentation of the first set of results. Three examples, including the two that exhibit finite-time blowup, are introduced and analyzed in section 5. Section 6 contains the second set of main results for stationary and slowly varying two-pulse solutions in systems that support more than a single stationary, homoclinic, one-pulse solution. Finally, in section 7, we discuss the extensions of the above results, including extensions to N-pulse solutions and generalizations to the broader class of coupled reaction-diffusion equations given by (1.2). We also discuss the stability of the solutions constructed here and the validity of the asymptotic

constructions.

Remark 1.1. For the PDEs (1.1) and (1.2), one may assume without loss of generality that the state $(U_0, V_0) = (0, 0)$ is a homogeneous steady state. Moreover, we assume that $(0, 0)$ is a linearly stable solution of the PDEs with $\mu > 0$. As shown in the appendix, this may entail a change of dependent variables such as is the case, for example, in the Gray–Scott model, which has $(1, 0)$ as a linearly stable homogeneous state. It may also entail ensuring that the solution components vanish at certain rates (see [7]), as is necessary for the generalized Gierer–Meinhardt equations, where $g(U) = 1/U$ as stated above, so that one needs V^2 to vanish more rapidly than U along solutions in order for the quotient to vanish.

Remark 1.2. Pulse interactions have also been analyzed in other contexts, and we mention some recent references without being exhaustive. In [38], one-dimensional (particle-like) pulses in the excitable regime of a coupled reaction-diffusion system of FitzHugh–Nagumo type are studied. The dispersion relation $c(d)$ for the speed of pulse-trains as a function of their wavelengths is presented, and a transition from long range dispersion (common in excitable systems) to anomalous dispersion, for large d, is found that is shown to be responsible for a long range attractive force between pulses. They also demonstrate the existence of stable bound states. [36] examines the dynamics of sharp interfaces in an activator-inhibitor model with global coupling in one and two spatial dimensions. It is shown that a pair of fronts that initially approach can reflect off of one another and then repel each other. Next, [13] reports on the discovery of pulse reflection in a two-component excitable system. Moreover, it is shown that reflections also occur for spirals and rings in two space dimensions, and the bifurcations are analyzed using center manifold theory for sufficiently small pulse velocities. Finally, two-dimensional pulses are also analyzed in an activator-inhibitor system with global coupling in [26]. The bifurcation of a stationary pulse to a traveling pulse is examined using interface dynamics, and the collision of two pulses is also observed in the presence of sufficiently strong global coupling.

2. Geometry of the invariant manifolds. In this section, we present the phase space geometry associated to the pulse solutions studied in this article. Let $c(t)$ denote the time-dependent velocity of a pulse, and let $x = \Gamma(t)$ denote the time-dependent position of the pulse's center, where

$$\Gamma(t) = \int_{t_0}^{t} c(s)ds. \tag{2.1}$$

It is also useful to introduce the moving coordinate, $\bar{x} = x - \Gamma(t)$, as well as a stretched moving coordinate, $\xi \equiv \bar{x}/\varepsilon$.

Looking ahead just a bit, we can also take advantage of the following two observations. First, the distinguished wave speed is small,

$$c(t) \equiv \varepsilon^3 \hat{c}(t), \tag{2.2}$$

where $\hat{c}(t)$ is $\mathcal{O}(1)$, and $|d\hat{c}/dt| \ll 1$. This agrees with the earlier analysis [3, 4] of self-replication and semistrong pulse interactions in the Gray–Scott model and will be shown to be a distinguished speed also for the general systems (1.1). See Remark 2.2 below. Second, it turns out that, for the solutions we study, the explicit time variation of U and V is also slow, i.e., $|\partial U/\partial t|, |\partial V/\partial t| \ll 1$. Therefore, we use a quasi-stationary approximation, which

entails treating $\hat{c}(t)$ as a slowly varying parameter that lies within a certain allowable range and ignoring the explicit partial derivatives with respect to t to leading order. Hence, to leading order, the solutions $(U(\xi(t),t),V(\xi(t),t))$ depend only on t through $\xi(t)$. (See also Remark 3.1 for a brief discussion of higher order terms and the validity of this quasi-stationary approximation.)

Within the context of this quasi-stationary approximation, the quasi-stationary solutions of the PDE (1.1) are described by the following system of four first order ODEs:

$$\left\{\begin{array}{rcl} \dot{u} &=& \varepsilon p, \\ \dot{p} &=& \varepsilon\left[-\varepsilon^5\hat{c}p+\varepsilon^2\mu u - f(u)v^2\right], \\ \dot{v} &=& q, \\ \dot{q} &=& -\varepsilon^2\hat{c}q+v-g(u)v^2. \end{array}\right. \tag{2.3}$$

Here, $u(\xi)$ and $v(\xi)$ are the leading order terms in the asymptotic expansions of the solutions $(U(\xi(t),t),V(\xi(t),t))$ of the PDE (1.1), and the overdot denotes derivatives with respect to ξ. Also, $\hat{c}=\hat{c}(t)$ is a parameter.

2.1. A normally hyperbolic invariant manifold for (2.3). One natural approach to determining the relevant geometrical features of the phase space of the ODE (2.3) is to identify the structures that exist when $\varepsilon=0$ and then to ask which of these persist for $0<\varepsilon\ll 1$.

For $\varepsilon=0$, the half-plane

$$\mathcal{M}=\{(u,p,v,q)|v=0,q=0,u>0\} \tag{2.4}$$

is a set of equilibrium points of (2.3). Off of $\mathcal{M}$, the system exhibits fast dynamics governed by the fast system

$$\left\{\begin{array}{rcl} \dot{v} &=& q, \\ \dot{q} &=& v-g(u)v^2, \end{array}\right. \tag{2.5}$$

where u is a fixed parameter. This fast system has the conserved quantity (energy)

$$K(v,q;u)=\frac{q^2}{2}-\frac{v^2}{2}+\frac{g(u)}{3}v^3. \tag{2.6}$$

Moreover, each fixed point $(u,p,0,0)$ on $\mathcal{M}$ is a saddle point under the flow (2.5). Hence the two-dimensional invariant plane $\mathcal{M}$ is normally hyperbolic. Finally, since $g(u)>0$ for $u>0$, each fixed point on $\mathcal{M}$ is connected to itself by a homoclinic orbit $(v_0(\xi),q_0(\xi))$, where

$$v_0(\xi)=\frac{3}{2g(u)}\mathrm{sech}^2\left(\frac{\xi}{2}\right), \tag{2.7}$$

of the fast system (2.5). Therefore, the three-dimensional stable and unstable manifolds of $\mathcal{M}$, denoted $W^S(\mathcal{M})$ and $W^U(\mathcal{M})$, coincide.

For $0<\varepsilon\ll 1$, the plane $\mathcal{M}$ is still an invariant set under the dynamics of (2.3), and it still has three-dimensional stable and unstable manifolds by the Fenichel theory of geometric singular perturbations; see [18] and [24]. On $\mathcal{M}$, the system (2.3) reduces to

$$\left\{\begin{array}{rcl} \dot{u} &=& \varepsilon p, \\ \dot{p} &=& \varepsilon^3(\mu u-\varepsilon^3\hat{c}p). \end{array}\right. \tag{2.8}$$

The equilibrium point S, $(u,p)=(0,0)$, of (2.8) on the boundary of $\mathcal{M}$ corresponds to the homogeneous state $(U_0,V_0)=(0,0)$ of (1.1), and it is a saddle equilibrium in the phase space of (2.8). Moreover, the stable and unstable manifolds of S restricted to $\mathcal{M}$, labeled ℓ^s and ℓ^u, are given to leading order by the lines

$$\ell^s,\ell^u: \qquad p=\mp\varepsilon\sqrt{\mu}u. \tag{2.9}$$

We are interested in that portion of $\mathcal{M}$ that lies in the wedge between (or on) ℓ^u and ℓ^s. In this wedge, the p-coordinates of all points are at most $\mathcal{O}(\varepsilon)$, and so it is useful to explicitly introduce $p=\varepsilon\hat{p}$, where $\hat{p}=\mathcal{O}(1)$.

Remark 2.1. For general $g(u)$, the boundary $u=0$ is excluded from the definition of $\mathcal{M}$, since $g(u)$ may have a pole there. In that case, one can still obtain results for those solutions along which v vanishes sufficiently rapidly so that the product $g(u)v^2$ vanishes. The method employed in [7] (see the end of section 2.4 there) to analyze semistrong pulses in the generalized Gierer–Meinhardt equation extends to system (1.1). Of course, if g is also smooth at $u=0$, such as in the Gray–Scott model (transformed as in the appendix), then one can include the boundary as part of the manifold.

2.2. Persistent orbits homoclinic to $\mathcal{M}$. For $0<\varepsilon\ll 1$, the stable and unstable manifolds of $\mathcal{M}$ no longer coincide. Instead, they are expected to intersect transversely in a two-dimensional surface, in which all orbits that are biasymptotic ($\xi\to\pm\infty$) to $\mathcal{M}$ lie. We find this surface in this section.

For points on $\mathcal{M}$, $K=0$. Hence orbits that are biasymptotic to $\mathcal{M}$ satisfy

$$\Delta K\equiv\int_{-\infty}^{\infty}\dot{K}(v(\xi),q(\xi);u(\xi))d\xi=0 \tag{2.10}$$

so that their α and ω limit sets are both on $\mathcal{M}$. By adiabatic Melnikov function theory [45], simple zeros of ΔK imply the existence of nearby transverse intersections of the manifolds $W^S(\mathcal{M})$ and $W^U(\mathcal{M})$.

Let $(u(\xi),p(\xi),v(\xi),q(\xi))$ denote any solution of (2.3) that, at $\xi=0$, passes through the hyperplane $\{q\equiv 0\}$ with $u,v>0$ and with p in the desired wedge. Denote such a point on $\{q\equiv 0\}$ by $(u,\varepsilon\hat{p},v(0),0)$. A straightforward calculation of $\dot{K}$ along these solutions of (2.3) yields

$$\dot{K}=\varepsilon^2\left(-\hat{c}q^2+\frac{\hat{p}}{3}g'(u)v^3\right). \tag{2.11}$$

Thus a solution $(u(\xi),p(\xi),v(\xi),q(\xi))$ of the type under consideration here is a persistent homoclinic orbit to $\mathcal{M}$ if

$$\varepsilon^2\int_{-\infty}^{\infty}\left(-\hat{c}q^2(\xi)+\frac{\hat{p}(\xi)}{3}g'(u(\xi))v^3(\xi)\right)d\xi=0. \tag{2.12}$$

For $0<\varepsilon\ll 1$, the persistent homoclinic orbits in the transverse intersection of $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$ lie close to their $\varepsilon=0$ counterparts. Specifically, persistence theory for invariant manifolds [18] establishes that solutions on $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$ are close to the unperturbed homoclinic orbit $(v_0(\xi),q_0(\xi)=\dot{v}_0(\xi),u,0)$ on appropriate time intervals, $(-\Xi,\xi_0)$ and (ξ_0,Ξ),

respectively, where Ξ may be taken arbitrarily large (but $\mathcal{O}(1)$) as long as ε is sufficiently small.

Therefore, asymptotically, we may approximate $u(\xi)$ by the constant u, $v(\xi)$ by $v_0(\xi)$ as given by (2.7), and $q(\xi)$ by $q_0(\xi)$. Hence, to leading order, condition (2.12) implies that either $\hat{c} = 0$ (which is the case when $g'(u) = 0$) or

$$\hat{p} = 2\hat{c}\frac{g(u)}{g'(u)} \tag{2.13}$$

(which is the case when $g'(u) \neq 0$ for all $u > 0$). Here, we used the intermediate results $\int_{-\infty}^{\infty} q_0^2 d\xi = 6/(5g^2(u))$ and $\int_{-\infty}^{\infty} v_0^3 d\xi = 36/(5g^3(u))$.

Geometrically, this result may be understood as follows. In the first case, there is a one-parameter family of orbits biasymptotic to $\mathcal{M}$, and these have the symmetry $(u, p, v, q) \to (u, -p, v, -q)$ about $\xi = 0$. In the second case, there is also a one-parameter family of orbits biasymptotic to $\mathcal{M}$. However, they are not symmetric. Instead, those orbits $(u(\xi), p(\xi), v(\xi), q(\xi))$ with $(u, \varepsilon\hat{p}, v(0), 0)$ at $\xi = 0$ and $v(0)$ near $v_0(0) = 3/(2g(u))$ that are biasymptotic to $\mathcal{M}$ satisfy the property that their u- and p-coordinates are related by (2.13) to leading order. Therefore, since $g(u) > 0$ by assumption, the sign of $\hat{p}$ is determined by the signs of $\hat{c}$ and $g'(u)$. This property will be essential throughout the analysis.

Finally, during the "time" interval when the homoclinic orbits in the transverse intersection of $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$ make their excursion through the fast field, i.e., when they lie outside of a fixed neighborhood of $\mathcal{M}$, their slow coordinates u and p undergo changes that are given to leading order by

$$\Delta u \equiv \int_{-\infty}^{\infty} \dot{u}(\xi) d\xi \qquad \text{and} \qquad \Delta p \equiv \int_{-\infty}^{\infty} \dot{p}(\xi) d\xi. \tag{2.14}$$

Now, since $p = \varepsilon\hat{p}$ and $\hat{p} = \mathcal{O}(1)$ for these orbits, we immediately see from (2.3) that

$$\Delta u = \mathcal{O}(\varepsilon^2). \tag{2.15}$$

Hence u is actually constant to a higher degree of approximation than originally stated. Moreover, with $u \sim$ constant, we find

$$\begin{aligned} \Delta p(u, \hat{c}) &= -\varepsilon \int_{-\infty}^{\infty} f(u) v^2 d\xi + \text{h.o.t.} \\ &= -\varepsilon f(u) \int_{-\infty}^{\infty} v_0^2 d\xi + \text{h.o.t.} \\ &= -6\varepsilon \frac{f(u)}{g^2(u)} + \text{h.o.t.} \end{aligned} \tag{2.16}$$

This change in p during the fast jump plays a central role in much of the geometry of the pulses and hence also in much of the analysis in this article.

Remark 2.2. That $c(t) = \varepsilon^3 \hat{c}(t)$ is a distinguished wave speed as introduced in (2.2) above can be seen directly from the adiabatic Melnikov condition (2.12). One wants to balance the first and second terms in order that ΔK can have zeros. The second term is $\mathcal{O}(\varepsilon^2)$, and hence

if one scales $c = \mathcal{O}(\varepsilon^3)$, then also the first term is $\mathcal{O}(\varepsilon^2)$. This is a distinguished limit; see [11].

Remark 2.3. In the special case in which $g(u) = h_2 u^{\alpha_2}$ with $h_2 > 0$ and $\alpha_2 < 0$, such as, for example, in the Gierer–Meinhardt equation and in its generalizations, we have $g(u)/g'(u) = u/\alpha_2$. Hence $\hat{p} = \hat{p}(u) = (2\hat{c}/\alpha_2)u$, a linear function.

Remark 2.4. We have focused exclusively on the geometry used to construct one-pulse solutions in this section. Multiple-pulse solutions that consist of two slow segments and one fast segment in which all of the pulses occur in rapid succession (i.e., they remain bounded away from $\mathcal{M}$ in between pulses) can also be constructed, but such stationary solutions are unstable solutions of the PDE; see [7].

2.3. The takeoff and touchdown curves for persistent homoclinic orbits. There are takeoff and touchdown curves on $\mathcal{M}$ that play a crucial role in the geometry of the pulse solutions. These are determined by the two-dimensional manifold that is the transverse intersection of $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$ identified in the previous section. The first intersection of $W^S(\mathcal{M})$ and $W^U(\mathcal{M})$ in the hyperplane $\{q = 0\}$ is a one-dimensional curve of points $(u, \varepsilon\hat{p}, v(0), 0)$, where u and $\hat{p}$ are related to leading order via (2.13) in this two-dimensional manifold. Moreover, through any such point there is an orbit $\gamma(\xi;\gamma_0)$ with phase γ_0 that approaches $\mathcal{M}$ as $\xi \to \pm\infty$. Fenichel theory [18] implies that for any $\gamma(\xi;\gamma_0)$ there are two orbits $\gamma^+_{\mathcal{M}} = \gamma^+_{\mathcal{M}}(\xi;\gamma_0^+) \subset \mathcal{M}$ and $\gamma^-_{\mathcal{M}}(\xi;\gamma_0^-) \subset \mathcal{M}$, respectively (where $\gamma^\pm_{\mathcal{M}}(0,\gamma_0^\pm) = \gamma_0^\pm \in \mathcal{M}$), such that $\|\gamma(\xi;\gamma_0) - \gamma^+_{\mathcal{M}}(\xi;\gamma_0^+)\|$ is exponentially small for positive values of ξ that satisfy $\xi \geq \mathcal{O}(\frac{1}{\varepsilon})$ and $\|\gamma(\xi;\gamma_0) - \gamma^-_{\mathcal{M}}(\xi;\gamma_0^-)\|$ is exponentially small for $\xi < 0$ satisfying $-\xi \geq \mathcal{O}(\frac{1}{\varepsilon})$. As a consequence,

$$d(\gamma(\xi;\gamma_0),\mathcal{M}) = \mathcal{O}\left(e^{-\frac{k}{\varepsilon}}\right) \quad \text{for} \quad |\xi| \geq \mathcal{O}\left(\frac{1}{\varepsilon}\right) \quad \text{and some } k > 0,$$

and the orbits $\gamma^\pm_{\mathcal{M}}(\xi;\gamma_0^\pm)$ on $\mathcal{M}$ determine the behavior of $\gamma(\xi;\gamma_0)$ near $\mathcal{M}$.

We define the curves $T_o(\hat{c}) \subset \mathcal{M}$ (takeoff) and $T_d(\hat{c}) \subset \mathcal{M}$ (touchdown) as

$$\begin{aligned} T_o(\hat{c}) &= \cup_{\gamma_0}\{\gamma_0^- = \gamma^-_{\mathcal{M}}(0;\gamma_0^-)\}, \\ T_d(\hat{c}) &= \cup_{\gamma_0}\{\gamma_0^+ = \gamma^+_{\mathcal{M}}(0;\gamma_0^+)\}, \end{aligned} \tag{2.17}$$

where the unions are over all γ_0 in $W^S(\mathcal{M}) \cap W^U(\mathcal{M}) \cap \{q = 0\}$. Therefore, the curves $T_o(\hat{c})$ and $T_d(\hat{c})$ are the collections of the basepoints of all of the fibers in $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$ (respectively) that lie in the transverse intersection of these two manifolds. See Figure 2.1.

Asymptotic formulas for the locations of $T_o(\hat{c})$ and $T_d(\hat{c})$ can be obtained as follows. The accumulated changes in the p-component of $\gamma(\xi)$ during two successive half-circuit excursions through the fast field are measured by

$$\int_{-\infty}^{0} \dot{p}d\xi \qquad \text{and} \qquad \int_{0}^{\infty} \dot{p}d\xi,$$

respectively, and these integrals may be calculated by the same method used to obtain (2.16). Moreover, at $\xi = 0$, the orbits have $\hat{p}$ given by (2.13). Hence we find

$$T_o(\hat{c}), T_d(\hat{c}): \qquad \hat{p} = 2\hat{c}\frac{g(u)}{g'(u)} \pm \frac{3f(u)}{g^2(u)}. \tag{2.18}$$

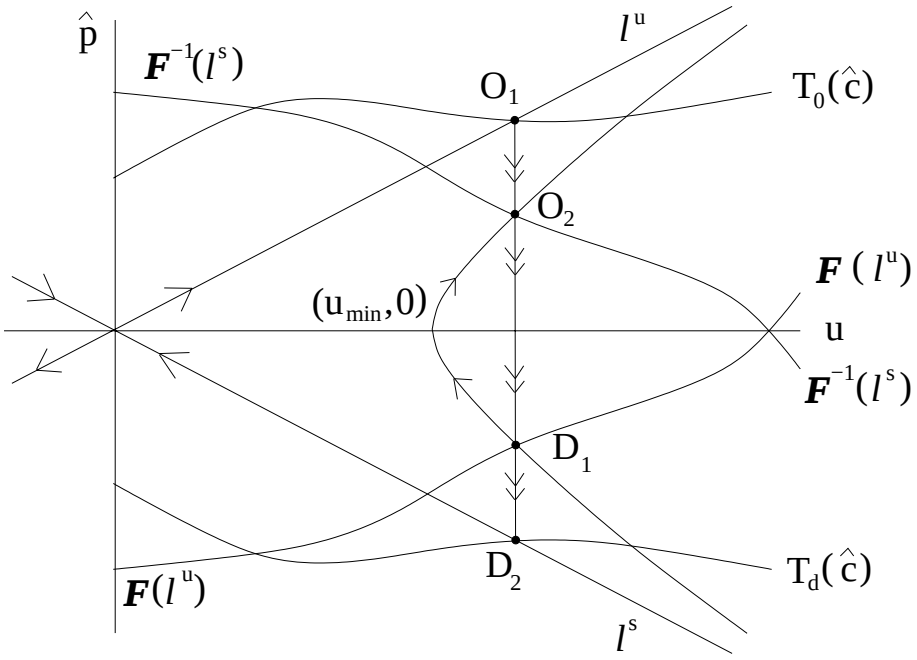


Figure 2.1. *A qualitative sketch of the curves $T_o(\hat{c})$, $T_d(\hat{c})$, $\mathcal{F}(\ell^u)$, and $\mathcal{F}^{-1}(\ell^s)$ on $\mathcal{M}$, as defined in section 2, as well as of a singular, slowly varying, two-pulse solution, whose construction is outlined in section 3. The sketch is for the symmetric case $\hat{c}_2 = -\hat{c}_1$. The successive slow, fast, and slow segments of the left pulse are from $(0,0)$ to O_1, O_1 to D_1, and D_1 to $(u_{\min},0)$, respectively. We remark that the line segments $\overline{O_1D_1}$ and $\overline{O_2D_2}$ are projections onto $\mathcal{M}$ of the fast near-homoclinic excursions in the $v-v_\xi$ directions. Also, we emphasize that $T_o(\hat{c})$ is a one-dimensional curve for each fixed value of $\hat{c}=\hat{c}_1$, and similarly for $T_d(\hat{c})$. By contrast, $\mathcal{F}(\ell^u)$ is a one-dimensional curve because it is the union over a certain range of $\hat{c}$ values of touchdown points—one for each value of $\hat{c}$. Similarly, $\mathcal{F}^{-1}(\ell^s)$ is a one-dimensional curve because it is the union of takeoff points—again one point for each $\hat{c}$ in a certain range. See the definitions in section 2.4.*

2.4. The images of the intersection points $\ell^u \cap T_o(\hat{c})$ and $\ell^s \cap T_d(\hat{c})$ under the fast jump. As a final preparatory step, we identify two other curves on $\mathcal{M}$ that will be useful throughout the analysis. In particular, for those values of $\hat{c}$ for which the intersections $\ell^u \cap T_o(\hat{c})$ and $\ell^s \cap T_d(\hat{c})$ exist, we are interested in the image of $\ell^u \cap T_o(\hat{c})$ and in the preimage of $\ell^s \cap T_d(\hat{c})$.

We showed in section 2.2 that, for fixed $\hat{c}$, a persistent homoclinic orbit to $\mathcal{M}$ that takes off from a point on $T_o(\hat{c})$ touches down on $\mathcal{M}$ at a point that has the same u-coordinate to leading order and whose p-coordinate is shifted by $\Delta p(u,\hat{c})$. This same conclusion holds, of course, for points that take off at the intersection point $\ell^u \cap T_o(\hat{c})$, if this intersection exists. The union over $\hat{c}$ of all of these touchdown points is denoted by $\mathcal{F}(\ell^u)$. See Figure 2.1. To leading order, $\mathcal{F}(\ell^u)$ is given by

$$\mathcal{F}(\ell^u): \qquad \hat{p} = \sqrt{\mu}u - \frac{6f(u)}{g^2(u)}. \tag{2.19}$$

Similarly, let $\mathcal{F}^{-1}(\ell^s)$ denote the set of points that are the preimages (takeoff points) under the fast jump of the touchdown points in $\ell^s \cap T_d(\hat{c})$ over all $\hat{c}$, when this intersection exists.

To leading order,

$$\mathcal{F}^{-1}(\ell^s): \quad \hat{p} = -\sqrt{\mu}u + \frac{6f(u)}{g^2(u)}. \tag{2.20}$$

These curves play a central role in the definition of the function H introduced above.

3. The basic construction of slowly varying two-pulse solutions. In this section, we present a brief introduction to slowly varying two-pulse solutions, focusing on the basic idea of how the geometry of the invariant manifolds described in the previous section is used to construct them. Specifically, we consider slowly varying two-pulse solutions with one pulse on $x \le x_0$ for some $x_0 \in \mathbb{R}$ that travels with speed $c_1(t)$ and the other on $x \ge x_0$ that travels with speed $c_2(t)$.

The moving coordinate $\bar{x} = x - \Gamma_1(t)$, with $\Gamma_1(t) = \int_{t_0}^{t} c_1(s)ds$, is used to analyze the left pulse, and hence the governing ODE is (2.3) with $\hat{c} = \hat{c}_1$ slowly changing in time. The results for the right pulse are obtained from the same ODE but with $\hat{c} = \hat{c}_2$ instead changing slowly in time, and the geometry for it is usually shown superimposed in the same phase space as for the left pulse. We refer the reader to Remark 3.1 and section 7 for a discussion of how these two components are hooked up smoothly and of the justification of this asymptotic construction.

The $\hat{c}_1$ values for which the two-pulse solutions can be constructed depend on the quantities in (i) and (ii), as we show in the next sections.

3.1. Constructing singular two-pulse solutions. The left pulse, centered at $x = \Gamma_1(t)$, is asymptotic to $(U, V) = (0, 0)$ as $x \to -\infty$ and satisfies a Neumann boundary condition $(U_x, V_x = 0)$ at $x = x_0$. In fact, the singular ($\varepsilon \to 0$) limit of the left pulse consists of the following three pieces (see Figure 2.1), two slow (outer) segments on $\mathcal{M}$ interspersed with a fast (inner) homoclinic orbit of the reduced fast system (2.5):

- The left slow segment corresponds to the interval $(-\infty, \Gamma_1(t))$ and is that portion of ℓ^u on $\mathcal{M}$ from $(0, 0)$ to the point, labeled O_1, at which $T_o(\hat{c}_1)$ and ℓ^u intersect transversely.
- The fast pulse is centered at $x = \Gamma_1(t)$ and is described in the stretched coordinate ξ introduced above. To leading order, it is given by $(v_0(\xi), q_0(\xi))$, see (2.7), with u constant. Over the fast pulse, the p-component makes a jump $\Delta p(u; \hat{c})$ of size $\mathcal{O}(\varepsilon)$ (see (2.16)). Therefore, the projection of this fast jump onto the $u-p$ plane is precisely the vertical line segment from O_1 to the point labeled D_1 on $\mathcal{F}(\ell^u)$.
- The right slow segment, which corresponds to $x \in (\Gamma_1(t), x_0]$, is that segment of the hyperbolic cosine orbit on $\mathcal{M}$ from D_1 to a point on $p = 0$ labeled $(u_{\min}, p = 0)$, so that the homogeneous Neumann boundary condition is satisfied.

The singular right pulse is now constructed in a similar manner on $x > x_0$. It has a left slow segment, which corresponds to $x \in [x_0, \Gamma_2(t))$, that is the continuation of the hyperbolic cosine solution on $\mathcal{M}$ (but for the ODE with $\hat{c} = \hat{c}_2$ now) from $(u_{\min}, 0)$ to a point on $\mathcal{F}^{-1}(\ell^s)$ labeled O_2. Note that the difference between $\hat{c} = \hat{c}_1$ and $\hat{c} = \hat{c}_2$ is a higher order effect in (2.3); see also Remark 3.1. The projection of the fast homoclinic orbit on $\mathcal{M}$ is the vertical line segment from O_2 to D_2, where $D_2 \in T_d(\hat{c}_2) \cap \ell^s$. Finally, the right slow segment, which corresponds to $x \in (\Gamma_2(t), \infty)$, is that portion of ℓ^s from D_2 to S; see Figure 2.1.

3.2. Slowly varying two-pulse solutions for $0 < \varepsilon \ll 1$ lie close to the singular two-pulse solutions. For $0 < \varepsilon \ll 1$ and $x < \Gamma_1(t) - \delta$ with $\delta > 0$ small but $\mathcal{O}(1)$, the left pulse lies exponentially close to ℓ^u on $\mathcal{M}$ and approaches the fixed point $(0,0)$ on $\partial\mathcal{M}$ as $x \to -\infty$. Moreover, as $x \to \Gamma_1(t)^-$, the left pulse solution leaves a fixed neighborhood of $\mathcal{M}$ near O_1. We recall from the above discussion that the point O_1 on $T_o(\hat{c}_1)$ corresponds precisely to the $\xi \to -\infty$ limit of the leading order fast (inner) solution v_0. Geometrically, it is the basepoint of a fast unstable fiber that lies in the transverse intersection of $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$, as was shown in section 2.3.

The $\xi \to \infty$ limit of the fast (inner) solution is the point on $T_d(\hat{c}_1)$ labeled D_1 above, and it is the basepoint of the fast stable fiber that lies in the transverse intersection of $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$. On all compact intervals in ξ, the left pulse of the slowly varying two-pulse solution lies close to (though not in) this transverse intersection.

The left and right pulses move when $\hat{c}_1(t)$ and $\hat{c}_2(t)$ are not identically zero, and hence their separation distance, $\Gamma_2(t) - \Gamma_1(t)$, changes in time. In particular, the given wave speed $\hat{c}_1$ determines the location of the point O_1, which in turn determines the points D_1, O_2, and O_2 sequentially, where

$$\begin{aligned} &O_1 : \ell^u \cap T_o(\hat{c}_1), \\ &D_1 : \mathcal{F}(\ell^u) \cap T_d(\hat{c}_1), \\ &O_2 : \mathcal{F}^{-1}(\ell^s) \cap T_o(\hat{c}_2), \\ &D_2 : \ell^s \cap T_d(\hat{c}_2). \end{aligned} \tag{3.1}$$

The ODE governing $\Gamma_2(t) - \Gamma_1(t)$ is then naturally obtained by imposing the leading order consistency condition that, at every instant of time, the difference $\Gamma_2(t) - \Gamma_1(t)$ equals the time of flight along the hyperbolic cosine orbit segment between the points D_1 and O_2 on $\mathcal{M}$.

3.3. Summary of the construction. In summary, the essential step of the analysis of the given reaction terms in (1.1) is to determine the geometry of $T_o(\hat{c})$, $T_d(\hat{c})$, ℓ^u, ℓ^s, $\mathcal{F}(\ell^u)$, and $\mathcal{F}^{-1}(\ell^s)$ on $\mathcal{M}$. With this in hand, one can then readily determine how many different types of pulses are possible and whether or not it is possible to construct repelling or attracting two-pulse solutions. We do this for various f and g in (1.1) in the next sections. Moreover, this method readily generalizes to asymmetric two-pulse solutions and to N-pulse solutions with $(N-1)$ segments near hyperbolic cosine solutions on $\mathcal{M}$, as is shown toward the end of this article.

Remark 3.1. A hard look at the leading order quasi-stationary solution constructed in this section reveals that it is not sufficiently smooth at $x = x_0$ due to the change in $\hat{c}$ at x_0 in (2.3). In particular, just as was observed in the construction of symmetric, slowly varying two-pulse solutions in the Gray–Scott model in [4], the third order (and higher order) derivatives of U do not vanish at $x = x_0$ as they should so that the full solution is symmetric. For example, the jump in U_{xxx} is $\mathcal{O}(\varepsilon^4)$ and hence truly of higher order. A similar smoothness problem is encountered for nonsymmetric two-pulse and N-pulse solutions at the points where the adjacent pulses are hooked up. However, we showed for the Gray–Scott model in section 3.4 of [4] that inclusion of the higher order terms beyond the leading order quasi-stationary approximation restores the smoothness in the third order derivatives and in all higher order odd derivatives. The lack of smoothness of the leading order quasi-stationary approximation

for the solutions studied here can be similarly ameliorated by inclusion of the higher order terms. We do not go into the details.

4. Symmetric, attracting, or repelling two-pulse solutions. In this section, we study symmetric, attracting, or repelling two-pulse solutions in systems of the form (1.1) for which

(i) $g'(u) \neq 0$ for all $u > 0$ and

(ii) the function $H(u) = 3f(u)/g^2(u) - \sqrt{\mu}u$ has one simple zero.

Under these conditions, the intersection of ℓ^u and $T_o(0)$ is transverse and consists of a single point; see (2.9) and (2.18). As a consequence, the PDE (1.1) possesses a stationary ($\hat{c} = 0$) one-pulse solution. See also section 7.1.

We construct slowly varying two-pulse solutions for which the slowly changing pulse velocities are equal in magnitude but opposite in sign at each instant of time t and for which the pulses either attract ($g'(u) > 0$) or repel ($g'(u) < 0$). In addition, we will see that there is an extremal allowable pulse speed (a minimum in the case of attracting pulses and a maximum in the case of repelling pulses) that arises naturally from the system geometry.

Conditions (i) and (ii) are formulated here mostly to facilitate the presentation of the geometric method. In the following sections, we will see that two-pulse solutions can also be constructed when $g'(u)$ has zeros, or when the function $H(u)$ has either no zeros or two (or more) zeros.

4.1. Singular two-pulse solutions. The singular two-pulse solutions are constructed following the blueprint developed in section 3.1. Consider the left pulse with peak at $x = \Gamma_1 < 0$ (where we have taken $x_0 = 0$ without loss of generality), and assume that the local maximum of u at the center of the pulse, denoted u^+, is given by some value satisfying

$$-\sqrt{\mu}u^+ < \sqrt{\mu}u^+ - \frac{6f(u^+)}{g^2(u^+)} < 0. \tag{4.1}$$

We note that Remark 4.1 provides the geometric reasons for the restriction (4.1).

The left pulse is backward asymptotic to $S = (0, 0, 0, 0)$. It has the following slow and fast segments (see also Figure 2.1):

- The left slow segment, which corresponds to the interval $(-\infty, \Gamma_1)$, is that portion of ℓ^u between S and the point O_1, which is the point in $\ell^u \cap T_o(\hat{c}_1)$ whose u-coordinate is given by the unique solution u^+ of

$$\sqrt{\mu}u = 2\hat{c}_1 \frac{g(u)}{g'(u)} + \frac{3f(u)}{g^2(u)}. \tag{4.2}$$

  Note that $\hat{c}_1$ is undetermined, as yet.
- The fast, homoclinic jump occurs at $x = \Gamma_1$, instantaneously on the outer scale x, and it is of infinite duration on the stretched ξ scale, with $\xi \in (-\infty, \infty)$. This fast jump begins at the point O_1 in $\ell^u \cap T_o(\hat{c}_1)$, where $O_1 = (u^+, p^+)$ and $p^+ = \varepsilon\sqrt{\mu}u^+$ by (2.9). The touchdown point of this fast jump is the point D_1 on $\mathcal{F}(\ell^u) \cap T_d(\hat{c}_1)$ with $u = u^+$. Specifically, $D_1 = (u^+, \varepsilon[\sqrt{\mu}u^+ - (6f(u^+)/g^2(u^+))])$, by (2.18) and (4.2) or by (2.19).
- The right slow segment, which corresponds to the interval $(\Gamma_1, 0]$, is the segment of hyperbolic cosine orbit on $\mathcal{M}$ that goes through D_1 up until it hits the u-axis at a point labeled $(u_{\min}, 0)$.

The phase space geometry reveals that, in this case, a good singular two-pulse solution can be formed by choosing the right pulse so that it is a reflection of the left pulse about $x = 0$. As a consequence, we have $\hat{c}_2 = -\hat{c}_1$. The singular solution for the right pulse has a left slow segment on $[0, \Gamma_2)$, where $\Gamma_2 = -\Gamma_1$, that is, the continuation of the above hyperbolic cosine orbit from $(u_{\min}, 0)$ to the point $O_2 = (u^+, \varepsilon[-\sqrt{\mu}u^+ + (6f(u^+)/g^2(u^+))])$, which is on both $T_o(\hat{c}_2)$ and $\mathcal{F}^{-1}(\ell^s)$; see (2.20). Finally, there is the fast jump at $x = \Gamma_2$ from the point O_2 to the point D_2 on ℓ^s in the phase space, followed by the right slow segment defined for (Γ_2, ∞) along ℓ^s from D_2 all the way back into S.

The results of section 3.2 now directly imply the existence of two-pulse solutions for $0 < \varepsilon \ll 1$, and these true solutions are close to the singular ($\varepsilon = 0$) two-pulse solutions just constructed.

Remark 4.1. The reason for the restriction (4.1) on the allowable u^+ values is intrinsic to the phase space geometry. The first landing point, D_1, which lies on $\mathcal{F}(\ell^u)$, must lie above ℓ^s and below the u-axis. Otherwise, the next slow segment lies outside of the wedge, and the solution cannot return to the $\{\hat{p} > 0\}$ half-plane, or the solution can never enter the $\{\hat{p} < 0\}$ half-plane. The restriction (4.1) states that u^+ should be such that $\mathcal{F}(\ell^u)$ is in between the positive u-axis and the ray $\hat{p} = -\sqrt{\mu}u$, which is ℓ^s, i.e., in that part of the wedge with negative $\hat{p}$-coordinate. As a consequence, D_1 must be a point on a hyperbolic cosine solution of the reduced slow system with negative $\hat{p}$-coordinate. It is also useful to note that, in the special case in which $f(0)/g^2(0) > 0$ and the intersection $\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s)$ is nonempty, the condition (4.1) can also be written as

$$\pi_u\left(\ell^u \cap T_o(0)\right) < u^+ < \pi_u(\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s)), \tag{4.3}$$

where π_u denotes the projection onto the u-coordinate.

4.2. The ODE for the pulse separation distance. In this section, we derive the implicit ODE for the separation distance between the pulses of the symmetric two-pulse solutions under consideration. From the definitions of the pulse positions, the distance between the two pulses in the lab frame (x-axis) is

$$\Delta\Gamma(t) = \Delta\Gamma(0) - 2\varepsilon^3 \int_0^t \hat{c}_1(s)ds, \tag{4.4}$$

where we have introduced $\Delta\Gamma(t) \equiv \Gamma_2(t) - \Gamma_1(t)$ as the distance between the two pulses and we have used $\hat{c}_2 = -\hat{c}_1$. Hence

$$\frac{d}{dt}\Delta\Gamma(t) = -2\varepsilon^3\hat{c}_1(t). \tag{4.5}$$

Moreover, instantaneously, at any t, the speed of the left pulse is given in terms of u^+ by

$$\hat{c}_1 = \left[\sqrt{\mu}u^+ - \frac{3f(u^+)}{g^2(u^+)}\right]\frac{g'(u^+)}{2g(u^+)}, \tag{4.6}$$

since the takeoff (or jump off) point O_1 for the fast jump lies in $\ell^u \cap T_o(\hat{c}_1)$; see (4.2). Therefore, plugging in the pulse speed (4.6) into the ODE (4.5), we find

$$\frac{d}{dt}\Delta\Gamma(t) = -\varepsilon^3\left[\sqrt{\mu}u^+ - \frac{3f(u^+)}{g^2(u^+)}\right]\frac{g'(u^+)}{g(u^+)}. \tag{4.7}$$

Next, we express the right member of this equation in terms of $\Delta\Gamma$, at least implicitly for the general $f(u)$ and $g(u)$ under consideration. From the geometry in the phase space, we see that the distance $\Delta\Gamma(t)$ equals the sum of the "times" of flight along the two hyperbolic cosine orbit segments on $\mathcal{M}$ between D_1 and $(u_{\min}, 0)$ and between $(u_{\min}, 0)$ and O_2, respectively. The leading order expression for this total "time" is found by examining the leading order ODEs for the dynamics on $\mathcal{M}$ written in terms of the fast "time" ξ,

$$\ddot{u} = \varepsilon^4 \mu u,$$

where we used (2.8) and dropped the last term because it is of higher order. Hence the general solution is

$$u(\xi) = Ae^{\varepsilon^2\sqrt{\mu}\xi} + Be^{-\varepsilon^2\sqrt{\mu}\xi}. \tag{4.8}$$

In turn, the coefficients A and B for the desired hyperbolic cosine orbit segment are determined by the condition that $(u(0), \dot{u}(0)/\varepsilon) = D_1$. Hence

$$\begin{aligned} A + B &= u^+, \\ \sqrt{\mu}\,(A - B) &= \sqrt{\mu}u^+ - \frac{6f(u^+)}{g^2(u^+)}, \end{aligned}$$

which implies that

$$A = u^+ - \frac{3f(u^+)}{\sqrt{\mu}g^2(u^+)} \quad \text{and} \quad B = \frac{3f(u^+)}{\sqrt{\mu}g^2(u^+)}. \tag{4.9}$$

The other condition that has to be satisfied is that, when the total "time" of flight ξ_T along these hyperbolic orbit segments satisfies $\varepsilon\xi_T = \Delta\Gamma$, the solution $u(\xi)$ has to be at the point O_2. Looking at $u_\xi(\xi_T)$, we find

$$\varepsilon^2\sqrt{\mu}(Ae^{\varepsilon^2\sqrt{\mu}\xi_T} - Be^{-\varepsilon^2\sqrt{\mu}\xi_T}) = \varepsilon^2\left(-\sqrt{\mu}u^+ + \frac{6f(u^+)}{g^2(u^+)}\right).$$

By setting $z = e^{\varepsilon^2\sqrt{\mu}\xi_T}$, rewriting this equation as a quadratic in z, using (4.9), and choosing the positive solution ($z_+ = B/A$, which is known in terms of u^+), we obtain

$$e^{-\varepsilon\sqrt{\mu}\Delta\Gamma(t)} = \frac{\sqrt{\mu}u^+ - \frac{3f(u^+)}{g^2(u^+)}}{\frac{3f(u^+)}{g^2(u^+)}}. \tag{4.10}$$

Finally, we use (4.10) to replace the term in square brackets in (4.7):

$$\frac{d}{dt}\Delta\Gamma(t) = -3\varepsilon^3\frac{f(u^+)g'(u^+)}{g^3(u^+)}e^{-\varepsilon\sqrt{\mu}\Delta\Gamma(t)}. \tag{4.11}$$

The ODE (4.11) is the desired ODE for the separation distance $\Delta\Gamma(t) = \Gamma_2(t) - \Gamma_1(t)$. The value of u^+ is given implicitly by (4.10), i.e., by

$$\mathcal{G}(u^+) = \frac{e^{-\varepsilon\sqrt{\mu}\Delta\Gamma} + 1}{\sqrt{\mu}}, \tag{4.12}$$

as long as this root exists, where $\mathcal{G}$ is the single-valued function defined by

$$\mathcal{G}(u^+) \equiv \frac{u^+g^2(u^+)}{3f(u^+)}. \tag{4.13}$$

4.3. Interpretation of the implicit ODE (4.11). Since f and g are strictly positive by assumption, the sign of $\frac{d}{dt}\Delta\Gamma$ is determined by that of $g'(u)$. Moreover, $g'(u)$ is sign-definite in the analysis of this section because we assumed that $g'(u) \neq 0$ for all $u > 0$. Therefore, we conclude from (4.11) that

$$\begin{aligned} &\text{if } g'(u) > 0, \quad \text{then } \frac{d}{dt}\Delta\Gamma < 0, \text{and the pulses attract;} \\ &\text{if } g'(u) < 0, \quad \text{then } \frac{d}{dt}\Delta\Gamma > 0, \text{and the pulses repel.} \end{aligned} \tag{4.14}$$

The ODE (4.11) also yields an extremal value of the pulse speed. In particular, the left member of (4.11) equals $-2\varepsilon^3\hat{c}_1$ by the definitions of Γ_1 and Γ_2 and by the fact that $\hat{c}_2 = -\hat{c}_1$ here. Therefore, since the right member has an extremum when $\Gamma_2 = \Gamma_1$, i.e., $\Delta\Gamma = 0$, we find that an extremum of $\hat{c}_1$ occurs at $\sqrt{\mu}u^+g'(u^+)/(4g(u^+))$. This extremum is a maximum when $g'(u) < 0$ and a minimum when $g'(u) > 0$. Also, we note that $\hat{c}(t)$ might have additional (local) extrema; see section 6.

Finally, in the limit $\Delta\Gamma \gg 1$, the ODE for the separation distance between the two pulses in this case is the same as that one would find from weak interaction theory; i.e., to leading order

$$\frac{d}{dt}\Delta\Gamma(t) = C_1 e^{-C_2\Delta\Gamma(t)}$$

for some positive constants C_1 and C_2; see [12, 14].

This completes our analysis of symmetric, slowly varying, two-pulse solutions of (1.1) under conditions (i) and (ii).

5. Examples. In this section, we consider three explicit examples of systems in which semistrong pulse interactions occur and analyze the dynamics of symmetric two-pulse solutions constructed by the methods developed in sections 3 and 4.

The first example has a purely mathematical character. We choose $f(U)$ and $g(U)$ in (1.1) such that the computational effort to determine an explicit version of the ODE (4.11) for $\Delta\Gamma(t)$ is minimal. The other two examples are more physically motivated. As stated in the introduction, both systems are closely related to (and contain as a special case) the well-studied Gierer–Meinhardt equation, and they both show that semistrong pulse interactions can generate unexpected phenomena, such as finite-time blowup.

5.1. An elementary mathematical example. The first example is provided by the elementary mathematical case of $f(u) = \frac{1}{3}g^2(u)$ and $g(u) = e^{\frac{2u}{\gamma}}$, with $\gamma \neq 0$. Here, $f(u)$ and $g(u)$ are chosen such that the quotients $g(u)/g'(u)$ and $3f(u)/g^2(u)$ that appear throughout the analysis of the general case reduce to constants. As a consequence, $T_o(0)$, $T_d(0)$, $\mathcal{F}(\ell^u)$, and $\mathcal{F}^{-1}(\ell^s)$ are all given by straight lines. The ODE (4.11) becomes

$$\frac{d}{dt}\Delta\Gamma = -\varepsilon^3\frac{2}{\gamma}e^{-\varepsilon\sqrt{\mu}\Delta\Gamma} \tag{5.1}$$

with the restriction that $0 < \gamma\hat{c}_1 \leq 1$. This restriction follows necessarily from the condition $u^+ < 2/\sqrt{\mu}$ that arises from the phase space geometry, where $2/\sqrt{\mu} = \pi_u(\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s))$; see (4.1) and Figure 2.1. It directly yields the maximum wave speed $|\hat{c}_1| \leq 1/|\gamma|$.

If $\gamma > 0$, it is clear from (5.1) that $\Delta\Gamma(t)$ decreases and the pulses attract. In addition, their speeds $\hat{c}_1 = -\hat{c}_2$ increase algebraically in time according to the ODE

$$\frac{d}{dt}\hat{c}_1 = 2\varepsilon^4\sqrt{\mu}\hat{c}_1^2$$

up to the maximum wave speed $\hat{c}_1 = 1/\gamma$. At this point, the interaction becomes strong since not even the V-component is near zero in between the pulses anymore, and hence the interaction can no longer be considered to be semistrong. If, on the other hand, $\gamma < 0$, then $\hat{c}_1 < 0$ so that the pulses are repelling. Moreover, $|\hat{c}_1|$ decreases to zero algebraically in time.

The ODE (5.1) for the semistrong interaction of pulses in this symmetric case is identical to that obtained in the weak interaction limit. However, the semistrong analysis also yields the upper bound, $|\hat{c}_1| \leq 1/|\gamma|$, and it is valid for shorter distances $\Delta\Gamma$ between the pulses.

5.2. A modified Gierer–Meinhardt equation with finite-time blowup. Next, we consider the more realistic example of $f(u) \equiv 1$ and $g(u) = (1/u) + \alpha$, with $\alpha \geq 0$ since $g(U)$ in (1.1) must be positive; see, however, Remark 5.2. Thus (1.1) becomes a modified Gierer–Meinhardt equation,

$$\left\{ \begin{array}{rcl} \varepsilon^2 U_t &=& U_{xx} - \varepsilon^2\mu U + V^2, \\ V_t &=& \varepsilon^2 V_{xx} - V + (\frac{1}{U} + \alpha)V^2, \end{array} \right. \tag{5.2}$$

in which α plays the role of (an additional) bifurcation parameter. This equation reduces to the classical Gierer–Meinhardt equation if $\alpha = 0$; recall [7, 21]. Moreover, $g'(u) < 0$ for all $u > 0$ so that the pulses in symmetric two-pulse solutions are repelling for any α, including the classical Gierer–Meinhardt case $\alpha = 0$, by the theory presented in section 4.

5.2.1. The ODE for $\Delta\Gamma(t)$. It follows from (2.18), (2.19), and (2.20) that

$$T_o(\hat{c};\alpha), T_d(\hat{c};\alpha): \quad \hat{p} = -2\hat{c}u(\alpha u + 1) \pm \frac{3u^2}{(\alpha u + 1)^2}$$

and

$$\mathcal{F}(\ell^u), \mathcal{F}^{-1}(\ell^s): \quad \hat{p} = \pm\sqrt{\mu}u \mp \frac{6u^2}{(\alpha u + 1)^2}.$$

First, we observe that this equation can have two stationary, homoclinic, one-pulse solutions since the equation $\alpha^2u^2 + (2\alpha - 3/\sqrt{\mu})u + 1 = 0$, which is obtained from $T_o(0;\alpha) \cap \ell^u$, has two (positive) solutions for

$$0 < \alpha < \alpha_{bu} = \frac{3}{4\sqrt{\mu}}. \tag{5.3}$$

See also Remark 5.3. As $\alpha \to 0$, one of these pulse solutions becomes unbounded, and the other one merges with the (uniquely determined) one-circuit homoclinic pulse solution of the (classical) Gierer–Meinhardt equation [7]. In the opposite limit as $\alpha \uparrow \alpha_{bu}$, the pulses merge in a saddle-node bifurcation of homoclinic orbits.

The intersection $\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s)$ also consists of up to two points, and hence the condition (4.1) can yield up to two intervals of existence. In this section, we focus on the interval that corresponds to two-pulse solutions that merge with two-pulse solutions of the Gierer–Meinhardt equation in the limit $\alpha \to 0$ since these two-pulse solutions can be expected to

be stable. (These two-pulse solutions are associated to the stationary pulse solution of the Gierer–Meinhardt equation, and these pulses are stable if μ is above a certain critical value associated to a Hopf bifurcation [7].) We shall see that two-pulse solutions exist beyond α_{bu}, i.e., for values of α for which there are no stationary pulses. This is not necessarily surprising since the construction of two-pulse solutions is based on the intersection(s) of $T_o(\hat{c};\alpha)$ and ℓ^u for $\hat{c} \neq 0$. This intersection can of course be nonempty even if $T_o(0;\alpha) \cap \ell^u = \emptyset$. Furthermore, $\mathcal{F}(\ell^u)$ lies in the $\{\hat{p} > 0\}$ half-plane for $\alpha > 3/(2\sqrt{\mu}) = 2\alpha_{bu}$; recall (5.3). Hence it is not possible to construct these types of two-pulse solutions for $\alpha > 2\alpha_{bu}$; see Remark 4.1.

The ODE for $\Delta\Gamma(t)$ in this example may be determined directly by substituting the choices of f and g into (4.11). We obtain

$$\frac{d}{dt}\Delta\Gamma = 3\varepsilon^3 \frac{u^+}{(\alpha u^+ + 1)^3} e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}, \tag{5.4}$$

where $0 < \alpha < 2\alpha_{bu}$, and we note that, by (4.12) and (4.13), u^+ is related to $\exp(-\varepsilon\sqrt{\mu}\Delta\Gamma)$ through

$$\frac{e^{-\varepsilon\sqrt{\mu}\Delta\Gamma} + 1}{\sqrt{\mu}} = \frac{(\alpha u^+ + 1)^2}{3u^+} = \mathcal{G}(u^+),$$

where we recall the definition of $\mathcal{G}$ given in (4.13). A little algebra yields

$$\frac{1}{\alpha u^+ + 1} = \frac{1}{2} - \frac{1}{2}\sqrt{1 - \frac{4\alpha}{3\mathcal{G}}}.$$

Hence the ODE (5.4) becomes an explicit ODE for $\Delta\Gamma(t)$,

$$\frac{d}{dt}\Delta\Gamma = \varepsilon^3\sqrt{\mu}\frac{\sqrt{1 - \frac{\alpha}{\alpha_{bu}} + e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}}}{(1 + e^{-\varepsilon\sqrt{\mu}\Delta\Gamma})^{3/2}} e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}. \tag{5.5}$$

Note that in the classical Gierer–Meinhardt equation, i.e., $\alpha = 0$ in (5.2), the ODE for the evolution of symmetric two-pulse solutions reduces to

$$\frac{d}{dt}\Delta\Gamma = \varepsilon^3\sqrt{\mu}\frac{e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}}{1 + e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}}. \tag{5.6}$$

5.2.2. Pulse dynamics governed by (5.5) and the blowup time. For parameter values $0 < \alpha < \alpha_{bu}$, the solutions $\Delta\Gamma(t)$ of the ODE (5.5) are defined for all $t > 0$ because $\exp(-\varepsilon\sqrt{\mu}\Delta\Gamma) \in (0,1)$. The ODE (5.5) describes a pair of symmetric pulses which always move away from each other with decreasing speed.

For parameter values $\alpha_{bu} \leq \alpha < 2\alpha_{bu}$, the ODE (5.5) also governs the evolution of symmetric two-pulse solutions even though there is no stationary, homoclinic, one-pulse solution for $\alpha \geq \alpha_{bu}$. However, the ODE is well defined now only for a finite time and only for $\Delta\Gamma(t)$ not too large. More precisely, in order for the ODE (5.5) to be well defined, $\Delta\Gamma(t)$ must be less than the critical pulse separation distance,

$$\Delta\Gamma_{bu} = \frac{-1}{\varepsilon\sqrt{\mu}} \ln\left(\frac{\alpha}{\alpha_{bu}} - 1\right),$$

which is the value of $\Delta\Gamma(t)$ when the term under the square root in (5.5) vanishes.

Let us take a more in-depth look at two pulses that, at $t = 0$, are less than a distance $\Delta\Gamma_{bu}$ apart, i.e., $\Delta\Gamma(0) < \Delta\Gamma_{bu}$. As just shown, the dynamics of these pulses is governed by (5.5). Nevertheless, (5.5) shows that these pulses move away from each other so that $\Delta\Gamma(t)$ increases. It follows that there is a time $t = t_{bu}$, which depends on the initial conditions, at which $\Delta\Gamma = \Delta\Gamma(t_{bu}) = \Delta\Gamma_{bu}$. Beyond t_{bu}, it is no longer possible to have two-pulse orbits (almost) homoclinic to the slow manifold $\mathcal{M}$. The methods developed in this article cannot be used to describe the pulses after the pulses reach the critical separation distance $\Delta\Gamma_{bu}$.

5.2.3. Finite-time blowup and a codimension two self-replication/blowup bifurcation point. Numerical simulations show that the theory works very well qualitatively and quantitatively. In Figure 5.1, we show the evolution of a symmetric two-pulse solution of (5.2) with $\varepsilon^2 = 0.01$, $\mu = 5$, and $\alpha = 0.342$; thus, by (5.3), $\alpha \in (\alpha_{bu}, 2\alpha_{bu})$, where $\alpha_{bu} \approx 0.335$. The evolution of the pulses is slow up to $t = t_{bu}$; in fact, the speed of the pulses reduces to zero as $t \to t_{bu}$ (5.2). Then, as soon as t passes through t_{bu}, one of the pulses starts to grow rapidly, both in its U-component and in its V-component. Both components blow up in finite time. The solution to (5.2) has become unbounded on a localized spatial domain for t slightly above t_{bu}.

This finite-time blowup behavior occurs generically in (5.2) for α and μ such that $\alpha \in (\alpha_{bu}, 2\alpha_{bu})$ with suitably chosen initial conditions—see below. Only one of the pulses blows up; the other one is almost immediately overtaken by the exploding one. The distinction between blowing up and being overtaken seems to depend on (extremely small) asymmetric effects in the sense that an initially (i.e., for $t < t_{bu}$) negligible difference between the amplitude of the two pulse is strongly magnified as soon as t passes through t_{bu}.

The roots of this dynamic finite-time blowup behavior lie in the bifurcation of the stationary, solitary, homoclinic, one-pulse solution to (5.2), which exists and is stable for $\alpha < \alpha_{bu}$. For $\alpha = \alpha_{bu} - \sigma$ with $0 < \sigma \ll 1$, the solitary, one-pulse solution is attracting, whereas for $\alpha = \alpha_{bu} + \sigma$, initial conditions coincident with—or sufficiently close to—the form of the one-pulse solution for $\alpha = \alpha_{bu} - \sigma$ blow up in finite time in (5.2). However, this is not the case for all μ. As soon as μ becomes $\mathcal{O}(1/\varepsilon^4)$—see [10] and Remark 5.4—the pulse might undergo a self-replication bifurcation instead of the finite-time blowup bifurcation; i.e., the pulse might split into a (symmetric) pair of traveling pulses that eventually split again, etc. Thus the self-replication bifurcation also occurs in the modified Gierer–Meinhardt equation (5.2). As explained in [10], self-replication of pulses is a generic phenomenon that is not unique to the Gray–Scott equation [3, 4, 5, 9, 28, 31, 34, 39, 40, 43, 44]. Both bifurcations determine a curve in the (μ, α) parameter plane. It is found (numerically) for $\varepsilon^2 = 0.1$ that these two curves intersect at a codimension two finite-time self-replication/blowup bifurcation point at $(\mu, \alpha) \approx (22, 0.06)$. This last statement is based on numerical observation; however, the methods developed in [7] can be used to obtain rigorous results. See also section 7.1.

Remark 5.1. The ODE (5.5) for $\Delta\Gamma(t)$ (in this case of symmetric pulses) has a natural counterpart in an ODE for $\hat{c} = \hat{c}(t) = \pm\hat{c}_{1,2}(t)$ that may be derived by substituting the general relation (4.5) between $\frac{d\Delta\Gamma}{dt}$ and $\hat{c}$ into the left member of (5.5). However, the expression for $\hat{c}$ determined in this way is in terms of $Z = \exp(-\varepsilon\sqrt{\mu}\Delta\Gamma)$ (recall (4.5)). Hence one first needs to determine an expression for Z in terms of $\hat{c}$ and then to invert that expression before one

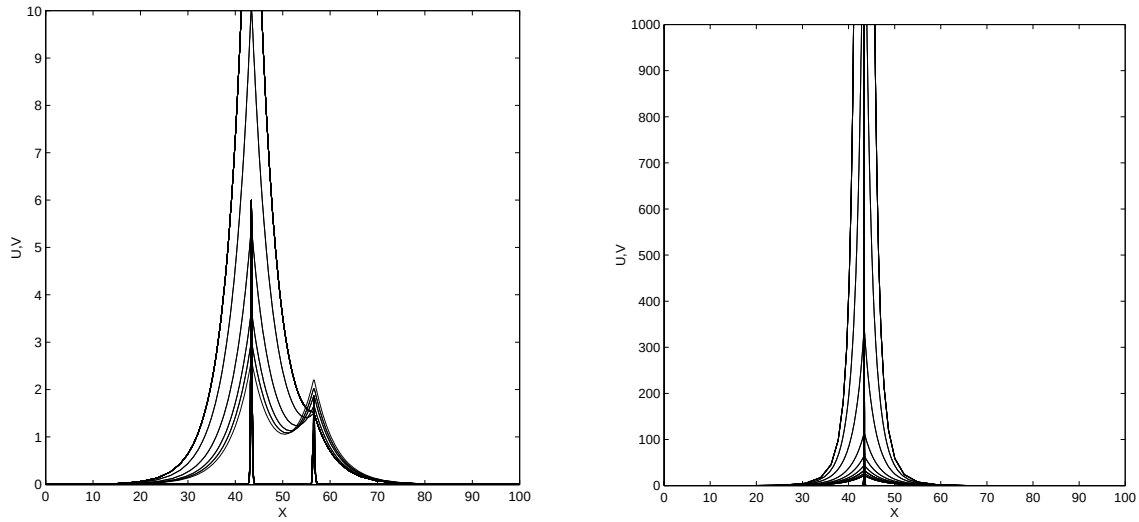


Figure 5.1. *Evolution of a symmetric two-pulse solution of the modified Gierer–Meinhardt equation* (5.2) *with* $\varepsilon = 0.1$, $\mu = 5$, *and* $\alpha = 0.342$. *Here,* $\alpha \in (\alpha_{bu}, 2\alpha_{bu})$, *but close to* $\alpha_{bu} \approx 0.335$; *see section* 5.3.2. *In the left frame, the sequence of times is* $t = 0, 40, 50, 55, 57, 57.6$, *where* $t = 0$ *corresponds (approximately) to the time* ($t = t_{bu}$) *at which the symmetric two-pulse solution reaches the critical separation distance* $\Delta\Gamma_{bu}$. *In the right frame, the results of continuing the same simulation are presented for the sequence of times* $t = 57.8, 57.9, 58.0, 58.1, 58.2, 58.3, 58.4, 58.45, 58.47$. *Blowup in both* U *and* V *occurs appears to occur at* $t \approx 58.472$ *(and of course the actual value may vary slightly depending on implementation and hardware). Again, we used* 201 *grid points and homogeneous Neumann boundary conditions.*

can derive the equation for $d\hat{c}/dt$ from (5.5). Thus, for general α, one has to solve a cubic equation in Z. This is an unpleasant task, and the outcome will not give much additional insight. However, there are two special values of α, the Gierer–Meinhardt case $\alpha = 0$ and the bifurcation case $\alpha = \alpha_{bu}$, for which the cubic equation reduces to a much simpler equation. For the Gierer–Meinhardt equation ($\alpha = 0$), we find that the equation for the speed of the pulses is given by

$$\frac{d}{dt}\hat{c} = 2\varepsilon^4(\sqrt{\mu} + 2\hat{c})\hat{c}^2.$$

Also, for the bifurcation case, $\alpha = \alpha_{bu}$, we find

$$\frac{d}{dt}\hat{c} = 2\varepsilon^4\mu^{1/6}(\mu^{1/3} - (2\hat{c})^{2/3})\hat{c}^2.$$

Remark 5.2. In this article, we assume that $g(U)$ in (1.1) is positive so that homoclinic solutions can exist in the fast reduced limit problem (2.5). This does not rule out examining (5.2) for $\alpha < 0$. In fact, for $\alpha < 0$ we can (only) consider U such that $g(U) \geq 0$; i.e., we consider U values that are small enough. In that case, it is also found that the stationary, solitary, homoclinic pulse of the Gierer–Meinhardt equation ($\alpha = 0$ in (5.2)) persists, as is the case for $\alpha > 0$. The fate of this pulse for decreasing α of course also depends on the other parameters (ε and μ). For instance, it is found numerically for $\varepsilon^2 = 0.01$ and $\mu = 5.0$ that the pulse remains stable up to $\alpha \approx -400$ (!). For such values of α, the pulse has indeed become

tiny ($0 < U < 0.0025$). As α decreases even further, the pulse undergoes the self-replication bifurcation, which leads to a spatially periodic array of "midget pulses."

Remark 5.3. Since $T_o(0;\alpha)$ and ℓ^u can have two intersection points, it is also possible to construct pairs of *asymmetric* pulses. The geometric construction is straightforward; however, the derivation of the ODE for the dynamics of the pulses is quite cumbersome and does not give additional insight. Interestingly, though, the asymmetric solution approaches the symmetric two-pulse solution as the parameter α approaches α_{bu}. So there may be a connection between the asymmetry that appears in the simulations reported here, which are for a value of α very close to, though above, α_{bu}. Finally, we do not expect that these asymmetric two-pulse solutions are stable.

Remark 5.4. It has been proved in [10] that there is a bifurcation in the Gierer–Meinhardt equation ($\alpha = 0$ in (5.2)) that annihilates the solitary homoclinic pulse for $\mu = \mathcal{O}(1/\varepsilon^4)$. This bifurcation initiates the self-replication process. A similar analytical result can be obtained for the modified Gierer–Meinhardt equation (5.2) introduced above and for the modified Gierer–Meinhardt equation (5.7) introduced below.

5.3. A modified Gierer–Meinhardt equation with asymptotically large pulses. Finally, we consider another small modification of the classical Gierer–Meinhardt equation

$$\left\{\begin{array}{rcl} \varepsilon^2 U_t &=& U_{xx} - \varepsilon^2 \mu U + V^2, \\ V_t &=& \varepsilon^2 V_{xx} - V + (\frac{1}{U} + \frac{\beta}{\sqrt{U}})V^2; \end{array}\right. \tag{5.7}$$

i.e., we have set $f(U) \equiv 1$ and $g(U) = (1/U) + (\beta/\sqrt{U})$ in (1.1), where $\beta \geq 0$ is the new bifurcation parameter. In the special case $\beta = 0$, (5.7) is again the Gierer–Meinhardt equation.

5.3.1. Blowup of stationary, homoclinic, one-pulse solutions. We immediately obtain

$$T_o(\hat{c};\beta), T_d(\hat{c};\beta): \quad \hat{p} = -2\hat{c}u\frac{1+\beta\sqrt{u}}{1+\frac{1}{2}\beta\sqrt{u}} \pm \frac{3u^2}{(1+\beta\sqrt{u})^2}$$

and the related expressions for $\mathcal{F}(\ell^u)$ and $\mathcal{F}^{-1}(\ell^s)$. For

$$0 \leq \beta < \beta_\infty = \sqrt{\frac{3}{\sqrt{\mu}}}, \tag{5.8}$$

there is a unique intersection of $T_o(0;\beta)$ and ℓ^u with u-coordinate

$$u = \frac{1}{(\beta_\infty - \beta)^2}. \tag{5.9}$$

Thus there is a uniquely determined stationary, homoclinic, one-pulse solution for $0 \leq \beta < \beta_\infty$. Moreover, the formula (5.9) reveals that the amplitude of this solution diverges as $\beta \uparrow \beta_\infty$; i.e., the blowup is determined *analytically* in this example. This situation contrasts with the situation for the previous example, where the amplitude approaches a finite number ($\frac{3}{4\sqrt{\mu}}$) as $\alpha \to \alpha_{bu}$ and the blowup was observed numerically.

Here the analysis is again confirmed by the numerical simulations. The homoclinic pulse that corresponds to $T_o(0;\beta) \cap \ell^u$ is asymptotically stable up to $\beta = \beta_\infty$. Its amplitude also

agrees (to leading order) with (5.9): for $\varepsilon = 0.1$ and $\mu = 5.0$, i.e., $\beta_\infty \approx 1.158$ (5.8), we observe, for instance, in the simulations that the amplitude of the U-component of the pulse is approximately 5.0 for $\beta = 0.7$, while (5.9) gives 4.76. Of course, the error grows as $\beta \to \beta_\infty$: for $\beta = 1.1$, i.e., $\beta_\infty - \beta \approx 0.058$, the height of the U-component of the pulse is ≈ 337.8 in the simulation (see also Figure 5.2), while (5.9) yields ≈ 294.2. Nevertheless, the relative error is still of $\mathcal{O}(\varepsilon)$ near the singular limit.

5.3.2. Symmetric two-pulse solutions: The ODE for $\Delta\Gamma(t)$ and an analytical expression for finite-time blowup. The derivation of the ODE for $\Delta\Gamma(t)$, the distance between the pulses of a two-pulse solution, proceeds directly along the lines of the general theory of section 4. We find

$$\frac{d}{dt}\Delta\Gamma = \varepsilon^3\sqrt{\mu}\frac{\left[\sqrt{1+e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}} - \frac{\beta}{2\beta_\infty}\right]}{(1+e^{-\varepsilon\sqrt{\mu}\Delta\Gamma})^{3/2}}e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}, \tag{5.10}$$

which also reduces to (5.6) in the Gierer–Meinhardt limit $\beta = 0$.

As in the previous example, it is possible to construct symmetric two-pulse solutions beyond the critical value β_∞. The upper bound on β is determined by the observation that two-pulse solutions can only be constructed for u values such that the corresponding point on $\mathcal{F}(\ell^u)$ lies below the u-axis. Hence it follows that two-pulse solutions exist for $0 \le \beta < \sqrt{2}\beta_\infty$. Note that this second critical value of β, $\sqrt{2}\beta_\infty$, does not show up explicitly in the ODE (5.10). We shall discuss the implications of this below.

The amplitude u^+ of the U-component of the pulse is given by

$$u^+ = \frac{1}{(\beta_\infty\sqrt{1+e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}} - \beta)^2}; \tag{5.11}$$

see (4.12) and (4.13). This relation confirms the necessity of the condition $\beta < \sqrt{2}\beta_\infty$. Thus u^+ remains bounded for all (finite) time as long as $\beta \le \beta_\infty$, and the ODE (5.10) is well defined for all time.

In contrast, for $\beta_\infty < \beta < \sqrt{2}\beta_\infty$, the ODE (5.10) is only well defined for

$$\Delta\Gamma < \Delta\Gamma_\infty = \frac{-1}{\varepsilon\sqrt{\mu}}\ln\left(\frac{\beta^2}{\beta_\infty^2} - 1\right).$$

Thus, for $\beta_\infty < \beta < \sqrt{2}\beta_\infty$, the ODE (5.10) governs the dynamics of two-pulse solutions that are close enough at $t = 0$; i.e., $\Delta\Gamma(0)$ must be less than $\Delta\Gamma_\infty$. Since $\Delta\Gamma(t)$ is an increasing function of time (5.10), there must be a critical time t_∞ at which $\Delta\Gamma(t_\infty) = \Delta\Gamma_\infty$. Moreover, it is clear from (5.11) that the amplitude of the pulse diverges as t approaches t_∞. Thus this second modified Gierer–Meinhardt equation (5.7) also exhibits finite-time ($t = t_\infty$) blowup, and in this example the blowup can be studied analytically. Finally, if $\Delta\Gamma(0) > \Delta\Gamma_\infty$, then the pulses blow up immediately by the same process that causes the blowup of the stationary pulse for $\beta \ge \beta_\infty$.

At first sight, the dynamics of pulses in (5.7) seems to be essentially the same as in (5.2). However, there are three crucial differences, two of which emerge from the analysis and a third one that is observed in the numerical simulations, which will be discussed below. First, as

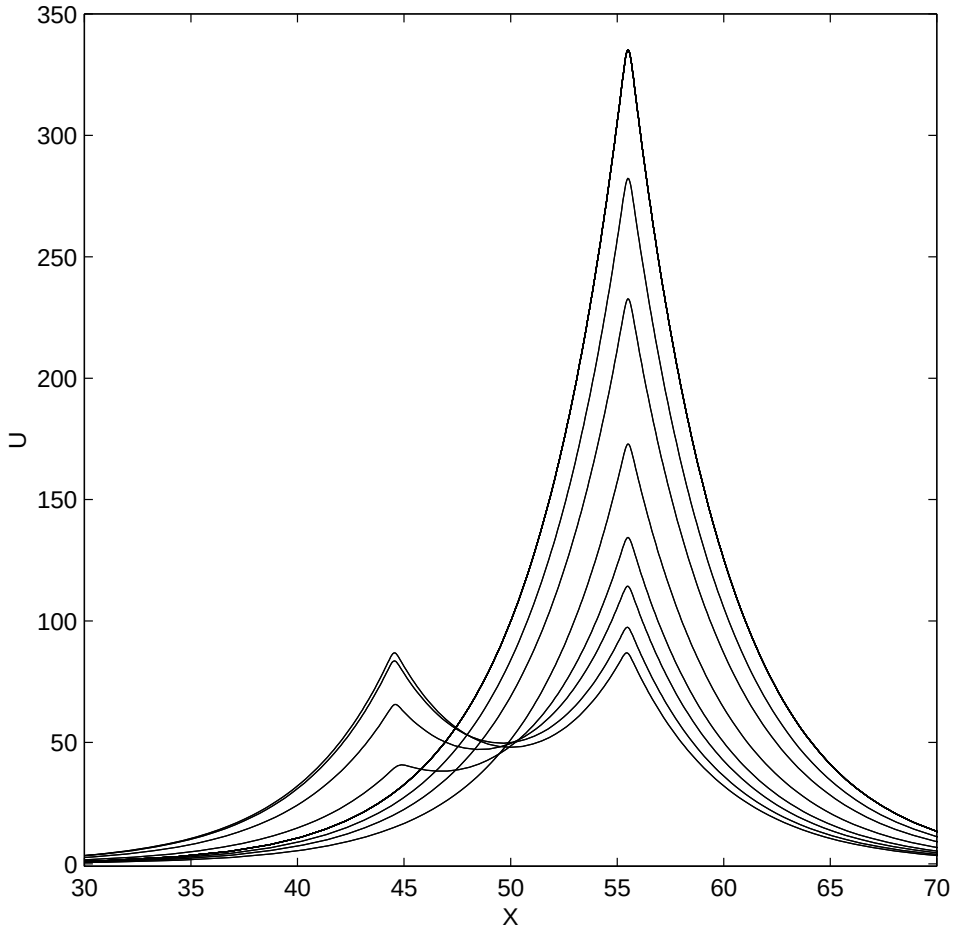


Figure 5.2. *Symmetry-breaking in the evolution of the U-component of an initially symmetric, slowly varying, two-pulse solution of the second modified Gierer–Meinhardt equation (5.7) with $\varepsilon^2 = 0.01$, $\mu = 5$, and $\beta = 1.1$. The symmetry-breaking bifurcation is observed in this simulation since $\Delta\Gamma(0) < \Delta\Gamma_{sb}$. See section 5.3.3. The time steps are $t = 0, 60, 68, 72, 76, 82, 90, \infty$, where again we set $t = 0$ at (approximately) the last time ($t = t_\infty$) at which a symmetric two-pulse solution exists. (The left pulse has its maximum value for the sequence of times shown here, while the right pulse has its minimum.) At $t = 60$, the asymmetry has become noticeable, and it is growing quickly. At $t = 72$, the left pulse has almost disappeared, and it is no longer visible at $t = 76$. The right pulse then grows even more quickly, and at $t = 90$ it is close to its asymptotic value ($t \to \infty$), which corresponds to the largest peak shown here. This simulation was also done with 201 grid points and homogeneous Neumann boundary conditions.*

was already noted, the finite-time blowup in this case can be fully understood by the analysis, unlike in the previous example. Here, the blowup occurs at $t = t_\infty$ so that the process can be described in full analytical detail, whereas in the previous example the finite-time blowup mechanism starts beyond the critical time t_{bu}. Second, there is a difference in the dynamics of the pulses at (or near) the critical time, t_{bu} or t_∞, respectively. In the previous example, the speed of the pulses approaches 0 as $t \to t_{bu}$. Here, the critical upper bound $\sqrt{2}\beta_\infty$ on β does not show up explicitly in the ODE (5.10). As a consequence, the speed of the pulses

approaches a well-defined limit as $t \to t_\infty$, namely,

$$|\hat{c}_\infty| = \frac{1}{4}\sqrt{\mu}\left(\frac{\beta^2}{\beta_\infty^2} - 1\right),$$

where we have used (4.5).

5.3.3. Stability and a symmetry-breaking bifurcation. We have based the two modified Gierer–Meinhardt examples of sections 5.2 and 5.3 on recently obtained results on the spectral stability of the homoclinic one-pulse solution in the Gierer–Meinhardt equation ((5.2), (5.7) with $\alpha, \beta = 0$); see [7]. Both pulses of the two-pulse solutions approach a single-pulse homoclinic solution of the Gierer–Meinhardt equation as $t \to \infty$ and $\alpha, \beta \to 0$. This gives a heuristic argument by which one can expect to see stable two-pulse solutions in the modified systems (5.2) and (5.7). It has been shown in [3, 4] that this argument works in the case of the Gray–Scott equation. Numerical simulations for the previous example suggest that the argument is also valid here. Moreover, the simulations also confirm the persistence of the stability beyond $\beta = 0$ for the stationary pulses. Of course, neither of these observations prove the stability.

On the contrary, the numerical simulations show that one cannot expect to fully understand the semistrong dynamics of the two-pulse solutions in (5.7) even with the combination of the existence analysis developed in this article and the stability results for the stationary homoclinic pulse.

We have performed numerical simulations of two-pulse solutions in (5.7) for $\varepsilon = 0.1$ and $\mu = 5.0$. (Note that the homoclinic pulse in the Gierer–Meinhardt ($\beta = 0$) equation has been proved to be stable for $\mu > \mu_{\text{Hopf}} \approx 0.36 + \mathcal{O}(\varepsilon)$.) Although we have already seen that the stationary pulses are stable up to $\beta = \beta_\infty$, which is approximately 1.158 by (5.8), we observed that the two-pulse solution undergoes a symmetry-breaking destabilizing bifurcation at $\beta = \beta_{sb} \approx 0.9 < \beta_\infty$. As long as $\beta < \beta_{sb}$, the two-pulse solution travels and grows according to (5.10) and (5.11). By contrast, for $\beta > \beta_{sb}$, the dynamics of the two-pulse solution depends crucially on the distance $\Delta\Gamma(0)$ between the pulses at $t = 0$. If $\Delta\Gamma(0)$ is large enough, i.e., if the system is close to the weak interaction limit and both pulses are sufficiently close to the solitary homoclinic pulse, then the pulses behave exactly as for $\beta < \beta_{sb}$. Note that this implies that the symmetry-breaking bifurcation is caused by the semistrong interactions. On the other hand, if $\Delta\Gamma(0)$ is below a certain critical value $\Delta\Gamma_{sb} = \Delta\Gamma_{sb}(\beta)$, then the pulses still evolve according to (5.10), but initially negligible differences between the heights of the two pulses increase (on a certain intermediate time scale). After a certain critical time, the smallest of the two pulses is overtaken completely by the larger one. The speed of the resulting pulse decreases to zero, and the end product of the symmetry-breaking bifurcation is a stable solitary pulse (see Figure 5.2). The numerics also show that the critical distance $\Delta\Gamma_{sb}(\beta)$ diverges rapidly as β approaches β_∞. The initial separation $\Delta\Gamma(0)$ must be so large for $\beta = 1.1$—recall that $\beta_\infty \approx 1.158$—that the interaction between the pulses indeed has become extremely weak. Thus we conclude that the predicted finite-time blowup does not occur due to the destabilizing symmetry-breaking effects. However, the appearance of the destabilizing bifurcation is expected to depend on the parameters μ and β. It is possible that the finite-time blowup behavior is stable for parameter combinations other than the ones used in the

numerical simulations reported here.

In summary, both the existence and the stability of two-pulse solutions depend crucially on the semistrong nature of the interaction.

5.4. Relation to literature on finite-time blowup. For both modified Gierer–Meinhardt equations (5.2) and (5.7), the associated ODEs that govern spatially homogeneous patterns have solutions with small initial conditions that grow arbitrarily large. Therefore, in a certain sense, the discovery of finite-time blowup for the PDEs is not a surprise. A more detailed examination of the mechanism(s) responsible here is under way and is beyond the scope of this article.

At first glance, for both (5.2) and (5.7), the spatially localized V profile acts as a source in the linear equation for U, and the amplitude of this source is $1/\varepsilon^2$, which is large compared to the $\mathcal{O}(1)$ decay rate (μ) of U. Hence unbounded growth in U is possible.

Now, for (5.7) with μ not too large, we observe the following interplay between the growth of U and V. The maximum value of V along the homoclinic orbit/pulse is given to leading order by $V_{\max} \sim 3/[2((1/U) + (\beta/\sqrt{U}))]$; see (2.7). Hence, as the (maximum) value of U at a pulse center increases, so does $V_{\max}$, which in turn fuels a further increase in U since V acts as a source term in the U equation, etc. In this sense, there appears to be a positive feedback loop, with the growth in the components reinforcing each other.

In addition, for (5.7), as reported above, there is a transition from blowup in finite time to self-replication at sufficiently large μ. In other words, as μ increases, there appears to be a transition to a regime in which the decay rate μ is strong enough to balance the large source term and hence prevent unbounded growth.

For (5.2), the mechanism causing the finite-time blowup appears to be different. In fact, as suggested by [46], the mechanism here may be similar to the known (see below) blowup results for self-similar solutions of scalar reaction-diffusion equations of the form $V_t = DV_{xx} - V + V^2$. More precisely, $g(U) \to \alpha$, a constant, as U gets large. Hence the V equation essentially decouples (for large U) and is of this classical form.

The literature on finite-time blowup in reaction-diffusion equations and systems is large. We cite three interesting examples of coupled reaction-diffusion equations in which the reactions are given by polynomials and in which blowup in finite time is known to occur. First, [37] establishes the blowup of the catalyst concentration b in the isothermal autocatalytic reaction $A \to B$ that has rate constant kab^n with $n \geq 1$ (i.e., in so-called nth order autocatalysis). In the limit that the ratio of the activator diffusivity to the inhibitor diffusivity is asymptotically small, it is shown that blowup occurs at a point in infinite time for $n \in [1, 2]$ and in finite time for $n > 2$. Second, [29] demonstrates the blowup of solutions of the coupled system $u_t = d_1\Delta u + u(a_1 - b_1u + c_1v)$, $v_t = d_2\Delta v + v(a_2 + b_2u - c_2v)$ with homogeneous Neumann boundary conditions and where the b_i, c_i, and d_i are all positive, where the a_i are real numbers, and where Δ denotes the Laplacian on R^N. Third, [17] presents "Fujita-type" blowup and global existence results for systems $u_t = \Delta u + u^{p_1}v^{q_1}$, $v_t = \Delta v + u^{p_2}v^{q_2}$, where the powers satisfy certain inequalities.

This third work, among others, extends the classical results of [20] for the scalar equations $u_t = \Delta u + f(u)$, where $f(u) > 0$ for $u > 0$. Fujita proved the existence of nonnegative solutions that decay sufficiently rapidly at infinity and that blow up in finite time for $f(u) = u^p$ with

$1 < p < 1 + (2/N)$, and he showed that, for p larger than critical, all solutions generated from small initial data exist for all time [20]. Hence one may say (see also [27]) that, in the former case, small data grows due to the reaction term and the diffusion is not strong enough to stabilize it, whereas, in the latter case, the diffusion is strong enough to prevent unbounded growth. Also, we refer to [19] for a seminal analysis of blowup in this equation on bounded domains and to the useful article [27] that reviews and presents various extensions of the classical results.

For completeness, we also mention that it has already been shown that diffusion can destabilize a globally stable equilibrium of the reaction kinetics in such a strong manner that there is blowup in finite time for certain pairs of weakly coupled systems. See [49] and the references there. Also, [41] presents new results showing that blowup in the L^∞ norm is possible for solutions of pairs of reaction-diffusion equations with mass dissipation even if there is an a priori bound on the solution in the L^1 norm. However, neither of our modified Gierer–Meinhardt models satisfies the hypotheses made in these works.

6. Attracting and repelling stationary two-pulse solutions. In this section, we study systems of the form (1.1) that satisfy the following:

(i) The function $g'(u)$ has one simple zero, say, at $u = \alpha$ (which differs from the α used in section 5.2).

(ii) The function $H(u)$ again has one simple zero.

Condition (ii) is the same as that imposed in section 4, where we recall that, geometrically, it implies that ℓ^u and $T_o(0)$ intersect transversely at a unique point (and similarly for ℓ^s and $T_d(0)$). However, condition (i) is different, and the presence of the zero of $g'(u)$ dramatically alters the locations of the takeoff and touchdown curves $T_o(\hat{c}_i)$ and $T_d(\hat{c}_i)$, $i = 1, 2$, respectively, and makes possible a richer pulse dynamics, as we will see below.

We assume, for ease of presentation, that the condition (4.3) holds, and we consider the following three cases:

$$\begin{aligned} &\text{Case I:} && 0 < \alpha < \pi_u(\ell^u \cap T_o(0)), \\ &\text{Case II:} && \pi_u(\ell^u \cap T_o(0)) < \alpha < \pi_u(\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s)), \\ &\text{Case III:} && \pi_u(\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s)) < \alpha, \end{aligned} \tag{6.1}$$

where we recall that π_u denotes the projection onto the u-coordinate. These three cases include all of the structurally different cases, so we see that the assumption (4.3) is not restrictive.

In all three cases, we consider symmetric singular two-pulse solutions (symmetric about $x_0 = 0$). These are constructed exactly as the two-pulse solutions in section 4 were constructed. Their left and right pulses have slow and fast segments that are defined in terms of the intersection points O_1, D_1, O_2, and D_2, although the locations of these points differ because the positions of the curves $T_o(\hat{c}_i)$, $T_d(\hat{c}_i)$, $\mathcal{F}(\ell^u)$, and $\mathcal{F}^{-1}(\ell^s)$ are different. Moreover, there are two sets of such points in case II, which is the feature that makes the richer pulse dynamics possible. See Figure 6.1.

The main outcome of the analysis in this section is that a pair of slowly varying two-pulse solutions does not necessarily reach either the weak interaction limit ($\Delta\Gamma \to \infty$) or the strong interaction limit ($\Delta\Gamma \downarrow 0$). Slowly varying two-pulse solutions may also converge to a stationary two-pulse solution of the semistrong type. We present the essence of the analysis

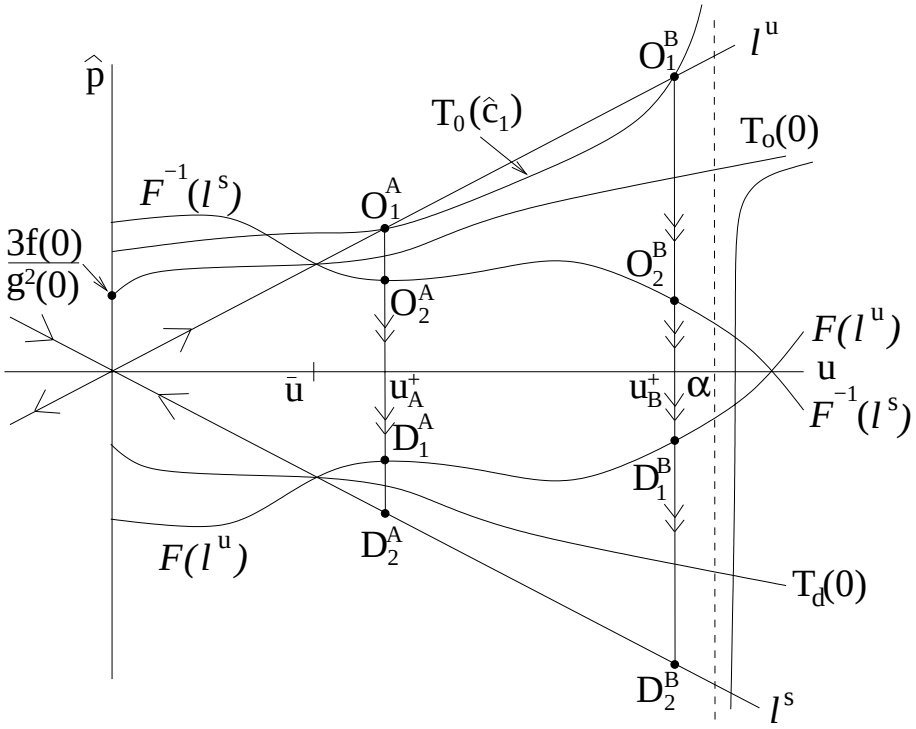


Figure 6.1. *The geometry of the curves $T_o(0)$, $T_o(\hat{c}_1)$, $T_d(0)$, ℓ^u, ℓ^s, $\mathcal{F}(\ell^u)$, and $\mathcal{F}^{-1}(\ell^s)$, along with their intersections, on the invariant manifold $\mathcal{M}$ for case* II *in the case studied in section* 6.1. *The curve $T_d(\hat{c}_2)$ is not shown in order not to clutter the figure. There are two sets of takeoff and touchdown points O_1^A, D_1^A, O_2^A, and D_2^A as well as O_1^B, D_1^B, O_2^B, and D_2^B, as shown in section* 6.1, *and hence two different singular, slowly varying, two-pulse solutions. The intersection $\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s)$ occurs at a point on the u-axis, denoted $\pi_u(\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s))$ in the text. Also, $\bar{u} = \pi_u(\ell^u \cap T_o(0))$.*

for the most general case since the details are similar to those in section 4. We do, however, give the details in the context of a specific example.

Remark 6.1. Throughout this section, we will assume, for the sake of definiteness, that $g'' < 0$ at $u = \alpha$. This implies that $g'(u) > 0$ for $u < \alpha$ and $g'(u) < 0$ for $u > \alpha$. The results for the opposite situation can be obtained in the same manner.

6.1. Case II. In case II, there exist two types of symmetric, slowly varying, two-pulse solutions. Moreover, there also exists a symmetric, stationary, two-pulse solution that attracts the slowly varying solutions for $g''(\alpha) < 0$ and repels them for $g''(\alpha) > 0$. We construct these slowly varying solutions and determine ODEs for the pulse separation distances.

Following the blueprint of section 3, we begin with the curves $T_o(\hat{c}_1)$, $T_d(\hat{c}_1)$, ℓ^u, ℓ^s, $\mathcal{F}(\ell^u)$, and $\mathcal{F}^{-1}(\ell^s)$. Recall from (2.18) that $T_o(\hat{c}_1)$ and $T_d(\hat{c}_1)$ are given by

$$\hat{p} = 2\hat{c}_1 \frac{g(u)}{g'(u)} \pm \frac{3f(u)}{g^2(u)}.$$

Hence, for $\hat{c}_1 > 0$, $T_o(\hat{c}_1)$ lies above $T_o(0)$ on $u < \alpha$, whereas it lies below $T_o(0)$ on $u > \alpha$

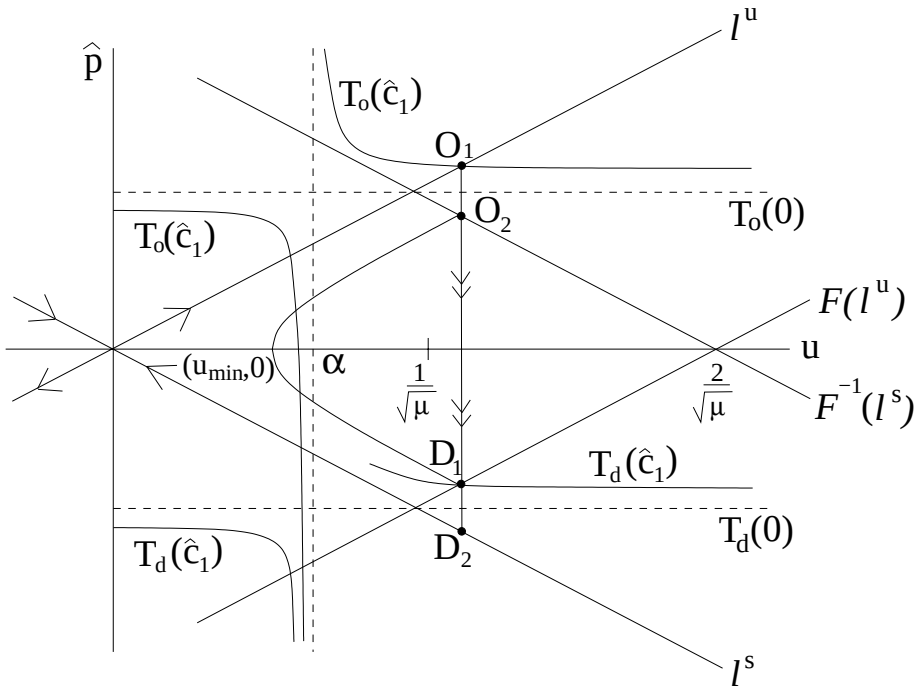


Figure 6.2. *The geometry on the invariant manifold $\mathcal{M}$ for case* I *studied in section* 6.2. *For definiteness, we have based some elements of the sketch on properties of the example in section* 6.3.

because of the assumptions (4.3) and $g''(\alpha) < 0$. Similarly, $T_d(\hat{c}_1)$ lies above $T_d(0)$ on $u < \alpha$ and below it on $u > \alpha$; see Figure 6.1. Moreover, these results would be reversed if one instead had $\hat{c}_1 < 0$.

We observe that, since the intersection $\ell^u \cap T_o(0)$ consists of a single point by assumption (ii), the curve $\mathcal{F}^{-1}(\ell^s)$ must intersect these two curves also at the same point. Similarly, ℓ^s, $T_d(0)$, and $\mathcal{F}(\ell^u)$ intersect in a point, symmetrically disposed about the u-axis with respect to the intersection point $\ell^u \cap T_o(0) \cap \mathcal{F}^{-1}(\ell^s)$; see Figures 6.1 and 6.2.

As illustrated in Figure 6.1, the geometry of these curves is such that, for an interval of positive $\hat{c}_1$ values, the intersection $T_o(\hat{c}_1) \cap \ell^u$ consists of at least two points. There could be others in between. That there exist at least two follows from the fact that $T_o(\hat{c}_1)$ can be taken to be close to $T_o(0)$ over as large a portion of the interval $(0, \alpha)$ as one pleases by choosing $\hat{c}_1$ small enough. Hence the first intersection point, labeled O_1^A, exists and lies close to the intersection point of ℓ^u and $T_o(0)$ for small $\hat{c}_1$. We denote its u-coordinate by u_A^+. The existence of the second intersection point, labeled O_1^B, then follows directly from the observation that, for each $\hat{c}_1$, $T_o(\hat{c}_1)$ has a vertical asymptote at $u = \alpha$ since $g'(u)$ has a simple zero there. The u-coordinate of O_1^B, which we label u_B^+, lies close to α for $\hat{c}_1$ small. Finally, there is a maximum value of $\hat{c}_1$ such that, at the maximum, $T_o(\hat{c}_1)$ has only a tangency to ℓ^u and, for values of $\hat{c}_1$ larger than the maximum, $T_o(\hat{c}_1)$ no longer intersects ℓ^u.

The singular solution for the A-pulse consists of the following segments:

- left slow on $(-\infty, \Gamma_1(t))$: $(0,0)$ to O_1^A along ℓ^u;

- left fast at $x = \Gamma_1(t)$: O_1^A to D_1^A;
- middle slow on $(\Gamma_1(t), 0] \cup [0, \Gamma_2(t))$: $D_1^A \to (u_{\min}, 0) \to O_2^A$;
- right fast at $x = \Gamma_2(t)$: O_2^A to D_2^A;
- right slow on $(\Gamma_2(t), \infty)$: D_2^A to $(0,0)$ along ℓ^s.

The singular B-pulse is the same type of solution, except with the superscripts A replaced by B.

Next, we turn to the ODE for the pulse separation distance $\Delta\Gamma(t) = \Gamma_2(t) - \Gamma_1(t)$. We recall from (4.11) that $\Delta\Gamma(t)$ satisfies

$$\frac{d}{dt}\Delta\Gamma = -3\varepsilon^3 \frac{f(u^+)g'(u^+)}{g^3(u^+)} e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}. \tag{6.2}$$

This ODE has a unique, nontrivial, fixed point at $u^+ = \alpha$ for all values of α in case II because α is the unique zero of $g'(u)$ by assumption (i). This value of u^+ corresponds to a symmetric, stationary, two-pulse solution for which the pulse separation distance is

$$\Delta\Gamma_{\text{stat}} = \frac{-1}{\varepsilon\sqrt{\mu}} \ln\left(\sqrt{\mu}\mathcal{G}(\alpha) - 1\right), \tag{6.3}$$

which is a strictly $\mathcal{O}(1/\varepsilon)$ quantity; see the definition (4.13) of $\mathcal{G}$, and see also (4.12). This fixed point is a two-circuit homoclinic solution of the stationary problem associated to the PDE (1.1), i.e., of the ODE (2.3) with $\hat{c} = 0$; see also Remark 6.2 below. It corresponds to a stable fixed point of (6.2) because we have assumed that $g'(u) > 0$ on $u < \alpha$ (and to an unstable point in the opposite case).

For the A-pulse, the ODE (6.2) with $u^+ = u_A^+$ governs the pulse separation distance, and for the B-pulse, the ODE is also (6.2) but with $u^+ = u_B^+$. Both solutions are attracted to the symmetric, homoclinic, two-pulse fixed point. However, the A- and B-type pulses are not necessarily different for all time: during the evolution, a pair of A-pulses may change into a pair of B-pulses. In other words, a pair of slowly varying pulses might start as being of type A, and then, after a certain time, when $\Delta\Gamma$ has decreased below a critical value or, equivalently, when $|\hat{c}|$ has reached a certain maximal value, the pulses become of type B. See section 6.3 for an explicit example.

Remark 6.2. The existence of the symmetric, stationary, two-pulse solution does not follow directly from the results of sections 2 and 3. Nevertheless, these solutions can be rigorously constructed in all three cases. A crucial observation is that (2.12) does not directly imply (2.13) but a slightly different version, namely, $\hat{p}g'(u) = \hat{c}g(u)$. Thus the intersection of $W^U(\mathcal{M})$ and $W^S(\mathcal{M})$ occurs for $\hat{c} = 0$ either at $\hat{p} = 0$ or at $u = \alpha$. The new two-circuit homoclinic orbit can now be constructed using the new $u = \alpha$ branch of $W^U(\mathcal{M}) \cap W^S(\mathcal{M})$. These homoclinic orbits are truly pulses of semistrong type, with the U-component slowly varying in between the pulses. This contrasts with the multipulse homoclinic orbits constructed in [9] for the Gray–Scott problem and in [7] for the more general class of problems of the type (1.1), for which the pulses are so close together that the U-component remains constant to leading order between them.

6.2. Cases I and III: Repelling or attracting two-pulse solutions. In case I, there exists a stationary, homoclinic, two-pulse solution and one type of symmetric, slowly varying, two-pulse solution in which the pulses repel, i.e., $\hat{c}_1 < 0$ and $\hat{c}_2 = -\hat{c}_1 > 0$. For $\hat{c}_1 < 0$, $T_o(\hat{c}_1)$ lies

below $T_o(0)$ on $u < \alpha$ and above it on $u > \alpha$ due to the assumption that $g''(\alpha) < 0$. Similarly, $T_d(\hat{c}_1)$ lies below $T_d(0)$ on $u < \alpha$ and above it on $u > \alpha$. Moreover, much is known for small values of $\hat{c}_1$ about the locations of these curves relative to the curves $T_o(0)$ and $T_d(0)$, just as in case II above. The unique, singular, two-pulse solution in case I is of the same type as the singular solutions constructed in case II, and it is sketched in Figure 6.2. Finally, from the ODE (6.2), we see that the two slowly varying pulses repel each other since $\alpha < \pi_u(\ell^u \cap T_0(0))$ here and since $u^+ > \alpha$ implies $g'(u^+) < 0$ by the assumption made on $g''(\alpha)$ throughout this section. See Figure 6.2. Thus, although there is a stationary two-pulse solution and a slowly varying two-pulse solution, these solutions cannot approach each other as time evolves due to geometric obstructions.

Finally, in case III, singular, slowly varying, two-pulse solutions can be constructed along the same lines as in the above cases. The pulses attract when $g''(\alpha) < 0$, which is the assumption made throughout this section, and they repel in the opposite case when $g''(\alpha) > 0$. Moreover, in the attracting case, there is an α_{crit} such that, if $\alpha > \alpha_{\text{crit}}$, then the pulses accelerate toward each other, whereas, if $\alpha < \alpha_{\text{crit}}$, then the rate of approach decreases. For general systems, α_{crit} is determined by requiring that the tangency between ℓ^u and $T_o(\hat{c}_1)$ occurs on the boundary at $u = \pi_u(\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s))$. See also the example in section 6.3.

Remark 6.3. In case I, there are other intersection points $\ell^u \cap T_0(\hat{c}_1)$; however, these do not lead to singular, slowly varying, two-pulse solutions. For example, as shown in Figure 6.2 with $\hat{c}_1 < 0$, a fast jump that starts at a takeoff point $\ell^u \cap T_0(\hat{c}_1)$ to the left of $u = \alpha$ lands on $\mathcal{F}(\ell^u)$ at a point below ℓ^s. Hence, after it lands, the singular solution is brought to the negative $\hat{p}$-axis by the slow flow, and it is not possible to complete a singular two-pulse solution of the type we study. Other possible intersection points share the same type of difficulty.

6.3. A fourth example. We briefly illustrate the results of this section on problems (1.1) for which $g(U) = e^{\gamma(U)}$, with $\gamma'(U) = 2(\alpha - U)/\beta$ and $\alpha, \beta > 0$, and for which $f(U) = g^2(U)/3$. Hence we model a general function g for which $g'(u)$ has one simple zero (at α), but we restrict the choice of f for the purpose of illustration, as was also done in the example of section 5.1. This example exhibits some interesting dynamics and illustrates various characteristics of the dynamics in cases I–III.

The important curves on $\mathcal{M}$ are branches of hyperbolas and lines, respectively,

$$\begin{aligned} T_o(\hat{c}), T_d(\hat{c}) &: \qquad \hat{p} = \frac{\hat{c}\beta}{\alpha - u} \pm 1, \\ \mathcal{F}(\ell^u), \mathcal{F}^{-1}(\ell^s) &: \qquad \hat{p} = \pm\sqrt{\mu}u \mp 2, \end{aligned}$$

so that $\ell^u \cap T_0(0)$ occurs at $u = 1/\sqrt{\mu}$ and $\mathcal{F}(\ell^u) \cap \mathcal{F}^{-1}(\ell^s)$ occurs at $u = 2/\sqrt{\mu}$. Moreover, the ODE (6.2) becomes

$$\frac{d\Delta\Gamma}{dt} = \frac{-2\varepsilon^3}{\sqrt{\mu}\beta}[\alpha\sqrt{\mu} - 1 - e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}]e^{-\varepsilon\sqrt{\mu}\Delta\Gamma}. \tag{6.4}$$

Note that the stationary, homoclinic, two-pulse solution corresponds to the fixed point of this ODE that is determined by $\Delta\Gamma_{\text{stat}} = (-1/\varepsilon\sqrt{\mu})\ln(\alpha\sqrt{\mu} - 1)$.

Case II corresponds to $\alpha \in (1/\sqrt{\mu}, 2/\sqrt{\mu})$. In this case, the pulses in a slowly varying two-pulse solution attract each other. The speed of attraction varies. From (4.5), we find

$$\frac{d^2}{dt^2}\Delta\Gamma = -2\varepsilon^3 \frac{d\hat{c}_1}{dt}.$$

Then differentiation of the ODE (6.4) yields

$$\frac{d\hat{c}_1}{dt} = \frac{\varepsilon}{\beta}[2e^{-\varepsilon\sqrt{\mu}\Delta\Gamma} - \alpha\sqrt{\mu} + 1]e^{\varepsilon\sqrt{\mu}\Delta\Gamma}\left(\frac{d}{dt}\Delta\Gamma\right). \tag{6.5}$$

Therefore, since $\frac{d}{dt}\Delta\Gamma < 0$ in case II and since $\beta > 0$ by assumption, there exists a critical separation distance,

$$\Delta\Gamma_\star = \frac{-1}{\varepsilon\sqrt{\mu}} \ln\left(\frac{1}{2}(\alpha\sqrt{\mu} - 1)\right), \tag{6.6}$$

such that $\frac{d}{dt}\hat{c}_1 = 0$ for $\Delta\Gamma = \Delta\Gamma_\star$. In addition, $\hat{c}_1(t)$ increases for $\Delta\Gamma > \Delta\Gamma_\star$, while $\hat{c}_1(t)$ decreases for $\Delta\Gamma_{\text{stat}} < \Delta\Gamma < \Delta\Gamma_\star$.

For the A-pulses, the pulses accelerate toward each other; i.e., $\hat{c}_1(t)$ increases, while, for the B-pulses, the rate of approach decreases in time. Moreover, there can be a dynamic change-over from type-A pulses to type-B pulses. This occurs, for example, with pulse pairs for which $\Delta\Gamma(0)$ exceeds $\Delta\Gamma_\star$. The pulses are first of the A type and accelerate toward each other. Then, when the separation distance has decreased to $\Delta\Gamma_\star$, the pulses change to being of type B, and their speeds decrease as the two-pulse solution approaches the stationary two-pulse solution.

In case III, the critical value of α at which the switch-over from deceleration to acceleration occurs is $\alpha_{\text{crit}} = 3/\sqrt{\mu}$. This may be determined as follows. We recall that $\mathcal{F}(\ell^u)$ and $\mathcal{F}^{-1}(\ell^s)$ intersect at $u = 2/\sqrt{\mu}$. Hence, by examining the intersection of $T_o(\hat{c})$ and ℓ^u at that u value, we find that the critical wave speed is

$$\hat{c}_{\text{crit}} = \frac{\alpha - \frac{2}{\sqrt{\mu}}}{\beta}.$$

Moreover, at this same value of u, the above intersection is a quadratic tangency, and hence, by equating the derivatives with respect to μ of the expressions for $T_o(\hat{c}_{\text{crit}})$ and ℓ^u, we find the above value of α_{crit}.

7. Discussion.

In this section, we discuss the stability of the pulse patterns and the validity of the asymptotic construction. Also, we discuss several extensions of the theory developed in this article, first to the larger class of systems (1.2) and then to a number of different types of N-pulse solutions for $N \geq 3$.

7.1. Stability and validity.

In this article, we considered the existence problem for interacting pulses; i.e., we have developed a method by which the dynamics of semistrong pulses can be described. We now briefly discuss their stability and the closely related issue of the validity of the asymptotic construction.

We begin with the stationary pulses and the pulses that are stationary in a comoving frame. A rigorous result on the existence of the homoclinic orbits associated to the intersections

$\ell^u \cap T_o(0)$ follows by geometric singular perturbation theory. It is a straightforward application of the methods developed in [7, 9]. (See also [8] for much more exotic singular patterns constructed by these methods.) Thus the validity of the asymptotic construction can be established also in this case. The validity of the more complex (stationary and traveling) multipulse patterns of sections 6 and 7, which are associated to zeros of g', can also be obtained along these lines (although here the procedure might be a little less straightforward; see also Remark 6.2).

In [5, 6, 7], a method has been developed by which the stability of stationary-pulse solutions of two-component, singularly perturbed, reaction-diffusion equations can be studied in full analytical detail. This method, the so-called nonlocal eigenvalue problem (NLEP) method, extends the Evans function approach of [1] to singularly perturbed systems of the same type as (1.1) and (1.2). A crucial ingredient of the NLEP method is the reduction of the eigenvalue problem, which is a coupled system of two second order equations, associated to the stability of a pulse to a single, second order, inhomogeneous equation of Sturm–Liouville type that has a nonlocal term.

The NLEP method can be applied directly to the stability problem for a stationary, homoclinic, one-pulse solution $(U_h(\xi), V_h(\xi))$ of (1.1). We give a brief sketch of the procedure and refer to [6, 7] for the details. In order to study the spectral stability of $(U_h(\xi), V_h(\xi))$, we set

$$(U(x,t), V(x,t)) = (U_h(\xi) + u(\xi)e^{\lambda t}, V_h(\xi) + v(\xi)e^{\lambda t}),$$

substitute this into (1.1), and linearize. As stated above, the coupled second order equations for u and v are then reduced to the corresponding NLEP problem,

$$(\mathcal{L}(\xi) - \lambda)v \equiv v_{\xi\xi} + [2g(U_h)V_h - (1+\lambda)]v = -V_h^2 g'(U_h) \tag{7.1}$$

with $v(\xi)$ and λ such that

$$t_2(\lambda) = 1 - \frac{1}{2\sqrt{\mu+\lambda}} \int_{-\infty}^{\infty} [f'(U_h)V_h^2 + 2f(U_h)V_h v] d\xi = 0. \tag{7.2}$$

Here, the expression for t_2 arises from an analysis of the u equation in both the fast and the slow limits. The function $t_2(\lambda)$ is the so-called slow transmission function; it is a factor in the Evans function $\mathcal{D}$ associated to the stability problem [6, 7].

For almost all λ (see [6, 7]), there exist bounded solutions to (7.1), and then the eigenvalues are selected by applying the additional constraint (7.2). So, it is convenient to express the system—equation plus constraint—as a single, nonlocal equation,

$$w_{\xi\xi} + [2g(U_h)V_h - (1+\lambda)]w = 2V_h^2 g'(U_h) \frac{\int_{-\infty}^{\infty} f(U_h)V_h w d\xi}{\int_{-\infty}^{\infty} f'(U_h)V_h^2 d\xi - 2\sqrt{\mu+\lambda}} \tag{7.3}$$

[5, 6, 7]. We briefly consider the NLEP (7.1), (7.3) here to indicate that the sign of g' plays a significant role in the stability analysis. The full stability problem is in itself the subject of an independent investigation.

First, (7.1) strongly suggests that the two- (and N-) pulse solutions that can be constructed when g' has zeros—see section 6—can be expected to be unstable. It follows from the character

of the pulse solutions that the homogeneous Sturm–Liouville problem $\mathcal{L}v = \lambda v$ associated to (7.1) must have an unstable eigenvalue $\lambda_0^{\text{fast}} > 0$ [7]. In general, this eigenvalue can be stabilized by the coupling of (7.1) to the slow nonlocal equation (7.2) [6, 7]. However, this coupling does not exist when $g'(U_h) = 0$, which is the case for the stationary two-pulse solutions. Thus it is to be expected that λ_0^{fast} is (to leading order) also an unstable eigenvalue of the full problem. Of course, this statement is far from rigorous; it is a fully heuristic argument that neglects several possible additional stabilizing effects. Nevertheless, the argument shows that the existence and stability analysis are intertwined, a feature that is common in the analysis of localized solutions.

Second, a preliminary analysis of (7.1) and (7.2) indicates that there is a regime in parameter space in which there are stable homoclinic pulses when $g' < 0$ (i.e., in the case in which the two-pulse solutions consist of repelling pulses). Here, the underlying mechanism is more subtle. The above-mentioned unstable eigenvalue λ_0^{fast} corresponds to an eigenvalue $\lambda_0(\mu)$ of the full equation in the sense that $\lim_{\mu\to 0} \lambda_0(\mu) = \lambda_0^{\text{fast}}$ [7]. In the Gray–Scott model studied in [6] and in the generalized Gierer–Meinhart equations studied in [7], this eigenvalue merges, as μ increases, with a second positive eigenvalue $\lambda_1(\mu)$ to form a complex conjugate pair of eigenvalues. This pair crosses the imaginary axis in the λ plane at a critical Hopf-bifurcation value of μ that stabilizes the pulse.

On the contrary, this same preliminary analysis suggests that the homoclinic pulses in systems for which $g' > 0$—i.e., in the case of attracting two-pulse solutions—will be unstable in general. The above-mentioned second positive eigenvalue $\lambda_1(\mu)$ cannot exist when $g' > 0$. Hence the pulse cannot be stabilized (at least not by the mechanism studied in [6, 7]).

It is clear that, on the one hand, the stability problem is far from being solved and, on the other hand, that the methods of [5, 6, 7] are strong enough to be applied here. However, these methods are developed in the context of stationary pulses. The slowly varying pulses constructed in this article are not of that type. Therefore, one has to extend the methods of [5, 6, 7] in order to be able to consider the interacting pulses. This is the essential next step in the analysis of semistrong pulse interactions. A first (and formal) result in that direction has been obtained in [3, 4]. Using the ideas of [5, 6], it was possible to derive slowly varying eigenvalues that governed the short time stability of the slowly varying pulses in the Gray–Scott equation. Again, the situation differs in an important aspect from that of the weak interaction case since the eigenvalues will undergo $\mathcal{O}(1)$ changes during the time evolution of the system.

Preliminary analysis indicates—under some additional conditions—that repelling pulse pairs can be stable (through the same stability mechanism that stabilizes the homoclinic pulses). Furthermore, we note that, in each of the systems in which $g'(u) < 0$ that we have simulated numerically, the numerical simulations also suggest that repelling-type two-pulse solutions can be stable on the time scales on which they exist. However, as we have seen in the example of section 5.3, the changes in the eigenvalues might be able to trigger bifurcations/destabilizations that are not present in the stationary (or weak interaction) limit. In addition, since the homoclinic pulses are not likely to be stable for $g' > 0$, we do not expect stable pairs of attracting pulses.

Finally, we come to the question of the validity of the asymptotic construction of the slowly varying two-pulse solutions. As explained in section 3.2, we can still evoke the concepts

of geometrical singular perturbation theory to conclude that the singular pulses do persist. However, the nonrigorous part of the construction is not during the fast jumps but in the point of symmetry between the pulses where the solution to the left is "glued" to its symmetrical counterpart on the right. There, we describe the solutions by ODEs (2.3) that have different values of $\hat{c}$, so the connection cannot be smooth or, in other words, cannot be a solution of the PDE—see section 3.1. As explained in Remark 3.1, it can be shown that this problem can be solved in a consistent way at the higher order level—see also [4]—but such a consistent higher order expansion of course does not settle the validity of the construction.

In the case of weak interactions, the validity question is tackled either by a center manifold approach [12, 14] or by a renormalization approach [42]. Both methods exploit the fact that the associated homoclinic pulses are stable/attracting. These methods cannot be used directly in the semistrong case. However, it is likely that it is possible to use either of these approaches in a modified fashion to settle the validity in the semistrong case for the stable two-pulse solutions. To do so, one has to have full control on the (slowly varying) eigenvalues of these two-pulse solutions. Thus the validity issue is closely intertwined with the stability question.

We conclude that the stability and validity analysis of the pulse solutions constructed in this article is a significant problem. It is the subject of work in progress.

Remark 7.1. There is an alternative approach to the stability analysis of stationary one-pulse solutions of a certain, slightly more restrictive version of the Gierer–Meinhardt equation; see [23, 33, 48]. As in [5, 6, 7], the stability problem is reduced to a system of the same type as (7.3). However, the corresponding equation is of a different nature.

7.2. A more general class of PDEs. In the preceding sections, we focused on the class of PDEs (1.1) that includes many named models. However, examination of the techniques employed in sections 2, 3, 4, 5, and 6 reveals that they can also be applied to the much broader class of coupled reaction-diffusion equations (1.2), which, we recall, are

$$\left\{ \begin{array}{rcl} \varepsilon^{2+\sigma} U_t &=& U_{xx} + \varepsilon^2 F_1(U) + V F_2(U,V), \\ V_t &=& \varepsilon^2 V_{xx} + G(U,V). \end{array} \right. \tag{7.4}$$

Here, $\sigma \geq 0$, $0 < \varepsilon \ll 1$, and the functions F_1, F_2, and G are smooth for $U > 0$ and $V \geq 0$ and admit at most poles at $U = 0$. They also satisfy additional assumptions. We assume that $(U_0, V_0) = (0,0)$ is a linearly stable, homogeneous steady state of (1.2) and use the stretched variable $\xi = \frac{x - \Gamma(t)}{\varepsilon}$ in the comoving frame. Finally, we assume that the traveling wave ODE has a normally hyperbolic invariant manifold $\mathcal{M}$ for $\varepsilon = 0$ and that, on the persistent slow manifold, which may coincide with the $\varepsilon = 0$ manifold as for systems (1.1) or which may be $\mathcal{O}(\varepsilon)$ close to the plane $\{V = 0, V_\xi = 0\}$, the flow is superslow, i.e., slow in the slow/outer comoving variable $x - \Gamma(t)$.

The first generalization is to systems of the form (1.2) for which $F_1(U) = -\mu U$, $F_2(U,V) = h_1 U^{\alpha_1} V^{\beta_1}$, and $G(U,V) = -V + h_2 U^{\alpha_2} V^{\beta_2}$ with $\mu > 0$, $h_1 > 0$, and $h_2 > 0$, namely, to

$$\begin{array}{rcl} \varepsilon^2 U_t &=& U_{xx} - \varepsilon^2 \mu U + h_1 U^{\alpha_1} V^{\beta_1}, \\ V_t &=& \varepsilon^2 V_{xx} - V + h_2 U^{\alpha_2} V^{\beta_2}. \end{array} \tag{7.5}$$

We require that $\alpha_1 > 1 + \frac{\alpha_2 \beta_1}{\beta_2 - 1}$, $\alpha_2 < 0$, $\beta_1 > 1$, and $\beta_2 > 1$. This class of equations includes the generalized Gierer–Meinhardt equations; see [7, 21, 23, 33]. In [7], we showed the existence

and analyzed the stability of stationary, homoclinic, one-pulse solutions for systems (7.5) and also of multipulse homoclinic solutions consisting of a finite number of rapid pulses in quick succession; see Remark 6.2. In addition, in [8] we established the existence of stationary spatially periodic solutions, which are known to be attractors in the self-replication region, and we showed that on a sufficiently large interval one can construct a solution for any arbitrary prespecified pulse sequence.

While the flow on $\mathcal{M}$ is still linear due to the choice of F_1, the functions F_2 and G are more general than those in (1.1). The function $G = -V + h_2 U^{\alpha_2} V^{\beta_2}$ is a more general choice for which the reduced fast system,

$$0 = v_{\xi\xi} + G(\bar{u}, v), \tag{7.6}$$

has a homoclinic orbit to the saddle point $(0,0)$, and the choice of F_2 reflects the competing goals of wanting to have the most general possible form for F_2 but at the same time retain the necessary property that the fast jump, Δp, bridges the correct takeoff and touchdown points on $\mathcal{M}$.

Two-pulse solutions of the different types constructed in sections 3, 4, 5, and 6 (and N-pulse solutions of the types studied below in section 7.3) can be constructed for systems (7.5) satisfying the above conditions following the same technique used above. Moreover, the system (7.5) is itself a normal form in the study of large-amplitude pulse solutions in coupled reaction-diffusion equations. See (1.1) and (1.2) in [7], and we refer the reader to the introduction and the appendix there for further details.

There is also a second level of generalization that we can make, and we conclude this section by briefly sketching the necessary ingredients. First we consider how general the choice of the function G can be. In particular, we look at all possible G for which the reduced fast system (7.6) supports a homoclinic orbit to the saddle fixed point $(v = 0, v_\xi = 0)$. Let $\bar{U}$ denote the value of the component U at the center of the pulse. This entails assuming that, for every $\bar{U} > 0$, $G(\bar{U}, 0) = 0$, $(\partial G/\partial V)(\bar{U}, 0) < 0$, $G(\bar{U}, V_1) = 0$ for some $V_1 > 0$, $(\partial G/\partial V)(\bar{U}, V_1) > 0$, and $G(\bar{U}, V) < 0$ for all $V \in (0, V_1)$. Second, we observe that there is a wide class of functions F_1 to which one can extend the results of this article. The leading order flow on the persistent slow manifold is given by

$$u_{xx} + \varepsilon^2 F_1(u) = 0, \tag{7.7}$$

which is automatically superslow. So, now, instead of choosing F_1 to be a linear function, as was done for both (1.1) and (7.5), it can be chosen to be any (nonlinear) function for which the fixed point of (7.7), which corresponds to the asymptotic state of the pulses, is a saddle fixed point, and for which the various takeoff and touchdown points on the persistent slow manifold can be connected via the jump, Δp determined by F_2, in the fast field. The requirement that the critical point $u, u_x = 0$, on the slow manifold is of saddle type is directly related to the (natural) assumption that the trivial background state $(U, V) \equiv (0, 0)$ is stable as a solution of the PDE (7.4). Examples of such nonlinear slow systems include planar Hamiltonian systems, such as the classical pendulum, the Duffing equation, and escape oscillators.

Remark 7.2. For the problem of resonance bands in Hamiltonian systems, one-pulse and multipulse (with alternating fast-slow segments) homoclinic orbits have been constructed in

[25]. There, the slow flow was given to leading order by the nonlinear pendulum with constant torque, adiabatic Melnikov theory was used to calculate the fast jumps, and geometric singular perturbation theory was used to assemble (and rigorously prove the existence of) the homoclinic orbits.

7.3. Beyond symmetric two-pulse solutions. In this last section, we give some straightforward extensions of the two-pulse results of the previous sections to establish the existence of stationary N-pulse solutions, traveling N-pulse solutions, and slowly varying N-pulse solutions for $N > 2$. These solutions are biasymptotic to $(U = 0, V = 0)$, just as was the case for $N = 2$, and they consist of N localized (fast) pulses in alternation with $N + 1$ slow segments. We sketch only the geometrical backbones of the constructions.

In order to construct some stationary N-pulse solutions, we focus on those PDEs (1.1) for which $g'(U)$ has one simple zero, again at $U = \alpha$, and for which $T_o(0)$ and ℓ^u intersect transversely in one point. Extensions to problems in which $g'(U)$ has more than one simple zero and/or in which $T_o(0)$ and ℓ^u intersect transversely in more than one point follow readily.

Singular, stationary, N-pulse solutions consist of the following segments, where $\hat{c} \equiv 0$ in the governing ODEs:

- left slow segment along ℓ^u from $(0, 0)$ to the takeoff point $O_1 \equiv \ell^u \cap \{u = \alpha\}$;
- left fast pulse from O_1 to the touchdown point $D_1 \equiv \mathcal{F}(\ell^u) \cap \{u = \alpha\}$;
- second slow segment along hyperbolic cosine orbit segment γ on $\mathcal{M}$ from D_1 to $O_2 \equiv \gamma \cap T_o(0)$;
- second fast pulse from O_2 to $D_2 \equiv \gamma \cap T_d(0)$;
- third through $(N-1)$th slow segments along γ from D_2 back to O_2, interspersed with the third through $(N-1)$th fast pulses, each of which connects O_2 to D_2;
- Nth slow segment along γ from D_2 to $O_N \equiv \mathcal{F}^{-1}(\ell^s) \cap \{u = \alpha\}$;
- Nth fast pulse from O_N to $D_N \equiv \ell^s \cap \{u = \alpha\}$;
- $(N+1)$th slow segment along ℓ^s from D_N to $(0, 0)$.

See Figure 7.1(a) for an illustration with $N = 5$. Moreover, the same type of geometric singular perturbation theory arguments (see also Remark 6.2) can be used here to show that there exists a true N-pulse solution for $0 < \varepsilon \ll 1$ near a singular ($\varepsilon = 0$) solution of the type just constructed. If we assume that (4.3) holds and that $\alpha > \pi_u(\ell^u \cap T_0(0))$, then the maximum values of u are highest at the tops of pulses $2, \ldots, N-1$ and slightly lower at the maxima of the first and Nth pulses; see Figure 7.1(a). Moreover, for each pulse in an N-pulse solution, the maximum of U exceeds that of the $(N = 1)$-pulse solution.

Next, we show how to construct N-pulse solutions that travel with constant wave speed, $\hat{c}(t) \equiv \hat{c}_N$. We do this for PDEs of the form (1.1) for which $g'(U)$ has one simple zero (again, at $U = \alpha$) but now for which $T_o(\hat{c}_1)$ and ℓ^u intersect transversely in at least two distinct points, one with U-coordinate less than α and the other with U-coordinate greater than α; see Figure 7.1(b). Here also, extensions are possible.

The ODE governing N-pulse classical traveling waves is (2.3) with $\hat{c}(t) \equiv \hat{c}_N$. We describe the singular solution for a traveling N-pulse solution with $N = 2$, which consists of the following segments:

- left slow segment along ℓ^u from $(0, 0)$ to the first takeoff point $O_1 \equiv \ell^u \cap T_0(\hat{c}_N)$, where $\pi_u(O_1) < \alpha$;

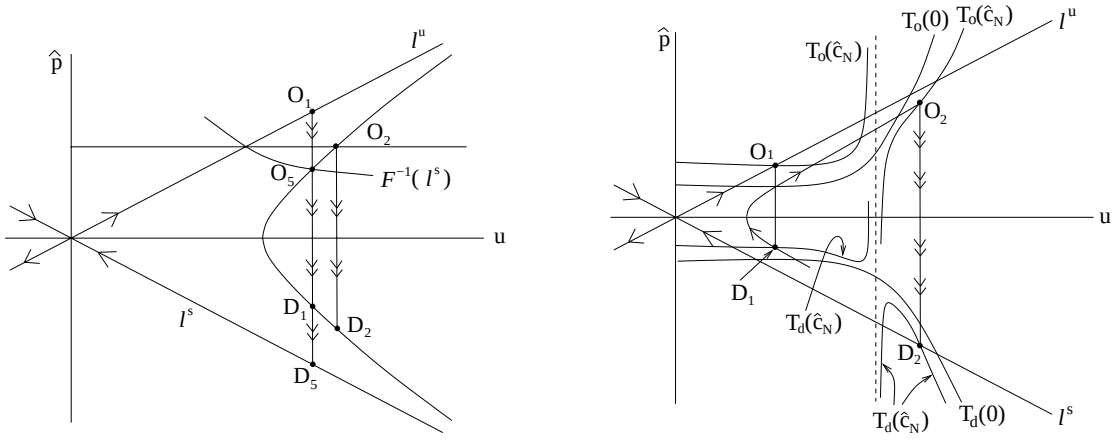


Figure 7.1. (a) *Sketch of a singular, stationary, N-pulse solution, with $N = 5$. The points O_3 and O_4 coincide with O_2, and the points D_3 and D_4 coincide with D_2.* (b) *Sketch of a singular, traveling, N-pulse solution, with $N = 2$.*

- left fast pulse from O_1 to the touchdown point $D_1 \equiv T_d(\hat{c}_N) \cap \{u = \pi_u(O_1)\}$;
- middle slow segment along a hyperbolic cosine orbit segment γ on $\mathcal{M}$ from D_1 to $O_2 \equiv \gamma \cap T_o(\hat{c}_N)$, where $\pi_u(O_2) > \alpha$;
- second fast pulse from O_2 to $D_2 \equiv \ell^s \cap T_d(\hat{c}_N)$;
- right slow segment along ℓ^s from D_2 back to $(0, 0)$.

See Figure 7.1(b). The generalization to $N > 2$ follows the same procedure. The wave speed $\hat{c}_N$ is selected by the relative positions of the various intersection points. (A small change in $\hat{c}_2$ in the construction in Figure 7.1(b) will remove the final touchdown point D_2 from ℓ^s.) Moreover, due to symmetries, if there exists a traveling N-pulse solution with wave speed $\hat{c}_N$, then there also exists a symmetrically disposed N-pulse traveling wave with speed $-\hat{c}_N$. Note that there are of course several conditions like (4.1) that have to be satisfied in order to be able to construct (stationary or traveling) multipulse solutions in the general PDE (1.1).

Finally, one can construct slowly varying N-pulse solutions by mimicking the results of section 4. For these solutions, the outermost pulses, i.e., the first and Nth pulses, move away from each other when $g'(U) < 0$ for all $u > 0$, and they move toward each other when $g'(U) > 0$ for all $u > 0$. Such an analysis would also be useful to further answer the questions of whether pulse-splittings are of the type 2^n or of the type for which just the edge pulses split. See [15] for the analysis of this question in the context of weak interactions.

Appendix. Coupled reaction-diffusion equations in general form. In this appendix, we briefly examine a number of named systems of coupled reaction-diffusion equations. We rescale the equations into a natural form associated to the singular character of the pulse solutions exhibited by these systems. We find that these natural forms are given by (1.1) or (1.2). Thus the methods developed in this article can be applied directly to any of these equations. Specifically, we examine the (classical and generalized) Gierer–Meinhardt equations, the Gray–Scott model, the Schnakenberg system, and the Thomas equations. The scaling analysis needed for the first two of these models has been derived earlier in [3, 4, 7, 10].

We begin with the classical Gierer–Meinhardt equations,

$$\begin{aligned} U_t &= U_{xx} - \mu U + V^2, \\ V_t &= DV_{xx} - V + \frac{V^2}{U}, \end{aligned} \tag{A.1}$$

where $\mu > 0$ and D is the ratio of the diffusivities, taken to be less than one following the usual Turing analysis. The scalings appropriate for large-amplitude pulse solutions are shown in [7, 10] to be

$$U = \frac{1}{\sqrt{D}}\hat{U}, \qquad V = \frac{1}{\sqrt{D}}\hat{V}, \qquad x = D^{1/4}\hat{x}.$$

Note that these scalings represent the fact that the amplitudes of both the U- and the V-components of the pulse solutions of (A.1) are $\mathcal{O}(1/\sqrt{D})$. The amplitudes of $\hat{U}$ and $\hat{V}$ will be $\mathcal{O}(1)$ in the natural scaled form (that will be of type (1.1)).

In terms of these scaled variables, the PDE (A.1) becomes

$$\begin{aligned} \sqrt{D}\hat{U}_t &= \hat{U}_{\hat{x}\hat{x}} - \mu\sqrt{D}\hat{U} + \hat{V}^2, \\ \hat{V}_t &= \sqrt{D}\hat{V}_{\hat{x}\hat{x}} - \hat{V} + \frac{\hat{V}^2}{\hat{U}}. \end{aligned}$$

Hence it is natural to introduce the parameter ε via

$$\varepsilon^2 = \sqrt{D}.$$

The scaled Gierer–Meinhardt equations therefore may be written in the normal form of the type (1.1), namely, as

$$\begin{aligned} \varepsilon^2\hat{U}_t &= \hat{U}_{\hat{x}\hat{x}} - \varepsilon^2\mu\hat{U} + \hat{V}^2, \\ \hat{V}_t &= \varepsilon^2\hat{V}_{\hat{x}\hat{x}} - \hat{V} + \frac{\hat{V}^2}{\hat{U}}. \end{aligned}$$

It was shown in [7] that stationary singular pulses exist for $\varepsilon^2 = \sqrt{D} \ll 1$ and that there is a Hopf bifurcation at μ_{Hopf} (≈ 0.36 to leading order) in the regime $\mu = \mathcal{O}(1)$ such that the pulses are stable for $\mu > \mu_{\text{Hopf}}$ and unstable for $\mu < \mu_{\text{Hopf}}$. In addition, it was shown in [10] that pulse splitting occurs in the regime in which $\mu = \mathcal{O}(1/\varepsilon^4)$. Finally, we remark that a similar analysis can be done for the generalized Gierer–Meinhardt equations, as shown in [7], provided that the conditions on the powers stated in [7] are satisfied.

Next, we turn our attention to singular-pulse solutions in the Gray–Scott model,

$$\begin{aligned} U_t &= U_{xx} - UV^2 + A(1-U), \\ V_t &= DV_{xx} + UV^2 - BV, \end{aligned} \tag{A.2}$$

where $A, B, D > 0$. Then the scaling introduced in [3, 4] (see especially equation (1.3) in [4]) is

$$x = \sqrt{\frac{D}{B}}\xi, \qquad U = \frac{B\sqrt{B}\sqrt{D}}{\sqrt{A}}\hat{U}, \qquad V = \sqrt{\frac{A}{BD}}\hat{V}.$$

The scaling of x reflects the narrowness of the pulses, and hence ξ is the appropriate stretched variable with which to analyze the localized pulses. The scalings of U and V reflect their amplitudes.

Substituting these scalings into the system (A.2), we obtain

$$\frac{D}{B}\hat{U}_t = \hat{U}_{\xi\xi} - \frac{A}{B^2}\hat{U}\hat{V}^2 + \frac{A\sqrt{A}\sqrt{D}}{B^2\sqrt{B}}\left(1 - \frac{B\sqrt{B}\sqrt{D}}{\sqrt{A}}\hat{U}\right),$$
$$\frac{1}{B}\hat{V}_t = \hat{V}_{\xi\xi} + \hat{U}\hat{V}^2 - \hat{V}.$$

Now, in order to put this scaled Gray–Scott PDE into the form of the general systems (1.2) studied here, one needs to look at this system in terms of the long (or slow) spatial variable and a scaled time,

$$\hat{x} = \frac{\sqrt{A}}{B}\xi, \qquad \hat{t} = Bt.$$

Also, the natural parameter groups that emerge are

$$\varepsilon^2 = \frac{A}{B^2}, \qquad \delta = \sqrt{BD}.$$

Furthermore, we introduce the parameter σ by

$$D = \varepsilon^{4+\sigma}. \tag{A.3}$$

Hence the Gray–Scott model may be written as

$$\varepsilon^{2+\sigma}\hat{U}_{\hat{t}} = \hat{U}_{\hat{x}\hat{x}} - \varepsilon^2\left[\frac{\delta}{\varepsilon}\left(\frac{\delta}{\varepsilon}\hat{U} - 1\right)\right] - \hat{U}\hat{V}^2,$$
$$\hat{V}_{\hat{t}} = \varepsilon^2\hat{V}_{\hat{x}\hat{x}} - \hat{V} + \hat{U}\hat{V}^2,$$

which is of the form of (1.1) modulo a simple linear shift in U if $\delta/\varepsilon = \mathcal{O}(1)$ and $\sigma = 0$, i.e., $D = \mathcal{O}(\varepsilon^4) = \mathcal{O}(A^2/B^4)$; recall (A.3). These conditions on the parameter play central roles in the analysis of the Gray–Scott equation [3, 4, 5, 9]. For instance, $\sigma = 0$ marks the transition (by a Hopf bifurcation) from unstable pulse patterns ($\sigma < 0$) to stable pulse patterns ($\sigma > 0$); these patterns include periodic patterns [5] and slowly varying two-pulse solutions [3, 4]. Note that this also implies that the condition $B^4D \ll A^2$ or $\sigma > 0$, i.e., the case in which the scaled Gray–Scott model is of the form (1.2), is a natural condition on the parameters A, B, and D in the original, unscaled, Gray–Scott equation (A.2). We refer the reader to [3, 4] for a detailed analysis of the existence, bifurcation, splitting, and stability of two-pulse solutions of the Gray–Scott model for the different parameter combinations.

Third, we look at the Schnakenberg model,

$$U_t = U_{xx} + \gamma(b - UV^2),$$
$$V_t = DV_{xx} + \gamma(a - V + UV^2), \tag{A.4}$$

where $a, b, \gamma \geq 0$, and we set $\gamma = 1$ for convenience. The reaction kinetics are very similar to those of the Gray–Scott model, except that there is a feed term for the activator and there is

no linear decay term for the inhibitor concentration. (Analysis of the Gray–Scott model with an activator-feed term is presented in [22, 40], and the analysis of the slowly varying two-pulse solutions presented in [3, 4] can be extended to this case.) There is a homogeneous steady state at the point $(b/(a+b)^2, b+a)$. See section 6.4 of [32] for analysis of the phase plane. While more general results can be found for $a > 0$, we analyze the special case $a = 0$ here for expedience.

Setting $D = \varepsilon^2$, we scale the variables as

$$U = \varepsilon^{\alpha}\hat{U}, \qquad V = \varepsilon^{-\alpha}\hat{V}, \qquad x = \varepsilon^{\alpha}\xi,$$

with $\alpha > 0$, since the u-component is small at the pulse center, whereas the v-component diverges as $\varepsilon \to 0$. Restricting ourselves to the interval $\alpha \in [1/2, 1)$, we find that the model (A.4) may be written as an equation of the form (1.1), (1.2)

$$\begin{aligned} \hat{\varepsilon}^{2+\sigma}\hat{U}_t &= \hat{U}_{\xi\xi} + \hat{\varepsilon}^2\hat{b} - \hat{U}\hat{V}^2, \\ \hat{V}_t &= \hat{\varepsilon}^2\hat{V}_{\xi\xi} - \hat{V} + \hat{U}\hat{V}^2, \end{aligned}$$

where we have scaled $b = \varepsilon^{2-3\alpha}\hat{b}$ and introduced $\hat{\varepsilon} = \varepsilon^{(1-\alpha)}$ and $\sigma = 4\alpha - 2$.

Note, however, that there is a certain difference here. The plane $V = 0, V_\xi = 0$ is again invariant and normally hyperbolic, although now there are no fixed points on it. So, instead of looking for one-pulse solutions of the homoclinic type that are biasymptotic to an equilibrium on the plane, we look for spatially periodic solutions that consist of periodic arrays of localized near-homoclinic pulses interspersed with slow segments exponentially close to solutions of the slow system on the plane. This can be done using the same machinery developed in [5] and further employed in [30] to study the spatially periodic solutions that are the attractors in the pulse-splitting regime of the Gray–Scott model, and here such periodic solutions are found to exist for $\alpha \in [2/3, 1)$.

Finally, we observe that a similar type of scaling analysis can be carried out on the Thomas equations,

$$\begin{aligned} U_t &= U_{xx} + \gamma\left(\alpha(b-U) - H(U,V)\right), \\ V_t &= DV_{xx} + \gamma\left(a - V + H(U,V)\right), \end{aligned}$$

where $H(U,V) = \rho UV/(1+V+KV^2)$, and all parameters are nonnegative; see Chapter 5 of [32]. We do not go into the details.

Acknowledgment. The authors thank the referees for comments that helped improve the presentation.

REFERENCES

[1] J. Alexander, R. A. Gardner, and C. K. R. T. Jones, *A topological invariant arising in the stability of traveling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] J. G. Blom and P. A. Zegeling, *Algorithm* 731: *A moving-grid interface for systems of one-dimensional time-dependent partial differential equations*, ACM Trans. Math. Software, 20 (1994), pp. 194–214.

[3] A. Doelman, W. Eckhaus, and T. J. Kaper, *Slowly modulated two-pulse solutions in the Gray–Scott model* I: *Asymptotic construction and stability*, SIAM J. Appl. Math., 61 (2000), pp. 1080–1102.

[4] A. Doelman, W. Eckhaus, and T. J. Kaper, *Slowly modulated two-pulse solutions in the Gray–Scott model* II: *Geometric theory, bifurcations, and splitting dynamics*, SIAM J. Appl. Math., 61 (2001), pp. 2036–2062.
[5] A. Doelman, R. A. Gardner, and T. J. Kaper, *Stability analysis of singular patterns in the* 1*-D Gray–Scott model: A matched asymptotics approach*, Phys. D, 122 (1998), pp. 1–36.
[6] A. Doelman, R. A. Gardner, and T. J. Kaper, *A stability index analysis of* 1*-D patterns of the Gray–Scott model*, Mem. Amer. Math. Soc., 155 (2002).
[7] A. Doelman, R. A. Gardner, and T. J. Kaper, *Large stable pulse solutions in reaction-diffusion equations*, Indiana Univ. Math. J., 50 (2001), pp. 443–507.
[8] A. Doelman, T. J. Kaper, and H. van der Ploeg, *Spatially periodic and aperiodic multi-pulse patterns in the one-dimensional Gierer-Meinhardt equations*, Methods Appl. Anal., 8 (2001), pp. 387–414.
[9] A. Doelman, T. J. Kaper, and P. Zegeling, *Pattern formation in the one-dimensional Gray-Scott model*, Nonlinearity, 10 (1997), pp. 523–563.
[10] A. Doelman and H. van der Ploeg, *Homoclinic stripe patterns*, SIAM J. Applied Dynamical Systems, 1 (2002), pp. 65–104.
[11] W. Eckhaus, *Asymptotic Analysis of Singular Perturbations*, North–Holland, Amsterdam, 1979.
[12] S. Ei, *The motion of weakly interacting pulses in reaction-diffusion systems*, J. Dynam. Differential Equations, 14 (2002), pp. 85–137.
[13] S. Ei, M. Mimura, and M. Nagayama, *Pulse-pulse interactions in reaction-diffusion systems*, Phys. D, 165 (2002), pp. 176–198.
[14] S. J. Ei, Y. Nishiura, and B. Sandstede, *Pulse-Interaction Approach to Self-Replicating Dynamics in Reaction-Diffusion Systems*, in preparation.
[15] S.-I. Ei, Y. Nishiura, and K.-I. Ueda, 2^n*-splitting or edge-splitting? A manner of splitting in dissipative systems*, Japan J. Indust. Appl. Math., 18 (2001), pp. 181–205.
[16] I. R. Epstein and J. A. Pojman, *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos*, Oxford University Press, Oxford, UK, 1998.
[17] M. Escobedo and H. A. Levine, *Critical blowup and global existence numbers for a weakly coupled system of reaction-diffusion equations*, Arch. Ration. Mech. Anal., 129 (1995), pp. 47–100.
[18] N. Fenichel, *Geometrical singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[19] A. Friedman and B. McLeod, *Blow-up of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425–447.
[20] H. Fujita, *On the blowing up of solutions of the Cauchy problem* $u_t = \Delta u + u^{1+\alpha}$, J. Fac. Sci. Univ. Tokyo Sect. I, 13 (1966), pp. 109–124.
[21] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik, 12 (1972), pp. 30–39.
[22] P. Gray and S. K. Scott, *Autocatalytic reactions in the isothermal, continuous stirred tank reactor: Oscillations and instabilities in the system* $A + 2B \to 3B$, $B \to C$, Chem. Engrg. Sci., 39 (1984), pp. 1087–1097.
[23] D. Iron, M. J. Ward, and J. Wei, *The stability of spike solutions of the one-dimensional Gierer-Meinhardt model*, Phys. D, 150 (2000), pp. 25–62.
[24] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems (Montecatini Terme, 1994), Lecture Notes in Math. 1609, R. Johnson, ed., Springer-Verlag, Berlin, 1995, pp. 44–118.
[25] T. J. Kaper and G. Kovacic, *Multi-bump orbits homoclinic to resonance bands*, Trans. Amer. Math. Soc., 348 (1996), pp. 3835–3887.
[26] S. Kawaguchi and M. Mimura, *Collision of traveling waves in a reaction-diffusion system with global coupling effect*, SIAM J. Appl. Math., 59 (1999), pp. 920–941.
[27] H. A. Levine, *The role of critical exponents in blowup theorems*, SIAM Rev., 32 (1990), pp. 262–288.
[28] K.-J. Lin, W. D. McCormick, J. E. Pearson, and H. L. Swinney, *Experimental observation of self-replicating spots in a reaction-diffusion system*, Nature, 369 (1994), pp. 215–218.
[29] Y. Lou, T. Nagylaki, and W.-M. Ni, *On diffusion-induced blowups in a mutualistic model*, Nonlinear Anal., 45 (2001), pp. 329–342.
[30] D. S. Morgan, A. Doelman, and T. J. Kaper, *Stationary periodic orbits in the* 1*-D Gray-Scott model*, Methods Appl. Anal., 7 (2000), pp. 105–150.

[31] D. S. Morgan and T. J. Kaper, *Annular Rings and Their Destabilization into Spots in the* 2*-D Gray-Scott Model*, in preparation.
[32] J. D. Murray, *Mathematical Biology*, Biomathematics Texts 19, Springer-Verlag, Berlin, Heidelberg, New York, 1989.
[33] W.-M. Ni, *Diffusion, cross-diffusion, and their spike-layer steady states*, Notices Amer. Math. Soc., 45 (1998), pp. 9–18.
[34] Y. Nishiura and D. Ueyama, *A skeleton structure for self-replication dynamics*, Phys. D, 130 (1999), pp. 73–104.
[35] Y. Nishiura and D. Ueyama, *Spatio-temporal chaos for the Gray-Scott model*, Phys. D, 150 (2001), pp. 137–162.
[36] T. Ohta, *Pulse dynamics in a reaction-diffusion system*, Phys. D, 151 (2001), pp. 61–72.
[37] J. Olwoch and D. J. Needham, *Diffusion-driven blow-up in a reaction-diffusion model for* $n \geq 1$*-th order autocatalysis*, Nonlinearity, 13 (2000), pp. 43–76.
[38] M. Or-Guil, I. G. Kevrekidis, and M. Bär, *Stable bound states of pulses in an excitable medium*, Phys. D, 135 (2000), pp. 154–174.
[39] J. E. Pearson, *Complex patterns in a simple system*, Science, 261 (1993), pp. 189–192.
[40] V. Petrov, S. K. Scott, and K. Showalter, *Excitability, wave reflection, and wave splitting in a cubic autocatalysis reaction-diffusion system*, Phil. Trans. Roy. Soc. London Ser. A, 347 (1994), pp. 631–642.
[41] M. Pierre and D. Schmitt, *Blowup in reaction-diffusion systems with dissipation of mass*, SIAM J. Math. Anal., 28 (1997), pp. 259–269.
[42] K. Promislow, *A renormalization method for modulational stability of quasi-steady patterns in dispersive systems*, SIAM J. Math. Anal., 33 (2002), pp. 1455–1482.
[43] W. N. Reynolds, J. E. Pearson, and S. Ponce-Dawson, *Dynamics of self-replicating patterns in reaction diffusion systems*, Phys. Rev. Lett., 72 (1994), pp. 2797–2800.
[44] W. N. Reynolds, S. Ponce-Dawson, and J. E. Pearson, *Self-replicating spots in reaction-diffusion systems*, Phys. Rev. E, 56 (1997), pp. 185–198.
[45] C. Robinson, *Sustained resonance for a nonlinear system with slowly varying coefficients*, SIAM J. Math. Anal., 14 (1983), pp. 847–860.
[46] J. B. van den Berg, *personal communication*, Free University of Amsterdam, Amsterdam, The Netherlands, 2002.
[47] D. Walgraef, *Spatio-Temporal Pattern Formation: With Examples from Physics, Chemistry, and Material Science*, Springer-Verlag, New York, 1997.
[48] J. Wei, *On single interior spike solutions of the Gierer-Meinhardt system: Uniqueness and spectrum estimates*, European J. Appl. Math., 10 (1999), pp. 353–378.
[49] H. Weinberger, *An example of blowup produced by equal diffusions*, J. Differential Equations, 154 (1999), pp. 225–237.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 2, pp. 97–143

Bifurcation on the Visual Cortex with Weakly Anisotropic Lateral Coupling*

Martin Golubitsky†, Lie June Shiau‡, and Andrei Török†

Abstract. Mathematical studies of drug induced geometric visual hallucinations include three components: a model (or class of models) that abstracts the structure of the primary visual cortex V1; a mathematical procedure for finding geometric patterns as solutions to the cortical models; and a method for interpreting these patterns as visual hallucinations.

Ermentrout and Cowan used the Wilson–Cowan equations to model the evolution of an activity variable $a(\mathbf{x})$ that represents, for example, the voltage potential a of the neuron located at point $\mathbf{x}$ in V1. Bressloff, Cowan, Golubitsky, Thomas, and Wiener generalize this class of models to include the orientation tuning of neurons in V1 and the Hubel and Wiesel hypercolumns. In these models, $a(\mathbf{x}, \phi)$ represents the voltage potential a of the neuron in the hypercolumn located at $\mathbf{x}$ and tuned to direction ϕ. The work of Bressloff et al. assumes that lateral connections between hypercolumns are *anisotropic*; that is, neurons in neighboring hypercolumns are connected only if they are tuned to the same orientation and then only if the neurons are oriented in the cortex along the direction of their cells' preference. In this work, we first assume that lateral connections are *isotropic*: neurons in neighboring hypercolumns are connected whenever they have the same orientation tuning. Wolf and Geisel use such a model to study development of the visual cortex. Then we consider the effect of perturbing the lateral couplings to be weakly anisotropic.

There are two common features in these models: the models are continuum models (neurons and hypercolumns are idealized as points and circles), and the models all have planar Euclidean $\mathbf{E}(2)$-symmetry (when cortical lateral boundaries are ignored). The approach to pattern formation is also common. It is assumed that solutions are spatially periodic with respect to a fixed planar lattice and that patterns are formed by symmetry-breaking bifurcations (corresponding to wave vectors of shortest length) from a spatially uniform state. There are also substantial differences. In the Ermentrout–Cowan model, $\mathbf{E}(2)$ acts in its standard representation on $\mathbf{R}^2$, whereas in the Bressloff et al. model, $\mathbf{E}(2)$ acts on $\mathbf{R}^2 \times \mathbf{S}^1$ via the *shift-twist* action. In our model, isotropic coupling introduces an additional $\mathbf{S}^1$-symmetry. Weak anisotropy is then thought of as a small forced symmetry-breaking from $\mathbf{E}(2)\dot{+}\mathbf{S}^1$ to $\mathbf{E}(2)$ in its shift-twist action.

The bifurcation analyses in each of these theories proceed along similar lines, but each produces different hallucinatory images—many of which have been reported in the psychophysics literature. The Ermentrout–Cowan model produces spirals and funnels, whereas the Bressloff et al. model produces in addition thin line images including honeycombs and cobwebs. Finally, our model produces three types of time-periodic states: rotating structures such as spirals, states that appear to rush

*Received by the editors June 18, 2002; accepted for publication (in revised form) by G. Kriegsmann December 12, 2002; published electronically May 8, 2003. This work was supported in part by NSF grant DMS-0071735, ARP grant 003652-0032-2001, and the Faculty Research Support Fund of UHCL (LJS).

http://www.siam.org/journals/siads/2-2/40988.html

†Department of Mathematics, University of Houston, Houston, TX 77204-3008 (mg@uh.edu, torok@math.uh.edu).

‡Department of Mathematics, University of Houston - Clear Lake, Houston, TX 77058 (shiau@cl.uh.edu).

SIAM J. Applied Dynamical Systems
Vol. 2, No. 2, pp. 97–143

into (or out from) a tunnel with its hole in the center of the visual field, and pulsating images. Although it is known that branches of time-periodic states can emanate from steady-state bifurcations in systems with symmetry, this model provides the first examples of this phenomena in a specific class of models.

Key words. bifurcation, pattern formation, symmetry, hallucinations, visual cortex, rotating waves

AMS subject classifications. 37G40, 34C23, 92C15

PII. S1111111102409882

1. Introduction and overview.

In the 1930's, Klüver classified geometric visual hallucinations into four groups of *form constants* (see [17, p. 66]): honeycombs, cobwebs, tunnels, and spirals. Klüver states "For the sake of analysis in terms of 'form,' we have ignored aspects of color, brightness, and movement" and, on p. 71, "We wish to stress merely one point, namely, that under diverse conditions the visual system responds in terms of a limited number of form constants."

Ermentrout and Cowan [9] pioneered an approach to the mathematical study of geometric patterns produced in drug induced hallucinations. They assumed that the drug uniformly stimulates an inactive cortex and produces, by spontaneous symmetry-breaking, a patterned activity state. The mind then interprets the pattern as a visual image—namely, the visual image that would produce the same pattern of activity on the primary visual cortex V1.[1] The Ermentrout–Cowan analysis assumes that a differential equation governs the symmetry-breaking transition from an inactive to an active cortex and then studies the transition abstractly using standard pattern formation arguments developed for reaction-diffusion equations. Their cortical patterns are obtained by thresholding. (Points where the solution is greater than some threshold are colored black, whereas all other points are colored white.) These cortical patterns are then transformed to retinal patterns[2] using the inverse of the retino-cortical map described below (see (1.4)), and these retinal patterns are similar to some of the geometric patterns of visual hallucinations, namely, tunnels and spirals.

It is now well established that neurons in V1 are sensitive to orientations in the visual field,[3] and it is mathematically reasonable to assign an orientation preference to each neuron in V1. Hubel and Wiesel [16] introduced the notion of a *hypercolumn*—a region in V1 containing

[1]The *primary visual cortex*, or V1, is the area of the visual cortex that receives electrical signals directly from the retina.

[2]The *retina* is the light-sensitive tissue lining the back of the eyeball that sends electrical impulses to the brain.

[3]Experiments show that most V1 cells signal the local orientation of a contrast edge or bar; these neurons are tuned to a particular local orientation. See [16, 12, 2, 4] and [6] for more discussion.

for each orientation at a single point in the visual field (a mathematical idealization) a neuron sensitive to that orientation.

More recently, Bressloff et al. [6] studied the geometric patterns of drug induced hallucinations by including orientation sensitivity. As before, the drug stimulation is assumed to induce spontaneous symmetry-breaking, and the analysis is local in the sense of bifurcation theory. There is one major difference between the approaches in [6] and [9]. Ignoring lateral boundaries, Ermentrout and Cowan [9] idealize the cortex as a plane, whereas Bressloff et al. [6] take into account the orientation tuning of cortical neurons and idealize the cortex as $\mathbf{R}^2 \times \mathbf{S}^1$. This approach leads to a method for recovering thin line hallucinations such as cobwebs and honeycombs in addition to the threshold patterns found in the Ermentrout–Cowan theory.

There are two types of connections between neurons in V1: local and lateral. Experimental evidence suggests that neurons within a hypercolumn are all-to-all connected, whereas neurons in different hypercolumns are connected in a very structured way. This structured lateral coupling is called *anisotropic*, and it is the bifurcation theory associated with anisotropic coupling that is studied in Bressloff et al. [6, 5].

In this paper, we study generic bifurcations when lateral coupling is weakly anisotropic. First, we study bifurcations in models that are isotropic, showing that these transitions lead naturally to a richer set of planforms than is found in [6, 5] and, in particular, to time-periodic states. (Isotropic models have an extra $\mathbf{S}^1$-symmetry and have been studied by Wolf and Geisel [26] as a model for the development of anisotropic lateral coupling.) There are three types of time dependent solutions: slowly *rotating* spiral and funnel shaped retinal images; *tunneling* images, where the retinal image appears to rush into or spiral into the center of the visual field; and *pulsating* images, where the spatial pattern of the solution changes periodically in time. Such images have been reported in the psychophysics literature; see Klüver [17, p. 24]. (Note that near death experiences are sometimes described as traveling down a tunnel toward a central area.) Second, we consider weak anisotropy as forced symmetry-breaking from isotropy and by doing so find two additional rotating time-periodic states.

We note that time-periodic, spatially periodic states have been studied previously in Tass [24] using averaging (Hopf bifurcation) applied to the Ermentrout–Cowan activator-inhibitor model [9], whereas our results lead to time-periodic states through symmetry-breaking steady-state bifurcations.

We mention two caveats associated with our methods. First, our analysis applies only to generic members of the general class of equations having symmetries associated to isotropic lateral coupling (and to weak symmetry-breakings that preserve the symmetry of anisotropic lateral coupling). It would be an extraordinarily difficult calculation to show, for example,

that every statement that we make is valid for some particular variant of the Wilson–Cowan equations, though most likely they are valid. Second, we assume that the new states are hyperbolic (which is true generically but is also quite difficult to verify in a particular model).

The remainder of this section is devoted to two primary issues: the background from previous studies that is needed to understand the current study and the comparison of our results (on the square lattice) with previous results. The section ends with a brief description of the (more complicated) hexagonal lattice results.

1.1. The continuum models and cortical planforms. The Ermentrout and Cowan [9] model of V1 consists of neurons located at each point $\mathbf{x}$ in $\mathbf{R}^2$. Their model equations, variants of the Wilson–Cowan equations [25], are written in terms of a real-valued *activity variable* $a(\mathbf{x})$, where a represents, say, the voltage potential of the neuron at location $\mathbf{x}$.

Bressloff et al. [6] incorporate the Hubel–Weisel hypercolumns [16] into their model of V1 by assuming that there is a hypercolumn centered at each location $\mathbf{x}$. Here a *hypercolumn* denotes a region of cortex that contains neurons sensitive to orientation ϕ for each direction ϕ. Their models, also adaptations of the Wilson–Cowan equations [25], are written in terms of a real-valued *activity variable* $a(\mathbf{x}, \phi)$, where a represents, say, the voltage potential of the neuron tuned to orientation ϕ in the hypercolumn centered at location $\mathbf{x}$. Note that angles ϕ and $\phi + \pi$ give the same orientation; so

$$a(\mathbf{x}, \phi + \pi) = a(\mathbf{x}, \phi).$$

The cortical planform associated to $a(\mathbf{x}, \phi)$ is obtained in a way that is different from the Ermentrout–Cowan approach. For each fixed $\mathbf{x} \in \mathbf{R}^2$, $a(\mathbf{x}, \cdot)$ is a function on the circle. The planform associated to a is obtained through a *winner-take-all* strategy. The neuron that is most active in its hypercolumn is presumed to suppress the activity of other neurons within that hypercolumn. The winner-take-all strategy chooses, for each $\mathbf{x}$, the directions ϕ that maximize $a(\mathbf{x}, \cdot)$ and results in a field of directions. The two approaches to creating planforms can be combined by assigning directions only to those locations $\mathbf{x}$ where the associated maximum of $a(\mathbf{x}, \cdot)$ is larger than a given threshold.

A possible justification for the continuum model that idealizes a hypercolumn at each cortex location is that each location is in fact surrounded by neurons sensitive to all of the possible orientations. This fact suggests that the signal read from the primary visual cortex V1 need not be limited to one orientation from each "physical" hypercolumn. In V1, there is a grid of physical hypercolumns that is approximately 36×36 in extent. (See [5] and references therein.) It is reasonable to suppose that other layers of the visual cortex receive much more information than a 36×36 matrix of orientation values.

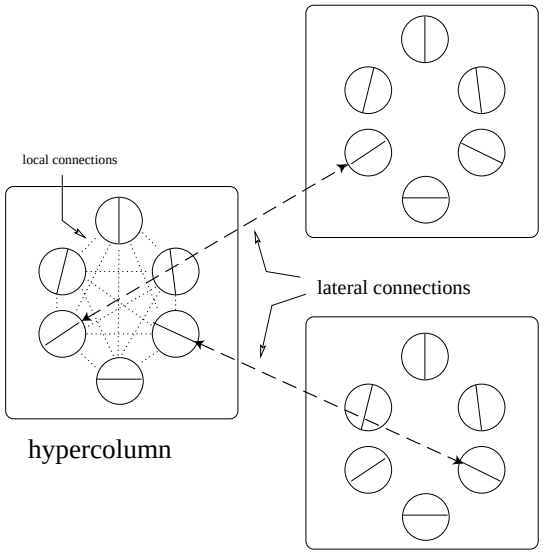


Figure 1. *Illustration of isotropic local and anisotropic lateral connection patterns.*

1.2. Euclidean symmetry. The Euclidean group $\mathbf{E}(2)$ is crucial to the analyses in both [9] and [6], but the way that group acts is different. In Ermentrout and Cowan, the Euclidean group acts on the plane by its standard action, whereas in Bressloff et al., the Euclidean group acts on $\mathbf{R}^2 \times \mathbf{S}^1$ by the so-called shift-twist representation, as we now explain.

Bressloff et al. [6] argue, based on experiments by Blasdel [2] and Eysel [10], that the lateral connections between neurons in neighboring hypercolumns are *anisotropic*. That anisotropy states that the *strength* of the connections between neurons in two neighboring hypercolumns depends on the orientation tuning of both neurons and on the relative locations of the two hypercolumns. Moreover, this anisotropy is idealized to the one illustrated in Figure 1, where only neurons with the same orientation selectivity are connected and then only neurons that are oriented along the direction of their cells preference are connected. These conclusions are based on work of Gilbert [12] and Bosking et al. [4]. In particular, the symmetries of V1 model equations are those that are consistent with the idealized structure shown in Figure 1.

The Euclidean group $\mathbf{E}(2)$ is generated by translations, rotations, and a reflection. The action of $\mathbf{E}(2)$ on $\mathbf{R}^2 \times \mathbf{S}^1$ that preserves the structure of lateral connections illustrated in Figure 1 is the *shift-twist* action. This action is given by

$$\begin{aligned} \mathcal{T}_{\mathbf{y}}(\mathbf{x}, \phi) &\equiv (\mathbf{x} + \mathbf{y}, \phi), \\ \mathcal{R}_{\theta}(\mathbf{x}, \phi) &\equiv (R_{\theta}\mathbf{x}, \phi + \theta), \\ \mathcal{M}_{\kappa}(\mathbf{x}, \phi) &\equiv (\kappa\mathbf{x}, -\phi), \end{aligned} \tag{1.1}$$

where $(\mathbf{x}, \phi) \in \mathbf{R}^2 \times \mathbf{S}^1$, $\mathbf{y} \in \mathbf{R}^2$, κ is the reflection $(x_1, x_2) \mapsto (x_1, -x_2)$, and $R_{\theta} \in \mathbf{SO}(2)$ is the rotation of the plane counterclockwise through angle θ.

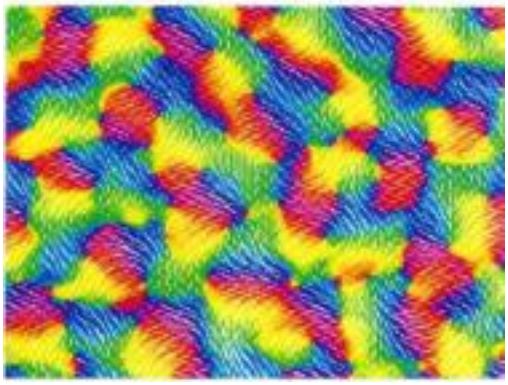


Figure 2. *Distribution of orientation preferences in the Macaque V1 obtained via optical imaging and using color to indicate iso-orientation patches. Redrawn from* [2].

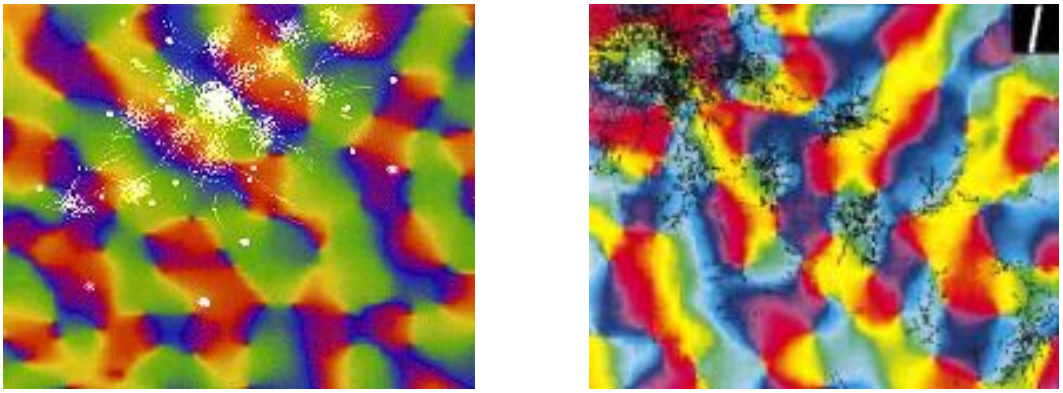


Figure 3. *Lateral connections made by cells in Macaque (left panel) and Tree Shrew (right panel) V1. A radioactive tracer is used to show the locations of all terminating axons from cells in a central injection site, superimposed on an orientation map obtained by optical imaging. Redrawn from* [23] *and* [4].

Work on optical imaging has made it possible to see how the orientation preferences of cells are actually distributed in V1 [2], and a variety of stains and labels have made it possible to see how they are interconnected [10, 4]. Figure 2 shows the distribution of orientation preferences in the Macaque. In particular, approximately every millimeter there is an *iso-orientation patch* of a given preference.

Recent optical imaging experiments combined with anatomical tracer injections suggest that there is a spatial anisotropy in the distribution of patchy horizontal connections, as illustrated in Figure 3. It will be seen from the right panel that the anisotropy is particularly pronounced in the tree shrew. The major axis of the horizontal connections tends to run parallel to the visuotopic axis of the connected cells' common orientation preference. There is also a clear anisotropy in the patchy connections of the Macaque, as seen in the left panel.

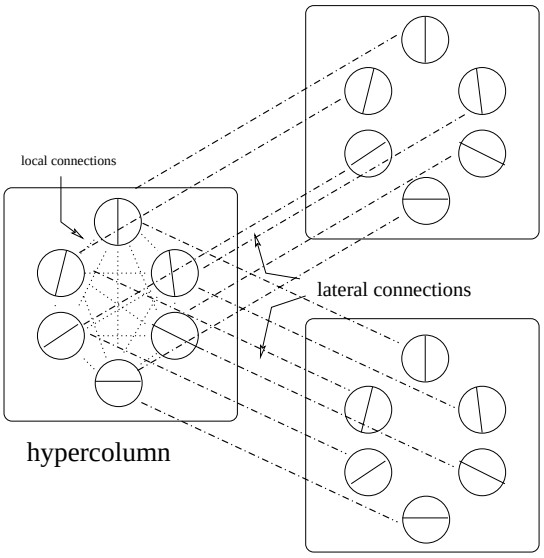


Figure 4. *Illustration of isotropic local and isotropic lateral connection patterns.*

However, in this case, most of the anisotropy can be accounted for by the fact that there is a stretching in the direction orthogonal to ocular dominance columns [1, 23]. It is possible that when this stretching is factored out, there remains a weak anisotropy correlated with orientation selectivity, but this remains to be confirmed experimentally. It should further be noted that feedback connections from higher cortical areas are also patchy and appear to link cells with similar functional features. Preliminary studies suggest that these feedback connections do seem to have a much stronger anisotropy than the lateral connections [1]. This is still a matter of ongoing investigation.

1.3. Isotropy of lateral connections. As noted in the previous paragraph, the anisotropy in lateral connections pictured in Figure 1 can be small in the following sense. We call the lateral connections between hypercolumns *isotropic*, as is done in Wolf and Geisel [26], if the strength of lateral connections between neurons in two neighboring hypercolumns depends only on the difference between the angles of the neurons' orientation sensitivity. Lateral connections in the isotropic model are illustrated in Figure 4. In this model, equations admit, in addition to Euclidean symmetry, the following $\mathbf{S}^1$-symmetry:

$$\mathcal{I}_{\widehat{\phi}}(\mathbf{x}, \phi) = (\mathbf{x}, \phi + \widehat{\phi}). \tag{1.2}$$

Note that $\widehat{\phi} \in \mathbf{S}^1$ commutes with $\mathbf{y} \in \mathbf{R}^2$ and $R_\theta \in \mathbf{SO}(2)$, but $\kappa\widehat{\phi} = (-\widehat{\phi})\kappa$.

The action of $\gamma \in \mathbf{E}(2)\dot{+}\mathbf{S}^1$ on the activity function a is given by

$$\gamma a(\mathbf{x}, \phi) = a(\gamma^{-1}(\mathbf{x}, \phi)).$$

For example, $R_\theta \in \mathbf{SO}(2)$ acts by

$$(\mathcal{R}_\theta a)(\mathbf{x}, \phi) = a(R_{-\theta}\mathbf{x}, \phi - \theta).$$

In this paper, we determine solutions to symmetry-breaking bifurcations in the isotropic case and then study how these solutions change when anisotropy is introduced as a small forced symmetry-breaking parameter.

1.4. Symmetry-breaking bifurcations on lattices. Spontaneous symmetry-breaking in the presence of a noncompact group such as the Euclidean group is far from completely understood. The standard approach is to reduce the technical difficulties by looking only for solutions that are spatially doubly periodic with respect to some planar lattice (see Golubitsky and Stewart [13]); this is the approach taken in [9, 6] and in this study. This approach is justified by the remarkable similarities between the geometric patterns obtained mathematically in [9, 6] and the hallucinatory images reported in the scientific literature [6, 7].

The first step in such an analysis is to choose a lattice type; in this paper, we describe transitions on both the square and hexagonal lattices. The second step is to decide on the size of the lattice. Euclidean symmetry guarantees that at bifurcation, critical eigenfunctions will have *plane wave* factors $e^{2\pi i \mathbf{k}\cdot\mathbf{x}}$ for some critical dual wave vector $\mathbf{k}$. See [5] or [13, Chapter 5]. Typically, the lattice size is chosen so that the critical wave vectors will be the vectors of shortest length in the dual lattice; that is, the lattice has the smallest possible size that can support doubly periodic solutions.

By restricting the bifurcation problem to a lattice, the group of symmetries is transformed to a compact group. First, translations in $\mathbf{E}(2)$ act modulo the spatial period (which we can take to be 1 on the square lattice) and thus act as a 2-torus $\mathbf{T}^2$. Second, only those rotations and reflections in $\mathbf{E}(2)$ that preserve the lattice (namely, the holohedry $\mathbf{D}_4$ for the square lattice) are symmetries of the lattice restricted problem. Thus the symmetry group of the square lattice problem is $\Gamma = \mathbf{D}_4 \dot{+} \mathbf{T}^2$. Recall that at bifurcation Γ acts on the kernel of the linearization, and a subgroup of Γ is *axial* if its fixed-point subspace in that kernel is one-dimensional. Solutions are guaranteed by the Equivariant Branching Lemma (see [14, 13]), which states the following: generically there are branches of equilibria to the nonlinear differential equation for every axial subgroup of Γ. The nonlinear analysis in [5, 9] proceeds in this fashion.

1.5. Previous results on the square lattice. In Ermentrout and Cowan [9], translation symmetry leads to eigenfunctions that are linear combinations of plane waves and, on the square lattice, to two axial planforms: stripes and squares. See Figure 5.

In Bressloff et al. [6, 5], translation symmetry leads to critical eigenfunctions that are linear combinations of functions of the form $u(\phi)e^{2\pi i \mathbf{k}\cdot\mathbf{x}}$. These eigenfunctions correspond to

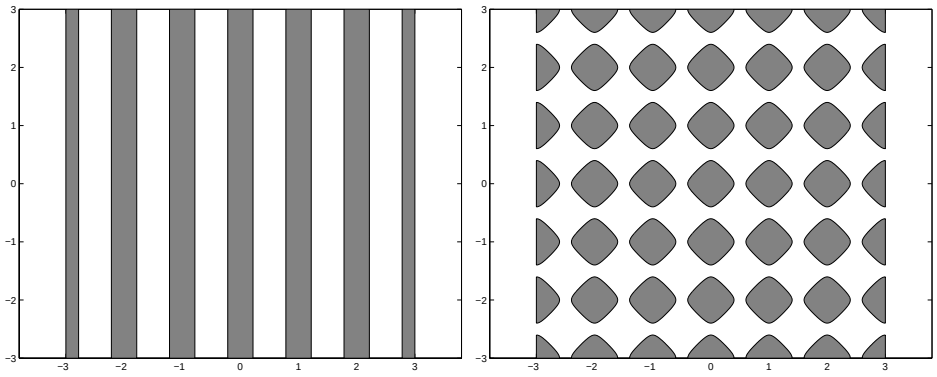


Figure 5. *Thresholding of eigenfunctions: (left) stripes; (right) squares.*

one of two types of representations of $\mathbf{E}(2)$ (restricted to the lattice): *scalar* (u even in ϕ) and *pseudoscalar* (u odd). The fact that two different representations of the Euclidean group can appear in bifurcations was first noted by Bosch Vivancos, Chossat, and Melbourne [3]. Bressloff et al. [6] also show that a trivial solution to the Wilson–Cowan equation will lose stability via a scalar or pseudoscalar bifurcation depending on the exact form of the lateral coupling. Thus each of these representations is, from a mathematical point of view, equally likely to occur. On the square lattice, [3, 5] show that there are two axial planforms each in the scalar and pseudoscalar cases: stripes and squares.

To picture the planforms in these cases, we must specify the function $u(\phi)$, and this can be accomplished by assuming that anisotropy is small. When anisotropy is zero, the $\mathbf{S}^1$-symmetry in (1.2) forces $u(\phi) = \cos(2m\phi)$ in the scalar case and $u(\phi) = \sin(2m\phi)$ in the pseudoscalar case. (This point will be discussed in more detail when we review representation theory in sections 2 and 4.) The assumptions in Bressloff et al. [6] imply that u is a small perturbation of sine or cosine. Note that the Ermentrout–Cowan planforms are recovered in the scalar case when $m = 0$; in this case, u is constant, and all directions are equally active. As often happens in single equation models, the first instability of a trivial (spatially constant) solution is to eigenfunctions with m small, and that is what occurs in certain models based on the Wilson–Cowan equation (see [6]). Planforms for the scalar and pseudoscalar planforms when $m = 1$ are shown in Figures 6 and 7.

1.6. New planforms when lateral connections are isotropic. In our analysis of the isotropic case ($\widetilde{\Gamma} = \Gamma\dot{+}\mathbf{S}^1$-symmetry), we find four axial subgroups (Σ_1–Σ_4) and one maximal isotropy subgroup Σ_5 with a two-dimensional fixed-point subspace. The axial subgroups lead to group orbits of equilibria. This fact must be properly interpreted to understand how the new planforms relate to the old. A phase shift of $\sin(2\phi)$ yields $\cos(2\phi)$. Thus the extra $\mathbf{S}^1$-symmetry based on isotropic lateral connections identifies scalar and pseudoscalar plan-

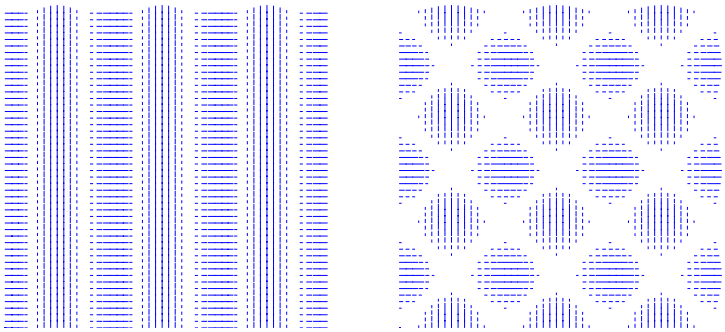


Figure 6. *Direction fields of scalar eigenfunctions: (left) stripes Σ_3; (right) squares Σ_1.*

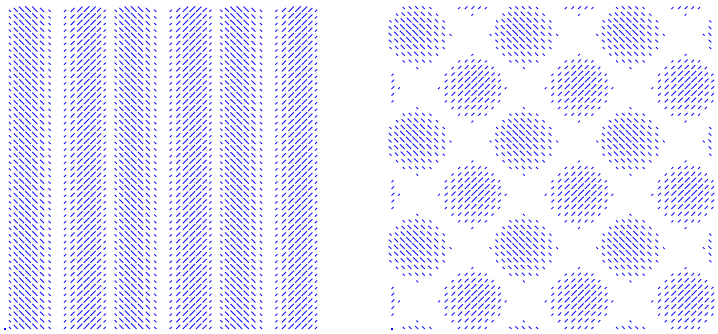


Figure 7. *Direction fields of pseudoscalar eigenfunctions: (left) stripes Σ_3; (right) squares Σ_1.*

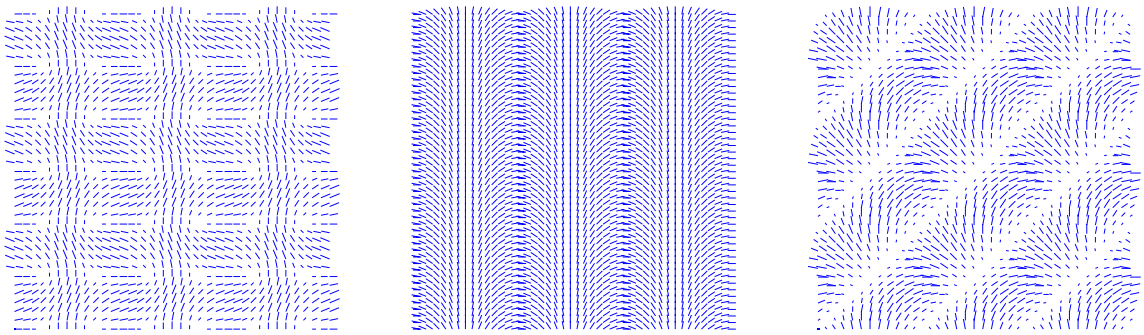


Figure 8. *Direction fields of new planforms in isotropic model: (left) axial planform Σ_2; (center) axial planform Σ_4; (right) rotating wave Σ_5 (direction of movement is up and to the left).*

forms; up to this new symmetry, the planforms are the same. Thus the axial subgroup Σ_3 corresponds to stripes (both scalar and pseudoscalar), and the axial subgroup Σ_1 corresponds to squares (both scalar and pseudoscalar). The axial subgroups Σ_2 and Σ_4 correspond to new types of planforms. Finally, the maximal isotropy subgroup Σ_5 with its two-dimensional fixed-point subspace leads to a time-periodic rotating wave whose frequency is zero at bifurcation. The planforms associated with these new types of solutions are pictured in Figure 8.

It is unusual for a steady-state bifurcation (eigenvalues of a linearization moving through 0) to lead to time-periodic states. It is well known that in systems without symmetry, time-periodic states will appear in unfoldings of codimension two Takens–Bogdanov singularities (a double zero eigenvalue with a nilpotent normal form). It is less well known that codimension one steady-state bifurcations with symmetry can also lead to time-periodic states. Field and Swift [11] were the first to find such a bifurcation (in a system with finite symmetry). Melbourne [21] was the first to find an example of a rotating wave in a steady-state bifurcation in a system with continuous symmetry. Nevertheless, the documented cases where time-periodic states occur in codimension one steady-state bifurcations are relatively rare, and our work provides the first example where this mathematical phenomenon appears in model equations.

1.7. Weak anisotropy in lateral connections. Next, we discuss what happens to the bifurcating solutions to the isotropic nonlinear equation when anisotropy is added as a small symmetry-breaking parameter. As was noted in Bressloff et al. [6], the linear effect of anisotropy is to split the eigenfunctions into scalar and pseudoscalar representations. The effect on solutions to the nonlinear equation can also be established using the methods of Lauterbach and Roberts [20]. This method is applied independently to each branch of (group orbits of) solutions found in the isotropic case. The results for square lattice solutions are easily described.

Generically, the dynamics on the $\widetilde{\Gamma}$ group orbit of equilibria corresponding to the axial subgroup Σ_3 has two (smaller Γ) group orbits of equilibria: scalar stripes and pseudoscalar stripes. There may be other equilibria coming from the $\widetilde{\Gamma}$ group orbit, but, at the very least, scalar and pseudoscalar stripes always remain as solutions.

Similarly, the dynamics on the group orbit of equilibria corresponding to the axial subgroup Σ_1 generically has two equilibria corresponding to scalar and pseudoscalar squares.

The dynamics on the group orbit of the axial subgroups Σ_2 and Σ_4 and the fifth maximal isotropy subgroup Σ_5 does not change substantially when anisotropy is added. These group orbits still remain as equilibria and rotating waves.

1.8. Retinal images. Finally, we discuss the geometric form of the cortical planforms in the visual field; that is, we try to picture the corresponding visual hallucinations. It is known that the density of neurons in the visual cortex is uniform, whereas the density of neurons in the retina falls off from the fovea[4] at a rate of $1/r^2$. Schwartz [22] observed that there is a unique conformal map taking a disk with $1/r^2$ density to a rectangle with uniform density, namely, the complex logarithm. This is also called the *retino-cortical* map. It is thought that using

[4]The *fovea* is the small central area of the retina that gives the sharpest vision.

the inverse of the retino-cortical map, the complex exponential, to push forward the activity pattern from V1 to the retina is a reasonable way to form the hallucination image—and this is the approach used in Ermentrout and Cowan [9] and in Bressloff et al. [6, 7]. Specifically, the transformation from polar coordinates (r, θ) on the retina to cortical coordinates (x, y) is given in Cowan [8] to be

$$\begin{aligned} x &= \tfrac{1}{\varepsilon} \ln\left(\tfrac{1}{\omega} r\right), \\ y &= \tfrac{1}{\varepsilon} \theta, \end{aligned} \tag{1.3}$$

where ω and ε are constants. See Bressloff et al. [7] for a discussion of the values of these constants. The inverse of the retino-cortical map (1.3) is

$$\begin{aligned} r &= \omega \exp(\varepsilon x), \\ \theta &= \varepsilon y. \end{aligned} \tag{1.4}$$

In our retinal images, we take

$$\omega = \frac{30}{e^{2\pi}} \quad \text{and} \quad \varepsilon = \frac{2\pi}{n_h},$$

where n_h is the number of hypercolumn widths in the cortex, which we take to be 36.

There are additional issues that need to be discussed.

1. What is the relationship between the spatial period of the planform and the size of a "physical" hypercolumn?
2. How many points within a hypercolumn should be used to create the visual image?
3. In our symmetry analysis, states are enumerated up to symmetry. Are the retinal images of symmetry-related cortical states the same?

We discuss each of these questions in turn.

1.8.1. Spatial period. The human visual cortex contains a grid of approximately 36×36 hypercolumns. Bressloff et al. [6, 7] argue that each spatial period is the size of two hypercolumns. This conclusion is based on the properties of reported visual hallucinations and on the responses of human subjects to perceived grating patterns.

1.8.2. Grid points per hypercolumn. There are regions within a fundamental square in which the line field varies continuously and curves across which discontinuities in line field direction appear. Discontinuities in the direction field follow from the winner-take-all strategy and cannot be avoided. Generally we find that to sample every region in which the line field is continuous, it is sufficient to evaluate the direction fields on a 4×4 array of points in each hypercolumn.

1.8.3. Retinal planforms from conjugate cortical planforms. We begin by justifying our choice of lattice orientation. Observe that vertical lines on the cortex are taken by (1.4) to circles in the retina. Therefore, the vertical y-direction on the cortex is a periodic direction, and it makes sense to align the square lattice so edges of the square are vertical and horizontal lines on the cortex. It is also sensible to align the hexagonal lattice so that one of the three sets of parallel lines defining a fundamental hexagon consists of horizontal lines.

Certain lattice symmetries do change the retinal planform, and some do not. Moreover, certain ways of representing the data graphically change with conjugacies, and others do not. To verify these points, consider a cortical pattern consisting of parallel stripes that is mapped by (1.4) to a concentric family of circles in the retina. Rotating this striped state on the cortex by $\pi/2$ leads to another square lattice state whose associated retinal pattern consists of radial lines. See Figures 10 (center, right) and 11 (center, right). Thus the transformation from cortex to retina given in (1.4) does not respect the symmetries of the cortex.

There are, however, certain cortical symmetries that do not change retinal planforms in any important way. First, each of the planforms that we draw has certain well-defined symmetries; indeed, our planforms are determined by these symmetries. So rotating a square symmetric cortical planform by $\pi/2$ will not change the cortical pattern and hence will not change the retinal planform. Second, translating a cortical pattern in the y-direction just rotates the retinal pattern, whereas translating a cortical pattern in the x-direction scales the retinal pattern in the radial direction. So cortical translations do not change retinal planforms in significant ways; in particular, the Klüver form constant of a retinal image is not changed by cortical translations.

Finally, note that we may draw patterns on the retina and cortex in two distinct ways. First, we may use the line fields obtained from the winner-take-all strategy; and, second, we may ignore the direction information in the line field and just fill in areas where the maximum value of $a(\mathbf{x}, \cdot)$ is greater than some threshold value, as is done in the Ermentrout–Cowan theory. The second method for representing patterns is also unchanged by symmetries in Γ that are projections of the isotropy group in $\widetilde{\Gamma}$.

The visual images that correspond to the previously derived planforms are reproduced here from Bressloff et al. [7]. Figure 9 illustrates the Ermentrout–Cowan planforms. The scalar and pseudoscalar planforms are shown in Figures 10 and 11. When these images are just thresholded, the scalar and pseudoscalar planforms are identical (since, because of $\mathbf{S}^1$-symmetry, the maximum value of $a(\mathbf{x}, \cdot)$ is the same in the corresponding planforms) and are shown in Figure 12. The new planforms found by assumption of isotropy in lateral connections are illustrated in Figures 13–15.

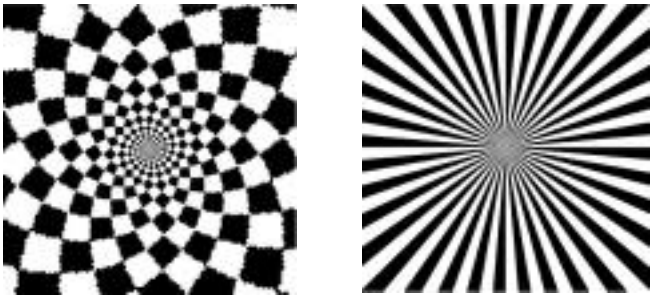


Figure 9. *Action of inverse retino-cortical map* (1.4) *on Ermentrout–Cowan noncontoured square lattice planforms:* (a) *squares;* (b) *stripes.*

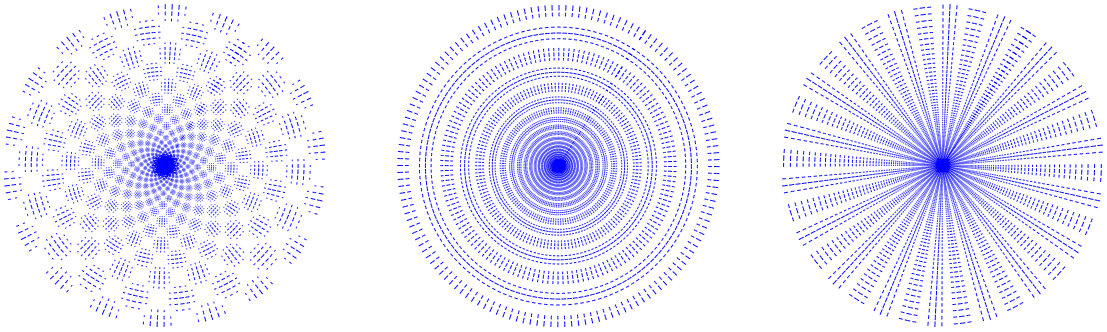


Figure 10. *Action of* (1.4) *on Bressloff* et al. *scalar square lattice planforms: (left) squares* Σ_1*; (center and right) stripes* Σ_3.

1.9. Hexagonal lattice planforms. In isotropic models on the hexagonal lattice, there are (at least) 12 maximal isotropy subgroups; nine of these are axial subgroups and lead to branches of group orbits of equilibria, and three of these have two-dimensional fixed-point spaces, two of which lead to rotating waves. The rotating waves look like rotating spirals in retinal coordinates. Depending on the specific model, the maximal isotropy subgroup corresponding to the third two-dimensional fixed-point space can lead to an equilibrium or a time-periodic state.

When anisotropy is added into the model, two of the nine axial solutions become time-periodic. The retinal planforms associated to these isotropy subgroups have a substantially different character—certain planforms move radially inward (or radially outward), and certain planforms spiral inward (or spiral outward). Thus rotating patterns appear in these models through spontaneous symmetry-breaking, whereas tunneling images appear through forced symmetry-breaking. All of the time-periodic patterns result from weak anisotropy, and the fact that they rotate or tunnel in the retinal image depends just on symmetry and not on the specific method by which the patterns are imaged.

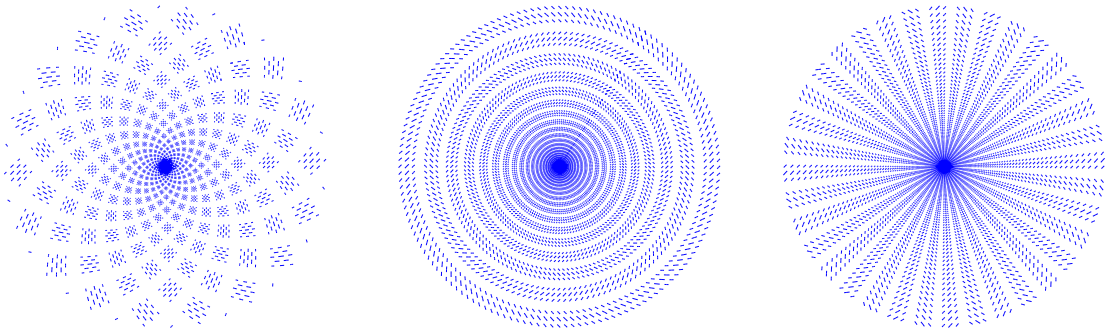


Figure 11. *Action of* (1.4) *on Bressloff et al. pseudoscalar square lattice planforms: (left) squares* Σ_1*; (center and right) stripes* Σ_3*.*

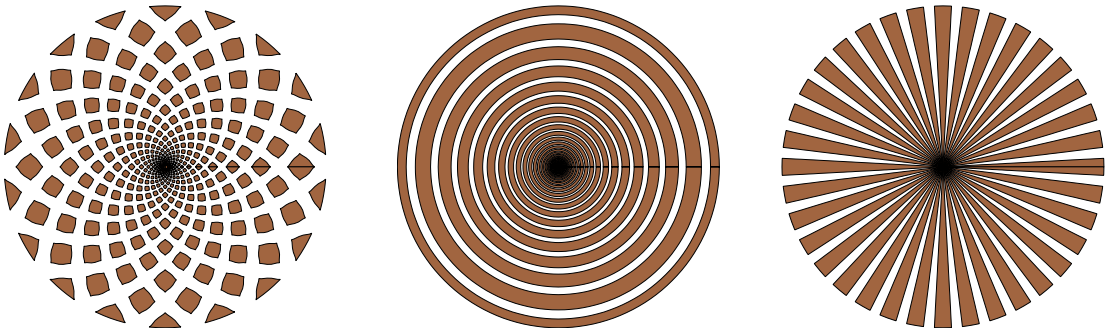


Figure 12. *Threshold retinal images on Bressloff et al. scalar and pseudoscalar square lattice planforms: (left) squares* Σ_1*; (center and right) stripes* Σ_3*.*

The detailed description of the square lattice results are given in section 2, and those for the hexagonal lattice are given in section 3. The proofs of the bifurcation theory statements made in these sections are deferred until section 4.

2. Square lattice planforms. In this section, we discuss the spatially doubly periodic solutions that must emanate from the simplest bifurcations of Euclidean invariant differential equations restricted to a *square* lattice. We assume that the Euclidean action on $\mathbf{R}^2 \times \mathbf{S}^1$ is the one given by (1.1) and that the extra $\mathbf{S}^1$ symmetries (1.2) associated with isotropy are present. Our findings include the following:

1. The simplest $\widetilde{\Gamma} = \Gamma \dot{+} \mathbf{S}^1$ bifurcations, where $\Gamma = \mathbf{D}_4 \dot{+} \mathbf{T}^2$, occur at irreducible representations of $\widetilde{\Gamma}$ that are the direct sum of the scalar and pseudoscalar bifurcations studied in [3, 5].
2. There are five branches of (group orbits of) solutions (corresponding to maximal isotropy subgroups of $\widetilde{\Gamma}$) that must bifurcate from a trivial equilibrium: four are equilibria, and one is a slowly traveling wave in cortical coordinates (and a rotating

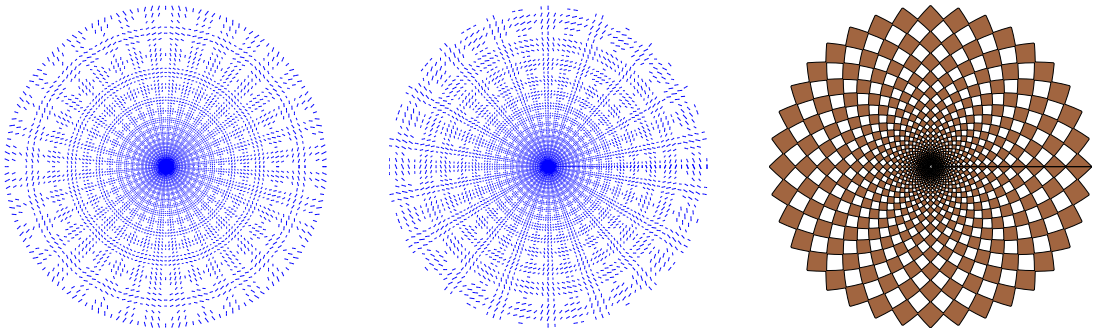


Figure 13. *New square axial planform Σ_2 in isotropic model: (left and center) conjugate line field images; (right) threshold image.*

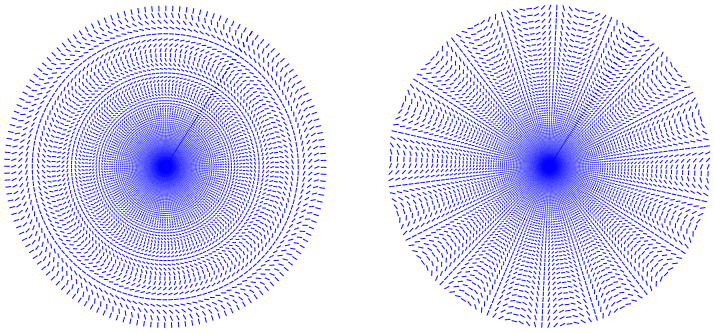


Figure 14. *New axial planform Σ_4 in isotropic model. Thresholding is not relevant for these planforms; see section 3.4.2.*

spiral wave in retinal coordinates).

3. When weak anisotropy in the lateral connections is assumed (that is, symmetry is broken from $\widetilde{\Gamma}$ to Γ in the model equations), one solution leads to both scalar and pseudoscalar rolls (and perhaps a third intermediate state), and a second leads to both scalar and pseudoscalar squares (and perhaps a third intermediate state). The two remaining equilibria persist as equilibria when symmetry is broken, and the rotating wave also persists. These last three solution types correspond to hallucinatory states that have not been discussed previously.

In our exposition, we describe the results in this section and refer to section 4 for the details of the proofs.

2.1. Representation theory of $\widetilde{\Gamma}$. Without loss of generality, we assume that the square lattice $\mathcal{L}$ consists of squares of unit length. Let $\mathcal{F}_{\mathcal{L}}$ be the space of functions $a(\mathbf{x}, \phi)$ that are doubly periodic with respect to translations in $\mathcal{L}$ and π-periodic in ϕ. The action of $\widetilde{\Gamma}$ on $\mathcal{F}_{\mathcal{L}}$ is the one induced from the action of $\mathbf{E}(2)\dot{+}\mathbf{S}^1$ on $\mathbf{R}^2 \times \mathbf{S}^1$ given in (1.1) and (1.2).

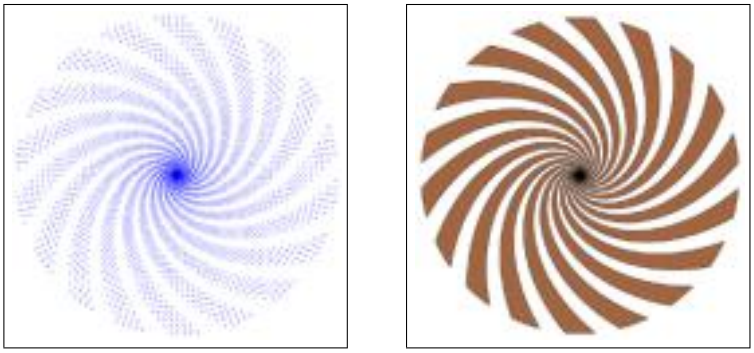


Figure 15. *Rotating planform Σ_5 in isotropic model; movies are associated to framed images.*

We expect the simplest square lattice bifurcations to be from equilibria whose linearizations have kernels that are irreducible subspaces of $\mathcal{F}_{\mathcal{L}}$, and we consider only bifurcations based on dual wave vectors of shortest (unit) length. It follows that we may assume that the critical eigenspace W_m consists of functions of the form

$$a(\mathbf{x},\phi) = \left(z_1 e^{2im\phi} + w_1 e^{-2im\phi}\right) e^{2\pi i \mathbf{k}_1\cdot\mathbf{x}} + \left(z_2 e^{2im(\phi-\pi/2)} + w_2 e^{-2im(\phi-\pi/2)}\right) e^{2\pi i \mathbf{k}_2\cdot\mathbf{x}} + c.c., \tag{2.1}$$

where $(z_1, w_1, z_2, w_2) \in \mathbf{C}^4$. Moreover, the action of $\widetilde{\Gamma}$ on this subspace is absolutely irreducible. These statements are verified in section 4.1. From now on, for $m \geq 1$, we will identify W_m with $\mathbf{C}^4$ through (2.1).

We show in section 4.2 that, from a bifurcation theoretic point of view, we may assume that $m = 0$ or $m = 1$. The case in which $m = 0$ was considered in [5]; so, in the current analysis, we assume that $m = 1$.

In this section, we describe planforms that generically appear in bifurcations with respect to the group $\widetilde{\Gamma}$ and discuss what happens to these planforms after symmetry is broken to $\Gamma \subset \widetilde{\Gamma}$. Therefore, we are interested in how a $\widetilde{\Gamma}$ representation decomposes into Γ representations. The proof of Lemma 4.1 leads to the following lemma.

Lemma 2.1. *The subspace $W_1 \subset \mathcal{F}_{\mathcal{L}}$ decomposes into two nonisomorphic absolutely irreducible representations of Γ: $W_1 = W_1^+ \oplus W_1^-$, where*

$$\begin{aligned} W_1^+ &= \{(u,u,v,v) \mid u,v \in \mathbf{C}\}, \\ W_1^- &= \{(u,-u,v,-v) \mid u,v \in \mathbf{C}\}. \end{aligned}$$

The representation W_1^+ is scalar, and the representation W_1^- is pseudoscalar.

Table 1

Group action on $(z_1, w_1, z_2, w_2) \in W_1 = \mathbf{C}^4$; $\xi, \kappa \in \mathbf{D}_4$, $0 \le \theta_1, \theta_2 < 1$, *and* $0 \le \widehat{\phi} < \pi$.

Generators	Action
ξ	$(\overline{w}_2, \overline{z}_2, z_1, w_1)$
κ	$(w_1, z_1, \overline{z}_2, \overline{w}_2)$
$[\theta_1, \theta_2, 0]$	$(e^{2\pi i\theta_1} z_1, e^{2\pi i\theta_1} w_1, e^{2\pi i\theta_2} z_2, e^{2\pi i\theta_2} w_2)$
$\left[0, 0, \widehat{\phi}\right]$	$(e^{-2i\widehat{\phi}} z_1, e^{2i\widehat{\phi}} w_1, e^{-2i\widehat{\phi}} z_2, e^{2i\widehat{\phi}} w_2)$

2.2. Group action on W_1. A calculation shows that when $m = 1$, $\widetilde{\Gamma}$ acts on $W_1 \cong \mathbf{C}^4$ in the way presented in Table 1. The action for general m is presented in Table 6. When $m = 1$, the action has a nontrivial kernel $\mathbf{Z}_2$ generated by $\left[\frac{1}{2}, \frac{1}{2}, \frac{\pi}{2}\right]$.

Table 2

Square lattice maximal isotropy subgroups of $\widetilde{\Gamma}$ *acting on* $\mathbf{C}^4$; $u \in \mathbf{C}$.

	Generators	Fixed subspace	Dim	Name
Σ_1	κ, ξ	$\mathbf{R}\{(1,1,1,1)\}$	1	squares
Σ_2	$\kappa, \left[\frac{3}{4}, \frac{1}{4}, \frac{\pi}{4}\right] \xi$	$\mathbf{R}\{(1,1,1,-1)\}$	1	
Σ_3	$\kappa, \xi^2, [0, \theta_2, 0]$	$\mathbf{R}\{(1,1,0,0)\}$	1	stripes
Σ_4	$\kappa\xi^2, [0, \theta_2, 0], [\theta_1, 0, \pi\theta_1]$	$\mathbf{R}\{(1,0,0,0)\}$	1	
Σ_5	$\kappa\xi, [\theta_1, \theta_1, \pi\theta_1]$	$\{(u, 0, \overline{u}, 0)\}$	2	rotating spirals

The relations among the generators are as follows:

$$\begin{aligned} \xi\kappa &= \kappa\xi^{-1}, \\ \kappa[\theta_1, \theta_2, \widehat{\phi}] &= [\theta_1, -\theta_2, -\widehat{\phi}]\kappa, \\ \xi[\theta_1, \theta_2, \widehat{\phi}] &= [-\theta_2, \theta_1, \widehat{\phi}]\xi. \end{aligned} \tag{2.2}$$

The list of maximal isotropy subgroups of $\widetilde{\Gamma}$ acting on $\mathbf{C}^4$ is given in Table 2. This list is a subset of the list of isotropy subgroups of $\widetilde{\Gamma}$ acting on $\mathbf{C}^4$ given in Table 7. We discuss only those planforms associated with maximal isotropy subgroups. Table 2 shows that there are four axial subgroups (Σ_1–Σ_4) of $\widetilde{\Gamma}$ acting on W_1 and one maximal isotropy subgroup with a two-dimensional fixed-point subspace Σ_5. Note that the normalizer of Σ_5 is generated by $[\theta_1, -\theta_1, 0]$ over Σ_5.

2.3. Solutions corresponding to maximal isotropy subgroups. The Equivariant Branching Lemma [14] proves the existence (generically) of equilibria corresponding to each axial subgroup. Generally, there is a branch of rotating waves corresponding to each maximal isotropy subgroup with a two-dimensional fixed-point subspace when the normalizer of that subgroup contains a circle group (see Melbourne [21, Theorem 2.4]). Moreover, the periods

of these solutions will tend to infinity at the bifurcation point. See section 4.4 for additional details.

Finally, we note that each of these solutions can be asymptotically stable to perturbations within the square lattice. The computation of linear stability is complicated and the calculations are not included in this paper. The results of these calculations are given in Table 8 in subsection 4.5.

2.4. The effects of weak anisotropy. We discuss how solutions corresponding to $\widetilde{\Gamma}$-bifurcations behave generically when the isotropy of the lateral connections is broken, that is, when the $\widetilde{\Gamma}$-equivariant vector field is perturbed to a Γ-equivariant field.

2.4.1. Squares: $\Sigma_1 = \mathbf{D}_4(\kappa, \xi)$. Generically, breaking the isotropy of the lateral connections leads to a bifurcation of the $\mathbf{D}_4$ steady state into scalar steady states of type Esquares (conjugate in Γ to $S_1 = (1,1,1,1)$) and pseudoscalar steady states of type Osquares (conjugate in Γ to $S_2 = (1,-1,1,-1)$). See [5, Tables 5, 8, and 11, and Figure 3]. In addition to these two types of steady states, it is possible for there to be other intermediate steady states. Details are found in section 4.6.1. See Figures 6 (right) and 7 (right) for the cortical planforms and Figures 10 (left) and 11 (left) for the associated retinal planforms.

2.4.2. $\Sigma_2 = \mathbf{D}_4(\langle \kappa, \left[\frac{3}{4}, \frac{1}{4}, \frac{\pi}{4}\right] \xi \rangle)$. The results for Σ_2 equilibria are similar to those for Σ_1 equilibria. See section 4.6.2 and Figure 13. None of these equilibria are scalar or pseudoscalar states.

2.4.3. Stripes: $\Sigma_3 = \langle \kappa, \xi^2, [0, \theta_2, 0] \rangle$. We show that generically, breaking the isotropy of the lateral connection leads to a bifurcation of this steady state into two states: a scalar steady state of type Erolls (conjugate in Γ to $R_1 = (1,1,0,0)$) and a pseudoscalar steady state of type Orolls (conjugate in Γ to $R_2 = (1,-1,0,0)$). See [5, Tables 5, 8, and 11, and Figures 2(c,d)]. In addition to these two types of steady states, it is possible for there to be other intermediate steady states. Details are found in section 4.6.3. See Figures 6 (left) and 7 (left) for the cortical planforms and Figures 10 (center, right) and 11 (center, right) for the associated retinal planforms.

2.4.4. $\Sigma_4 = \langle \kappa\xi^2, [0, \theta_2, 0], [\theta_1, 0, \pi\theta_1] \rangle$. Steady states corresponding to Σ_4 persist as steady states under symmetry-breaking perturbations of the system and are conjugate to $T_1 = (1,0,0,0)$. Details are found in section 4.6.4. See Figure 8 (center) for the cortical planform and Figure 14 for the associated retinal planforms.

2.4.5. Rotating spirals: $\Sigma_5 = \langle \kappa\xi, [\theta_1, \theta_1, \pi\theta_1] \rangle$. In section 4.4, we show that generically a branch of time-periodic rotating waves bifurcates in the fixed-point subspace of isotropy subgroup Σ_5. These rotating waves $Z(t)$ persist when symmetry is broken to Γ and up to

conjugacy have the form

$$Z(t) = [t, -t, 0](1, 0, 1, 0).$$

Details are found in section 4.6.5. See Figure 8 (right) for the cortical planform and Figure 15 for the associated retinal planform. This picture is a static image of a time-periodic rotating spiral.

3. Hexagonal lattice planforms. In this section, we discuss the spatially doubly periodic solutions that must emanate from the simplest (shortest wave vector) bifurcations of Euclidean invariant differential equations restricted to a *hexagonal* lattice. Our findings include the following:

1. There are (at least) 12 maximal isotropy subgroups in the isotropic case.
2. Three of these 12 have two-dimensional fixed-point subspaces.
3. Two of those three lead generically to rotating waves, and the third one can lead either to equilibria or to time-periodic states.
4. Weak anisotropy forces two of the nine axial solutions to be time-periodic.
5. Pseudoscalar hexagons do not appear naturally as solutions on the hexagonal lattice when weak anisotropy is present.

The group action in the smallest wave vector isotropic case is on a 12-dimensional space as shown in Table 3. The maximal isotropy subgroups are listed in Table 4.

3.1. Representation theory of $\widetilde{\Gamma}$. Without loss of generality, we assume that the hexagonal lattice $\mathcal{L}$ is generated by vectors

$$\boldsymbol{\ell}_1 = \left(1, \frac{1}{\sqrt{3}}\right) \quad \text{and} \quad \boldsymbol{\ell}_2 = \left(0, \frac{2}{\sqrt{3}}\right).$$

Generators for the dual lattice $\mathcal{L}^*$ are the unit length vectors

$$\mathbf{k}_1 = (1, 0) \quad \text{and} \quad \mathbf{k}_2 = \frac{1}{2}(-1, \sqrt{3}).$$

Let

$$\mathbf{k}_3 = -(\mathbf{k}_1 + \mathbf{k}_2) = \frac{1}{2}(-1, -\sqrt{3}).$$

Let $\mathcal{F}_{\mathcal{L}}$ be the space of functions $a(\mathbf{x}, \phi)$ that are doubly periodic with respect to translations in $\mathcal{L}$ and π-periodic in ϕ. The full symmetry group whose bifurcations we analyze is

$$\widetilde{\Gamma} = (\mathbf{D}_6 \dot{+} \mathbf{T}^2) \dot{+} \mathbf{S}^1.$$

The action of $\widetilde{\Gamma}$ on $\mathcal{F}_{\mathcal{L}}$ is the one induced from the action of $\mathbf{E}(2)\dot{+}\mathbf{S}^1$ on $\mathbf{R}^2 \times \mathbf{S}^1$ given in (1.1) and (1.2).

We expect the simplest hexagonal lattice bifurcations to be from equilibria whose linearizations have kernels that are irreducible subspaces of $\mathcal{F}_{\mathcal{L}}$, and we consider only bifurcations based on dual wave vectors of shortest (unit) length. It follows from these assumptions that we may assume that the critical eigenspace W_m consists of eigenfunctions of the form

$$\begin{aligned} a(\mathbf{x},\phi) &= \left(z_1 e^{2im\phi} + w_1 e^{-2im\phi}\right) e^{2\pi i \mathbf{k}_1\cdot\mathbf{x}} \\ &+ \left(z_2 e^{2im(\phi-2\pi/3)} + w_2 e^{-2im(\phi-2\pi/3)}\right) e^{2\pi i \mathbf{k}_2\cdot\mathbf{x}} \\ &+ \left(z_2 e^{2im(\phi+2\pi/3)} + w_2 e^{-2im(\phi+2\pi/3)}\right) e^{2\pi i \mathbf{k}_3\cdot\mathbf{x}} + c.c., \end{aligned} \tag{3.1}$$

where $(z_1, w_1, z_2, w_2, z_3, w_3) \in \mathbf{C}^6$. On this subspace, the action of the group $\widetilde{\Gamma}$ is absolutely irreducible. This statement is verified in a manner entirely analogous to that of (2.1). As with the square lattice, we assume that $m = 1$ in our bifurcation analysis. The cases in which $m > 1$ are identical once one divides by the kernel of the representation, and the case in which $m = 0$ was considered in [5]. From now on, for $m \geq 1$, we will identify W_m with $\mathbf{C}^6$ through (3.1).

Lemma 3.1. *The subspace $W_1 \subset \mathcal{F}_{\mathcal{L}}$ decomposes into two nonisomorphic absolutely irreducible representations of Γ: $W_1 = W_1^+ \oplus W_1^-$, where*

$$\begin{aligned} W_1^+ &= \{(u,u,v,v,w,w) \mid u,v \in \mathbf{C}\}, \\ W_1^- &= \{(u,-u,v,-v,w,-w) \mid u,v \in \mathbf{C}\}. \end{aligned}$$

The representation W_1^+ is scalar, and the representation W_1^- is pseudoscalar. The proof is similar to that of Lemma 2.1.

3.2. Group action on W_1. A calculation leads to the group action on W_1 given in Table 3. Note that the action has a trivial kernel.

Table 3

Group action on $(z_1, w_1, z_2, w_2, z_3, w_3) \in W_1 = \mathbf{C}^6$; $\xi, \kappa \in \mathbf{D}_6$, $0 \leq \theta_1, \theta_2 < 1$, and $0 \leq \widehat{\phi} < \pi$.

Generators	Action
ξ	$(\overline{w}_2, \overline{z}_2, \overline{w}_3, \overline{z}_3, \overline{w}_1, \overline{z}_1)$
κ	$(w_1, z_1, w_3, z_3, w_2, z_2)$
$[\theta_1, \theta_2, 0]$	$(e^{-2\pi i\theta_1} z_1, e^{-2\pi i\theta_1} w_1, e^{-2\pi i\theta_2} z_2, e^{-2\pi i\theta_2} w_2, e^{2\pi i(\theta_1+\theta_2)} z_3, e^{2\pi i(\theta_1+\theta_2)} w_3)$
$\left[0, 0, \widehat{\phi}\right]$	$(e^{-2i\widehat{\phi}} z_1, e^{2i\widehat{\phi}} w_1, e^{-2i\widehat{\phi}} z_2, e^{2i\widehat{\phi}} w_2, e^{-2i\widehat{\phi}} z_3, e^{2i\widehat{\phi}} w_3)$

Table 4

Hexagonal lattice maximal isotropy subgroups: $u \in \mathbf{C}$; $v = e^{i\frac{\pi}{6}}$. *The general fixed-point subspace of* k*-groups is* $\mathbf{R}\{(v^{3k}, v^{3k}, v^{7k}, v^{11k}, v^{11k}, v^{7k})\}$.

	Generators	Fixed subspace	Dim	Normalizer
$k=0$	κ, ξ	$\mathbf{R}\{(1,1,1,1,1,1)\}$	1	$[0,0,\frac{\pi}{2}]$
$k=1$	$\kappa, [0,0,\frac{\pi}{6}]\xi$	$\mathbf{R}\{(i,i,-v,\overline{v},\overline{v},-v)\}$	1	$[0,0,\frac{\pi}{2}]$
$k=2$	$\kappa, [0,0,\frac{\pi}{3}]\xi$	$\mathbf{R}\{(1,1,-v^2,iv,iv,-v^2)\}$	1	$[0,0,\frac{\pi}{2}]$
$k=3$	$\kappa, [0,0,\frac{\pi}{2}]\xi$	$\mathbf{R}\{(i,i,i,i,i,i)\}$	1	$[0,0,\frac{\pi}{2}]$
$k=4$	$\kappa, [0,0,\frac{2\pi}{3}]\xi$	$\mathbf{R}\{(1,1,iv,-v^2,-v^2,iv)\}$	1	$[0,0,\frac{\pi}{2}]$
$k=5$	$\kappa, [0,0,\frac{5\pi}{6}]\xi$	$\mathbf{R}\{(i,i,\overline{v},-v,-v,\overline{v})\}$	1	$[0,0,\frac{\pi}{2}]$
7	$\kappa\xi^3, [0,\theta_2,0], [\theta_1,0,-\pi\theta_1]$	$\mathbf{R}\{(1,0,0,0,0,0)\}$	1	$[\frac{1}{2},0,0]$
8	$\kappa, \xi^3, [0,\theta_2,0], [\frac{1}{2},0,\frac{\pi}{2}]$	$\mathbf{R}\{(1,1,0,0,0,0)\}$	1	$[\frac{1}{2},0,0]$
9	$\xi^2, \kappa\xi, [\frac{1}{3},\frac{1}{3},\frac{2\pi}{3}]$	$\mathbf{R}\{(1,0,1,0,1,0)\}$	1	$[0,0,\frac{\pi}{2}]$
10	$\kappa\xi, [\theta_1,\theta_1,-\pi\theta_1]$	$(u,0,\overline{u},0,0,0)$	2	$[\theta_1,-\theta_1,0]$
11	$\kappa\xi^4, [\theta_1,-\theta_1,-\pi\theta_1]$	$(u,0,0,u,0,0)$	2	$[\theta_1,\theta_1,0]$
12	$\kappa\xi, \xi^3, [\frac{1}{2},\frac{1}{2},\frac{\pi}{2}]$	$(u,\overline{u},\overline{u},u,0,0)$	2	$[0,-\frac{1}{2},\frac{\pi}{4}]$

The relations among the generators are as follows:

$$\begin{aligned} \kappa\xi\kappa &= \xi^{-1}, \\ \kappa[\theta_1,\theta_2,\widehat{\phi}] &= [\theta_1,-\theta_1-\theta_2,-\widehat{\phi}]\kappa, \\ \xi[\theta_1,\theta_2,\widehat{\phi}] &= [-\theta_2,\theta_1+\theta_2,\widehat{\phi}]\xi. \end{aligned} \tag{3.2}$$

Remark 3.1. *It follows from the action of* $[0,0,\widehat{\phi}]$ *and Lemma* 3.1 *that the action of* $\widetilde{\Gamma}$ *on* $\mathbf{C}^6$ *is absolutely irreducible.*

3.3. Solutions corresponding to maximal isotropy subgroups. Up to conjugacy there are (at least) 12 maximal isotropy subgroups of the action of $\widetilde{\Gamma}$ on $W_1 \cong \mathbf{C}^6$, and these are listed in Table 4. We believe that it is unlikely that there are additional maximal isotropy subgroups, but we have not been able to give a complete proof of this conjecture (see section 4.7). Nine of the maximal isotropy subgroups are axial (including a family of six that are isomorphic to $\mathbf{D}_6$); hence the Equivariant Branching Lemma [14] proves the existence (generically) of a branch of equilibria for each of them.

The other three maximal isotropy subgroups have a two-dimensional fixed-point space. Two of them have as their normalizer a circle group, and generically they lead to a rotating wave just as in the square lattice Σ_4 case. See section 4.4. The remaining one has $\mathbf{Z}_4$ as its normalizer and can lead either to equilibria or to time-periodic discrete rotating waves, as discussed in section 3.4.7.

The cortical and retinal planforms associated to each of these maximal isotropy subgroups are presented in Figures 16–28. In these figures, we use the eigenfunctions associated to points listed in Table 4. In addition, where conjugate points lead to different retinal images, we

Table 5

Conjugacy classes of symmetry-broken states represented by the point $V \in \mathbf{C}^6$, where $v = e^{i\pi/6}$. Scalar and pseudoscalar refer to states in [5]. $\Sigma_V \subset \Gamma$ is the isotropy group of the point V. ℓ denotes the number of conjugacy classes in Γ of each state.

k	V	Comment	Σ_V	ℓ
0	$(1,1,1,1,1,1)$	scalar equilibrium	$\mathbf{D}_6(\kappa,\xi)$	1
	$(-1,-1,-1,-1,-1,-1)$	scalar equilibrium	$\mathbf{D}_6(\kappa,\xi)$	1
1	$(i,i,-v,\overline{v},\overline{v},-v)$	rotating wave	$\mathbf{D}_1(\kappa)$	6
	$(v^2,iv,-1,1,-iv,-v^2)$	rotating wave	$\mathbf{D}_1(\kappa\xi)$	6
2	$(1,1,-v^2,iv,iv,-v^2)$	equilibrium	$\mathbf{D}_2(\kappa,\xi^3)$	3
	$(-1,-1,v^2,-iv,-iv,v^2)$	equilibrium	$\mathbf{D}_2(\kappa,\xi^3)$	3
3	(i,i,i,i,i,i)	scalar equilibrium	$\mathbf{D}_3(\kappa,\xi^2)$	2
	$(1,-1,1,-1,1,-1)$	pseudoscalar equilibrium	$\mathbf{D}_3(\kappa\xi,\xi^2)$	2
4	$(1,1,iv,-v^2,-v^2,iv)$	equilibrium	$\mathbf{D}_2(\kappa,\xi^3)$	3
	$(-1,-1,-iv,v^2,v^2,-iv)$	equilibrium	$\mathbf{D}_2(\kappa,\xi^3)$	3
5	$(i,i,\overline{v},-v,-v,\overline{v})$	rotating wave	$\mathbf{D}_1(\kappa)$	6
	$(v^2,iv,-iv,-v^2,-1,1)$	rotating wave	$\mathbf{D}_1(\kappa\xi^5)$	6

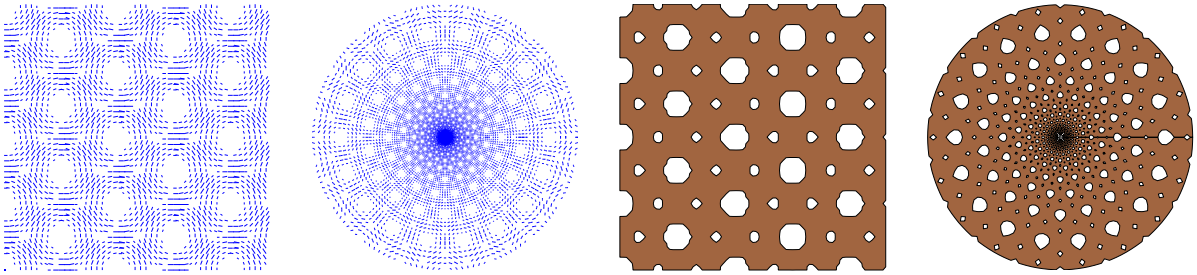


Figure 16. *Line field/thresholding of cortex/retinal images for planform $k = 0$.*

graph the planforms associated to these conjugate points. (In the figure captions, we indicate which group element produces the conjugacy.) Recall the discussion about retinal planforms and conjugate cortical planforms in section 1. Finally, we note that movies (time-periodic solutions) are associated to framed retinal images in Figures 17, 21, and 25–28.

3.4. The effects of weak anisotropy. Next we assume that the system is weakly anisotropic. That assumption is equivalent to assuming weak symmetry-breaking in the equations and generates more complicated dynamical descriptions corresponding to each maximal isotropy subgroup. We list these in turn.

3.4.1. The family $\Sigma_{k+1} = \mathbf{D}_6(\kappa, [0,0,\frac{k\pi}{6}]\xi)$, $k = 0,\dots,5$. The results for these states are summarized in Table 5, and the details are found in section 4.8.1. See Figures 16–21.

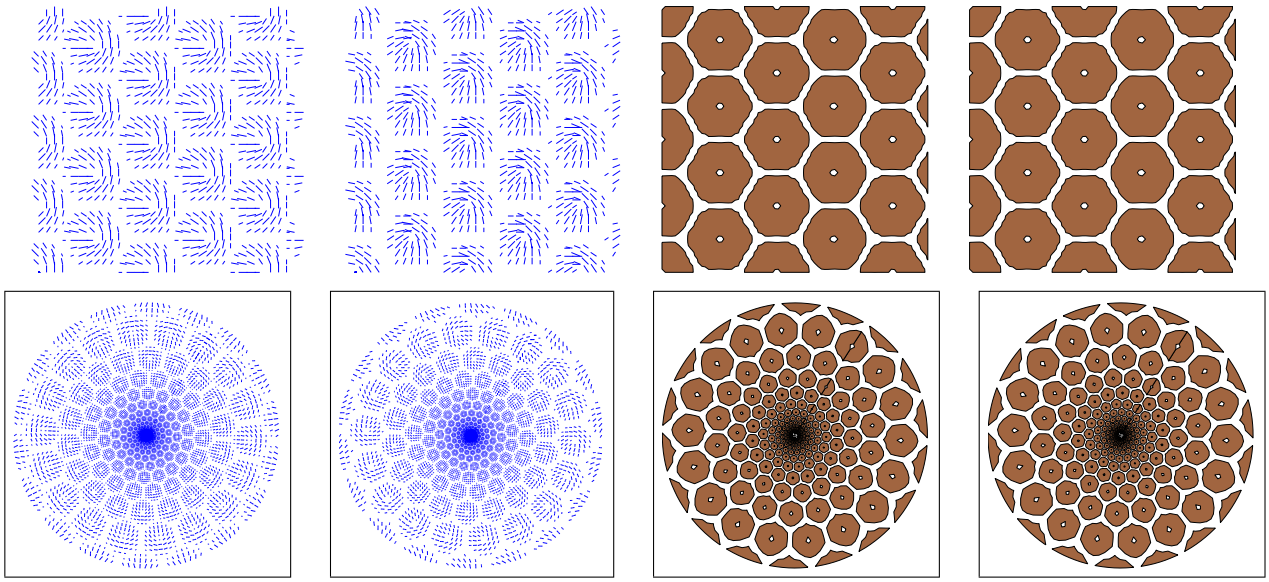


Figure 17. *Line field/thresholding of cortex/retinal images for planform $k = 1$. Second and fourth columns with ξ action; note that thresholded images with and without ξ action are identical. Movies are associated to framed images.*

3.4.2. $\Sigma_7 = \langle \kappa\xi^3, [0, \theta_2, 0], [\theta_1, 0, -\pi\theta_1] \rangle$. Steady states corresponding to Σ_7 persist as steady states under symmetry-breaking perturbations of the system and are conjugate to $T_1 = (1, 0, 0, 0, 0, 0)$. Details are found in section 4.8.2. Note that the eigenfunction for planform 7 is

$$\cos(2\phi + 2\pi \mathbf{k}_1 \cdot \mathbf{x}),$$

whose maximum value is 1 for every $\mathbf{x}$. It follows that sensible contouring cannot be obtained directly from the linear eigenfunction, and we present nonthresholded line field pictures in Figure 22. The same point is valid for the Σ_4 solutions on the square lattice.

3.4.3. $\Sigma_8 = \langle \kappa, \xi^3, [0, \theta_2, 0], [\frac{1}{2}, 0, \frac{\pi}{2}] \rangle$. In the anisotropic case, there are two conjugacy classes of equilibria corresponding to type Σ_8, namely, $(1, 1, 0, 0, 0, 0)$ and $(1, -1, 0, 0, 0, 0)$. These correspond to scalar and pseudoscalar rolls. Generically, there may be intermediate equilibria, and all dynamics on this group orbit converge to one of these equilibria. Details are found in section 4.8.3. See Figure 23.

3.4.4. $\Sigma_9 = \langle \xi^2, \kappa\xi, [\frac{1}{3}, \frac{1}{3}, \frac{2\pi}{3}] \rangle$. The dynamics on this group orbit is complicated to describe. We prove that there are at least two conjugacy classes of equilibria corresponding to $\pm(1, 0, 1, 0, 1, 0)$. Details are found in section 4.8.4. See Figure 24.

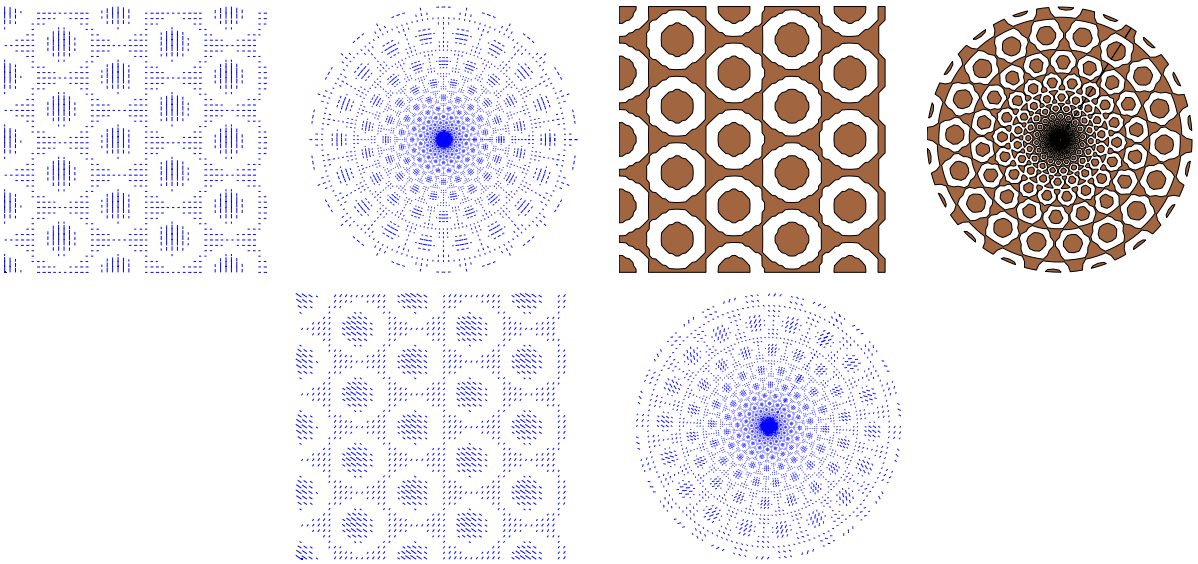


Figure 18. *Line field/thresholding of cortex/retinal images for planform $k = 2$. (Bottom) line fields with ξ action.*

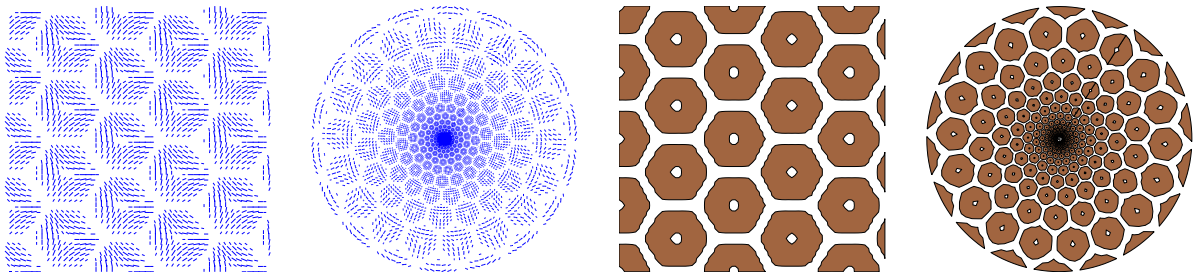


Figure 19. *Line field/thresholding of cortex/retinal images for planform $k = 3$.*

3.4.5. $\Sigma_{10} = \langle \kappa\xi, [\theta_1, \theta_1, -\pi\theta_1] \rangle$. In the isotropic case, solutions corresponding to the maximal isotropy subgroup of type Σ_{10} can be expected to be a rotating wave, and the same is true in the anisotropic case. Up to conjugacy the rotating wave is $[\theta_1, -\theta_1, 0](1, 0, 1, 0, 0, 0)$. Details are found in section 4.8.5. See Figure 25.

3.4.6. $\Sigma_{11} = \langle \kappa\xi^4, [\theta_1, -\theta_1, -\pi\theta_1] \rangle$. In the isotropic case, solutions corresponding to the maximal isotropy subgroup of type Σ_{11} can be expected to be a rotating wave, and the same is true in the anisotropic case. Up to conjugacy the rotating wave is $[\theta_1, \theta_1, 0](1, 0, 0, 1, 0, 0)$. Details are found in section 4.8.6. See Figure 26.

3.4.7. $\Sigma_{12} = \langle \kappa\xi, \xi^3, [\frac{1}{2}, \frac{1}{2}, \frac{\pi}{2}] \rangle$. When symmetry breaks, equilibria and discrete rotating waves are possible. See Krauskopf [18, 19] and sections 4.7.3 and 4.8.7. See Figures 27 and 28 for possible equilibrium planforms.

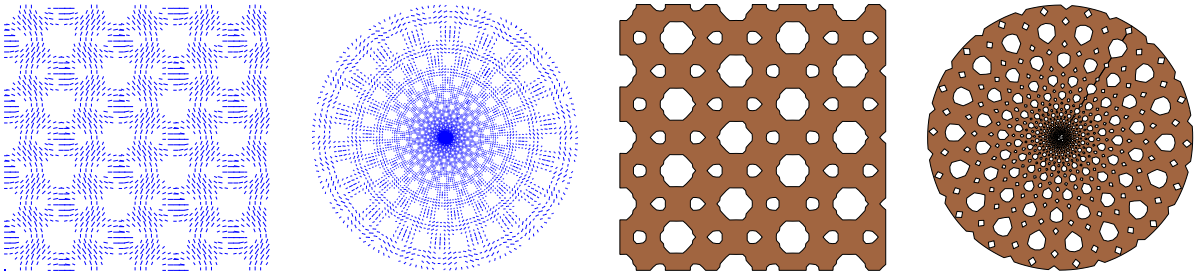


Figure 20. *Line field/thresholding of cortex/retinal images for planform $k = 4$.*

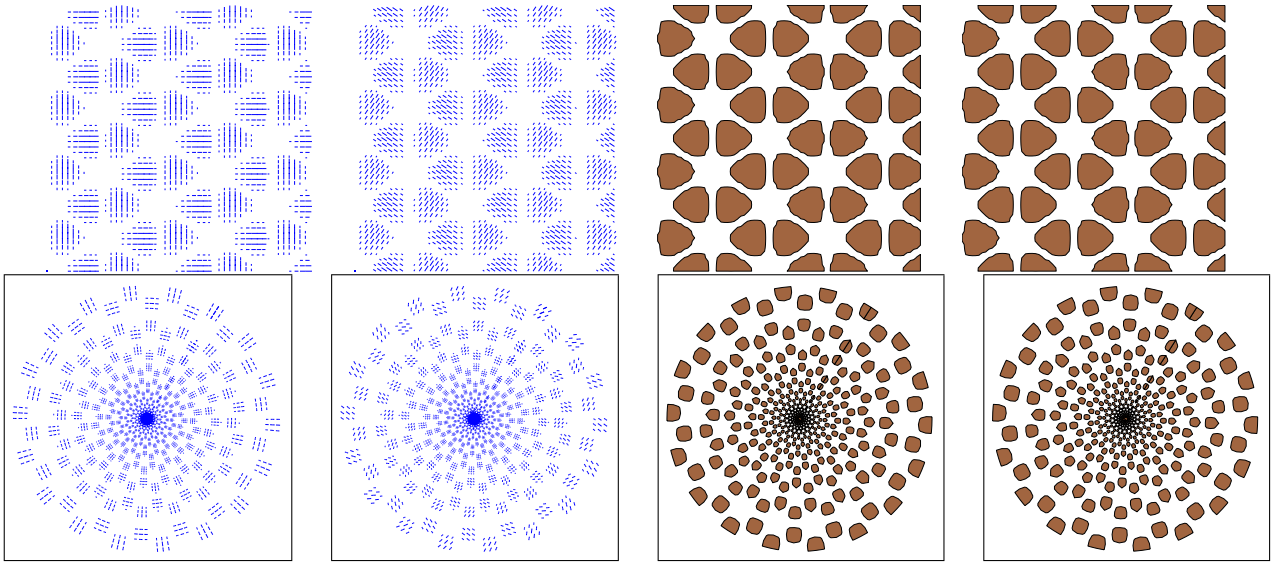


Figure 21. *Line field/thresholding of cortex/retinal images for planform $k = 5$. Second and fourth columns with ξ action; note that thresholded images with and without ξ action are identical. Movies are associated to framed images.*

Note that in a system of PDEs (or in an integro-differential equation such as a Wilson–Cowan equation), a typical solution will be a function of both space and time. In such systems, a discrete rotating wave is a solution whose shape in space changes periodically in time; that is, the associated planform at different times need not be symmetry-related. The qualitative features of a movie of a discrete rotating wave are quite different from those of a (continuous) rotating wave.

4. Proofs of lattice results.

In this section, we verify the results stated in sections 2 and 3.

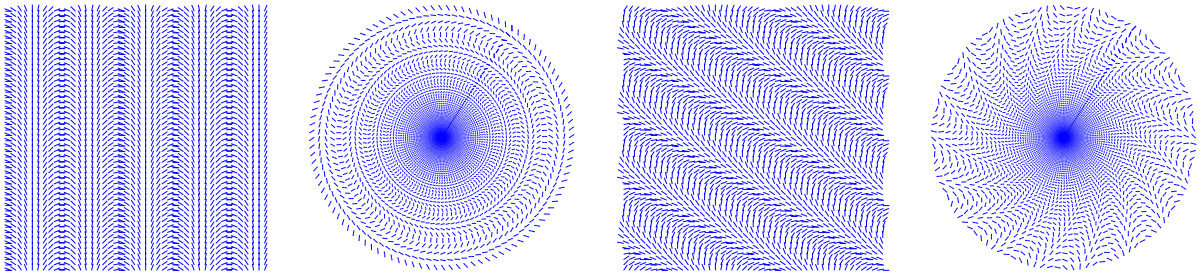


Figure 22. *Line fields of cortex/retinal images for planform* 7. *(Second and fourth panels) line fields with* ξ *action. The maximum value of* $a(\mathbf{x}, \cdot)$ *is constant in* $\mathbf{x}$ *for these planforms, so thresholding is not relevant.*

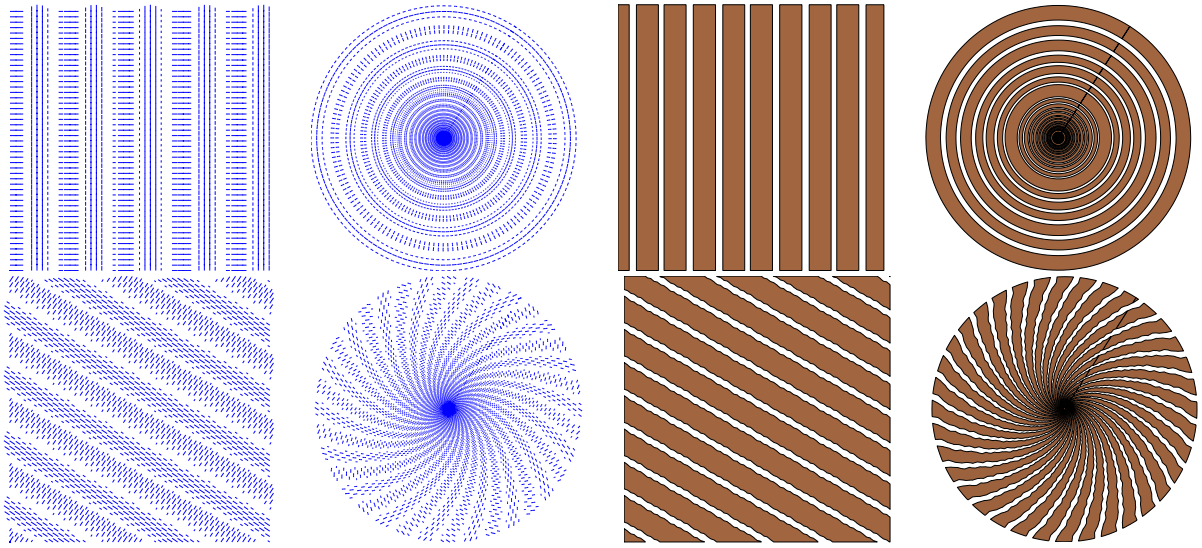


Figure 23. *Line field/thresholding of cortex/retinal images for planform* 8. *Bottom with* ξ *action.*

4.1. Verification of (2.1). Without loss of generality, we may assume that the planar square lattice $\mathcal{L}$ is the integer lattice and that the dual lattice $\mathcal{L}^*$ is generated by

$$\mathbf{k}_1 = (1, 0) \quad \text{and} \quad \mathbf{k}_2 = (0, 1).$$

Using Fourier series, we may write each function $f \in \mathcal{F}_{\mathcal{L}}$ as

$$f(\mathbf{x}, \phi) = \sum_{m,\mathbf{k}} z_{m,\mathbf{k}} e^{2im\phi} e^{2\pi i \mathbf{k}\cdot\mathbf{x}} + c.c.,$$

where $m \geq 0$, $\mathbf{k} \in \mathbf{Z}^2$, and $z_{m,\mathbf{k}} \in \mathbf{C}$. The functions corresponding to wave vectors of constant length k, namely,

$$V_k = \left\{ \sum_{m,|\mathbf{k}|=k} z_{m,\mathbf{k}} e^{2im\phi} e^{2\pi i \mathbf{k}\cdot\mathbf{x}} + c.c. \right\},$$

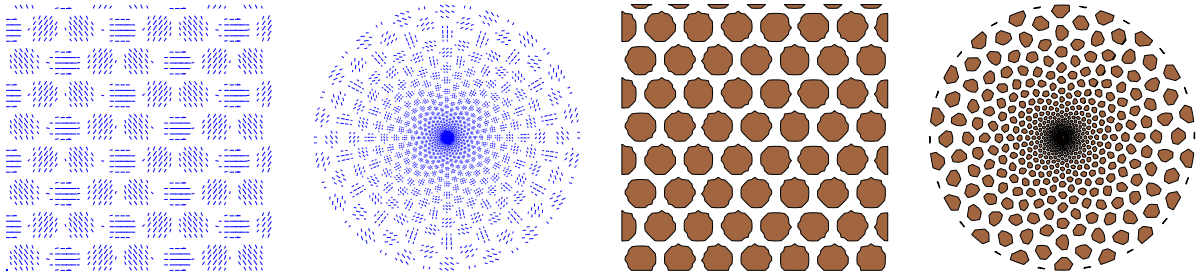


Figure 24. *Line field/thresholding of cortex/retinal images for planform* 9.

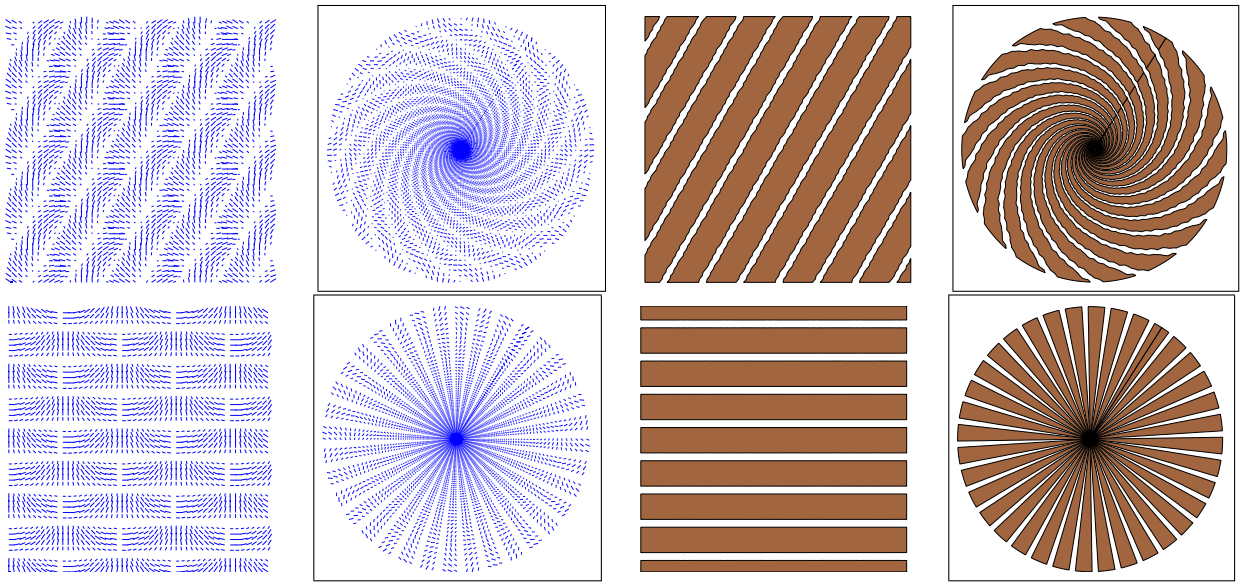


Figure 25. *Line field/thresholding of cortex/retinal images for planform* 10. *Bottom with* ξ^2 *action. Movies are associated to framed images.*

are $\widetilde{\Gamma}$-invariant subspaces. In this paper, we consider only steady-state bifurcations whose critical eigenspaces are irreducible subspaces in V_1, that is, those eigenfunctions corresponding to wave vectors of shortest length.

4.1.1. The irreducible subspaces of $V_1 \subset \mathcal{F}_{\mathcal{L}}$. The group $\widetilde{\Gamma}$ contains a 3-torus $\mathbf{T}^3$ generated by translations $\mathbf{y} \in \mathbf{T}^2$ and rotations $\widehat{\phi} \in \mathbf{S}^1$. For each $\mathbf{k} \in \mathcal{L}^*$ and $m \in \mathbf{Z}$, the two-dimensional subspaces

$$W_{\mathbf{k},m} = \{ze^{2im\phi}e^{2\pi i\mathbf{k}\cdot\mathbf{x}} + c.c. : \ z \in \mathbf{C}\} \subset \mathcal{F}_{\mathcal{L}}$$

are distinct irreducible representations of $\mathbf{T}^3$. Counterclockwise rotation ξ through angle $\pi/2$ and reflection κ across the horizontal axis generate $\widetilde{\Gamma}$ over $\mathbf{T}^3$, and

$$\theta W_{\mathbf{k},m} = W_{\theta\mathbf{k},m} \quad \text{and} \quad \kappa W_{\mathbf{k},m} = W_{\kappa\mathbf{k},-m}$$

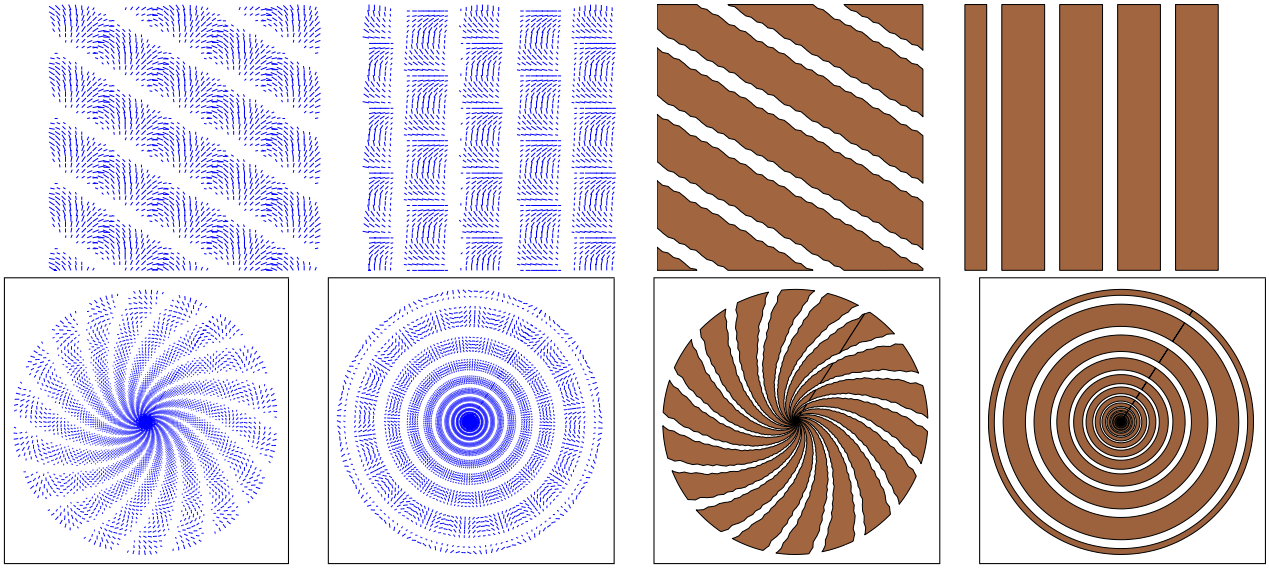


Figure 26. *Line field/thresholding of cortex/retinal images for planform* 11. *Second and fourth columns with* ξ^2 *action. Movies are associated to framed images.*

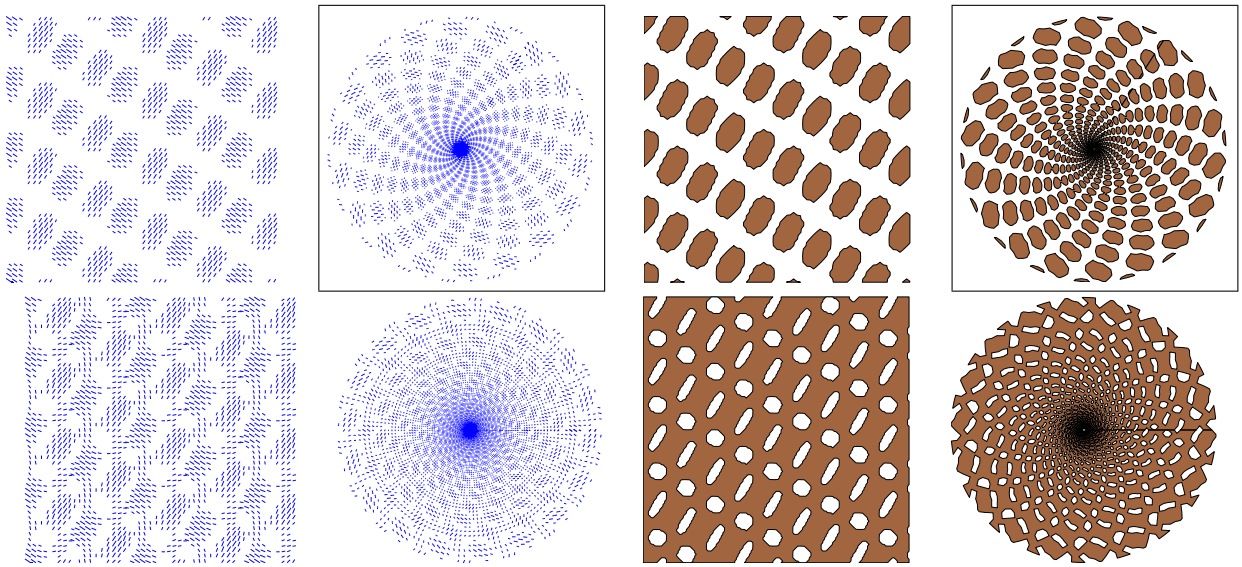


Figure 27. *Line field/thresholding of cortex/retinal images for planform* 12*: (Top)* $u = 1$ *and (bottom)* $u = 0.5i$*. Movies are associated to framed images.*

for any $\theta \in \mathbf{SO}(2)$. Finally,

$$W_{-\mathbf{k},-m} = W_{\mathbf{k},m}.$$

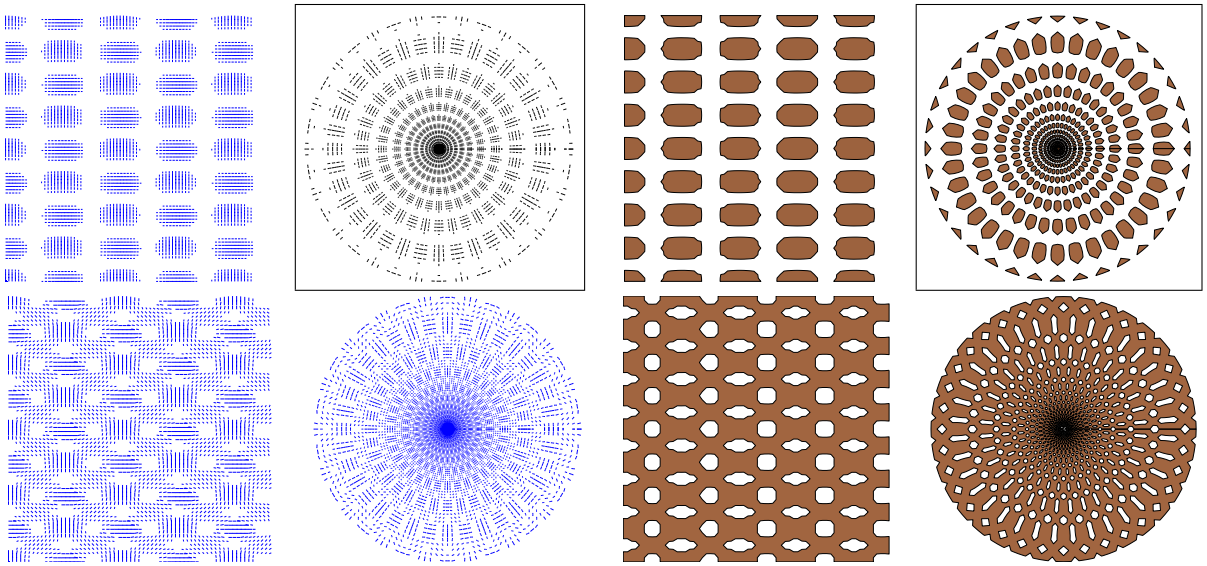


Figure 28. *Line field/thresholding of cortex/retinal images for planform* 12 *with* ξ^2 *action: (Top)* $u = 1$ *and (bottom)* $u = 0.5i$*. Movies are associated to framed images.*

Now fix $\mathbf{k} = \mathbf{k}_1$, and denote

$$W_m = \sum_{\gamma \in \mathbf{D}_4} \gamma W_{\mathbf{k}_1, m},$$

which is $\widetilde{\Gamma}$-invariant. The subspace W_0 is four-dimensional, and the subspaces W_m are eight-dimensional when $m > 0$. Indeed, W_m consists of all functions of the form (2.1). For $m > 0$, we identify $a(\mathbf{x}, \phi) \in W_m$ with $(z_1, w_1, z_2, w_2) \in \mathbf{C}^4$.

Lemma 4.1. *The subspace* $W_m \subset \mathcal{F}_{\mathcal{L}}$ *is an absolutely irreducible representation of* $\widetilde{\Gamma}$.

Proof. The case in which $m = 0$ is established by Lemma 3.1 in [5]. For $m > 0$, write $W_m = W_m^+ \oplus W_m^-$ as a sum of even and odd functions of ϕ (that is, $\cos(2m\phi)$ and $\sin(2m\phi)$). The summands are absolutely irreducible and nonisomorphic (look at the action of κ) for $\mathbf{D}_4 \dot{+} \mathbf{T}^2$ by Lemma 3.1 in [5]. Thus any matrix commuting with $\widetilde{\Gamma}_{\mathcal{L}}$ on W_m is block diagonal, with each diagonal block being a real multiple of the identity. Now the extra $\mathbf{S}^1$-symmetry forces the two diagonal blocks to be equal. (Consider specifically $[0, 0, \frac{\pi}{4}]$, which interchanges W_m^+ and W_m^-.) ∎

4.2. Group action and isotropy subgroups. Fix $m > 0$. The action of $\widetilde{\Gamma}$ is given in Table 6, where we denote elements in the 2-torus $\mathbf{T}^2$ of translations by $[\theta_1, \theta_2]$. Recall that $0 \le \theta_j < 1$, since the lattice is 1-periodic, $\widehat{\phi}$ is π-periodic, and $a \in W_m$ given by (2.1) is identified with $(z_1, w_1, z_2, w_2) \in \mathbf{C}^4$.

Table 6

Group action on $(z_1, w_1, z_2, w_2) \in W_m = \mathbf{C}^4$; $\xi, \kappa \in \mathbf{D}_4$, $0 \le \theta_1, \theta_2 < 1$, *and* $0 \le \widehat{\phi} < \pi$.

Generators	Action
ξ	$(\overline{w}_2, \overline{z}_2, z_1, w_1)$
κ	$(w_1, z_1, \overline{z}_2, \overline{w}_2)$
$[\theta_1, \theta_2, 0]$	$(e^{2\pi i\theta_1} z_1, e^{2\pi i\theta_1} w_1, e^{2\pi i\theta_2} z_2, e^{2\pi i\theta_2} w_2)$
$[0, 0, \widehat{\phi}]$	$(e^{-2im\widehat{\phi}} z_1, e^{2im\widehat{\phi}} w_1, e^{-2im\widehat{\phi}} z_2, e^{2im\widehat{\phi}} w_2)$

Table 7

$u, v \in \mathbf{C}$, $a, b \in \mathbf{R}$.

	Normal form	Generators	Fixed subspace	Dim
0	$(0,0,0,0)$	$\widetilde{\Gamma}$	0	0
1	$(1,1,1,1)$	κ, ξ	$\mathbf{R}\{(1,1,1,1)\}$	1
2	$(1,1,1,-1)$	$\kappa, \left[\frac{3}{4}, \frac{1}{4}, \frac{\pi}{4}\right] \xi$	$\mathbf{R}\{1,1,1,-1\}$	1
3	$(1,1,0,0)$	$\kappa, \xi^2, [0, \theta_2, 0]$	$\mathbf{R}\{(1,1,0,0)\}$	1
4	$(1,0,0,0)$	$\kappa\xi^2, [0, \theta_2, 0], [\theta_1, 0, \pi\theta_1]$	$\mathbf{R}\{(1,0,0,0)\}$	1
5	$(1,0,1,0)$	$\kappa\xi, [\theta_1, \theta_1, \pi\theta_1]$	$\{(u, 0, \overline{u}, 0\}$	2
6	$(1,1,b,b)$	κ, ξ^2	$\{(a,a,b,b)\}$	2
7	$(1,1,-b,b)$	$\kappa, \left[0, \frac{1}{2}, 0\right] \xi^2$	$\{(a,a,-b,b)\}$	2
8	$(i,-i,-i,i)$	$\kappa\xi, \xi^2$	$\{(u,\overline{u},\overline{u},u)\}$	2
9	$(1,b,0,0)$	$\kappa\xi^2, [0, \theta_2, 0]$	$\{(a,b,0,0)\}$	2
10	$(1,1,b,b)$	ξ^2	$\{(u,\overline{u},v,\overline{v})\}$	4
11	$(1,1,a,b)$	κ	$\{(u,u,a,b)\}$	4
12	$(1,b,1,b)$	$\kappa\xi$	$\{(u,v,\overline{u},\overline{v})\}$	4
13	$(1,0,b,0)$	$[\theta_1, \theta_1, \pi\theta_1]$	$\{(u,0,v,0)\}$	4
14		**1**	$\mathbf{C}^4$	8

Note that the action always has a nontrivial kernel $\mathbf{Z}_{4m}$ generated by the element

$$\left[\theta_1, \theta_2, \widehat{\phi}\right] = \left[\frac{1}{2}, \frac{1}{2}, \frac{\pi}{2m}\right].$$

In order to simplify the presentation, we will not include this kernel explicitly in the isotropy subgroups. Indeed, the bifurcation analysis is identical for all $m \ge 1$. (Just factor out the kernel.) The case in which $m = 0$ was considered in [5]. Hence, without loss of generality, we assume that $m = 1$. The assumption that $m = 1$ does make a difference in the planforms—but in a very controllable way. The case in which $m \ge 2$ differs from $m = 1$ by the fact that the activity function $a(\mathbf{x}, \cdot)$ restricted to the unit circle at $\mathbf{x}$ has period π/m instead of π. Therefore, the maxima of this activity variable occur simultaneously at m points on that circle rather than at a single point.

Using the group action given in Table 6, we can compute the lattice of isotropy subgroups

(up to conjugacy) of the action of $\widetilde{\Gamma}$ on W_1. The results for $m = 1$ are given in Table 7.

We verify that up to conjugacy the maximal subgroups are those shown in Table 7. Let $z = (z_1, w_1, z_2, w_2)$. Our strategy is based on the fact that conjugate points have conjugate isotropy subgroups. Moreover, it is convenient to note that multiplying z by a nonzero real number does not change its isotropy subgroup. So we can conjugate z by elements of $\widetilde{\Gamma}$ and scale z to put z into a "normal form." Once we have the normal form, we compute generators for the isotropy subgroups of that normal form.

There is a useful remark that concerns conjugating z with 3-torus $\mathbf{T}^3 = \{[\theta_1, \theta_2, \widehat{\phi}]\}$ elements: z can be conjugated by an element in $\mathbf{T}^3$ so that any three of its nonzero coordinates are real and positive.

We discuss only maximal isotropy subgroups: cases **0** to **5**. First, we classify the isotropy subgroups of z when z has some of its coordinates equal to zero. If $z = 0$, then its isotropy subgroup is $\widetilde{\Gamma}$, which is case **0**. If three of the coordinates of z are zero, then z is conjugate to $(1, 0, 0, 0)$, and we have case **4**. If two of the coordinates of z are zero, then we can assume after conjugacy that the other two coordinates are real and positive. After conjugacy, by elements in $\mathbf{D}_4$ we can assume that the fourth coordinate and either the second or the third coordinate is zero. Thus, after scaling, z has the normal form $(1, 0, b, 0)$ or $(1, b, 0, 0)$, where $b \geq 1$. If $b = 1$, then we have cases **5** and **3**. (If $b > 1$, then we have cases **13** and **9**, which are not maximal.)

Next, we assume that all coordinates of z are equal in modulus. After scaling, we assume that all coordinates have modulus one, and after conjugating by an element in $\mathbf{T}^3$ we can assume that $z = (1, 1, 1, e^{2\pi i\rho})$. Observe that $\kappa z = (1, 1, 1, e^{-2\pi i\rho})$. After conjugacy, it follows that we may assume $0 \leq \rho \leq \frac{1}{2}$. We consider three possibilities: $\rho = 0$, $\rho = \frac{1}{2}$, and $0 < \rho < \frac{1}{2}$. Note that $\rho = 0$ is $z = (1, 1, 1, 1)$, which is case **1**. When $\rho = \frac{1}{2}$, then $z = (1, 1, 1, -1)$; the isotropy subgroup is $\langle \kappa, [\frac{3}{4}, \frac{1}{4}, \frac{\pi}{4}]\xi \rangle \cong \mathbf{D}_4$, which is case **2**. Finally, when $0 < \rho < \frac{1}{2}$; the isotropy subgroup is $\langle [0, \rho, 0]\xi^2, [\frac{\rho}{2}, \frac{\rho}{2}, \frac{\pi\rho}{2}]\kappa\xi \rangle$, which is conjugate to $\langle \kappa\xi, \xi^2 \rangle$ by $[0, -\frac{\rho}{2}, -\frac{\pi\rho}{4}]$. This is case **8**, which is not maximal.

Finally, all z whose coordinates are nonzero and not of equal modulus have isotropy subgroups that are not maximal.

4.3. Equivariants of $\widetilde{\Gamma}$ acting on W_1 in square lattice. Let $F : \mathbf{C}^4 \to \mathbf{C}^4$ be a $\widetilde{\Gamma}$-equivariant polynomial mapping. We can write the form of F in terms of invariant generators using standard invariant theory, and we do so in Theorem 4.1. The proof of this theorem uses standard techniques and is not presented here. For background, see [14].

In complex coordinates, we can write F as

$$F = (Z_1, W_1, Z_2, W_2). \tag{4.1}$$

Using the $\mathbf{D}_4$-equivariance of F (specifically $\kappa, \xi^3, \kappa\xi^3$), we see that

$$\begin{aligned} W_1(z_1, w_1, z_2, w_2) &= Z_1(w_1, z_1, \overline{z}_2, \overline{w}_2), \\ Z_2(z_1, w_1, z_2, w_2) &= Z_1(z_2, w_2, \overline{w}_1, \overline{z}_1), \\ W_2(z_1, w_1, z_2, w_2) &= Z_1(w_2, z_2, w_1, z_1). \end{aligned} \tag{4.2}$$

In addition, Z_1 must satisfy $\mathbf{T}^3$-invariance conditions

$$Z_1(e^{2\pi i\theta_1} z_1, e^{2\pi i\theta_1} w_1, z_2, w_2) = e^{2\pi i\theta_1} Z_1(z_1, w_1, z_2, w_2), \tag{4.3}$$

$$Z_1(z_1, w_1, e^{2\pi i\theta_2} z_2, e^{2\pi i\theta_2} w_2) = Z_1(z_1, w_1, z_2, w_2), \tag{4.4}$$

$$Z_1(z_1, e^{2\pi i\widehat{\phi}} w_1, z_2, e^{2\pi i\widehat{\phi}} w_2) = Z_1(z_1, w_1, z_2, w_2) \tag{4.5}$$

as well as

$$Z_1(\overline{z}_1, \overline{w}_1, w_2, z_2) = \overline{Z_1(z_1, w_1, z_2, w_2)}. \tag{4.6}$$

Here we take as $\mathbf{T}^3$ generators the circles $(\theta_1, 0, 0)$, $(0, \theta_2, 0)$, and $(\widehat{\phi}/2, \widehat{\phi}/2, \pi\widehat{\phi}/2)$.

Theorem 4.1. *The invariance conditions* (4.3)–(4.5) *and* (4.6) *imply that the polynomial* Z_1 *has the form*

$$\begin{aligned} Z_1 = {} & C(|z_1|^2, |w_1|^2, |z_2|^2 + |w_2|^2, i(|z_2|^2 - |w_2|^2), z_1\overline{w}_1\overline{z}_2 w_2) z_1 \\ & + D(|z_1|^2, |w_1|^2, |z_2|^2 + |w_2|^2, i(|z_2|^2 - |w_2|^2), \overline{z}_1 w_1 z_2 \overline{w}_2) w_1 z_2 \overline{w}_2, \end{aligned}$$

where C *and* D *are polynomials with real coefficients. Moreover, this form is unique.*

4.4. Existence of time-periodic rotating waves. We claim that generically there is a branch of time-periodic rotating wave solutions corresponding to the maximal isotropy subgroup Σ_5; moreover, the period of these solutions tends to infinity at the bifurcation point. To verify these statements, we must use the $\widetilde{\Gamma}$-equivariant polynomials on $\mathbf{C}^4$.

We assume that (4.1), with bifurcation parameter λ, is obtained by a center manifold reduction so that the asymptotic dynamics of the differential equation is reproduced by (4.1). Observe that (4.1) restricted to $\mathrm{Fix}(\Sigma_5)$ is given by

$$\begin{aligned} \frac{du}{dt} &= Z_1(u, 0, \overline{u}, 0, \lambda) \\ &= C(|u|^2, 0, |u|^2, i|u|^2, 0, \lambda) u \\ &= \left(c_\lambda \lambda + (c_1 + c_3)|u|^2 + c_4 i|u|^2 + \cdots\right) u, \end{aligned} \tag{4.7}$$

where

$$C = c_1|z_1|^2 + c_2|w_1|^2 + c_3(|z_2|^2 + |w_2|^2) + c_4 i(|z_2|^2 - |w_2|^2) + c_\lambda \lambda + \cdots \tag{4.8}$$

and $c_1, c_2, c_3, c_4, c_\lambda$ are real constants. Write $u = \rho e^{i\tau}$ in polar coordinates, and write (4.7) as phase amplitude equations (to third order) obtaining

$$\frac{d\rho}{dt} = (c_\lambda \lambda + (c_1 + c_3)\rho^2)\rho,$$
$$\frac{d\tau}{dt} = c_4 \rho^2.$$

Assume that the *eigenvalue crossing condition*

$$c_\lambda \neq 0 \tag{4.9}$$

is valid. Then there exists an invariant circle for the dynamics (given by a zero to the amplitude equation) at

$$\lambda = -\frac{c_1 + c_3}{c_\lambda}\rho^2,$$

and the invariant is a periodic solution if the nondegeneracy condition

$$c_4 \neq 0$$

holds. Note that the frequency of the periodic solution is $c_4\rho^2/2\pi$, which goes to zero at the bifurcation point.

4.5. Stability of maximal isotropy solutions. The following computations were done using Mathematica for the equivariant F determined by Theorem 4.1, where

$$\begin{aligned} &C(|z_1|^2, |w_1|^2, |z_2|^2 + |w_2|^2, i(|z_2|^2 - |w_2|^2), z_1\overline{w}_1\overline{z}_2 w_2, \lambda) \\ &\quad = c_1|z_1|^2 + c_2|w_1|^2 + c_3(|z_2|^2 + |w_2|^2) + c_4 i(|z_2|^2 - |w_2|^2) + c_\lambda\lambda + \cdots \\ &D(|z_1|^2, |w_1|^2, |z_2|^2 + |w_2|^2, i(|z_2|^2 - |w_2|^2), \overline{z}_1 w_1 z_2 \overline{w}_2, \lambda) \\ &\quad = d_0 + d_\lambda\lambda + \cdots . \end{aligned} \tag{4.10}$$

Table 8 describes necessary and sufficient conditions for the orbital asymptotic stability of $\widetilde{\Gamma}$-orbits of fixed points and periodic orbits corresponding to the maximal isotropy subgroups of the action on the square lattice near bifurcation. This stability refers to perturbations in the class of spatially periodic functions on the square lattice. These conditions were obtained by computing the eigenvalues of the differential of a $\widetilde{\Gamma}$-equivariant vector field (to lowest order in the bifurcation parameter) and requiring that those eigenvalues that are not forced to be zero by symmetry have negative real part.

Table 8

Stability of maximal isotropy types for the square lattice; $dZ/dt = F(Z,\lambda)$, with F determined by (4.10). Numbering of isotropy groups is as in Table 7.

Group	Solution	Stability conditions			
1	equilibrium	$c_1 - c_2 - d_0 < 0$	$c_1 + c_2 \pm (2c_3 + d_0) < 0$	$d_0 > 0$	
2	equilibrium	$c_1 - c_2 + d_0 < 0$	$c_1 + c_2 \pm (2c_3 - d_0) < 0$	$d_0 < 0$	
3	equilibrium	$c_1 \pm c_2 < 0$	$-c_1 - c_2 + 2c_3 \pm d_0 < 0$		
4	equilibrium	$c_2 - c_1 < 0$	$c_3 - c_1 < 0$	$c_1 < 0$	
5	periodic	$c_2 - c_1 \pm d_0 < 0$	$c_1 \pm c_3 < 0$	$c_\lambda > 0$	$c_4 \neq 0$

4.6. Effect of weak anisotropy on square lattice. The study of forced symmetry-breaking requires the following general result.

Lemma 4.2. *Let $K \subset H$ be compact subgroups in $\mathbf{O}(n)$. Assume that the H-equivariant vector field $F : \mathbf{R}^n \to \mathbf{R}^n$ has a normally hyperbolic, flow-invariant, H-invariant compact manifold S contained in* $\mathrm{Fix}(K)$. *Let $F' : \mathbf{R}^n \to \mathbf{R}^n$ be a small H-equivariant perturbation of F.*

Then F' has a unique flow-invariant manifold S' near S that is also H-invariant, contained in $\mathrm{Fix}(K)$, *diffeomorphic to S, and normally hyperbolic for F'.*

Proof. Normal hyperbolicity implies that any small perturbation of F (equivariant or not) will have a unique flow-invariant set S' that is close to (and diffeomorphic to) S. (See Hirsch, Pugh, and Shub [15].)

Since the subspace $\mathrm{Fix}(K)$ is preserved by all H-equivariant vector fields, we can apply the previous result to the vector fields restricted to $\mathrm{Fix}(K)$ and conclude that there is an F'-invariant set $S'' \subset \mathrm{Fix}(K)$ near S. Uniqueness forces $S' = S''$. Similarly, uniqueness forces S' to be H-invariant. Let $h \in H$. Then $h(S')$ is also F'-invariant and close to S (because $h(S) = S$); hence $h(S')$ must coincide with S'. ∎

4.6.1. Squares: $\Sigma_1 = \mathbf{D}_4(\kappa, \xi)$. Let $S_1 = (1,1,1,1)$. The isotropy subgroup Σ_1 of S_1 is axial, with fixed-point subspace $V_0 = \mathbf{R}\{S_1\}$. The Equivariant Branching Lemma implies that a generic $\widetilde{\Gamma}$ bifurcation with kernel W_1 has a branch of solutions $Z(\lambda) = u(\lambda)S_1$, $u(\lambda) > 0$, in V_0 with symmetry Σ_1. When discussing forced symmetry-breaking of this equilibrium, which we assume to be normally hyperbolic, we can assume that $u(\lambda) = 1$.

The equilibria that are conjugate by a $\widetilde{\Gamma}$-symmetry to S_1 fill out the 3-torus

$$\mathbf{T}^3 = \{(z_1, w_1, z_2, w_2) : |z_1| = |w_1| = |z_2| = |w_2|, z_1\overline{w}_1\overline{z}_2 w_2 = 1\},$$

and each point on $\mathbf{T}^3$ has the form $[\theta_1, \theta_2, \widehat{\phi}]S_1$. Recall that the isotropy subgroup of σv is

$\Sigma_{\sigma v} = \sigma \Sigma_v \sigma^{-1}$. A short calculation shows that

$$[\theta_1, \theta_2, \widehat{\phi}]\xi = \xi[\theta_2, -\theta_1, \widehat{\phi}],$$
$$[\theta_1, \theta_2, \widehat{\phi}]\kappa = \kappa[\theta_1, -\theta_2, -\widehat{\phi}].$$

It follows that

$$[\theta_1, \theta_2, \widehat{\phi}]\xi[-\theta_1, -\theta_2, -\widehat{\phi}] = \xi[\theta_2 - \theta_1, -\theta_2 - \theta_1, 0],$$
$$[\theta_1, \theta_2, \widehat{\phi}]\kappa[-\theta_1, -\theta_2, -\widehat{\phi}] = \kappa[0, -2\theta_2, -2\widehat{\phi}].$$

Thus, when symmetry breaks from $\widetilde{\Gamma}$ to Γ, the isotropy subgroup of $[\theta_1, \theta_2, \widehat{\phi}]S_1$ is isomorphic to $\mathbf{Z}_4(\xi)$ unless $\widehat{\phi} = 0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}$. The reason that $\widehat{\phi} = \pm\frac{\pi}{4}$ are possibilities is that $[\frac{1}{2}, \frac{1}{2}, \frac{\pi}{2}]$ is in the kernel of $\widetilde{\Gamma}$ acting on $\mathbf{C}^4$. In these exceptional cases, the isotropy subgroup of $[\theta_1, \theta_2, \widehat{\phi}]S_1$ is isomorphic to $\mathbf{D}_4(\kappa, \xi)$.

Since $\dim \mathrm{Fix}(\mathbf{D}_4) = 1$, it follows from Lemma 4.2 that both S_1 and

$$S_2 \equiv \left[\frac{1}{4}, \frac{1}{4}, \frac{\pi}{4}\right] S_1 = (1, -1, 1, -1)$$

remain as equilibria when small anisotropy is assumed. Points conjugate to S_1 in Γ are scalar squares (Esquares in [5]), and points conjugate to S_2 in Γ are pseudoscalar squares (Osquares in [5]).

The fixed-point subspace of $\mathbf{Z}_4(\xi)$ is two-dimensional and consists of vectors of the form $(u, \overline{u}, u, \overline{u})$. The intersection of this fixed-point subspace with the torus $\mathbf{T}^3$ is the circle where $|u| = 1$. It follows by conjugacy that the flow of the vector field restricted to $\mathbf{T}^3$ when the isotropy subgroup is conjugate to $\mathbf{Z}_4$ is on circles. Moreover, each circle has four equilibria, points conjugate to $\pm S_1$ in Γ and points conjugate to $\pm S_2$. Depending on the exact form of the anisotropy in model equations, it is possible for there to be additional equilibria on these circles.

4.6.2. $\Sigma_2 = \mathbf{D}_4(\kappa, [\frac{3}{4}, \frac{1}{4}, \frac{\pi}{4}]\,\xi)$. The verification that forced symmetry-breaking leads to equilibria in this case is similar in spirit to that for Σ_1 in section 4.6.1, although the details are somewhat different. We can show that all points on the $\widetilde{\Gamma}$ group orbit 3-torus of equilibria have isotropy subgroup in Γ containing a group element conjugate to ξ^2. This fact forces the Γ-equivariant dynamics on the torus to be constrained to circles given by the action of $\widehat{\phi}$. Moreover, on each of these circles, there are eight points whose Γ isotropy subgroup contains group elements conjugate to κ or $\kappa\xi$. These points are forced to remain as equilibria for Γ-equivariant perturbations. Thus, as in the Σ_1 case, the trajectories of the forced symmetry-broken flow converge to equilibria.

4.6.3. Rolls: $\Sigma_3 = \langle \kappa, \xi^2, [0, \theta_2, 0] \rangle$. The analysis of Σ_3 steady states is similar to that done in section 4.6.1 for Σ_1; therefore, we just sketch the details.

Let $R_1 = (1, 1, 0, 0)$. The isotropy subgroup Σ_3 is axial, with fixed-point space $V_0 = \mathbf{R}\{R_1\}$; hence generically a branch of steady states $Z(\lambda) = u(\lambda)(1, 1, 0, 0)$ exists. Again, when discussing forced symmetry-breaking, we may assume that $u(\lambda) = 1$.

The steady states conjugate to R_1 in $\widetilde{\Gamma}$ form two 2-tori that are conjugate by ξ. Since this symmetry is not lost when we break symmetry from $\widetilde{\Gamma}$ to Γ, it is enough to analyze one of these tori. We focus on the torus given by the orbit of R_1 under $[\theta_1, 0, \widehat{\phi}]$, the connected component of $\widetilde{\Gamma}$. This torus is contained in $\mathrm{Fix}([0, \theta_2, 0])$ and has the form

$$\mathbf{T}^2 = \{(z_1, w_1, 0, 0) : |z_1| = |w_1| = 1\}.$$

The isotropy subgroups of these points are conjugate to either Σ_3 or the index two subgroup $\langle \xi^2, [0, \theta_2, 0] \rangle$. The former leads to equilibria and the latter to flow-invariant circles on $\mathbf{T}^2$. The equilibria are conjugate either to R_1 (scalar rolls) or to $R_2 = (1, -1, 0, 0)$ (pseudoscalar rolls).

As before, each circle has at least four equilibria, points conjugate in Γ to $\pm R_1$ and $\pm R_2$. Depending on the exact form of the anisotropy in model equations, it is possible for there to be additional equilibria on these circles.

4.6.4. $\Sigma_4 = \langle \kappa\xi^2, [0, \theta_2, 0], [\theta_1, 0, \pi\theta_1] \rangle$. We claim that steady states corresponding to Σ_4 persist as steady states under symmetry-breaking perturbations of the system. Let $T_1 = (1, 0, 0, 0)$. The isotropy subgroup Σ_4 of T_1 is axial and, by the Equivariant Branching Lemma, leads generically to a branch of steady states $Z(\lambda) = u(\lambda)T_1$ with symmetry Σ_4.

The connected group orbit through T_1 is the circle $[\theta_1, 0, 0]T_1$. Since

$$\mathrm{Fix}(\langle \kappa\xi^2, [0, \theta_2, 0] \rangle) = \{(a, b, 0, 0) : a, b \in \mathbf{R}\}$$

intersects the circle in two points, these points are equilibria. Since points on this circle are conjugate in Γ, it follows that all points on the circle are equilibria. There are three conjugate circles of steady states obtained by applying κ, ξ, and $\xi\kappa$.

4.6.5. Rotating spirals: $\Sigma_5 = \langle \kappa\xi, [\theta_1, \theta_1, \pi\theta_1] \rangle$. In subsection 4.4, we showed that generically a branch of time-periodic rotating waves bifurcates in the fixed-point subspace of isotropy subgroup Σ_5. We show that these rotating waves persist when symmetry is broken to Γ.

Note that the connected group orbit that contains the rotating waves is the 2-torus $[\theta_1, \theta_2, 0](1, 0, 1, 0)$. Since the fixed-point subspace of $\kappa\xi$ intersects the torus in a circle, it follows that the flow on this torus is restricted to circles. Finally, the fact that $[\theta_1, -\theta_1, 0]$

normalizes the isotropy subgroup of $(1,0,1,0)$ forces the flow on the circles to be a rotating wave. Since these statements rely only on elements in Γ, it follows that the solutions remain rotating waves even after a symmetry-breaking perturbation.

4.7. Maximal isotropy groups for the hexagonal lattice. In this section, we will find maximal isotropy groups for the hexagonal lattice action. Namely, we will describe vectors $Z = (z_1, w_1, z_2, w_2, z_3, w_3) \in \mathbf{C}^6$ such that Σ_Z is a maximal isotropy group.

We will look at the cases determined by the number of nonzero elements of Z. Since we are interested only in the maximal isotropy groups, the following lemma is useful.

Lemma 4.3. *If Σ_Z is a maximal isotropy group, then there is a vector Z' such that $\Sigma_Z = \Sigma_{Z'}$ and all nonzero entries of Z' have the same absolute value.*

Proof. Denote $\ell = \max\{|z_i|, |w_i|\} > 0$, and let Z' be the element obtained from Z by setting to zero all entries whose absolute value is not equal to ℓ. Then, each element $\sigma \in \widetilde{\Gamma}$ that fixes Z fixes Z' as well (since the action of $\widetilde{\Gamma}$ at most permutes the set of absolute values of the entries of Z; see Table 3); that is, $\Sigma_Z \subset \Sigma_{Z'}$. If Σ_Z is maximal, then the inclusion has to be equality. ∎

Denote by $N(Z)$ the number of nonzero entries of Z. By the previous lemma, we can assume that each nonzero element has absolute value equal to 1. Notice that the action of $\mathbf{D}_6$ either preserves the z_i- and w_i-positions or interchanges them. Moreover, the pairs (z_i, w_i) are never "broken up."

By the notation $t\sigma$ for an element of $\widetilde{\Gamma}$, we mean that $t = [\theta_1, \theta_2, \widehat{\phi}] \in \mathbf{T}^3$ and $\sigma \in \mathbf{D}_6$. We use the numbering of isotropy groups given in Table 4.

N(Z) = 1. Use $\mathbf{D}_6$ to make $z_1 \neq 0$, and then apply θ_1 to obtain $z_1 > 0$. This gives case **7**.

N(Z) = 2. Up to the action of $\mathbf{D}_6$, there are three possibilities for the position of the nonzero elements. We then use $\mathbf{T}^3$ to make the entries positive (hence equal):

(a) $Z = (1,1,0,0,0,0)$ is case **8**.
(b) $Z = (1,0,1,0,0,0)$ is case **10**.
(c) $Z = (1,0,0,1,0,0)$ is case **11**.

N(Z) = 3. Again, we use the action of $\mathbf{D}_6$ to bring the nonzero elements to the "leftmost" positions, followed by $\mathbf{T}^3$ to make them positive (hence equal). There are four cases:

(a) $Z = (1,0,1,0,1,0)$ is case **9**.
(b) $Z = (1,0,1,0,0,1)$: let $t\sigma \in \Sigma_Z$. Note that the only element of $\mathbf{D}_6$ that preserves the nonzero positions is the identity; hence $\sigma = 1$. The only element $t \in \mathbf{T}^3$ that preserves these positions is the identity as well; hence Σ_Z is the trivial group.
(c) $Z = (1,1,1,0,0,0)$: let $t\sigma \in \Sigma_Z$. Then $\sigma = 1$, and $t \in \mathbf{T}^3$ has to be $t_1 = [\frac{1}{2}, \frac{1}{2}, \frac{\pi}{2}]$ or the identity; hence $\Sigma_Z = \langle t_1 \rangle \subset \Sigma_{12}$ (from case **12**), and hence Σ_Z is not maximal.

(d) $Z = (1,1,0,1,0,0)$: let $t\sigma \in \Sigma_Z$. Then $\sigma = 1$, and $t \in \mathbf{T}^3$ has to be the identity as well.

N(Z) = 4. After applying $\mathbf{D}_6$, there are three possibilities for the nonzero positions. We can also make three entries equal to 1 by using a $\mathbf{T}^3$-element.

(a) $Z = (1,1,1,e^{2\pi i\rho},0,0)$: if $t\sigma \in \Sigma_Z$, then σ is one of ξ^3, $\kappa\xi$, $\kappa\xi^4$, or 1. We conclude that $\Sigma_Z = \langle [\frac{1}{2},\frac{1}{2},\frac{\pi}{2}], [0,-\rho,0]\xi^3, [-\frac{\rho}{2},-\frac{\rho}{2},\frac{\pi\rho}{2}]\kappa\xi\rangle$. This is conjugated to case **12** by $[0,-\frac{\rho}{2},\frac{\pi\rho}{4}]$.

(b) $Z = (1,0,1,0,1,e^{2\pi i\rho})$: if $t\sigma \in \Sigma_Z$, then $\sigma = \kappa\xi$ or 1. $\kappa\xi Z = (1,0,1,0,1,e^{-2\pi i\rho})$, and, if Z has a nontrivial stabilizer, then t has to be a multiple of $[\frac{1}{3},\frac{1}{3},-\frac{\pi}{3}]$. Thus $\Sigma_Z \subset \Sigma_{10}$ (from case **10**) and hence is not maximal.

(c) $Z = (1,0,0,1,1,e^{2\pi i\rho})$: if $t\sigma \in \Sigma_Z$m, then $\sigma = \kappa\xi^4$ or 1, and $t = [\theta_1,-\theta_1,-\pi\theta_1]$ for a particular value of θ_1; hence $\Sigma_Z \subset \Sigma_{11}$ (from case **11**).

N(Z) = 5. Such points have a trivial isotropy group.

N(Z) = 6. Since such points have a trivial stabilizer in $\mathbf{T}^3$, the projection of Σ_Z to $\mathbf{D}_6$ is one-to-one. Thus there is a group isomorphism $\sigma \in G \mapsto \tilde{\sigma} = t_\sigma\sigma \in \Sigma_Z$, where $G \subset \mathbf{D}_6$.

Note that if $s \in \mathbf{T}^3$, then $s(t_\sigma\sigma)s^{-1} = (t_\sigma s(\sigma s^{-1}\sigma^{-1}))\sigma$. Thus we can reduce t_σ by any element in the range of $s \in \mathbf{T}^3 \mapsto s(\sigma s^{-1}\sigma^{-1}) \in \mathbf{T}^3$. The action of $\mathbf{D}_6$ by the conjugation of $\mathbf{T}^3$ is given by relations (3.2).

We classify first the isomorphisms $\mathbf{D}_6 \hookrightarrow \widetilde{\Gamma}$, up to conjugacy. Since the action of ξ on $[\theta_1,\theta_2]$ does not have 1 as an eigenvalue, one can reduce t_ξ to $[0,0,\widehat{\phi}]$. Thus $\tilde{\xi} = [0,0,\widehat{\phi}]\xi$. The action of κ has eigenvalue -1 in the $\widehat{\phi}$-direction. Hence, by conjugating with an element of $\mathbf{T}^3$ in this eigenspace (which commutes with ξ), we can simultaneously reduce t_κ to $[\theta_1,\theta_2,0]$.

Let us now impose the relations of $\mathbf{D}_6$. $\tilde{\xi}^6 = 1$ implies $6\widehat{\phi} = 0 \pmod{\pi}$, and the relation $\tilde{\kappa}^2 = 1$ implies $\theta_1 = 0 \pmod 1$ because $\tilde{\kappa}^2 = [2\theta_1,-\theta_1,0]$. Finally, $\tilde{\kappa}\tilde{\xi}\tilde{\kappa}\tilde{\xi} = 1$ implies $\theta_2 = 0 \pmod 1$ because $\tilde{\kappa}\tilde{\xi}\tilde{\kappa}\tilde{\xi} = [-\theta_2,\theta_2,0]$. In conclusion, up to conjugacy, there are six inclusions $\mathbf{D}_6 \hookrightarrow \widetilde{\Gamma}$, those given in cases **k=0,** ... **,5**. (These cannot be conjugated by an element of $\mathbf{T}^3$ because the $\widehat{\phi}$-component of $\tilde{\xi}$ is not altered by such a conjugacy.)

It remains to establish that there are no finite maximal isotropy subgroups that are isomorphic to a proper subgroup of $\mathbf{D}_6$ and whose fixed-point subspace consists of vectors all of whose entries are nonzero. We conjecture that this statement is true, but we have not been able to prove the result.

4.7.1. Some cubic $\widetilde{\Gamma}$-equivariants. The computation of the general $\widetilde{\Gamma}$-equivariant on $\mathbf{C}^6$ is quite complicated. Below we show that generically the dynamics on the two-dimensional fixed-point subspaces of Σ_{10} and Σ_{11} involves rotating waves and that the dynamics on $\mathrm{Fix}(\Sigma_{12})$ can involve discrete rotating waves. To verify these statements, we need only show that certain cubic order terms in the Taylor expansion of a typical $\widetilde{\Gamma}$-equivariant are nonzero. In

this subsection, we list explicitly a few of the cubic $\widetilde{\Gamma}$-equivariants, namely, those that are sufficient to verify these statements.

Since the action of $\widetilde{\Gamma}$ on $\mathbf{C}^6$ is absolutely irreducible (see Remark 3.1), the only linear $\widetilde{\Gamma}$-equivariant is the identity map. Since $\left[0,0,\frac{\pi}{2}\right]$ acts as $-I$ on $\mathbf{C}^6$, there are no nonzero quadratic $\widetilde{\Gamma}$-equivariants.

Since $\mathbf{D}_6$ acts transitively on the coordinates of $\mathbf{C}^6$, it follows that (cubic) equivariants are determined by their first coordinates. We claim that

$$\begin{array}{llll}(|z_2|^2+|z_3|^2)z_1, & i(|z_2|^2-|z_3|^2)z_1, & (z_2\overline{w}_2+z_3\overline{w}_3)w_1, & |z_1|^2z_1\\ (|w_2|^2+|w_3|^2)z_1, & i(|w_2|^2-|w_3|^2)z_1, & i(z_2\overline{w}_2-z_3\overline{w}_3)w_1 & \end{array} \tag{4.11}$$

are the first coordinates of $\widetilde{\Gamma}$-equivariants on $\mathbf{C}^6$. This assertion can be verified using the explicit action given in Table 3. Note that it is only necessary to check equivariance with respect to the action of $[\theta_1,\theta_2,\widehat{\phi}]$ and $\kappa\xi^3$.

4.7.2. Rotating waves. The verification that the maximal isotropy subgroups Σ_{10} and Σ_{11} lead to rotating waves on the hexagonal lattice is identical to the verification that the isotropy subgroup Σ_5 on the square lattice leads to rotating waves, as shown in section 4.4. All that needs to be shown is that there is a cubic equivariant whose restriction to $\mathrm{Fix}(\Sigma_{10})$ is $i|u|^2u$. For example, we may take the equivariant whose first coordinate is $i(|z_2|^2-|z_3|^2)z_1$. Similarly, the equivariant whose first coordinate is $i(|w_2|^2-|w_3|^2)z_1$ restricts to $i|u|^2u$ on $\mathrm{Fix}(\Sigma_{11})$. The group orbit along which the rotating wave travels is given by the normalizer of the isotropy subgroup Σ, and that information is given in Table 4.

4.7.3. Discrete rotating wave: $\Sigma_{12}=\langle\kappa\xi,\xi^3,[\frac{1}{2},\frac{1}{2},\frac{\pi}{2}]\rangle$. Any $\widetilde{\Gamma}$-equivariant vector field restricted to $\mathrm{Fix}(\Sigma_{12})\cong\mathbf{C}$ is $\mathbf{Z}_4$-equivariant since $N(\Sigma_{12})/\Sigma_{12}\cong\mathbf{Z}_4$. See Table 4. It follows that the restricted vector field has the form

$$f(u)=A(|u|^2,u^4)u+B(|u|^2,u^4)\overline{u}^3, \tag{4.12}$$

where A and B are complex-valued functions. To cubic order

$$f(u,\lambda)=(\lambda+a|u|^2)u+b\overline{u}^3,$$

where $\lambda,a,b\in\mathbf{C}$. In fact, absolute irreducibility of the action of $\widetilde{\Gamma}$ implies that $\lambda\in\mathbf{R}$. For $\lambda\neq0$, we rescale time t so that $\lambda=1$. If $b\neq0$, then after rescaling u to αu, where $\alpha\in\mathbf{C}$, we can assume that $b=1$ and that λ is unchanged. Krauskopf [18, 19] classified the dynamical states that the cubic truncation can exhibit for all b and λ; these states include equilibria and time-periodic orbits even when $\lambda=1$. For periodic states, see [18, p. 1086, Figure 12 (between regions 14 and 15)]. Normal hyperbolicity of these solutions guarantees that the equilibria

and periodic states persist as solutions in (4.12). Note that $|z_1|^2 z_1$ and $i(|z_2|^2 - |z_3|^2)z_1$ are the first coordinates of $\widetilde{\Gamma}$-equivariants on $\mathbf{C}^6$ whose restrictions to $\mathrm{Fix}(\Sigma_{12})$ are $|u|^2 u$ and $i|u|^2 u$. To complete this proof, we need to find $\widetilde{\Gamma}$-equivariant cubics that restrict to $\overline{u}^3$ and $i\overline{u}^3$ on $\mathrm{Fix}(\Sigma_{12})$. The first coordinates of such cubics are $(z_2\overline{w}_2 + z_3\overline{w}_3)w_1$ and $i(z_2\overline{w}_2 - z_3\overline{w}_3)w_1$.

4.8. Effect of weak anisotropy on hexagonal lattice. We now present the details of the effect of forced symmetry-breaking on each of the 12 types of solutions that correspond to maximal isotropy subgroups on the hexagonal lattice.

The strategy is the same as in section 2.4: given a (relative) equilibrium of a $\widetilde{\Gamma}$-equivariant flow, consider its orbit under the connected component of $\widetilde{\Gamma}$, a 3-torus. This set (which is a torus of dimension at most 3) is invariant under the flow. Assuming that this set is normally hyperbolic, Lemma 4.2 guaranties that small Γ-equivariant perturbations of the flow still admit a nearby invariant set, diffeomorphic to that of the unperturbed flow. Since this diffeomorphism can be made Γ-equivariant, we can analyze the effect of breaking the $\widehat{\phi}$-symmetry on the unperturbed torus itself. We base our analysis on symmetry considerations, namely, intersections of this torus with fixed-point spaces. These determine flow-invariant sets: when these are points, they are equilibria, whereas if they are higher-dimensional, then one expects that generically the flow is nontrivial. This approach is used by Lauterbach and Roberts [20] in their analysis of forced symmetry-breaking.

We use two assumptions to draw our conclusions: the flow-invariant toral sets described above are generically normally hyperbolic for a $\widetilde{\Gamma}$-equivariant flow, and, after breaking the isotropy, the Γ-invariant flow is generically nontrivial whenever the Γ-symmetry *on the invariant torus* allows it.

Note that normal hyperbolicity is generic. For the case of the square lattice, we computed the general form of a $\widetilde{\Gamma}$-equivariant flow and concluded that, for various parameter choices, the invariant sets corresponding to the five maximal isotropy subgroups can each be attractive. See section 4.5. We expect the same to be true in the case of the hexagonal lattice. However, the computations are tedious, and we did not perform them.

To address the other assumption, we must either compute the general form of the Γ-equivariant perturbation of a $\widetilde{\Gamma}$-equivariant vector field and show that it permits vector fields that are nonzero on the sets under consideration or show that certain Γ-equivariant vector fields on the toral sets introduced above can be extended to Γ-equivariant vector fields on $\mathbf{C}^6$.

4.8.1. The family $\Sigma_{k+1} = \mathbf{D}_6(\kappa, [0, 0, \frac{k\pi}{2}]\xi)$, $k = 0, \ldots, 5$. Each of these subgroups is axial, with the fixed-point subspace spanned, respectively, by

$$v_k = (v^{3k}, v^{3k}, v^{7k}, v^{11k}, v^{11k}, v^{7k}), \quad v = e^{i\frac{\pi}{6}}.$$

In the isotropic case, the flow-invariant group orbit is a 3-torus $\mathbf{T}^3_k$ obtained by applying $[\theta_1, \theta_2, \widehat{\phi}]$ to the point v_k. When the anisotropy is small, there still exists a flow-invariant set close to the 3-torus described above. The flow on this set can be understood by analyzing the isotropic case. In the following, we consider Γ-equivariant flows on this 3-torus. To simplify notation, we sometimes drop the reference to k.

Each of these 3-tori is foliated by 2-tori $T_{\widehat{\phi}} = T_{k,\widehat{\phi}}$ that are swept out by the action of $[\theta_1, \theta_2, 0]$ on $[0, 0, \widehat{\phi}]v_k$. By $[\theta_1, \theta_2]$-equivariance, the flow on these 2-tori is parallel. Certain of the 2-tori $T_{\widehat{\phi}}$ are forced by Γ-symmetry to be flow-invariant; we call them *critical*. Generically, other than the critical 2-tori, there are only finitely many other 2-tori that are flow-invariant. We expect the flow between the invariant 2-tori to be transverse to the noninvariant $T_{\widehat{\phi}}$'s.

If $k \neq 1, 5$, the invariant 2-tori consist of fixed points, and on the 3-torus T^3_k the flow is along the curves $[\theta_1, \theta_2] =$ constant. For $k = 1, 5$, the flow on the invariant 2-tori is generically expected to be nontrivial (and parallel, as discussed above). For each critical torus, the direction of the flow can be determined explicitly. See Table 9.

Table 9

Conjugacy classes of critical $T_{\widehat{\phi}}$'s and flows on them. When flow is nontrivial, the direction of flow $[\theta_1, \theta_2]$ is given. When equilibrium is scalar or pseudoscalar, that is denoted by $W_1^{\pm}$. ℓ denotes the number of conjugacy classes in Γ of each critical torus, and $v = [0, 0, \widehat{\phi}]v_k$ is the point where the isotropy subgroup $\Sigma_v \subset \Gamma$ is computed.

k	$\widehat{\phi}$	$[\theta_1, \theta_2]$ or $W_1^{\pm}$	Σ_v	$\widehat{\phi}$	$[\theta_1, \theta_2]$ or $W_1^{\pm}$	Σ_v	ℓ
0	0	W_1^+	$\mathbf{D}_6(\kappa, \xi)$	$\frac{\pi}{2}$	W_1^+	$\mathbf{D}_6(\kappa, \xi)$	1
1	0	$[-2\theta_2, \theta_2]$	$\mathbf{D}_1(\kappa)$	$\frac{\pi}{12}$	$[-\theta_2, \theta_2]$	$\mathbf{D}_1(\kappa\xi)$	6
2	0		$\mathbf{D}_2(\kappa, \xi^3)$	$\frac{\pi}{2}$		$\mathbf{D}_2(\kappa, \xi^3)$	3
3	0	W_1^+	$\mathbf{D}_3(\kappa, \xi^2)$	$\frac{\pi}{4}$	W_1^-	$\mathbf{D}_3(\kappa\xi, \xi^2)$	2
4	0		$\mathbf{D}_2(\kappa, \xi^3)$	$\frac{\pi}{2}$		$\mathbf{D}_2(\kappa, \xi^3)$	3
5	0	$[-2\theta_2, \theta_2]$	$\mathbf{D}_1(\kappa)$	$\frac{\pi}{12}$	$[\theta_1, 0]$	$\mathbf{D}_1(\kappa\xi^5)$	6

We now present the details needed to derive the information in Table 9. For $\sigma \in \Gamma$, we are interested in the intersection of $\mathrm{Fix}(\sigma)$ with the 3-torus $\mathbf{T}^3_k$ swept out by v_k under the action of $[\theta_1, \theta_2, \widehat{\phi}]$.

This intersection is nontrivial only if the isotropy subgroup of a point in $\mathbf{T}^3_k$ contains σ. However, these isotropy subgroups are easy to compute because we know the isotropy subgroup of v_k:

$$\Sigma_{v_k} = \langle \kappa, \tilde{\xi} \rangle \cong \mathbf{D}_6,$$

where $\tilde{\xi} = \left[0, 0, \frac{k\pi}{6}\right] \xi$.

Hence the isotropy subgroup of $[\theta_1, \theta_2, \widehat{\phi}]v_k$ is given by $[\theta_1, \theta_2, \widehat{\phi}]\Sigma_{v_k}[-\theta_1, -\theta_2, -\widehat{\phi}]$. It

follows from (3.2) that for $\tilde{\sigma} = \kappa^{\varepsilon}\tilde{\xi}^{p} \in \Sigma_{v_k}$ ($\varepsilon \in \{0,1\}$, $p = 0,\dots,5$),

$$[\theta_1,\theta_2,\widehat{\phi}]\tilde{\sigma}[-\theta_1,-\theta_2,-\widehat{\phi}] = \left[(\mathrm{I}-A_\sigma)\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}, 2\varepsilon\widehat{\phi} + (-1)^{\varepsilon}\frac{pk\pi}{6}\right]\sigma, \tag{4.13}$$

where $\sigma = \kappa^{\varepsilon}\xi^{p}$ and A_σ is the action by conjugation of $\sigma \in \mathbf{D}_6$ on $[\theta_1,\theta_2]$ (see (3.2)), and

$$\kappa[\theta_1,\theta_2] = [\theta_1, -\theta_1-\theta_2]\kappa,$$
$$\xi[\theta_1,\theta_2] = [-\theta_2, \theta_1+\theta_2]\xi.$$

Hence

$$A_\kappa = \begin{pmatrix} 1 & 0\\ -1 & -1\end{pmatrix}, \qquad A_\xi = \begin{pmatrix} 0 & -1\\ 1 & 1\end{pmatrix}.$$

For $\sigma = \kappa^{\varepsilon}\xi^{p}$ and $0 \le \eta_1,\eta_2 \le 1$, we conclude that

$$\mathrm{Fix}([\eta_1,\eta_2,0]\sigma) \cap \mathbf{T}^3_k = \{[\theta_1,\theta_2,\widehat{\phi}]v_k : \theta_1,\theta_2,\widehat{\phi} \text{ satisfies (4.15), (4.16)}\}, \tag{4.14}$$

where

$$(\mathrm{I}-A_\sigma)\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} = \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix}, \tag{4.15}$$

$$2\varepsilon\widehat{\phi} + (-1)^{\varepsilon}\frac{pk\pi}{6} = 0 \pmod{\pi}. \tag{4.16}$$

Each of these intersections is flow-invariant and one-dimensional unless it is empty or $\mathbf{T}^3_k$ (as one can check by computing the A_σ's). We now consider separately the two cases, determined by whether or not k is relatively prime to 6.

***k* = 1, 5.** Equation (4.16) cannot be satisfied if $\varepsilon = 0$; hence the intersections are trivial for $\sigma = \xi^p$, $p \neq 0$. For $\varepsilon = 1$, (4.16) gives $2\widehat{\phi} = \frac{pk\pi}{6} \pmod{\pi}$; note that different values of p give different values of $\widehat{\phi}$, and hence different elements of Γ produce disjoint intersections with $\mathbf{T}^3_k$. These are the critical 2-tori. Since $\mathrm{I}-A_{\kappa\xi^p}$ has rank one, (4.15) gives a flow-invariant line in the corresponding $T_{\widehat{\phi}}$.

***k* = 0, 2, 3, 4.** Setting $\varepsilon = 1$ and $p = 0,\dots,5$, we find the finite set of critical $\widehat{\phi}$-values from (4.16). Choose a value $p \neq 0$ such that $pk \equiv 0 \pmod{6}$. Then, for $\sigma = \xi^p$, the intersection (4.14) is given by circles with constant $[\theta_1,\theta_2]$, because the rank of $\mathrm{I}-A_{\kappa\xi^p}$ is equal to 2. Since these flow-invariant circles intersect each $T_{\widehat{\phi}}$ transversely, the flow has to be trivial on the invariant $T_{\widehat{\phi}}$'s. Note that A_σ gives the only faithful two-dimensional representation of $\mathbf{D}_6$. This explains the values of $\mathrm{rank}(\mathrm{I}-A_\sigma)$.

We decide now which critical 2-tori are conjugate by Γ. (Since the field is Γ-equivariant, the action of Γ permutes the invariant 2-tori $T_{\widehat{\phi}}$.) Relation (4.13) shows that

$$\kappa^{\varepsilon}\xi^{p'}T_{\widehat{\phi}} = T_{\hat{\psi}}, \quad \text{where} \quad \hat{\psi} = (-1)^{\varepsilon}\left(\widehat{\phi} - \frac{p'k\pi}{6}\right) \pmod{\pi}.$$

Solving (4.16) with $\varepsilon = 1$ for the critical values of $\widehat{\phi}$, we see that for each k there are two Γ-conjugacy classes of critical tori $T_{\widehat{\phi}}$ in $\mathbf{T}^3_k$.

4.8.2. $\Sigma_7 = \langle \kappa\xi^3, [0, \theta_2, 0], [\theta_1, 0, -\pi\theta_1]\rangle$. Equilibria of type Σ_7 appear in circles. When symmetry breaks, we see that $(\pm 1, 0, 0, 0, 0, 0)$ must remain as equilibria, since these points form the intersection of $\mathrm{Fix}(\kappa\xi^3)$ with the circle $(z_1, 0, 0, 0, 0, 0)$, where $|z_1| = 1$. Since $[\theta_1, 0, 0]$ acts transitively on the circle, all points on the circle are equilibria.

4.8.3. $\Sigma_8 = \langle \kappa, \xi^3, [0, \theta_2, 0], [\frac{1}{2}, 0, \frac{\pi}{2}]\rangle$. Equilibria of type Σ_8 lie on the 2-torus $(z_1, w_1, 0, 0, 0, 0)$, where $|z_1| = |w_1| = 1$. The following group elements act on this torus:

$$\begin{aligned}
[\theta_1, 0, 0](z_1, w_1) &= (e^{-2\pi i\theta_1}z_1, e^{-2\pi i\theta_1}w_1),\\
\kappa(z_1, w_1) &= (w_1, z_1),\\
\xi^3(z_1, w_1) &= (\overline{w}_1, \overline{z}_1),\\
\kappa\xi^3(z_1, w_1) &= (\overline{z}_1, \overline{w}_1).
\end{aligned}$$

Note that $\mathrm{Fix}(\mathbf{D}_2)$ intersects the 2-torus at two points $(1, 1)$ and $(-1, -1)$. Thus these two points are fixed in the anisotropic case. Moreover, $\mathrm{Fix}(\kappa)$ is the circle $z_1 = w_1$ on the torus, and it must be flow-invariant. Since $[\theta_1, 0, 0]$ acts on that circle, the points on the circle are conjugate and must also be fixed points of the flow.

Next note that $\mathrm{Fix}(\kappa\xi^3)$ consists of four points $(\pm 1, \pm 1)$. Thus the circle generated by $[\theta_1, 0, 0]$ through $(1, -1)$ also consists of fixed points. Finally, note that $\mathrm{Fix}(\xi^3)$ is a circle perpendicular to the diagonal (z_1, z_1) and is also flow-invariant. Thus $[\theta_1, 0, 0]$-symmetry implies that the dynamics on the 2-torus is along circles perpendicular to the diagonal and that each such circle has four equilibria (two pairs of conjugate equilibria).

We can assume that (up to symmetry) in the anisotropic case there are two kinds of equilibria corresponding to type Σ_8, namely, $(1, 1, 0, 0, 0, 0)$ and $(1, -1, 0, 0, 0, 0)$. These correspond to scalar and pseudoscalar rolls.

4.8.4. $\Sigma_9 = \langle \xi^2, \kappa\xi, [\frac{1}{3}, \frac{1}{3}, \frac{2\pi}{3}]\rangle$. The group orbits for equilibria of type Σ_9 are 3-tori modeled by $|z_1| = |z_2| = |z_3| = 1$. In the anisotropic case, there are seven flow-invariant circles on this $\mathbf{T}^3$ and two fixed points. The fixed points are given by $\mathrm{Fix}(\mathbf{D}_3(\xi^2, \kappa\xi)) = \pm(1, 1, 1)$. $\mathrm{Fix}(\xi^2)$ gives one of the circles $z_1 = z_2 = z_3$, and each of $\mathrm{Fix}(\kappa\xi)$, $\mathrm{Fix}(\kappa\xi^3)$, and $\mathrm{Fix}(\kappa\xi^5)$ gives two invariant circles. Although the dynamics on this group orbit seems complicated to describe, we know that we will get at least two fixed points corresponding to $\pm(1, 0, 1, 0, 1, 0)$.

4.8.5. $\Sigma_{10} = \langle \kappa\xi, [\theta_1, \theta_1, -\pi\theta_1]\rangle$. In the isotropic case, solutions corresponding to the maximal isotropy subgroup of type Σ_{10} can be expected to be a rotating wave, since its fixed-point subspace is two-dimensional and the normalizer of the isotropy subgroup acts as a circle on this fixed-point subspace. See section 4.7.2. We claim that the same is true in the anisotropic case.

Observe that the group orbit of solutions is the 2-torus $(z_1, 0, z_2, 0, 0, 0)$, where $|z_1| = 1 = |z_2|$. Observe that

$$\kappa\xi(z_1, z_2) = (\overline{z}_2, \overline{z}_1).$$

Note that $[\theta_1, \theta_2, 0]$ acts transitively on the 2-torus so that the flow on the 2-torus is a linear flow. Finally, note that $\mathrm{Fix}(\kappa\xi) = \{(z_1, \overline{z}_1)\}$ is a circle $\rho_1 + \rho_2 = 0$, where $z_j = e^{2\pi i\rho_j}$. So this circle (and all circles parallel to it on the 2-torus) are flow-invariant and rotating waves.

4.8.6. $\Sigma_{11} = \langle \kappa\xi^4, [\theta_1, -\theta_1, -\pi\theta_1]\rangle$. The analysis of the maximal isotropy subgroup of type Σ_{11} is identical to that of Σ_{10}, and, in the anisotropic case, group orbits are 2-tori foliated by flow-invariant traveling waves. Note that the invariant circles $\mathrm{Fix}(\kappa\xi^4)$ are parallel to the main diagonal $z_1 = w_2$.

4.8.7. $\Sigma_{12} = \langle \xi^3, \kappa\xi, [\frac{1}{2}, \frac{1}{2}, \frac{\pi}{2}]\rangle$. We now describe what happens to the discrete rotating wave associated to the isotropy subgroup Σ_{12} (discussed in section 4.7.3) when symmetry-breaking terms are added. The main mathematical issue is that $\mathrm{Fix}(\Sigma_{12})$ is no longer flow-invariant when symmetry-breaking terms are added. Since the symmetry group of the discrete rotating wave is finite, the action of the connected component of $\widetilde{\Gamma}$ on this periodic solution yields an invariant 4-torus, which is preserved by normal hyperbolicity. However, the three-dimensional invariant subspace

$$W = \mathrm{Fix}(\langle \xi^3, \kappa\xi\rangle) = \{(u, \bar{u}, \bar{u}, u, x, x) \mid u \in \mathbf{C}, x \in \mathbf{R}\} \supset \mathrm{Fix}(\Sigma_{12})$$

survives symmetry-breaking. It is straightforward to check that the intersection of the 4-torus with W is the original periodic solution. Therefore, the periodic state will survive small symmetry-breaking terms and remain in W.

Acknowledgments. We are grateful to Paul Bressloff and Jack Cowan for many helpful discussions. The motivation to consider the nonlinear consequences of weak anisotropy was provided in a talk given by Fred Wolf at the 2001 SIAM Snowbird Dynamical Systems meeting.

REFERENCES

[1] A. Angelucci, J. B. Levitt, and J. S. Lund, *Anatomical origins of the classical receptive field and modulatory surround field of single neurons in macaque visual cortical area*, in Changing Views of Cajal's Neuron, Progress in Brain Research 136, Elsevier, Amsterdam, 2002, pp. 373–388.

[2] G. G. Blasdel, *Orientation selectivity, preference, and continuity in monkey striate cortex*, J. Neurosci., 12 (1992), pp. 3139–3161.

[3] I. Bosch Vivancos, P. Chossat, and I. Melbourne, *New planforms in systems of partial differential equations with Euclidean symmetry*, Arch. Rational Mech. Anal., 131 (1995), pp. 199–224.

[4] 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.

[5] P. C. Bressloff, J. D. Cowan, M. Golubitsky, and P. J. Thomas, *Scalar and pseudoscalar bifurcations motivated by pattern formation on the visual cortex*, Nonlinearity, 14 (2001), pp. 739–775.

[6] P. C. Bressloff, J. D. Cowan, M. Golubitsky, P. J. Thomas, and M. C. Wiener, *Geometric visual hallucinations, Euclidean symmetry, and the functional architecture of striate cortex*, Phil. Trans. Royal Soc. London B, 356 (2001), pp. 299–330.

[7] P. C. Bressloff, J. D. Cowan, M. Golubitsky, P. J. Thomas and M. C. Wiener, *What geometric visual hallucinations tell us about the visual cortex*, Neural Computation, 14 (2002), pp. 473–491.

[8] J. D. Cowan, *Some Remarks on channel bandwidths for visual contrast detection*, Neurosciences Research Program Bull., 15 (1977), pp. 492–517.

[9] G. B. Ermentrout and J. D. Cowan, *A mathematical theory of visual hallucination patterns*, Biol. Cybernetics, 34 (1979), pp. 137–150.

[10] U. Eysel, *Turning a corner in vision research*, Nature, 399 (1999), pp. 641–644.

[11] M. Field and J. W. Swift, *Stationary bifurcation to limit cycles and heteroclinic cycles*, Nonlinearity, 4 (1991), pp. 1001–1043.

[12] C. D. Gilbert, *Horizontal integration and cortical dynamics*, Neuron, 9 (1992), pp. 1–13.

[13] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, Progr. Math., 200, Birkhäuser, Basel, 2002.

[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] M. Hirsch, C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math 583, Springer-Verlag, New York.

[16] D. H. Hubel and T. N. Wiesel, *Sequence regularity and geometry of orientation columns in the monkey striate cortex*, J. Comp. Neurol., 158 (1974), pp. 267–294; *Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor*, J. Comp. Neurol., 158 (1974), pp. 295–306; *Ordered arrangement of orientation columns in monkeys lacking visual experience*, J. Comp. Neurol., 158 (1974), pp. 307–318.

[17] H. Klüver, *Mescal and Mechanisms of Hallucinations*, University of Chicago Press, Chicago, 1966.

[18] B. Krauskopf, *Bifurcation sequences at* 1:4 *resonance: An inventory*, Nonlinearity, 7 (1994), pp. 1073–1091.

[19] B. Krauskopf, *The bifurcation set for the* 1:4 *resonance problem*, Experiment. Math., 3 (1994), pp. 107–128.

[20] R. Lauterbach and M. Roberts, *Heteroclinic cycles in dynamical systems with broken spherical symmetry*, J. Differential Equations, 100 (1992), pp. 22–48.

[21] I. Melbourne, *Maximal isotropy subgroups for absolutely irreducible representations of compact Lie groups*, Nonlinearity, 7 (1994), pp. 1385–1393.

[22] E. Schwartz, *Spatial mapping in the primate sensory projection: Analytic structure and relevance to projection*, Biol. Cybernetics, 25 (1977), pp. 181–194.

[23] L. C. Sincich and G. G. Blasdel, *Oriented axon projections in primary visual cortex of the monkey*, J. Neurosci., 21 (2001), pp. 4416–4426.

[24] P. Tass, *Oscillatory cortical activity during visual hallucinations*, J. Biol. Phys., 23 (1997), pp. 21–66.

[25] H. R. Wilson and J. D. Cowan, *Excitatory and inhibitory interactions in localized populations of model neurons*, Biophys. J., 12 (1972), pp. 1–24.

[26] F. Wolf and T. Geisel, *Spontaneous pinwheel annihilation during visual development*, Nature, 395 (1998), pp. 73–78.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 2, pp. 144–170

Optimal Low Thrust Trajectories to the Moon*

John T. Betts† and Sven O. Erb‡

Abstract. The direct transcription or collocation method has demonstrated notable success in the solution of trajectory optimization and optimal control problems. This approach combines a sparse nonlinear programming algorithm with a discretization of the trajectory dynamics. A challenging class of optimization problems occurs when the spacecraft trajectories are characterized by thrust levels that are very low relative to the vehicle weight. Low thrust trajectories are demanding because realistic forces, due to oblateness, and third-body perturbations often dominate the thrust. Furthermore, because the thrust is so low, significant changes to the orbits require very long duration trajectories. When a collocation method is applied to a problem of this type, the resulting nonlinear program is very large, because the trajectories are long, and very nonlinear because of the perturbing forces.

This paper describes the application of the transcription method to compute an optimal low thrust transfer from an Earth orbit to a specified lunar mission orbit. It is representative of the SMART-1 or "Small Missions for Advanced Research in Technology" of the ESA scientific program [J. Schoenmaekers, J. Pulido, and R. Jehn, Tech. report S1-ESC-RP-5001, European Space Agency, 1998]. The spacecraft is deployed from an Ariane-5 into an elliptic Earth centered park orbit. The goal is to insert the spacecraft into a lunar orbit that is polar and elliptic and has its pericenter above the south pole. The spacecraft utilizes a solar electric propulsion system. The goal is to compute the optimal steering during the orbit transfer, which takes over 200 days, so that the fuel consumption is minimized.

The vehicle dynamics are defined using a modified set of equinoctial coordinates, and the trajectory modeling is described using these dynamics. A solution is presented that requires the solution of a sparse optimization problem with 211031 variables and 146285 constraints. The trajectory we present requires two very long thrust arcs, and, consequently, the overall mission duration is much shorter than multiburn trajectories. Issues related to the numerical conditioning and problem formulation are discussed.

Key words. optimal control, sparse nonlinear programming, orbit transfer

AMS subject classifications. 49J15, 90C30, 90C90, 70M20, 70F15

PII. S1111111102409080

1. Overview of the problem.

The European Space Agency (ESA) has introduced a project called "Small Missions for Advanced Research in Technology," or SMART-1, as a part of their program [17]. A prime objective of SMART-1 is to demonstrate the use of solar electric propulsion (SEP) as a key technology for scientific deep space missions. In addition to the significant hardware challenges associated with this technology, low thrust trajectories lead to challenging mission design problems. Our goal in this paper is to describe an ap-

*Received by the editors June 5, 2002; accepted for publication (in revised form) by J. Marsden November 22, 2002; published electronically May 8, 2003.
http://www.siam.org/journals/siads/2-2/40908.html
†Mathematics and Computing Technology, The Boeing Company, P.O. Box 3707, MS 7L-21, Seattle, WA 98124-2207 (John.T.Betts@boeing.com).
‡Institut für Flugmechanik und Flugregelung, Universität Stuttgart, Pfaffenwaldring 7a, 70550 Stuttgart, Germany (sven.erb@ifr.uni-stuttgart.de).

proach for solving low thrust trajectory problems in general and illustrate the technique for the SMART-1 mission.

The motion of a spacecraft can be described by a system of second order ordinary differential equations,

$$\ddot{\mathbf{r}} + \mu \frac{\mathbf{r}}{r^3} = \mathbf{a}_d, \tag{1.1}$$

where the radius, $r = \|\mathbf{r}\|$, is the magnitude of the inertial position vector $\mathbf{r}$, and μ is the gravitational constant. In this formulation, we define the vector $\mathbf{a}_d$ as the *disturbing acceleration*. This representation for the equations of motion is referred to as Gauss's form of the variational equations. The Gauss form of the equations of motion isolates the disturbing acceleration from the central force gravitational acceleration. Note that when the disturbing acceleration is zero, $\|\mathbf{a}_d\| = 0$, the fundamental system (1.1) is just a two-body problem. The solution of the two-body problem can of course be stated in terms of the constant orbital elements. For low thrust trajectories, this formulation is appealing because we expect $\|\mathbf{a}_d\|$ to be "small," and, consequently, we expect that the solution can be described in terms of "almost constant" orbital elements. In order to exploit the benefits of the variational form of the differential equations (1.1), it is necessary to transform the Cartesian state into an appropriate set of orbit elements. Classical elements exhibit singularities for zero eccentricity and inclinations of 0° and 90°. Furthermore, the semimajor axis changes discontinuously for orbits with eccentricity equal to one. To eliminate these deficiencies, a modified set of equinoctial orbit elements is described in [6] based on the work in [19]. The equinoctial dynamics are summarized in section 2.1.

Equation (1.1) formalizes a key modeling concept that can be exploited for our application. Specifically, we consider that the *dominant* acceleration $\mu\mathbf{r}/r^3$ can be attributed to the gravitational acceleration of the *primary* body. All other quantities are considered *disturbing accelerations* and are included in $\mathbf{a}_d$. In our example,

$$\mathbf{a}_d = \boldsymbol{\delta}\mathbf{g} + \boldsymbol{\delta}\mathbf{q} + \boldsymbol{\delta}T, \tag{1.2}$$

where $\boldsymbol{\delta}\mathbf{g}$ is the acceleration caused by Earth oblateness (section 2.2), $\boldsymbol{\delta}\mathbf{q}$ is the acceleration attributed to *secondary* bodies (section 2.3), and $\boldsymbol{\delta}T$ is the acceleration produced by the thrust (section 2.4). For our application, it is important to switch from an Earth centered (geocentric) coordinate frame to a Moon centered (selenocentric or lunicentric) system.

The complete Earth-Moon trajectory can be modeled using three distinct *phases* as illustrated in Figure 1 and summarized as follows:

Phase 1: *Geocentric burn.* Earth as the primary body, Sun and Moon as secondary bodies, including Earth oblateness, and thrust, i.e., $\boldsymbol{\delta}\mathbf{g} \neq 0$, $\boldsymbol{\delta}\mathbf{q} \neq 0$, and $\boldsymbol{\delta}T \neq 0$.

Phase 2: *Geocentric coast.* Earth as the primary body, Sun and Moon as secondary bodies, including Earth oblateness, with no thrust, i.e., $\boldsymbol{\delta}\mathbf{g} \neq 0$, $\boldsymbol{\delta}\mathbf{q} \neq 0$, and $\boldsymbol{\delta}T = 0$.

Phase 3: *Selenocentric burn.* Moon as the primary body, Sun and Earth as secondary bodies, ignoring Earth oblateness, with thrust, i.e., $\boldsymbol{\delta}\mathbf{g} = 0$, $\boldsymbol{\delta}\mathbf{q} \neq 0$, and $\boldsymbol{\delta}T \neq 0$.

It should be emphasized that all three phases include the gravitational acceleration for the Earth, Sun, and Moon. It is not necessary to quantify the distinction between a "perturbing" force and the "dominant" force—something that is difficult to achieve in the Earth-Moon system.

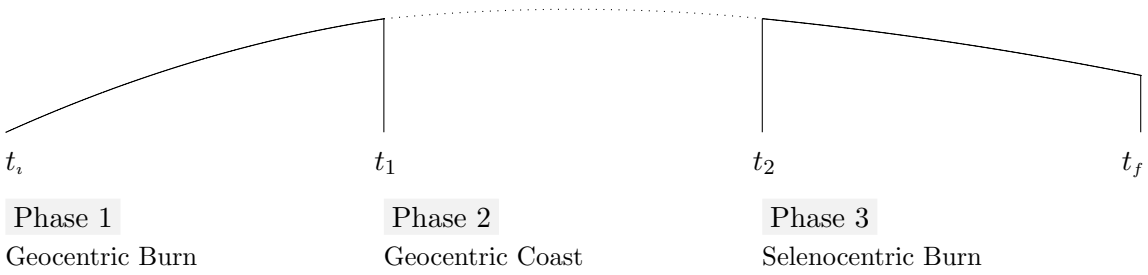


Figure 1. *Earth-Moon transfer trajectory.*

2. Equations of motion.

2.1. Translational dynamics in modified equinoctial coordinates. The dynamics of the system can be described in terms of the state variables

$$\mathbf{z}^{\mathsf{T}} = [\mathbf{y}^{\mathsf{T}}, w] = [p, f, g, h, k, L, w]\,, \tag{2.1}$$

where $\mathbf{y}$ is the vector of equinoctial coordinates and w is the weight. The control variables

$$\mathbf{u}^{\mathsf{T}} = [u_r, u_\theta, u_h] \tag{2.2}$$

define the orientation of the thrust in a rotating radial frame given by (2.16).

Using the modified equinoctial elements, we can state the equations of motion for a vehicle with constant thrust as

$$\dot{\mathbf{y}} = \mathbf{A}(\mathbf{y})\boldsymbol{\Delta} + \mathbf{b}, \tag{2.3}$$

$$\dot{w} = -T/I_{sp}, \tag{2.4}$$

$$0 = \|\mathbf{u}\| - 1. \tag{2.5}$$

The equinoctial dynamics are defined by the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & \frac{2p}{\xi}\sqrt{\frac{p}{\mu}} & 0 \\ \sqrt{\frac{p}{\mu}}\sin L & \sqrt{\frac{p}{\mu}}\frac{1}{\xi}\{(\xi+1)\cos L + f\} & -\sqrt{\frac{p}{\mu}}\frac{g}{\xi}\{h\sin L - k\cos L\} \\ -\sqrt{\frac{p}{\mu}}\cos L & \sqrt{\frac{p}{\mu}}\frac{1}{\xi}\{(\xi+1)\sin L + g\} & \sqrt{\frac{p}{\mu}}\frac{f}{\xi}\{h\sin L - k\cos L\} \\ 0 & 0 & \sqrt{\frac{p}{\mu}}\frac{s^2\cos L}{2\xi} \\ 0 & 0 & \sqrt{\frac{p}{\mu}}\frac{s^2\sin L}{2\xi} \\ 0 & 0 & \sqrt{\frac{p}{\mu}}\frac{1}{\xi}\{h\sin L - k\cos L\} \end{bmatrix} \tag{2.6}$$

and the vector

$$\mathbf{b}^{\mathsf{T}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \sqrt{\mu p}\left(\frac{\xi}{p}\right)^2 \end{bmatrix}, \tag{2.7}$$

where

$$\xi = 1 + f\cos L + g\sin L, \tag{2.8}$$

$$r = \frac{p}{\xi}, \tag{2.9}$$

$$\alpha^2 = h^2 - k^2, \tag{2.10}$$

$$\chi = \sqrt{h^2 + k^2}, \tag{2.11}$$

$$s^2 = 1 + \chi^2. \tag{2.12}$$

The equinoctial coordinates $\mathbf{y}$ are related to the Cartesian state $(\mathbf{r}, \mathbf{v})$ according to the expressions

$$\mathbf{r}(\mathbf{y}) = \begin{bmatrix} \frac{r}{s^2}\left(\cos L + \alpha^2\cos L + 2hk\sin L\right) \\ \frac{r}{s^2}\left(\sin L - \alpha^2\sin L + 2hk\cos L\right) \\ \frac{2r}{s^2}\left(h\sin L - k\cos L\right) \end{bmatrix}, \tag{2.13}$$

$$\mathbf{v}(\mathbf{y}) = \begin{bmatrix} -\frac{1}{s^2}\sqrt{\frac{\mu}{p}}\left(\sin L + \alpha^2\sin L - 2hk\cos L + g - 2fhk + \alpha^2 g\right) \\ -\frac{1}{s^2}\sqrt{\frac{\mu}{p}}\left(-\cos L + \alpha^2\cos L + 2hk\sin L - f + 2ghk + \alpha^2 f\right) \\ \frac{2}{s^2}\sqrt{\frac{\mu}{p}}\left(h\cos L + k\sin L + fh + gk\right). \end{bmatrix}. \tag{2.14}$$

As a result of this transformation, the disturbing acceleration vector $\mathbf{a}_d$ in (1.1) is replaced by

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}_g + \boldsymbol{\Delta}_q + \boldsymbol{\Delta}_T \tag{2.15}$$

with a contribution due to oblate Earth effects $\boldsymbol{\Delta}_g$, another caused by the secondary bodies $\boldsymbol{\Delta}_q$, and a third caused by thrust $\boldsymbol{\Delta}_T$. The disturbing acceleration is expressed in a rotating radial frame whose principle axes are defined by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{i}_r & \mathbf{i}_\theta & \mathbf{i}_h \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{r}}{\|\mathbf{r}\|} & \frac{(\mathbf{r}\times\mathbf{v})\times\mathbf{r}}{\|\mathbf{r}\times\mathbf{v}\|\|\mathbf{r}\|} & \frac{\mathbf{r}\times\mathbf{v}}{\|\mathbf{r}\times\mathbf{v}\|} \end{bmatrix}. \tag{2.16}$$

As stated, (2.3)–(2.5) are perfectly general and describe the motion of a point mass when subject to the disturbing acceleration vector $\boldsymbol{\Delta}$. Notice that when the disturbing acceleration is zero, $\boldsymbol{\Delta} = 0$, the first five equations are simply $\dot{p} = \dot{f} = \dot{g} = \dot{h} = \dot{k} = 0$, which implies that the elements are constant. During a coast phase when there is no thrust, $\boldsymbol{\Delta}_T = \mathbf{0}$, (2.4) and (2.5) are not needed to model the dynamics, although we do include the other perturbations, i.e., $\boldsymbol{\Delta} \neq 0$. A more complete derivation of the equinoctial dynamics can be found in [6].

2.2. Gravitational disturbing acceleration. Oblate gravity models are typically defined in a local horizontal reference frame, that is,

$$\boldsymbol{\delta}\mathbf{g} = \delta g_n \mathbf{i}_n - \delta g_r \mathbf{i}_r, \tag{2.17}$$

where

$$\mathbf{i}_n = \frac{\mathbf{e}_n - (\mathbf{e}_n^\top \mathbf{i}_r)\mathbf{i}_r}{\|\mathbf{e}_n - (\mathbf{e}_n^\top \mathbf{i}_r)\mathbf{i}_r\|} \tag{2.18}$$

defines the local North direction with $\mathbf{e}_n = (0, 0, 1)$. A reasonably accurate model is obtained if the tesseral harmonics are ignored and only the first four zonal harmonics are included in the geopotential function. In this case, the oblate Earth perturbations to the gravitational acceleration are given by

$$\delta g_n = -\frac{\mu \cos\phi}{r^2} \sum_{k=2}^{4} \left(\frac{a_e}{r}\right)^k P_k' J_k, \tag{2.19}$$

$$\delta g_r = -\frac{\mu}{r^2} \sum_{k=2}^{4} (k+1) \left(\frac{a_e}{r}\right)^k P_k J_k, \tag{2.20}$$

where ϕ is the geocentric latitude, a_e is the equatorial radius of the Earth, $P_k(\sin\phi)$ is the kth order Legendre polynomial with corresponding derivative P_k', and the zonal harmonic coefficients are given by J_k.

Finally, to obtain the gravitational perturbations in the rotating radial frame, it follows that

$$\boldsymbol{\Delta}_g = \mathbf{Q}^\mathsf{T} \boldsymbol{\delta g}. \tag{2.21}$$

2.3. Secondary body acceleration. It has been demonstrated in [6] that secondary bodies introduce the perturbing acceleration

$$\boldsymbol{\delta q} = -\sum_{j=1}^{n} \mu_j \left[\frac{\mathbf{d}_j}{d_j^3} + \frac{\mathbf{s}_j}{s_j^3}\right], \tag{2.22}$$

where $\mathbf{s}_j$ is a vector from the primary body to the secondary body j, with gravitational constant μ_j, and

$$\mathbf{d}_j = \mathbf{r} - \mathbf{s}_j \tag{2.23}$$

is a vector from the secondary body to the vehicle. The calculation of the gravitational disturbances due to the other bodies can be achieved directly from (2.22). However, Battin [2] has shown that this calculation is prone to cancelation because of the significantly different size of the terms involved. Instead, he suggests defining the function

$$F(q_k) = q_k \left[\frac{3 + 3q_k + q_k^2}{1 + (\sqrt{1+q_k})^3}\right], \tag{2.24}$$

where

$$q_k = \frac{\mathbf{r}^{\mathsf{T}}(\mathbf{r} - 2\mathbf{s}_k)}{\mathbf{s}_k^{\mathsf{T}}\mathbf{s}_k}. \tag{2.25}$$

Substituting (2.24)–(2.25) into (2.22) yields

$$\boldsymbol{\delta}\mathbf{q} = -\sum_{k=1}^{n} \frac{\mu_k}{d_k^3} \left[\mathbf{r} + F(q_k)\mathbf{s}_k\right]. \tag{2.26}$$

Finally, to obtain the acceleration of the perturbing bodies in the rotating radial frame, it follows that

$$\boldsymbol{\Delta}_q = \mathbf{Q}^{\mathsf{T}}\boldsymbol{\delta}\mathbf{q}. \tag{2.27}$$

The calculation of the n-body perturbing forces requires an ephemeris which specifies the position and velocity of the Moon and all planetary objects as a function of time. For this purpose, we have used the Jet Propulsion Laboratory (JPL) Ephemeris [18], DE 405, supplied electronically via the URL http://ssd.jpl.nasa.gov/horizons_doc.html. The JPL ephemeris is provided as a set of FORTRAN subroutines that reads a large tabular data file. For improved computational efficiency, a Hermite cubic interpolant was created to approximate the ephemeris for the duration of the mission. This Hermite evaluation approach is approximately three times faster than the JPL evaluation procedure because it avoids repeated input/output (I/O) operations on the tabular data. This computational speedup is extremely important because the ephemeris is evaluated so often (3,604,637,090 times for our solution). We also ensure continuity and differentiability in the planetary position and velocity.

2.4. Thrust acceleration—burn arcs. To this point, the discussion has concentrated on incorporating perturbing forces due to secondary bodies and oblate Earth effects. Of course, the third major perturbation is the thrust acceleration defined by

$$\boldsymbol{\Delta}_T = \frac{g_o T}{w}\mathbf{u}, \tag{2.28}$$

where T is the maximum thrust. In general, the direction of the thrust acceleration vector, which is defined by the time varying control vector $\mathbf{u}(t) = (u_r, u_\theta, u_h)$, can be chosen arbitrarily as long as the vector has unit length at all points in time, which is achieved using the path constraint (2.5). The magnitude of the thrust is of course related to the vehicle weight according to (2.4), where g_0 is the mass to weight conversion factor, and the specific impulse of the motor is denoted by I_{sp}. Defining the thrust direction using the vector $\mathbf{u}(t)$ and path constraint $\|\mathbf{u}(t)\| = 1$ is particularly well suited for missions that involve steering over large portions of the trajectory, as illustrated in [7], because ambiguities in the pointing direction are avoided. Specifying the thrust direction by two angles (e.g., yaw and pitch) which are treated as control variables is not unique since the angles $\alpha = \alpha_0 \pm 2k\pi$ all yield the *same* direction. In contrast, there is a *unique* set of control variables $\mathbf{u}$ corresponding to any thrust direction.

3. The boundary conditions.

3.1. Initial conditions. The nominal launch window opens on December 20, 2002 at 23:18:48. For our trajectory, we define a reference epoch at midnight on December 20, 2002, which corresponds to a Julian date of 2452629.5. More precisely, we define all trajectory times with respect to this reference epoch and, in particular, the beginning of phase 1 according to

$$\mathcal{T} = 2452629.5 + t_{_i}. \tag{3.1}$$

The spacecraft mass at the beginning of the mission is 350 kg, and it utilizes a single PPS-1350 Hall-plasma thruster with a force of 73.19 mN and an exhaust velocity of 16.434 km/s. The parameters of the nominal geocentric transfer orbit (GTO) established after deployment from an Ariane-5 are

semimajor axis (km)	$\underline{a}$	24661.144
eccentricity	$\underline{e}$	.716227925
inclination (deg)	$\underline{\imath}$	7.0
argument of perigee (deg)	$\underline{\omega}$	178.
ascending node		free.

Since the classical orbital elements are related to the modified equinoctial elements (cf. [6]), we can impose equivalent initial conditions directly in terms of the equinoctial elements as follows:

$$p_{_i} = \underline{p}, \tag{3.2}$$

$$\sqrt{f_{_i}^2 + g_{_i}^2} = \underline{e}, \tag{3.3}$$

$$\sqrt{h_{_i}^2 + k_{_i}^2} = \tan\left[\frac{\underline{\imath}}{2}\right], \tag{3.4}$$

$$g_{_i}h_{_i} - f_{_i}k_{_i} = \underline{e}\sin\left[\underline{\omega}\right]\tan\left[\frac{\underline{\imath}}{2}\right], \tag{3.5}$$

$$f_{_i}h_{_i} + g_{_i}k_{_i} \leq 0, \tag{3.6}$$

where $\underline{p} = \underline{a}\left(1 - \underline{e}^2\right)$ and quantities evaluated at the initial time are denoted by the subscript "$_{_i}$" as in $p_{_i} = p(t_{_i})$.

3.2. Final conditions. The target orbit around the Moon is polar (in the lunar coordinate system) and elliptic and has its pericenter above the south pole. The parameters of the desired lunar orbit are

radius of perilune (km)	$\overline{r}_p$	2738.
radius of apolune (km)	$\overline{r}_a$	11738.
inclination (deg)	$\overline{\imath}$	90.
argument of perilune (deg)	$\overline{\omega}$	270.
ascending node		free.

As before, the classical orbital elements define a corresponding set of final conditions on

the equinoctial elements

$$p_f = \overline{p}, \tag{3.7}$$

$$\sqrt{f_f^2 + g_f^2} = \overline{e}, \tag{3.8}$$

$$\sqrt{h_f^2 + k_f^2} = \tan\left[\frac{\overline{\imath}}{2}\right], \tag{3.9}$$

$$g_f h_f - f_f k_f \leq 0, \tag{3.10}$$

$$f_f h_f + g_f k_f = \overline{e}\cos\left[\overline{\omega}\right]\tan\left[\frac{\overline{\imath}}{2}\right], \tag{3.11}$$

where $\overline{p} = \overline{a}\left(1 - \overline{e}^2\right)$, $\overline{a} = \frac{1}{2}\left(\overline{r}_p + \overline{r}_a\right)$, and $\overline{e} = \left(\overline{r}_a - \overline{a}\right)/\overline{a}$. We assume that the final time t_f for phase 3 (and the mission) is free, and quantities evaluated at the final time are denoted by the subscript "$_f$" as in $p_f = p(t_f)$.

3.3. Intermediate conditions. Because the complete transfer is modeled using three phases, conditions must be imposed at the phase boundaries. Clearly, we must have continuity in the state across the phase boundaries. For reasons to be described in section 6, we impose these continuity conditions in Cartesian coordinates. For the trajectory illustrated in Figure 1, we denote the left side of the phase 1 boundary by t_1^-. Similarly, the right side of the boundary is denoted by t_1^+. Thus we require

$$\mathbf{r}[\mathbf{y}(t_1^-)] = \mathbf{r}[\mathbf{y}(t_1^+)], \tag{3.12}$$

$$\mathbf{v}[\mathbf{y}(t_1^-)] = \mathbf{v}[\mathbf{y}(t_1^+)], \tag{3.13}$$

where $\mathbf{r}[\mathbf{y}(t_1^-)]$ is the Cartesian position vector evaluated at the *end* of phase 1, and $\mathbf{r}[\mathbf{y}(t_1^+)]$ is the position vector at the *beginning* of phase 2. The Cartesian position and velocity vectors can be evaluated using the relations (2.13) and (2.14).

In like fashion, we also require continuity in the position and velocity across the phase 2 boundary. However, at the phase 2 boundary the dynamics also switch from a geocentric to a selenocentric frame. First, we define the orientation of the lunar Cartesian frame with respect to the Earth centered inertial (ECI) system. We have defined the lunar z-axis perpendicular to the lunar orbit plane; the lunar x-axis is along the intersection of the lunar orbit with the equatorial plane, and the y-axis forms a right-hand system. The transformation is defined by

$$\mathbf{Q}_{em} = \left[\begin{array}{ccc} \mathbf{i}_m & \mathbf{j}_m & \mathbf{k}_m \end{array}\right], \tag{3.14}$$

where

$$\mathbf{k}_m = \frac{\mathbf{r}_m \times \mathbf{v}_m}{\|\mathbf{r}_m \times \mathbf{v}_m\|}, \tag{3.15}$$

$$\mathbf{i}_m = \frac{\mathbf{k}_m \times \mathbf{k}_e}{\|\mathbf{k}_m \times \mathbf{k}_e\|}, \tag{3.16}$$

$$\mathbf{j}_m = \frac{\mathbf{k}_m \times \mathbf{i}_m}{\|\mathbf{k}_m \times \mathbf{i}_m\|}, \tag{3.17}$$

and $\mathbf{k}_e^{\mathsf{T}} = (0, 0, 1)$ defines the ECI z-axis. The position $\mathbf{r}_m$ and velocity $\mathbf{v}_m$ of the Moon in the ECI frame are defined at the reference epoch. Now if we apply this transformation to the definition (2.23), the linkage conditions across the phase 2 boundary are

$$\mathbf{Q}_{em}^{\mathsf{T}} \left\{ \mathbf{r}[\mathbf{y}(t_2^-)] - \mathbf{s}(t_2^-) \right\} = \mathbf{r}[\mathbf{y}(t_2^+)], \tag{3.18}$$

$$\mathbf{Q}_{em}^{\mathsf{T}} \left\{ \mathbf{v}[\mathbf{y}(t_2^-)] - \dot{\mathbf{s}}(t_2^-) \right\} = \mathbf{v}[\mathbf{y}(t_2^+)]. \tag{3.19}$$

The position and velocity of the Moon relative to the Earth, namely, $\mathbf{s}(t_2^-)$ and $\dot{\mathbf{s}}(t_2^-)$, can be computed from the ephemeris. As before, the Cartesian coordinates of the vehicle can be computed from the equinoctial state using (2.13) and (2.14).

And finally, since there is no weight change during the coast phase, we also require

$$w(t_1^-) = w(t_2^+). \tag{3.20}$$

4. Optimal control algorithm. The method used to solve the optimal control problem is referred to as a *collocation* or direct transcription algorithm [3, 4, 5, 6, 7, 8], as implemented in the $\mathbb{SOCS}$ software [1] (http://www.boeing.com/phantom/socs/). There are three basic operations performed in the method:

Direct transcription. *Transcribe* the optimal control problem into a nonlinear programming (NLP) problem by discretization.

Sparse NLP. Solve the sparse NLP using sequential quadratic programming (QP); i.e.,

1. solve a sparse QP to estimate the NLP solution;
2. if the solution is acceptable, terminate; otherwise, update the NLP solution estimate, and solve a new QP subproblem.

Mesh refinement. Assess the accuracy of the approximation (i.e., the finite dimensional problem), and if necessary refine the discretization, and then repeat the optimization steps.

The basic idea of a transcription method is to replace the optimal control problem by its finite dimensional counterpart via discretization. All approaches divide a phase into n_s segments

$$t_I = t_1 < t_2 < \cdots < t_M = t_F, \tag{4.1}$$

where the points are referred to as node, mesh, or grid points. Define the number of mesh points as $M \equiv n_s + 1$. Let us introduce the notation $\mathbf{z}_k \equiv \mathbf{z}(t_k)$ to indicate the value of the state variable at a grid point. In like fashion, denote the control at a grid point by $\mathbf{u}_k \equiv \mathbf{u}(t_k)$. In addition, the control variable at the midpoint of an interval is denoted by $\bar{\mathbf{u}}_k \equiv \mathbf{u}(\bar{t})$ with $\bar{t} = \frac{1}{2}(t_k + t_{k-1})$. To be consistent, we also denote the right-hand sides of the differential equations by $\mathbf{f}_k \equiv \mathbf{f}\left[\mathbf{z}(t_k), \mathbf{u}(t_k), \mathbf{p}, t_k\right]$. For the Hermite–Simpson discretization, the NLP variables are

$$\mathbf{x}^{\mathsf{T}} = (\mathbf{z}_1, \mathbf{u}_1, \bar{\mathbf{u}}_2, \ldots, \bar{\mathbf{u}}_M, \mathbf{z}_M, \mathbf{u}_M). \tag{4.2}$$

The state equations (2.3)–(2.4) are approximately satisfied by setting the *defects*

$$\boldsymbol{\zeta}_k = \mathbf{z}_{k+1} - \mathbf{z}_k - \frac{h_k}{6}\left(\mathbf{f}_k + 4\bar{\mathbf{f}}_{k+1} + \mathbf{f}_{k+1}\right), \tag{4.3}$$

where

$$\overline{\mathbf{z}}_{k+1} = \frac{1}{2}(\mathbf{z}_k + \mathbf{z}_{k+1}) + \frac{h_k}{8}(\mathbf{f}_k - \mathbf{f}_{k+1}), \tag{4.4}$$

$$\overline{\mathbf{f}}_{k+1} = \mathbf{f}\left(\overline{\mathbf{z}}_{k+1}, \overline{\mathbf{u}}_{k+1}, t_k + \frac{h_k}{2}\right), \tag{4.5}$$

to zero for $k = 1, \ldots, n_s$. The right-hand sides of the differential equations (2.3)–(2.4) are given by $\mathbf{f}_k \equiv \mathbf{f}[\mathbf{z}(t_k), \mathbf{u}(t_k), \mathbf{p}, t_k]$. The step size is denoted by

$$h_k = \tau_k(t_F - t_I) = \tau_k \Delta t, \tag{4.6}$$

where $\Delta t \equiv (t_F - t_I)$ with constants $0 < \tau_k < 1$ chosen so that the grid points are located at fixed fractions of the total phase duration.

5. Objective function. The overall mission objective is to maximize the vehicle weight in the lunar mission orbit, which is equivalent to minimizing the fuel consumed during the transfer. One major hurdle of a low thrust trajectory is just satisfying the constraints, and this subject will be discussed more fully in section 6. A technique for dealing with this is to pose a more "relaxed" optimization problem. First, we relax the constraint on the semiparameter of the final orbit which determines the orbit energy; i.e., instead of (3.7) we impose

$$p_f \geq \overline{p}. \tag{5.1}$$

We also introduce an impulsive velocity change at the end of phases 1 and 2. Let us introduce new variables $\boldsymbol{\varrho}$ and $\boldsymbol{\varphi}$ and then define

$$\boldsymbol{\Delta}\mathbf{v}_1 = \boldsymbol{\varrho}_1 - \boldsymbol{\varphi}_1, \tag{5.2}$$

$$\boldsymbol{\Delta}\mathbf{v}_2 = \boldsymbol{\varrho}_2 - \boldsymbol{\varphi}_2, \tag{5.3}$$

subject to the bounds

$$\boldsymbol{\varrho}_1 \geq \mathbf{0}, \tag{5.4}$$

$$\boldsymbol{\varrho}_2 \geq \mathbf{0}, \tag{5.5}$$

$$\boldsymbol{\varphi}_1 \geq \mathbf{0}, \tag{5.6}$$

$$\boldsymbol{\varphi}_2 \geq \mathbf{0}. \tag{5.7}$$

It is straightforward to incorporate the impulsive velocity change into the trajectory definition by simply altering the phase linkage conditions (3.13) and (3.19), respectively, as follows:

$$\begin{aligned} \mathbf{v}[\mathbf{y}(t_1^-)] &= \mathbf{v}[\mathbf{y}(t_1^+)] - \boldsymbol{\Delta}\mathbf{v}_1 \\ &= \mathbf{v}[\mathbf{y}(t_1^+)] - (\boldsymbol{\varrho}_1 - \boldsymbol{\varphi}_1), \end{aligned} \tag{5.8}$$

$$\begin{aligned} \mathbf{Q}_{em}^{\mathsf{T}}\left\{\mathbf{v}[\mathbf{y}(t_2^-)] - \dot{\mathbf{s}}(t_2^-)\right\} &= \mathbf{v}[\mathbf{y}(t_2^+)] - \boldsymbol{\Delta}\mathbf{v}_2 \\ &= \mathbf{v}[\mathbf{y}(t_2^+)] - (\boldsymbol{\varrho}_2 - \boldsymbol{\varphi}_2). \end{aligned} \tag{5.9}$$

Recalling that $\|\mathbf{x}\|_1 = \sum_k |x_k|$, observe that the magnitude of the velocity change is

$$\|\boldsymbol{\Delta}\mathbf{v}_1\|_1 + \|\boldsymbol{\Delta}\mathbf{v}_2\|_1 = \mathbf{e}^{\mathsf{T}}\left[\boldsymbol{\varrho}_1 + \boldsymbol{\varphi}_1 + \boldsymbol{\varrho}_2 + \boldsymbol{\varphi}_2\right], \tag{5.10}$$

where $\mathbf{e}^{\mathsf{T}} = (1, 1, 1)$. We are now ready to define the *composite objective* function:

$$
\begin{aligned}
J &= -\frac{w_f}{w_i} + \rho \left[\|\Delta \mathbf{v}_1\|_1 + \|\Delta \mathbf{v}_2\|_1 + \frac{1}{\overline{p}}(p_f - \overline{p}) \right] \\
&= -\frac{w_f}{w_i} + \boldsymbol{\rho}^{\mathsf{T}} \left[\boldsymbol{\varrho}_1 + \boldsymbol{\varphi}_1 + \boldsymbol{\varrho}_2 + \boldsymbol{\varphi}_2 \right] + \frac{\rho}{\overline{p}}(p_f - \overline{p}),
\end{aligned} \tag{5.11}
$$

where $\boldsymbol{\rho}^{\mathsf{T}} = \rho \mathbf{e}^{\mathsf{T}}$, $w_i = w(t_i)$, $w_f = w(t_f)$, and $p_f = p(t_f)$.

It is worth emphasizing the important features of the problem formulation. First, the "real" problem requires maximizing the final weight w_f subject to the "real" constraints (3.7), (3.13), and (3.19). Clearly, the "real" constraints will be satisfied by the relaxed formulation provided $\|\Delta \mathbf{v}_1\|_1 = \|\Delta \mathbf{v}_2\|_1 = 0$ and $p_f = \overline{p}$. The composite objective function includes these contributions as a penalty term. Since the composite objective utilizes a penalty which is "exact" [11], the "real" problem and the "relaxed" problem have the same solution provided the (finite) scalar penalty weight $\rho \geq \|\boldsymbol{\lambda}\|_\infty$, where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers for the sparse NLP subproblem. For our results, we set $\rho = 10^5$. However, why not treat $\Delta \mathbf{v}_1$ and $\Delta \mathbf{v}_2$ as variables directly? By introducing the slack variables $\boldsymbol{\varrho}$ and $\boldsymbol{\varphi}$ along with the nonnegativity bounds (5.4)–(5.7), we treat the constraints $\|\Delta \mathbf{v}_1\|_1 = \|\Delta \mathbf{v}_2\|_1 = 0$ without introducing a nondifferentiable function! Of course, the NLP algorithm could treat *all* constraints this way. The SNOPT algorithm [12] uses this technique, where $\boldsymbol{\varrho}$ and $\boldsymbol{\varphi}$ are called *elastic variables*. However, because the number of constraints is large for our application, we have chosen to introduce slack variables in a very selective manner. Finally, since all of the penalty terms appear *linearly* in the composite objective function, the Hessian matrix is not altered from the original problem. Fortunately, this formulation also leads to a very natural physical interpretation which we can exploit when constructing an initial guess as described in section 6.

At this point, it is worth discussing the issue of *local* versus *global* solutions. It is well known that orbit transfers often have many locally optimal solutions. For example, in Zondervan, Wood, and Caughy [21], optimal trajectories that are "geometrically similar" to impulsive solutions are obtained using many low thrust burns, and comparable results are described by Redding and Breakwell [15]. The multiple burn concept for this mission has been investigated by Schoenmaekers, Horas, and Pulido [16]. In essence it is more fuel efficient to use many (short duration) burns than one long duration burn. Unfortunately, there is a penalty for enhanced fuel efficiency—mission duration. Multiple burn transfers take longer to perform, and this can have significant impact on other aspects of the mission design. Thus, to be more precise about our results, the trajectory we present is a minimum fuel transfer *using two burns.* In fact, we also demonstrate that it is close to a minimum time *two burn* transfer.

6. Getting started. As with any iterative method, it is necessary to supply an initial guess. For a collocation method this means the analyst must supply the NLP variables (4.2) as well as the associated grid distribution (4.1). In general the direct transcription method outlined in section 4 is quite robust, and this is one of the more desirable features of the algorithm. In fact, for many problems it suffices to supply a guess that is a simple linear function between the boundary conditions. Unfortunately, this is not true for our application. The dynamics are very nonlinear, the problem size is large, and there are many local solutions.

Thus it is critical to construct a guess which is "in the ball-park." We describe how an initial guess was constructed for our results. It should be emphasized that there is nothing unique about our heuristics, and it is quite possible that better alternatives could be used to get started.

The initial guess procedure assumes that the thrust should be along the velocity vector in either a posigrade or retrograde direction. The posigrade direction is given by

$$\widehat{\mathbf{u}} = \mathbf{Q}^{\mathsf{T}} \frac{\mathbf{v}(\mathbf{y})}{\|\mathbf{v}(\mathbf{y})\|}. \tag{6.1}$$

Using this approximation, the disturbing acceleration caused by the thrust from (2.28) is just

$$\widehat{\boldsymbol{\Delta}}_T(\mathbf{y}) = \frac{g_o T}{w} \widehat{\mathbf{u}}, \tag{6.2}$$

which can be substituted into the original system of DAEs (2.3)–(2.5) to yield a system of ordinary differential equations:

$$\dot{\mathbf{z}} = \begin{bmatrix} \dot{\mathbf{y}} \\ \dot{w} \end{bmatrix} = \begin{bmatrix} \mathbf{A}(\mathbf{y})\widehat{\boldsymbol{\Delta}}(\mathbf{y}) + \mathbf{b} \\ -T/I_{sp} \end{bmatrix} = \mathbf{f}(\mathbf{z}). \tag{6.3}$$

For our application, it is likely that the orbits will be highly eccentric during portions of the trajectory, and this has implications for the construction of an initial guess. In particular, since the velocity near periapsis of an orbit is much larger than near apoapsis, we would like to choose an initial guess with grid points distributed to reflect this asymmetry. In other words, we would like a guess with many grid points near periapsis and few near apoapsis. To achieve this goal, let us treat the variable L in (2.1) as the *independent* variable. Since

$$\dot{\mathbf{z}} = \frac{d\mathbf{z}}{dL}\frac{dL}{dt} = \mathbf{f}(\mathbf{z}), \tag{6.4}$$

we can write

$$\frac{d\mathbf{z}}{dL} = \mathbf{f}(\mathbf{z}) \left[\frac{dL}{dt}\right]^{-1} = \frac{1}{f_6(\mathbf{z})}\mathbf{f}(\mathbf{z}), \tag{6.5}$$

and this transformation is valid provided $\frac{dL}{dt} > 0$. Since the true longitude L is an angle in the orbit plane, it is reasonable to choose grid points such that the angles are equidistributed; i.e.,

$$L_k = L_\imath + \frac{(k-1)}{(M-1)}(L_f - L_\imath) \tag{6.6}$$

for $k = 1, \ldots, M$, where $L_\imath$ and L_f are the initial and final values for L, and M is the number of grid points on the phase. Observe that grid points with equidistributed angles do not produce grid points equidistributed in time. In fact, this approach will cluster many grid points near periapsis, with few near apoapsis. Figure 12 illustrates the initial asymmetric grid distribution on each phase as a solid region. Let us now describe the steps used to construct the initial guess.

Step 1. Guess a value for the parameter $r_a(t_1)$, the apogee radius at the end of phase 1. Numerically integrate the equations of motion (6.3) from $t_i = 0$ with the park orbit initial conditions given in section 3.1. Terminate phase 1 when the apogee radius is equal to $r_a(t_1)$. Propagate the coast phase for one year (i.e., to $\hat{T} = t_i + 1$ (yr.)). The goal is to locate a point on this trajectory such that the spacecraft is "close" to the Moon and is likely to be captured by the lunar gravity. To quantify "close," we simply measure the distance from the spacecraft to the Moon. For a capture condition, we would like the velocity vector to be directed toward the Moon with a small value for the lunar eccentricity. So, every hour during the coast, compute the distance to the Moon d_m, the eccentricity relative to the Moon e_m, and the flight path angle relative to the Moon. If the trajectory is approaching the Moon (has a negative flight path angle), then save the time $\hat{t}_2$, with the smallest value for the quantity $\phi = w_1 e_m^2 + w_2(d_m - \hat{d}_m)^2$. The weights w_1 and w_2 and target distance $\hat{d}_m$ are not critical since our goal is to achieve a "good" trajectory that goes near the Moon. We repeat this process for a series of different target radius values $r_a(t_1)$. This "brute force" sweep over the parameter space can be automated easily and requires very little computation time. For our example, we found $r_a(t_1) \approx 437500$(km) at a burnout time of $\hat{t}_1 = 159.999$(days), and the closest approach time was $\hat{t}_2 = 170.791$(days). A variable order, variable stepsize, Adams predictor-corrector integration algorithm was used to propagate the trajectory. When step 1 is completed, we have a guess for the phase times $\hat{t}_1$ and $\hat{t}_2^-$ as well as the corresponding states $\widehat{\mathbf{z}}(\hat{t}_1)$ and $\widehat{\mathbf{z}}(\hat{t}_2^-)$.

Step 2. Next we focus entirely on the selenocentric portion of the trajectory—phase 3. Here we assume that the thrust is strictly retrograde; i.e., we replace (6.1) by $\widetilde{\mathbf{u}} = -\widehat{\mathbf{u}}$. Using the specified lunar orbit conditions given in section 3.2, we integrate backward from the final time to solve the ordinary differential equations (6.3). Specifically, we solve a small optimization problem with four variables and one constraint. The final values for weight, time, longitude of the ascending node, and true anomaly are treated as variables; i.e., we define an NLP with $\mathbf{x}^\mathsf{T} = (w_f, t_f, \Omega_f, \nu_f)$. The trajectory is propagated backward from t_f to $\hat{t}_2^+$. Of course the time at the beginning of phase 3 is the same as the time at the end of phase 2; i.e., $\hat{t}_2^+ = \hat{t}_2^-$. The weight computed from the reverse trajectory propagation must match the value at the end of phase 1, leading to the single constraint $c(\mathbf{x}) = w(\hat{t}_2^+) - w(\hat{t}_1^-) = 0$. We would like to choose the variables so that the state at the end of phase 2 is close to the state at the beginning of phase 3; consequently, we minimize

$$\phi = \left[\mathbf{r}(\hat{t}_2^-) - \mathbf{r}(\hat{t}_2^+)\right]^\mathsf{T} \left[\mathbf{r}(\hat{t}_2^-) - \mathbf{r}(\hat{t}_2^+)\right] + \left[e(\hat{t}_2^-) - e(\hat{t}_2^+)\right]^2. \tag{6.7}$$

Just as in step 1, this small optimization problem can be solved quickly, and it defines a guess for the third phase which is a good approximation to the selenocentric spiral trajectory. Note that this iteration defines a reasonable orientation for the mission orbit.

Step 3. A reasonable guess for phase 1 is constructed in step 1, and step 2 yields a good guess for phase 3. Finally, let us construct a guess that couples phases 1 and 3. Here we treat the position and time at the end of phase 1 and the beginning of phase 3 as fixed quantities. Again, define a small problem with three variables and three constraints. Specifically, treat the velocity at the beginning of phase 2 as variables; i.e., $\mathbf{x} = \mathbf{v}(\hat{t}_1^+)$. We choose these values so that the position at the end of phase 2 matches the initial position for phase 3; in other words, we impose the constraints $\mathbf{c}(\mathbf{x}) = \mathbf{r}(\hat{t}_2^-) - \mathbf{r}(\hat{t}_2^+) = \mathbf{0}$. This (Lambert's) problem has

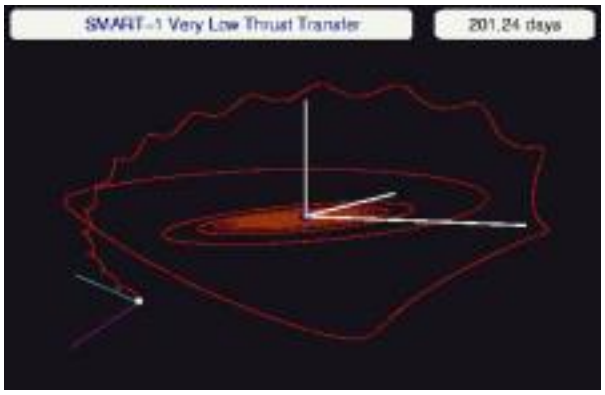


Figure 2. *Optimal trajectory animation.*

a solution provided we allow impulsive velocity increments at the phase boundary. It also is consistent with the composite objective function defined by (5.11).

Step 4. After completing the first three steps of our initialization procedure, reasonable values for the state and control variables are available at all phase boundaries; i.e., we have computed guesses for $(\mathbf{y}, w, \mathbf{u}, t)$ at t_i, $\hat{t}_1^-$, $\hat{t}_1^+$, $\hat{t}_2^-$, $\hat{t}_2^+$, and t_f. The final step in the initialization process requires computing values for these quantities at grid points between the phase boundaries. For this we numerically integrate the transformed equations (6.5) using a variable order, variable stepsize integrator to construct values that are equidistributed with respect to L, as in (6.6). For our numerical results, we constructed an initial grid with 18 points per revolution during phase 1 and 54 points per revolution during phases 2 and 3. The total number of grid points for the initial guess (prior to mesh refinement) was 3310.

7. Numerical results.

7.1. Minimum fuel two burn transfer. Figure 2 presents an animation of the optimal trajectory.

Figure 3 illustrates the optimal trajectory in an Earth centered Cartesian coordinate system. The position of the Moon at the end of phase 1 is illustrated with a solid circle, and the corresponding position of the vehicle at the same time is shown with a solid triangle. At the end of phase 2, the Moon and vehicle are plotted with an open circle and a triangle, respectively. The orbit of the Moon during the third phase is plotted with a dashed line, and the final position is shown with a diamond symbol. For comparison, Figure 4 illustrates the trajectory used as an initial guess. The coarse discretization grid is clearly evident in this plot. Figure 5 shows the trajectory during the third (selenocentric) phase. Because this figure illustrates the motion *relative* to the Moon, it clearly shows the spiral nature of the lunar orbit insertion. Figure 6 illustrates the solution in a rotating selenocentric frame. Specifically, this frame is centered at the Moon with the x-axis pointed toward the Earth and the z-axis normal the lunar orbit. The individual equinoctial elements are plotted in Figure 7, and the corresponding values of the classical orbit elements are illustrated in Figure 8. Figure 9 illustrates the three components of the control vector $\mathbf{u}$ which define the optimal steering used

Table 1
Optimal trajectory at phase boundaries.

	t_i	t_1	t_2	t_f
t(day)	-.442651	160.969	167.799	201.284
$\|\mathbf{r}_E\|$(km)	9522.7	313142.6	423616.2	369095.6
$\|\mathbf{r}_L\|$(km)	387718.2	126246.0	103270.7	4267.2
w(kg)	350.000	287.891	287.891	275.006

Table 2
Mesh refinement summary.

k	M	NGC	NHC	NFE	NRHS	ϵ	Time (sec)
1	3310	413	239	89481	592095777	5.0012×10^{-2}	4.2000×10^{4}
2	4940	837	835	293454	2898445158	2.7758×10^{-3}	6.8547×10^{4}
3	5162	14	12	4290	44277090	2.0359×10^{-4}	1.6547×10^{3}
4	5384	6	4	1481	15942965	4.5207×10^{-5}	9.7732×10^{2}
5	6949	6	4	1481	20578495	3.2154×10^{-5}	1.6215×10^{3}
6	12427	4	2	779	19358929	1.7178×10^{-7}	5.9152×10^{3}
7	16285	3	1	428	13938676	1.3623×10^{-8}	7.4684×10^{3}
Total	16285	1283	1097	391394	3604637090		1.2818×10^{5}

during the transfer. In Figure 10, we plot the angle

$$\theta = \arccos\left[\mathbf{u}^{\mathsf{T}}\widehat{\mathbf{u}}\right], \tag{7.1}$$

where the posigrade direction is given by (6.1). This shows that the optimal steering deviates significantly from the posigrade direction in phase 1 and significantly from the retrograde direction in phase 3. This figure also demonstrates how much the solution differs from the initial guess. Figure 11 illustrates the transition from the Earth's sphere of influence to the Moon's. We show the distance from the Earth, the distance from the Moon, and the orbital energy. The distribution of the grid points for the initial guess is illustrated in Figure 12 with a solid gray region, and the final mesh distribution is plotted with a solid line. The final weight was 275.006 kg, which was achieved at insertion into the lunar orbit at $t_f = 201.284$ days. It is interesting to observe that the optimal duration of the coast was 6.83 days, presumably to achieve the correct phasing for lunar orbit insertion. Table 1 summarizes the optimal trajectory parameters.

The performance of the $\mathbb{SOCS}$ algorithm on this problem is presented in Table 2. The overall problem required seven mesh refinement iterations, with each iteration summarized by a single row in Table 2. The initial coarse grid had 3310 grid points. The first NLP solution was obtained after 413 gradient evaluations (NGC), and 239 Hessian evaluations (NHC). The first and second derivatives were computed using sparse finite differences and required a total of 89481 function evaluations (NFE). Since each function evaluation requires the calculation of the right-hand sides of the DAEs (2.3)–(2.5) at the grid points, the total number of right-hand side evaluations (NRHS) was 592,095,777. After the first NLP subproblem is solved, the discretization accuracy ϵ is computed, and in this case it was 5.0012×10^{-2}. The CPU time

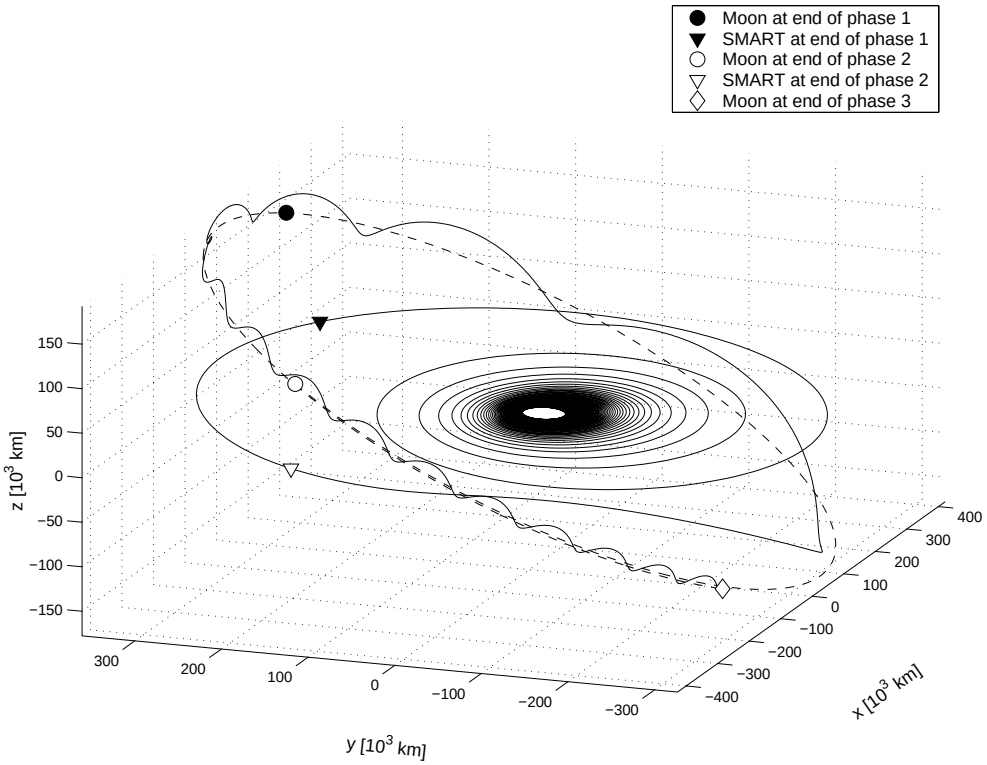


Figure 3. *Optimal trajectory ECI frame.*

required to solve the first NLP was 4.2000×10^4 seconds on a DEC Alpha with 8 Gbytes of memory. Since the discretization accuracy was not sufficiently small, the mesh was refined by adding grid points using the procedure described in [7]. This larger problem had 4940 grid points, and the NLP problem was then solved in an additional 6.8547×10^4 CPU seconds. This mesh refinement procedure was repeated seven times with the final grid containing 16285 points. Notice that as the mesh is refined the solution of the NLP subproblems becomes easier even though the problem size increases. There are two major reasons for this. First, the initial guess for each problem is constructed by interpolating the solution from the previous (coarse) grid. Consequently, as the mesh is refined the NLP initial guess becomes better and better. Furthermore, the sparse NLP algorithm utilizes a full finite difference approximation to the Hessian matrix, and consequently the NLP exhibits quadratic convergence. Quadratic convergence is extremely important because the final NLP has 211031 variables and 146285 constraints active at the solution, which means that the size of the projected Hessian matrix is 64746. Because a full Newton method is used by $\mathbb{SOCS}$, the final large NLP was solved with only one QP subproblem (NHC = 1). In contrast, a quasi-Newton method would require over 64746 iterations to solve an NLP subproblem of this size.

7.2. Minimum time two burn transfer. The preceding section presented the minimum fuel two burn transfer. For comparison we also computed a minimum time two burn transfer. It is straightforward to modify the formulation since the composite objective (5.11) is replaced

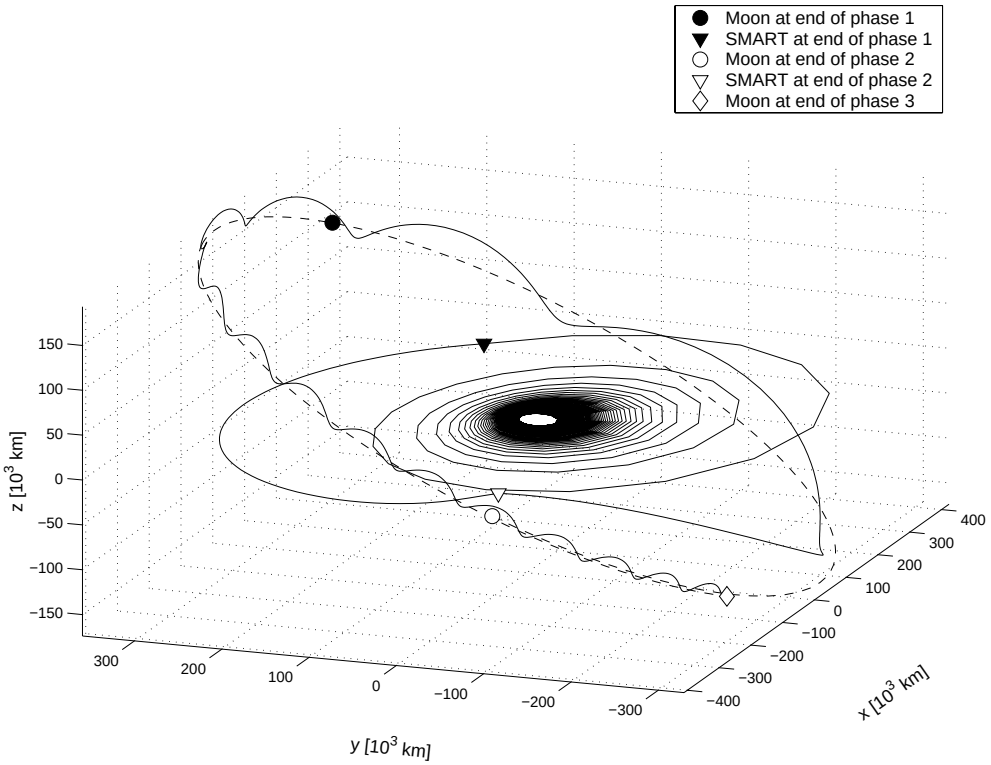


Figure 4. *Initial guess trajectory ECI frame.*

Table 3
Minimum fuel versus minimum time solutions.

		t_i	t_1	t_2	t_f
Min fuel	t(day)	-.442651	160.969	167.799	201.284
Min time	t(day)	.445787	163.015	165.171	198.384
Min fuel	w(kg)	350.000	287.891	287.891	275.006
Min time	w(kg)	350.000	287.445	287.445	274.665

Table 4
Mesh refinement summary (minimum time transfer).

k	M	NGC	NHC	NFE	NRHS	ϵ	Time (sec)
1	3310	277	226	80840	534918280	3.4766×10^{-2}	1.3512×10^{4}
2	4736	34	32	11331	107293239	5.6377×10^{-4}	4.0090×10^{3}
3	5064	691	689	242192	2452194000	3.6902×10^{-4}	6.7454×10^{4}
4	6045	8	6	2183	26385921	9.0421×10^{-5}	2.0105×10^{3}
5	6105	4	2	779	9509253	1.7918×10^{-5}	8.4666×10^{2}
6	10726	5	3	1130	24237370	6.5944×10^{-7}	3.6719×10^{3}
7	10958	70	68	23958	524991654	6.5509×10^{-6}	2.9613×10^{4}
8	11884	4	2	779	18512935	1.8837×10^{-7}	4.0347×10^{3}
9	14611	3	1	428	12505732	3.7902×10^{-8}	4.7744×10^{3}
Total	14611	1096	1029	363620	3710548384		1.2993×10^{5}

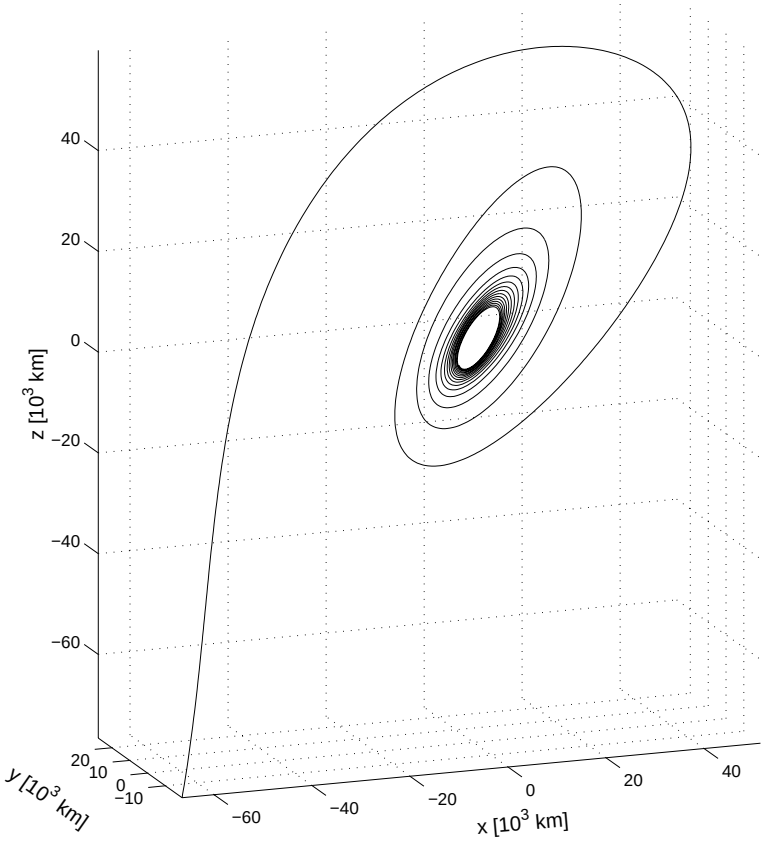


Figure 5. *Phase 3 of optimal trajectory, rotated selenocentric frame.*

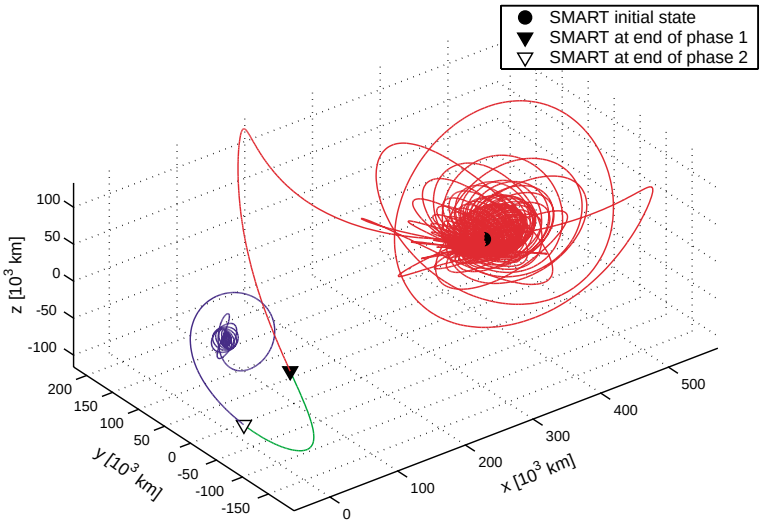


Figure 6. *Optimal trajectory in rotating selenocentric frame.*

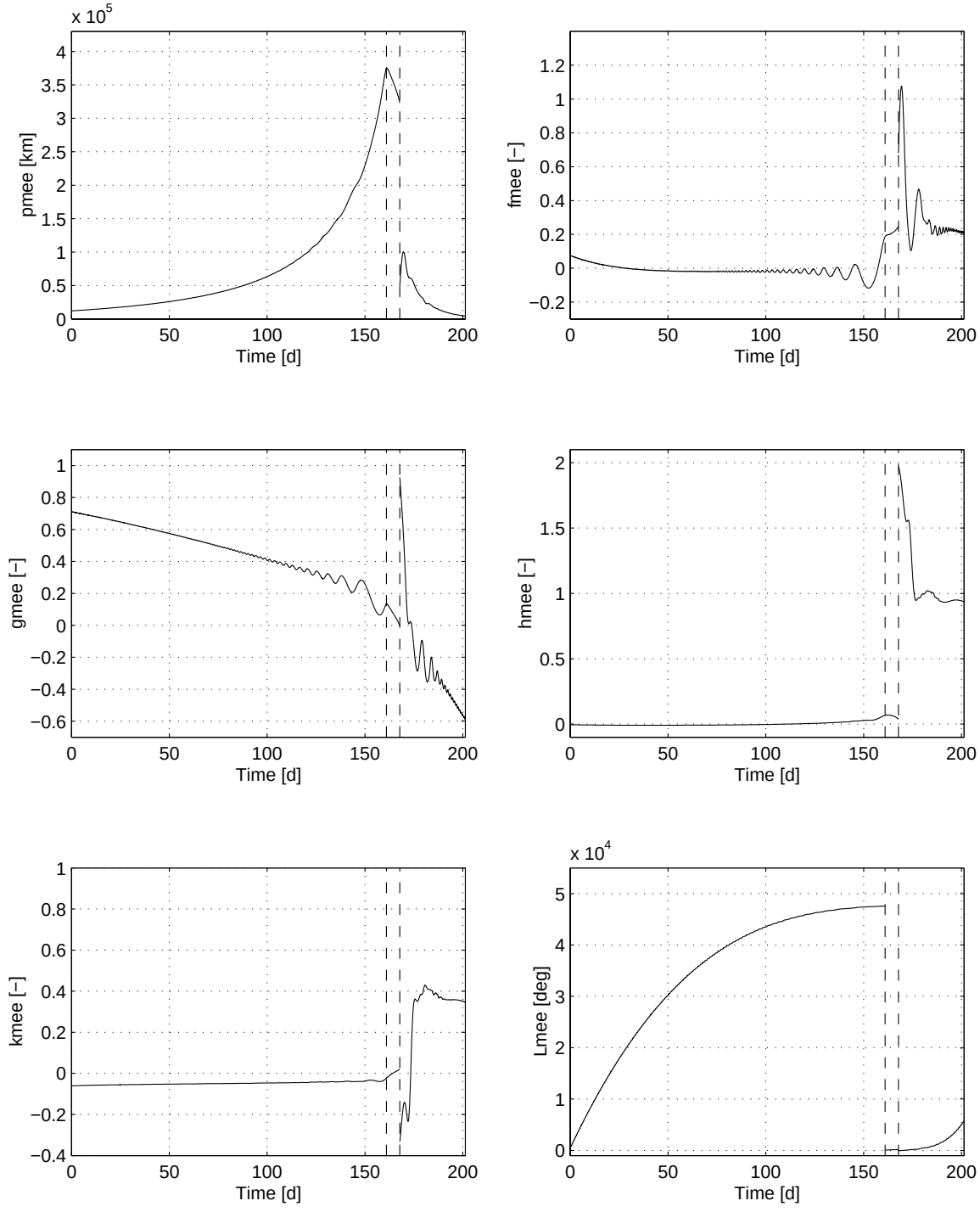


Figure 7. *Equinoctial elements during transfer.*

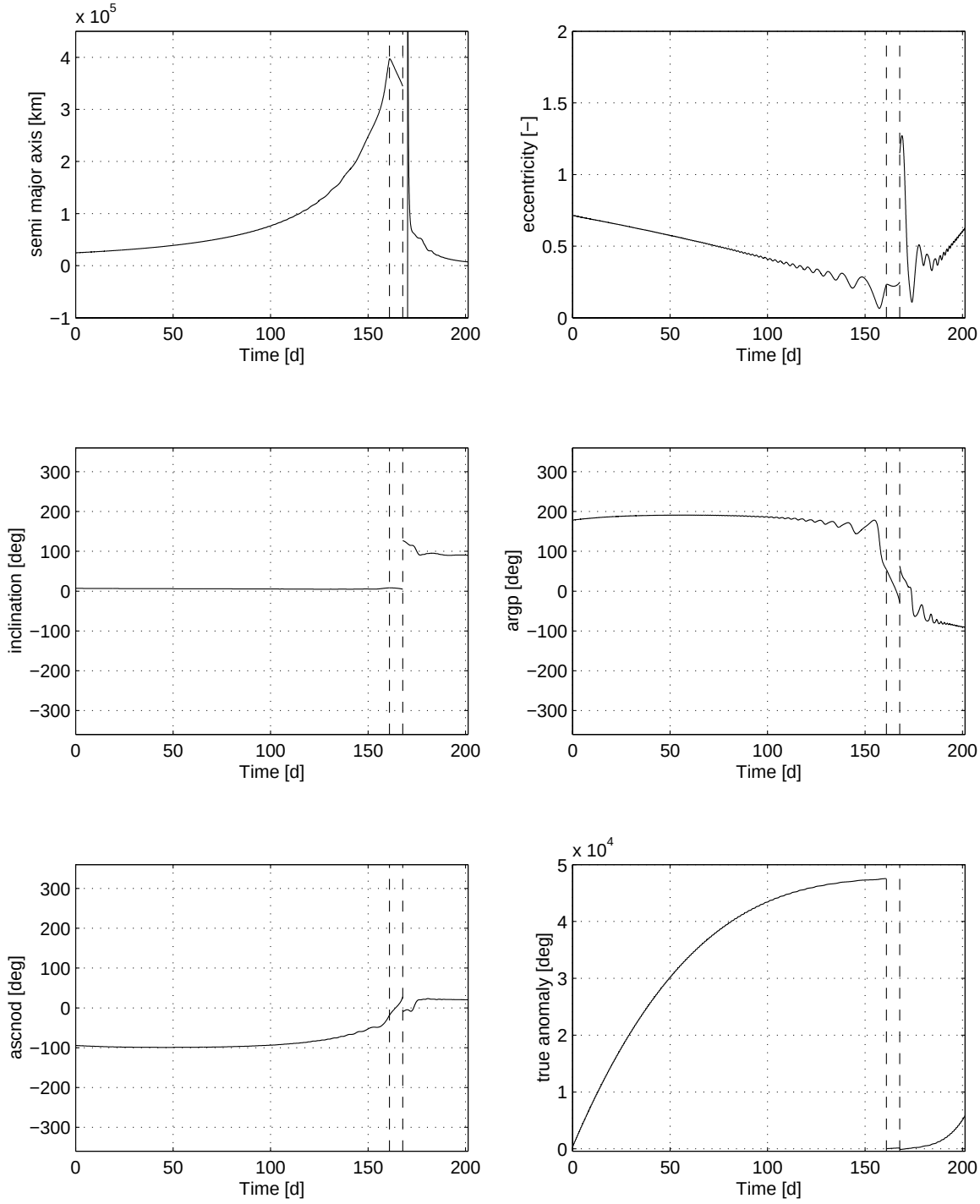


Figure 8. *Classical elements during transfer.*

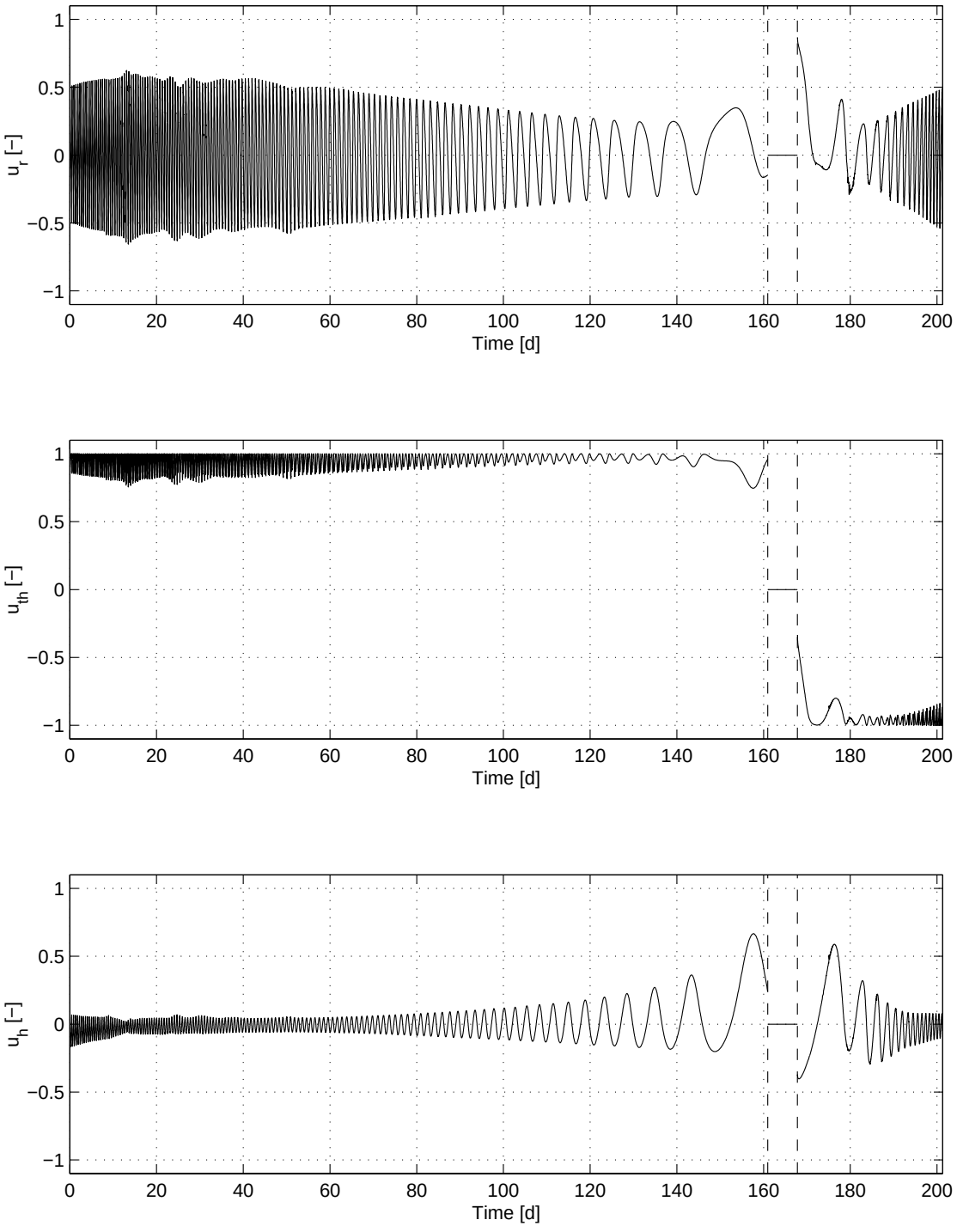


Figure 9. *Optimal control history.*

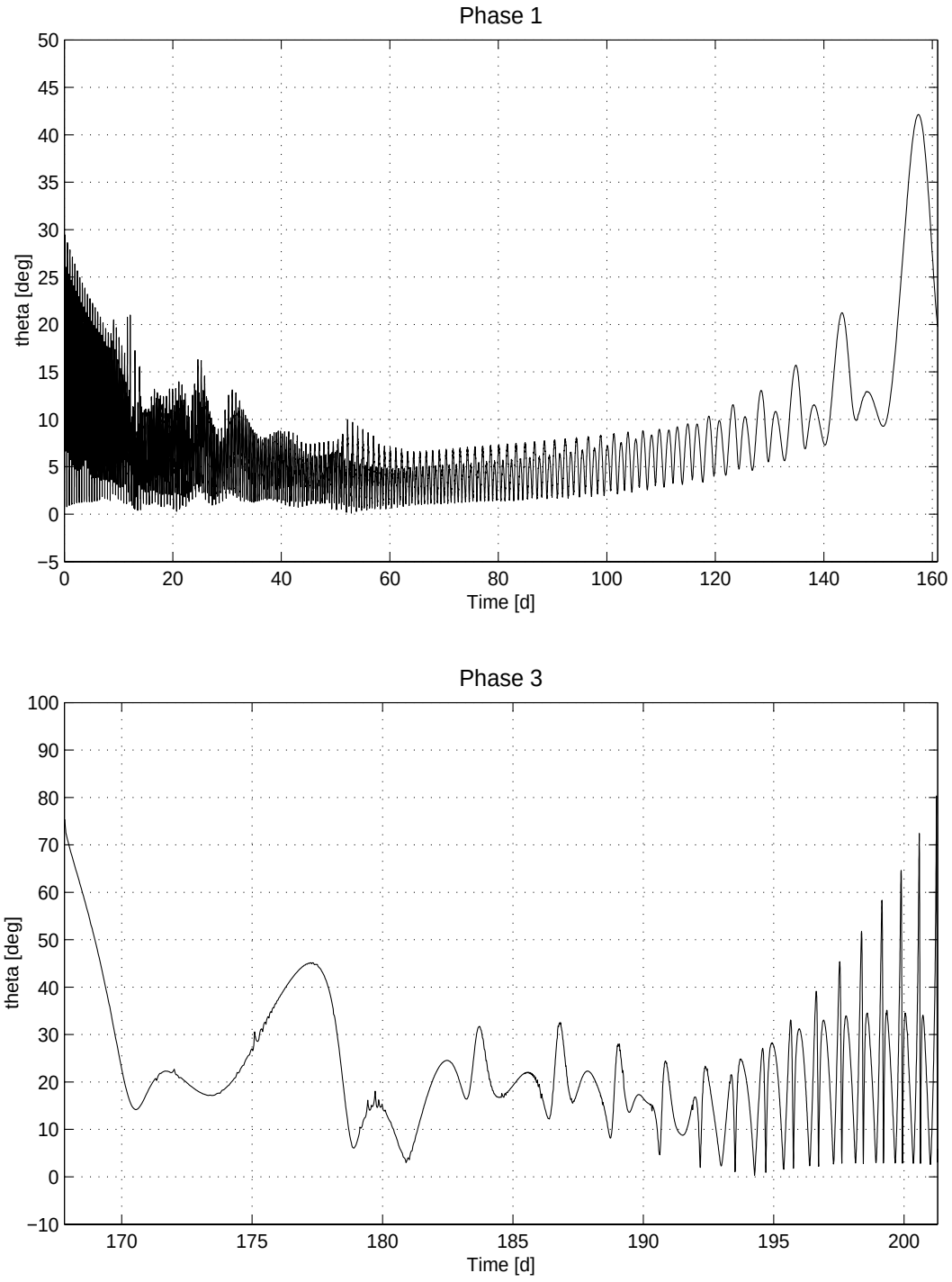


Figure 10. *Optimal control deviation from posigrade/retrograde.*

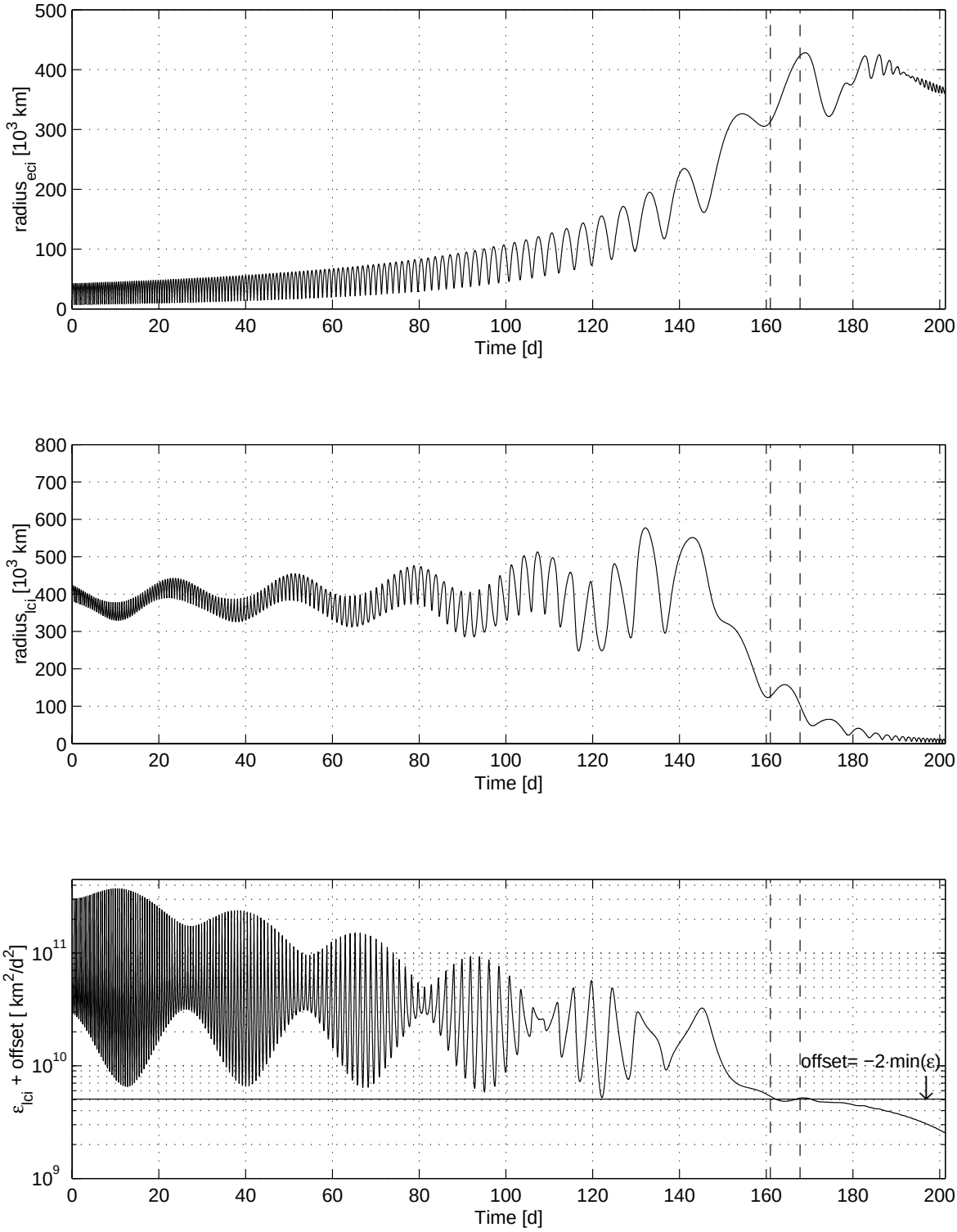


Figure 11. *Distance to Earth (top), distance to Moon (center), orbit energy (bottom).*

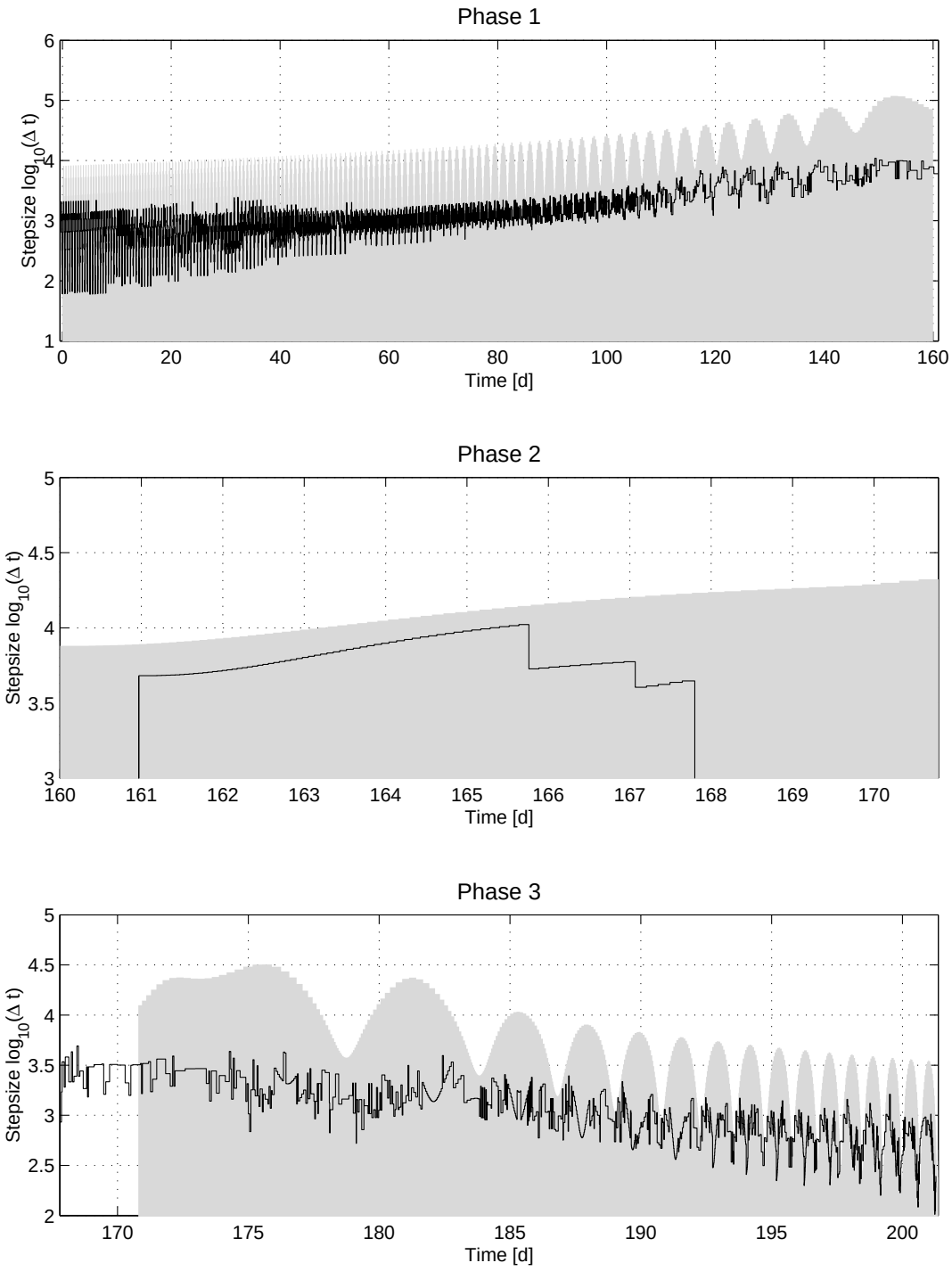


Figure 12. *Discretization grid history.*

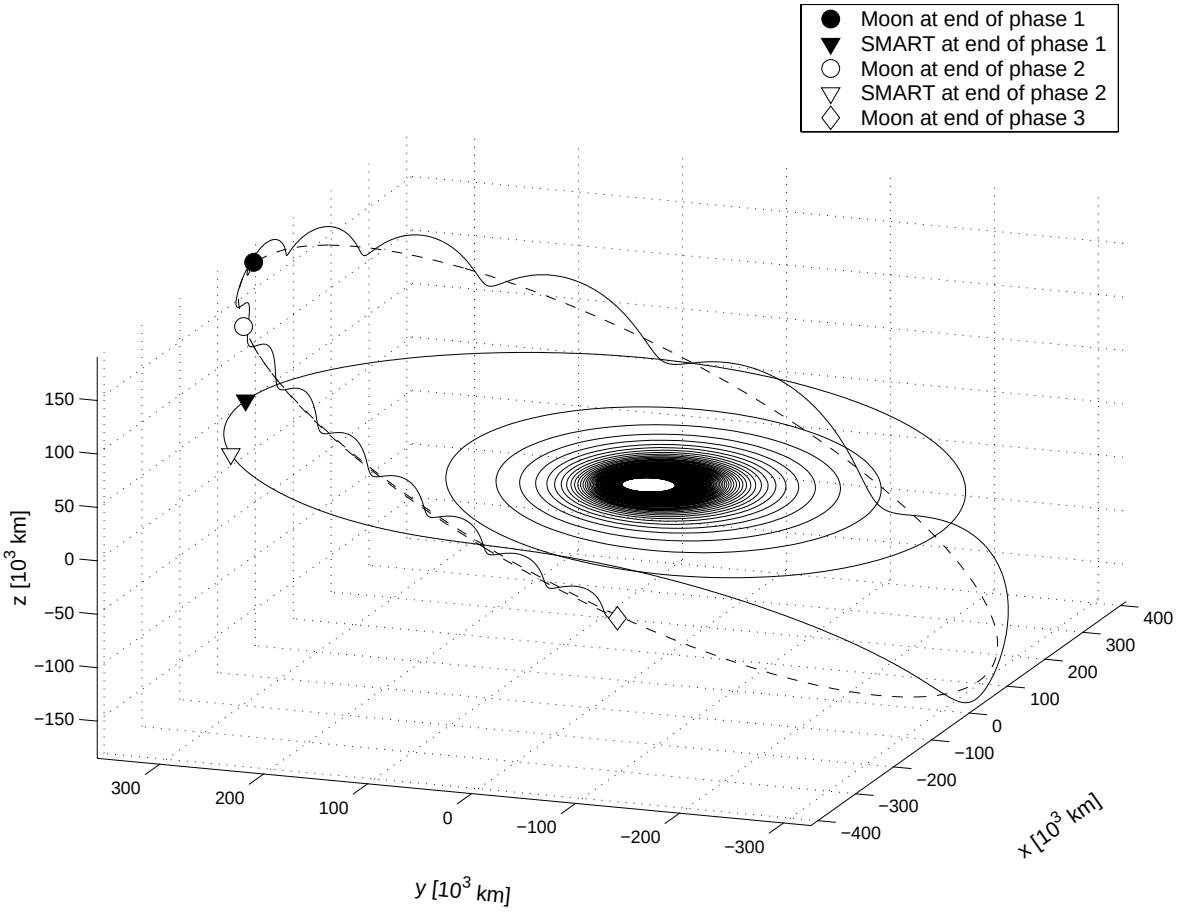


Figure 13. *Minimum time trajectory ECI frame.*

by

$$J = \rho_0(t_f - t_i) + \boldsymbol{\rho}^{\mathsf{T}}\left[\boldsymbol{\varrho}_1 + \boldsymbol{\varphi}_1 + \boldsymbol{\varrho}_2 + \boldsymbol{\varphi}_2\right] + \frac{\rho}{\overline{p}}(p_f - \overline{p}), \tag{7.2}$$

where ρ_0 is a scale factor which we set to 1/200 (days). All other equations remain unchanged.

In Table 3, the results of the minimum fuel transfer are compared with those for a minimum time transfer. Observe that the total mission duration was reduced from 201.727 days to 197.938 days, a reduction of 3.789 days. Most of this time saving can be attributed to the shorter coast (phase 2). Figure 13 illustrates the minimum time trajectory. It is also interesting to note that the minimum time transfer is accomplished with a rather small penalty in fuel consumed, i.e., an additional .341 kg. Table 4 summarizes the performance of the $\mathbb{SOCS}$ algorithm for this minimum time transfer.

8. Remarks. It is worthwhile comparing the results and methods described in this paper with those appearing elsewhere. In [16], Schoenmaekers, Horas, and Pulido summarize their results:

> Using a pragmatic engineering approach without applying optimization techniques, fuel efficient transfer trajectories have been constructed for the SMART-1 Moon mission, by combining low thrust and multiple Moon swing-by's. The transfer duration is about 17 months. The fuel demand for a 350 kg launch mass ranges from 54.3 kg to 60.9 kg depending on the launch date.

Our optimal solution requires 14–21 kg more fuel but is considerably shorter in duration (10.3 months). This difference can be attributed to the many burns used in the trajectories in [16]. Clearly other mission design considerations must be used to trade fuel consumption for mission duration. The propulsive force we used does not include thrust degradation caused by limitations on the spacecraft power system. Although this information was not available to us, it could easily be incorporated into the trajectory dynamics and would alter the results. In short, our results differ from [16] because we

- used large-scale optimization methods but
- did not use multiple burns and
- did not model thrust degradation.

On the other hand, the method we have described can easily accommodate multiple burns and thrust degradation.

There are a number of key points with regard to the method itself. A number of authors have used equinoctial coordinates to define the dynamics (cf. [2, 9, 10, 13, 14, 19, 20]). Furthermore, there are many trajectory optimization tools in use, including, CHEBYTOP, VARITOP, SEPTOP, SECKSPOT, and HillTop. In contrast to these tools, the optimization method we described is a direct method (i.e., does not require adjoint equations) and is quadratically convergent for very large scale problems (i.e., $n \geq 100000$). The approach we have presented is unique because it combines

- a large-scale, quadratically convergent, direct optimization method with
- nonlinear dynamics using equinoctial coordinates.

9. Summary and conclusions. This paper describes the solution of an optimal low thrust trajectory problem from an Earth to lunar orbit. The dynamics incorporate gravitational perturbations from the Sun, Moon, and Earth as well as oblate Earth effects. Furthermore, the method we present can readily accommodate other significant modeling effects (e.g., thrust degradation) without difficulty. The numerical method exploits the use of a large sparse NLP algorithm in conjunction with a collocation technique to approximate the differential equations. The overall approach is quadratically convergent, robust, and applicable to a wide variety of nonlinear dynamic systems.

REFERENCES

[1] *OPTLIB Optimization and Optimal Control Software Library*, Mathematics and Computing Technology Report M & CT-TECH-01-014, Boeing Information and Support Services, The Boeing Company, Seattle, WA, 2001.

[2] R. H. Battin, *An Introduction to the Mathematics and Methods of Astrodynamics*, AIAA Ed. Ser., AIAA, Washington, DC, 1987.

[3] J. T. Betts, *Trajectory optimization using sparse sequential quadratic programming*, in Optimal Control, R. Bulirsch, A. Miele, J. Stoer, and K. H. Well, eds., Internat. Ser. Numer. Math. 111, Birkhäuser, Basel, 1993, pp. 115–128.

[4] J. T. Betts, *Using sparse nonlinear programming to compute low thrust orbit transfers*, J. Astronaut. Sci., 41 (1993), pp. 349–371.

[5] J. T. Betts, *Issues in the direct transcription of optimal control problems to sparse nonlinear programs*, in Computational Optimal Control, R. Bulirsch and D. Kraft, eds., Internat. Ser. Numer. Math. 115, Birkhäuser, Basel, 1994, pp. 3–18.

[6] J. T. Betts, *Optimal interplanetary orbit transfers by direct transcription*, J. Astronaut. Sci., 42 (1994), pp. 247–268.

[7] J. T. Betts, *Practical Methods for Optimal Control Using Nonlinear Programming*, Adv. Des. Control 3, SIAM, Philadelphia, PA, 2001.

[8] J. T. Betts and W. P. Huffman, *Path constrained trajectory optimization using sparse sequential quadratic programming*, AIAA J. Guidance Control Dynam., 16 (1993), pp. 59–68.

[9] R. A. Broucke and P. J. Cefola, *On equinoctial orbit elements*, Celestial Mechanics, 5 (1972), pp. 303–310.

[10] T. N. Edelbaum, L. L. Sackett, and H. L. Malchow, *Optimal low thrust geocentric transfer*, in Proceedings of the AIAA 10th Electric Propulsion Conference, AIAA 73-1074, Lake Tahoe, NV, 1973.

[11] R. Fletcher, *Practical Methods of Optimization, Vol.* 2*, Constrained Optimization*, John Wiley and Sons, New York, 1985.

[12] P. E. Gill, W. Murray, and M. A. Saunders, *SNOPT: An SQP Algorithm for Large-Scale Constrained Optimization*, Tech. report SOL 97-3, Department of Operations Research, Stanford University, Stanford, CA, 1997.

[13] J. A. Kechichian, *Equinoctial Orbit Elements: Application to Optimal Transfer Problems*, in Proceedings of the AIAA/AAS Astrodynamics Specialist Conference, AIAA 90-2976, Portland, OR, 1990.

[14] J. A. Kechichian, *Trajectory Optimization with a Modified Set of Equinoctial Orbit Elements*, in Proceedings of the AIAA/AAS Astrodynamics Specialist Conference, AAS 91-524, Durango, CO, 1991.

[15] D. Redding and J. V. Breakwell, *Optimal low-thrust transfers to synchronous orbit*, AIAA J. Guidance Control Dynam., 7 (1984), pp. 148–155.

[16] J. Schoenmaekers, D. Horas, and J. A. Pulido, *SMART*-1 *with Solar Electric Propulsion to the Moon*, in Proceedings of the 16th International Symposium on Space Flight Dynamics, Pasadena, CA, 2001, pp. 1–14.

[17] J. Schoenmaekers, J. Pulido, and R. Jehn, *SMART*-1 *Mission Analysis: Moon Option*, Tech. Report S1-ESC-RP-5001, European Space Agency, Directorate of Technical and Operational Support, Ground Systems Engineering Department, ESOC European Space Operations Centre, Noordwijk, The Netherlands, 1998.

[18] E. M. Standish, *JPL Planetary and Lunar Ephemerides, DE*405*/LE*405, Interoffice memorandum IOM 312.F-98-048, Jet Propulsion Laboratory, Pasadena, CA, 1998.

[19] M. J. H. Walker, B. Ireland, and J. Owens, *A set of modified equinoctial orbit elements*, Celestial Mechanics, 36 (1985), pp. 409–419.

[20] T. Yee and J. A. Kechichian, *On the dyanmic modeling in optimal low-thrust orbit transfer*, in Proceedings of the AAS/AIAA Spaceflight Mechanics Meeting, AAS 92-177, Colorado Springs, CO, 1992.

[21] K. P. Zondervan, L. J. Wood, and T. K. Caughy, *Optimal low-thrust, three-burn transfers with large plane changes*, J. Astronaut. Sci., 32 (1984), pp. 407–427.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 2, pp. 171–186

Computing the Dynamics of Complex Singularities of Nonlinear PDEs*

J. A. C. Weideman†

Abstract. A two-step strategy is proposed for the computation of singularities in nonlinear PDEs. The first step is the numerical solution of the PDE using a Fourier spectral method; the second step involves numerical analytical continuation into the complex plane using the epsilon algorithm to sum the Fourier series. Test examples include the inviscid Burgers and nonlinear heat equations as well as a transport equation involving the Hilbert transform. Numerical results, including Web animations that show the dynamics of the singularities in the complex plane, are presented.

Key words. nonlinear PDEs, complex singularities, Fourier spectral methods, Padé approximation, epsilon algorithm

AMS subject classifications. 65M70, 65B10, 34M99

PII. S1111111102398305

1. Introduction.

Many nonlinear PDEs have solutions that become singular in finite time. Examples are the inviscid Burgers equation,

$$u_t + \frac{1}{2}(u^2)_x = 0, \tag{1.1}$$

which exhibits shocks [6], and the nonlinear heat equation,

$$u_t - u_{xx} - u^2 = 0, \tag{1.2}$$

which exhibits a pole-type blow-up [3].

A less familiar equation that also displays blow-up is

$$u_t - \nu u_{xx} + (H(u)u)_x = 0, \tag{1.3}$$

where ν is a positive constant and H is the Hilbert transform. (A singular integral operator, the Hilbert transform may also be defined implicitly by (2.5) below. For more details, as well as the physical significance of (1.3), we refer to [2].)

In the last decade or two, it has been recognized that the analytic structure of the solutions $u(x+iy,t)$ in the complex plane may aid in the understanding of the formation of singularities. Among the first authors to explore this were Bessis and Fournier [6], who have studied (1.1) subject to a generic initial condition (given in (4.1) below). In particular, they showed that the solution has branch point singularities that move in the complex plane. The shock is formed when these singularities reach the real axis. This work was extended in various directions in

*Received by the editors September 6, 2002; accepted for publication by A. Stuart January 15, 2003; published electronically May 8, 2003. This work was supported by NRF grant 2039549.
http://www.siam.org/journals/siads/2-2/39830.html
†Department of Applied Mathematics, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa (weideman@dip.sun.ac.za).

[15]. Similarly, the blow-up in (1.3) can be explained by complex poles moving onto the real axis [2]. We do not know of any similar study of (1.2).

This paper addresses the numerical computation of the singularity structure of PDEs such as (1.1)–(1.3). The procedure we propose here is intended primarily as an aid to visualize the dynamics of the singularities. (Do they move at a constant speed toward the real axis, or do they accelerate? What trajectories do they follow in the complex plane?) The secondary goal, not achieved quite yet, is to classify the singularities. (Are they poles, and if so, of what order? Branch points, perhaps? Essential singularities?)

The proposed strategy is as follows.

Step 1. Use a numerical method to solve the PDE up to a time near breakdown.

Step 2. Then use numerical analytic continuation to extend the computed solution into the complex plane.

In the first step, we use the Fourier spectral method for the space discretization combined with a method-of-lines integration with respect to the time variable. We apply the second step at each level of the time integration; i.e., given the numerical solution at a particular time-step, we continue that solution into the complex domain. This allows us to trace the movement of the singularities as a sequence of snapshots of the complex plane. The analytic continuation of Step 2 is done via the epsilon algorithm, which is an efficient way of computing certain Padé approximations, as will be discussed below.

Both spectral methods and Padé approximations are well-established computational tools, and we do not claim to make any new contributions to either the theory or the application of these techniques. The *combination* of these two tools, however, seems not to have been explored to good effect in the field of PDEs. We believe it is this combination that allows one to see singularity dynamics in a problem such as (1.2) that we suspect may not have been noted before.

We briefly contrast our procedure with some approaches found in the literature.

There exists, of course, a large body of work devoted to the computation of singular solutions of PDEs. In the case of the Burgers equation (1.1), one thinks of shock-capturing and shock-fitting techniques [10, sect. 8.4–8.5], and in the case of nonlinear diffusion equations such as (1.2), there exist various adaptive techniques based on scale invariance [5] or moving grids [9]. The methods discussed in this paper do not aim to compete with these methods. (For example, to compute blow-up in (1.2), the method described in [5] is designed to track the blow-up solution until it reaches values of 10^{12} or more—we do not go much beyond 10^2 in this paper.) Instead, our interest here is in the behavior of singularities in the complex plane and particularly in the manner in which they approach the real axis.

Regarding computational studies of complex singularities of ODEs, we refer to [8] and the references therein. As for PDEs, a common approach is to complexify the time variable and then look for singularities in the complex t-domain. One example is the study reported in [14], in which Padé approximation was used to study singularity formation in a fluid flow problem. Another study of this type is [17], which includes a test problem similar to (1.2).

The tracing of singularities in the complex x-domain was reported in [15], with reference

to the viscous and dispersive Burgers equations. This was done by numerically integrating a Calogero-type dynamical system, an approach more problem-specific than the general procedure suggested here.

The paper with the philosophy most similar to ours is probably [16]. These authors use, as we do, a Fourier spectral method to solve the PDE. To obtain information on the nature and location of the singularity, however, they examine the rate of decay of the Fourier coefficients. By fitting an exponential model to these coefficients, they were able to correctly characterize the singularities that give rise to shock formation in (1.1). However effective this procedure may be, it enables one to characterize only the nearest singularity to the real axis. It reveals no information on any singularities farther out, which our procedure will do if the singularity structure is not too complicated. Also, since our procedure computes the solution directly in the complex domain, we consider it a more appropriate tool for visualizing the dynamics of the singularities.

The outline of the paper is as follows. In the next section, we describe the computational details of our procedure. In sections 3 and 4, we apply the procedure, respectively, to the model problems (1.3) and (1.1). Here we use explicit theoretical solutions to test the reliability of the procedure. Then, in section 5, we tackle a problem involving (1.2) with apparently no known explicit solution.

2. Numerical procedure. We shall consider solutions of (1.1)–(1.3) that are periodic in x. (Other types of boundary conditions, which will not be addressed here, will require different techniques.) A natural choice for solving (1.1)–(1.3) is therefore the Fourier spectral method. Assuming 2π-periodicity in x, the method is based on the truncated Fourier series

$$u(x,t) \approx \sum_{n=-N}^{N} a_n(t) e^{inx}. \tag{2.1}$$

The Galerkin approximation of (1.1), for example, is then given by the following nonlinear dynamical system that governs the evolution of the Fourier coefficients [10]:

$$\frac{da_n}{dt} + \frac{1}{2} in \sum_{\substack{j+k=n \\ |j|,|k|\le N}} a_j a_k = 0, \qquad |n| \le N. \tag{2.2}$$

The corresponding discretizations of (1.2) and (1.3) are, respectively,

$$\frac{da_n}{dt} + n^2 a_n - \sum_{\substack{j+k=n \\ |j|,|k|\le N}} a_j a_k = 0 \tag{2.3}$$

and

$$\frac{da_n}{dt} + \nu n^2 a_n - n \sum_{\substack{j+k=n \\ |j|,|k|\le N}} \operatorname{sgn}(j)\, a_j a_k = 0. \tag{2.4}$$

In deriving (2.4), we have used the fact that the Hilbert transform satisfies [2]

$$H(e^{inx}) = i \operatorname{sgn}(n)\, e^{inx}. \tag{2.5}$$

Systems (2.2)–(2.4) can be integrated by most ODE solvers. (This is the so-called method of lines.) The solution provides approximations to $u(x,t)$ in the form of the truncated Fourier series (2.1).

A more efficient alternative to the spectral Galerkin method for nonlinear problems is the spectral collocation or pseudospectral method; see [10]. As efficiency was not a concern here—the numerical analytic continuation procedure described below is more expensive than the solution of the PDE—we used the full spectral method.

Relatedly, nothing prevents one from using discretizations based on finite difference or finite element methods. When approximate values $u(x_j,t)$ are available at equidistant nodes x_j at any time t, an inverse discrete Fourier transform (DFT) will provide the approximate coefficients $a_n(t)$ in (2.1). These coefficients are all we need for the next phase of the procedure.

Once the approximation (2.1) is available at a given t, the next step is to continue this solution analytically into the complex plane. A direct summation is doomed to fail for two reasons. First, the truncated Fourier series (2.1) is an entire function and will not reveal any singularity structure. Second, numerical analytic continuation is ill conditioned. Indeed, with $z = x + iy$,

$$\sum_{n=-N}^{N} a_n(t)e^{inz} = \sum_{n=-N}^{N} a_n(t)e^{-ny}e^{inx}, \tag{2.6}$$

and any error in the coefficients $a_n(t)$ gets amplified exponentially as $|y| \longrightarrow \infty$.

Both of these problems may be overcome to a degree by converting the truncated Fourier series (2.1) to a Fourier–Padé approximant [1, p. 382]. We digress briefly to give some details of Padé approximation, first for power series and then for Fourier series.

Consider the power series

$$f(z) = \sum_{k=0}^{\infty} c_k z^k. \tag{2.7}$$

The $[L,M]$ Padé approximation to this series, if it exists, is defined as the rational function

$$r_{L,M}(z) = \frac{a_0 + a_1 z + \cdots + a_L z^L}{1 + b_1 z + \cdots + b_M z^M} \tag{2.8}$$

with the property that

$$f(z) - r_{L,M}(z) = O(z^{L+M+1}).$$

Here we shall use only approximants with $L = M$.

The same idea may be applied to Fourier series [1, p. 382]. Consider

$$u(z) \approx \sum_{n=-N}^{N} a_n e^{inz};$$

when $z = x$ is real, this would be an approximate solution to a PDE such as (1.1)–(1.3) at a specific time t. By defining $w = e^{iz}$ and $v = e^{-iz}$, the Fourier series on the right may be expressed as

$$u(z) \approx \sum_{n=0}^{N} a_n w^n + \sum_{n=0}^{N} a_{-n} v^n - a_0.$$

Both power series on the right may now be converted to Padé approximants. (Note that, since z may be complex, the two series are not necessarily conjugate, and they have to be converted independently.) Assuming N is even, this gives

$$u(z) \approx r_{\frac{1}{2}N,\frac{1}{2}N}(w) + s_{\frac{1}{2}N,\frac{1}{2}N}(v) - a_0, \tag{2.9}$$

where $r_{\frac{1}{2}N,\frac{1}{2}N}(z)$ and $s_{\frac{1}{2}N,\frac{1}{2}N}(z)$ are both rational functions of the form (2.8) that satisfy

$$r_{\frac{1}{2}N,\frac{1}{2}N}(w) = \sum_{n=0}^{N} a_n w^n + O(w^{N+1}), \quad s_{\frac{1}{2}N,\frac{1}{2}N}(v) = \sum_{n=0}^{N} a_{-n} v^n + O(v^{N+1}).$$

Expression (2.9) is the formula we shall use to extend (2.1) into the complex plane.

In connection with computational matters, we note that computing the coefficients in the numerator/denominator polynomials of (2.8) is an ill-conditioned problem [1, p. 38]. Computing function values of $r_{M,M}(z)$ for a given z, however, seems to be better conditioned. The latter may be done efficiently with Wynn's epsilon algorithm, the details of which may be found in [1, p. 73] and the references therein. (We remark that the epsilon algorithm may too become unstable near poles of intermediate rational functions computed as part of the process. Remedies were suggested by Wynn, but we found no need to implement them in our computations.)

The epsilon algorithm is also known as a sequence accelerator theoretically equivalent to the Shanks transformation and the repeated application of Aitken's Δ^2 extrapolation process [1, p. 67]. It may be used as a series accelerator as well when applied to the sequence of partial sums of series such as (2.7). In this manner, it can be used to sum a power series outside its circle of convergence [1, p. 75]. Indeed, when we apply the epsilon algorithm to a Fourier series, our goal is to evaluate it outside its strip of analyticity.

It remains to investigate the singularity structure of (2.9). For qualitative purposes, a visual representation of $u(z)$ in the complex plane often suffices—see, for example, Figures 3.1, 4.1, and 5.2. Here we should warn, however, against the possibility of spurious poles and defects in Padé approximation; see [1, sect. 2.2]. However, as these authors also point out, these phenomena are transient in nature, so as N is increased, they tend to appear and disappear. In our situation, it may also happen that a "pole" appears at a particular time level t but then disappears at the next level. This is a clue that it is almost surely a spurious phenomenon that can be ignored.

A quantitative investigation requires that one is able to locate and characterize the singularities. What we propose next assumes that the singularities are poles. The modification for algebraic branch points will be discussed below, but other types of singularities (logarithmic branch points, essential singularities) will require different techniques.

To locate the pole, we use brute force, by applying a numerical maximization search to the objective function

$$f(z) = \log|u(z)|. \tag{2.10}$$

Plots such as Figures 3.1 and 5.2 may be used to determine good initial guesses for the search.

To compute the order of the pole, the principle of the argument may be used [13, p. 278]:

$$\frac{1}{2\pi i}\int_C \frac{u'(z)}{u(z)}\,dz = Z(u) - P(u). \tag{2.11}$$

Here it is assumed that u is analytic and nonzero at each point of a simple closed positively oriented contour C, and inside C the only singularities of u are poles. $Z(u)$ and $P(u)$ are, respectively, the number of zeros and poles (counting multiplicity) of u inside C.

For computational purposes, we choose C to be the circle

$$C: \ z = z_0 + re^{i\theta}, \qquad 0 \le \theta \le 2\pi,$$

where the center, z_0, is the computed location of a pole. Assuming that the radius, r, is sufficiently small to exclude all zeros and all other poles of u inside C, we may compute the order of the pole at z_0 according to

$$P(u) = -\frac{r}{2\pi}\int_0^{2\pi} g(z_0 + re^{i\theta})e^{i\theta}\, d\theta.$$

Here we have defined the logarithmic derivative of u by $g(z) = u'(z)/u(z)$.

The integral may be approximated by the K-panel trapezoidal rule (which, for periodic integrands such as this, is the same as a rectangle rule and a very accurate method for evaluating the integral [18])

$$P(u) \approx -\frac{r}{K}\sum_{j=1}^{K} g(z_0 + re^{2\pi ij/K})e^{2\pi ij/K}. \tag{2.12}$$

In practice, some trial and error is required to choose K and particularly r. (If r is too small and z_0 is not very accurate, the circle C may exclude the actual pole, and z_0 may be classified, incorrectly, as a point of analyticity. On the other hand, if r is too large, a zero or a different pole of u may be enclosed by C.) For the computations reported below, we have settled, after experimenting on known functions, on the values $K = 128$, $r = 0.05$.

The above discussion applies to poles. If the singularity is an algebraic branch point, i.e., $u(z) \sim C(z - z_0)^\alpha$ as $z \longrightarrow z_0$, the integral on the left side of (2.11) may also be used to compute α. In this case, the right side of (2.11) should be interpreted as a winding number. For further details, we refer to [13, Ch. 4], where alternative methods for computing winding numbers are also suggested.

The last point to discuss is the computation of the logarithmic derivative g. There seem to be three possible methods. The first is to continue into the complex plane the function $u(z)$, defined by its truncated Fourier series (2.1), and independently also its derivative

$$u'(z) \approx \sum_{n=-N}^{N} (in)a_n e^{inz}.$$

The function $g(z) = u'(z)/u(z)$ can then be computed pointwise at each required z.

A better alternative seems to be to express $g(x)$ (on the real axis) as

$$\frac{u'(x)}{u(x)} \approx \sum_{n=-N}^{N} c_n e^{inx}. \tag{2.13}$$

The left side can be evaluated at equidistant nodes x_j, and hence the c_n can be computed with an inverse DFT. Once the c_n are known, the right side can be continued analytically into the complex plane by formula (2.9). In this method, only one Fourier series is analytically continued—not two. We have experimented with test functions such as $u(z) = (a - \sin z)^\alpha$, and the second method was more accurate in predicting the value of α. Therefore, in the experiments reported below, the second method was used.

The third alternative is similar to the second method. Instead of using the inverse DFT to compute the c_n from (2.13), this may be done by deconvolution. As this approach avoids aliasing errors, it might be more accurate than the second method, but we have not tested it.

We now turn to numerical tests of these algorithms. The details of our implementations are as follows.

The epsilon algorithm was implemented in MATLAB 6. To allow efficient interactive experimentation, it was necessary to optimize the code. In MATLAB, this means vectorization; i.e., when the epsilon algorithm is applied to a series such as (2.7), it is not done pointwise for each z, but instead our code handles z as an array.

The systems of ODEs (2.2)–(2.4) were implemented with $N = 50$ throughout. This means that the rational functions used in (2.9) were of degree 25 in both numerator and denominator. The convolution sums in (2.2)–(2.4) were computed with the MATLAB function `conv`. The time integration was done with `ode45`, a Runge–Kutta algorithm with adaptive time-steps. The optimization of (2.10) was done with the function `fminsearch`, which is based on a Nelder–Mead simplex (direct search) method. In all experiments, we used default tolerance parameters in both `ode45` and `fminsearch`.

3. Numerical experiment I. Since an explicit solution was readily available for (1.3), we chose it as our first test. This solution is

$$u(x,t) = \sigma + \nu \frac{1 - \varrho^2 e^{2\sigma t}}{1 + \varrho^2 e^{2\sigma t} - 2\varrho e^{\sigma t} \cos x}, \tag{3.1}$$

with σ and ϱ arbitrary positive constants [2]. The singularities are simple poles, located at

$$z = 2n\pi \pm i(\sigma t + \log \varrho), \qquad n = 0, \pm 1, \pm 2, \ldots. \tag{3.2}$$

When $\varrho < 1$, these poles move toward the real axis, which they reach at $t = -(\log \varrho)/\sigma$. At that instant, the solution blows up on the real axis.

The first row of graphs in Figure 3.1 shows the evolution (3.1) for parameter values $\sigma = 1$, $\varrho = 1/e$, and $\nu = 0.1$; times $t = 0.3, 0.3, 0.9$; and space interval $x \in [-\pi, \pi]$. Blow-up occurs at $t = 1$.

In the first step of our procedure, we have integrated (2.4) using the same set of parameter values and picking the initial condition from (3.1) by setting $t = 0$. The numerical solution (not shown) was virtually indistinguishable from the theoretical one displayed in the first row of Figure 3.1.

Next, we continued the numerical solution into the complex plane with the epsilon algorithm as described in section 2. The second row of figures in Figure 3.1 shows the computed values of $|u(z,t)|$ in the complex domain $z = x + iy$, $x \in [-\pi, \pi]$, $y \in [0, 1]$. (Here and in

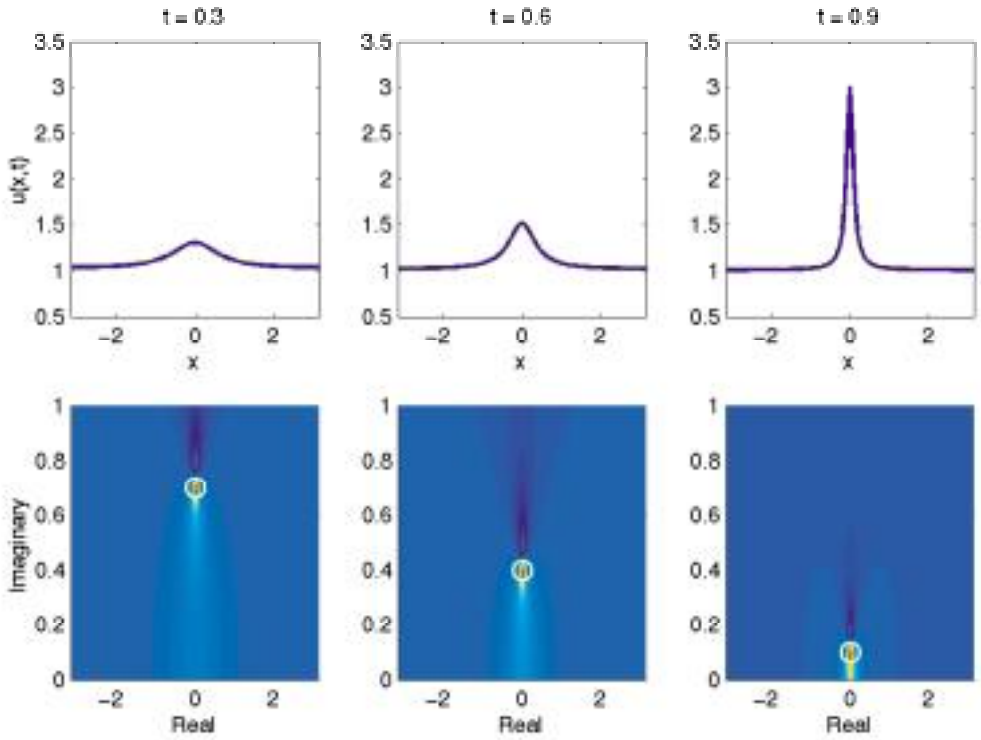


Figure 3.1. *Evolution of* (1.3) *subject to the initial condition* (3.1) *(with* $t = 0$*). The first row shows* $|u(x,t)|$ *for various* t*, and the second row shows* $|u(z,t)|$ *in the upper half-plane for the same values of* t*, as computed numerically with the procedures suggested in this paper. The white circles represent the actual locations of the poles as computed from the theoretical solution* (3.1)*.*
[Animation (2.3MB)]

Table 3.1
Location and classification of the pole shown in Figure 3.1*.*

Time	Exact location of pole	Computed location of pole	$P(u)$
0.3	$0.7i$	$-2.9604\text{e-}08 + 7.0000\text{e-}01i$	$9.9995\text{e-}01 + 1.8376\text{e-}05i$
0.6	$0.4i$	$6.8126\text{e-}09 + 4.0000\text{e-}01i$	$9.9999\text{e-}01 + 5.1559\text{e-}06i$
0.9	$0.1i$	$-1.1751\text{e-}11 + 1.0000\text{e-}01i$	$9.9994\text{e-}01 + 2.6968\text{e-}05i$

the figures below, interpolated shading on a 51×51 grid was used.) Since the solution is symmetric with respect to the real axis, only the upper half-plane is shown.

In the color map of the figure, red indicates large values and blue small. The red dot therefore represents the pole, which can be seen moving downward. Its conjugate (not shown) moves in the upward direction, and the two poles coalesce on the real axis at $t = 1$. This is when the solution $u(x,t)$ blows up on the real axis.

To see how well the epsilon algorithm succeeds in locating the poles, we drew white circles, the centers of which represent the actual position of the poles computed by (3.2). Clearly there is excellent agreement.

For a quantitative investigation into the nature of the singularity, we used the classification

algorithm described by (2.10)–(2.12). In Table 3.1, we give the approximate position of the singularity as well its estimated order, $P(u)$. These results confirm that the singularity is indeed a pole of order one and that it travels at the right speed.

The next example provides a more stringent numerical test. It involves a branch point singularity.

4. Numerical experiment II. In [6], the singularity structure of the Burgers equation (1.1) was investigated with reference to the pedagogical initial condition

$$u(x,0) = x^3 - x/t_*, \tag{4.1}$$

with t_* a positive constant.

The authors showed that $u(z,t)$ has two square root singularities, placed symmetrically on the imaginary axis with respect to the origin. These singularities move from $z = \pm i\infty$ at $t = 0$ toward the real axis, where they meet at $t = t_*$; this is when the shock occurs. The authors proceeded to characterize a three-sheeted Riemann surface that provides an analytic path across the jump of the shock.

As the initial condition (4.1) is not periodic in x, the results of [6] cannot be used as a check on our scheme. Instead, we have derived what appears to be a new periodic solution to (1.1). We have not seen this solution, which involves Lambert's W-function, mentioned anywhere in the literature.[1]

Consider the initial condition

$$u(x,0) = e^{ix}. \tag{4.2}$$

By using the implicit solution to (1.1), namely, $u = f(x - ut)$, one gets

$$u = e^{ix}e^{-iut},$$

or, by defining $w = iut$,

$$we^w = ite^{ix}.$$

By comparison, the Lambert function, $w = W(z)$, solves the equation $we^w = z$. One concludes that

$$u = (1/it)W(ite^{ix}), \tag{4.3}$$

which is the desired 2π-periodic explicit solution.

In the computations below, the W-function was computed with MATLAB's Symbolic Toolbox, which is based on the Maple computer algebra system. (It can also be computed with Mathematica, where the W-function is known as `ProductLog`.)

The singularity structure of the W-function has been analyzed in [12], and we summarize as follows. The function $W(z)$ is multivalued, with an infinite number of branches. Only one of these, designated the principal branch, is analytic at $z = 0$. The principal branch and two other branches share an order 2 branch point at $z = -1/e$. The branch cut dividing these branches is the subset of the real line from $-\infty$ to $-1/e$. We refer to [7] for a picture of the three-sheeted Riemann surface as well as the Maple code that will generate it.

[1] Reference [12] reviews several applications in which the W-function features, but (1.1) is not among them. In addition, an Internet search in MathSciNet using the keywords "Burgers" and "Lambert" yielded no meaningful hits. Neither did we find our solution in the older table [4].

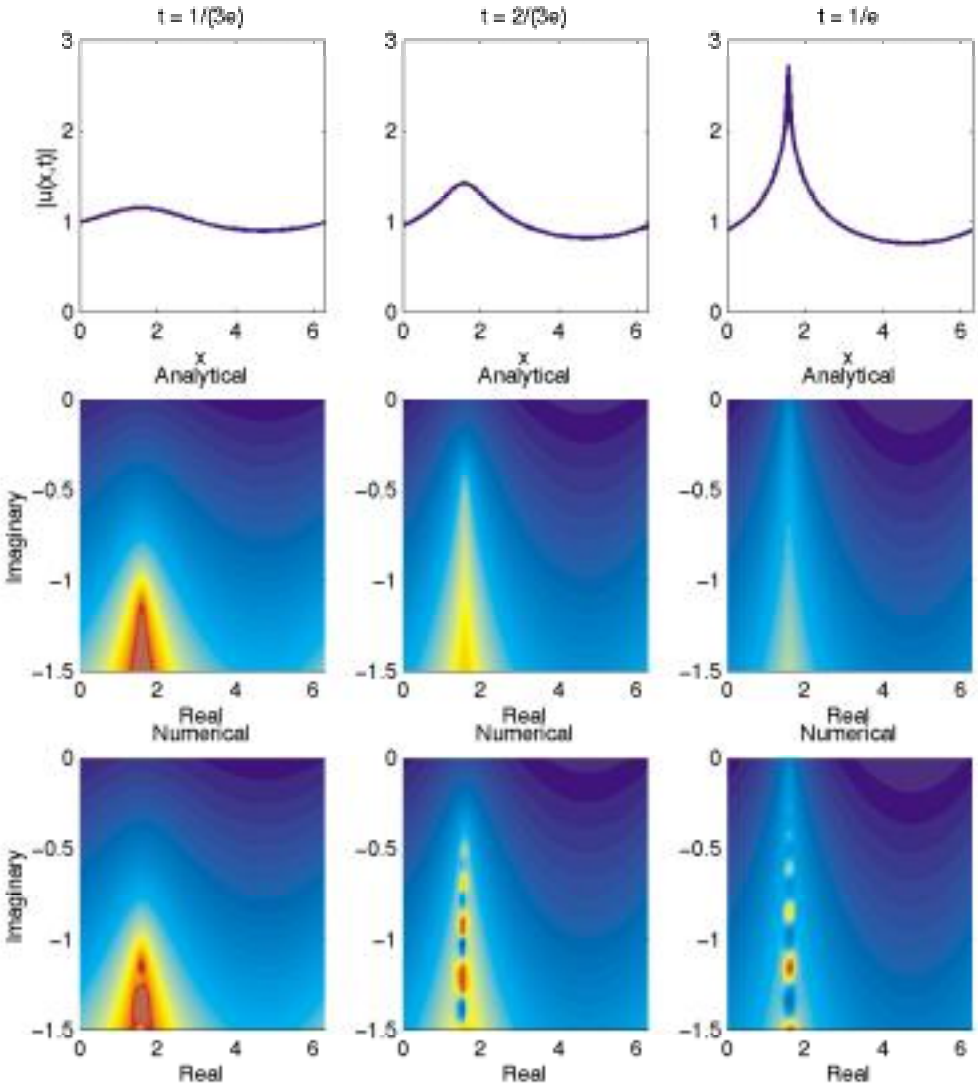


Figure 4.1. *Evolution of the Burgers equation* (1.1) *subject to the initial condition* (4.2). *The first row shows* $|u(x,t)|$ *for various* t, *and the second row shows* $|u(z,t)|$ *in the lower half-plane for the same values of* t, *as computed from the explicit solution* (4.3). *The third row shows numerical approximations to the middle row using the procedure suggested in this paper.*
[Animation (2.2MB)]

Transplanting this information to the solution of the Burgers equation, we conclude that in order to satisfy the initial condition (4.2), the principal branch of W should be selected in the solution (4.3). This solution has branch points at z, where $ite^{iz} = -1/e$, or

$$z = (4n+1)\tfrac{\pi}{2} + i(1+\log t), \qquad n = 0, \pm 1, \pm 2, \ldots .$$

Unlike the situation corresponding to the initial condition (4.1), these singularities are of logarithmic and not algebraic type. The corresponding branch cuts are lines parallel to the

imaginary axis, extending from the branch points down to infinity. The branch points approach the real axis from below and with negative acceleration. The axis is reached at $t = 1/e$, at which time the solution to the Burgers equation becomes singular at $x = \pi/2 \,(\mathrm{mod}\, 2\pi)$ on the real line.

The first row of Figure 4.1 shows the evolution of $|u(x,t)|$, computed according to (4.3) on $x \in [0, 2\pi]$ for $t = 1/(3e)$, $2/(3e)$, $1/e$. The latter time corresponds to the appearance of a singularity, as can be seen in the top right figure.

The second row of figures in Figure 4.1 shows the corresponding values $|u(z,t)|$ in the complex domain $z = x+iy$, $x \in [0, 2\pi]$, $y \in [-1.5, 0]$. Note that the top boundary in each figure represents the real axis: in the upper half-plane, the solution is analytic and uninteresting and is therefore not displayed here. The needle-like structure at $x = \pi/2$ represents the branch cut that moves upward. When it hits the real axis, the singularity is formed.

The first two rows of figures in Figure 4.1 were computed using the theoretical expression (4.3). Turning to numerical approximations, we integrated (2.2) using the Fourier spectral method. The numerical solution (not shown) was virtually indistinguishable from the analytical solution shown in the top row of Figure 4.1 (even at $t = 1/e$, the time the singularity appears on the real axis).

Next, the numerical solution was extended into the complex plane with the epsilon algorithm. The result is shown in the third row of figures in Figure 4.1.

As expected, the epsilon algorithm has difficulty approximating the branch cut. The singularity appears as a string of poles located where the branch cut should have been.[2] This is typical behavior for Padé approximants; see, for example, [1, p. 47]. Nonetheless, the singularity is located on the correct line ($x = \pi/2$, $y < 0$) and approaches the real axis at the right speed.[3]

5. Numerical experiment III. Our final set of numerical experiments comes from the class of nonlinear heat equations. Let σ and ϱ be constants, and consider the function

$$u(x,t) = \frac{e^{5/3\,(\sigma-t)+2\,ix}}{\left(ie^{5/6\,(\sigma-t)+ix} + \varrho\right)^2}, \tag{5.1}$$

which is 2π-periodic and solves the PDE

$$u_t - \frac{1}{6}u_{xx} + u^2 + u = 0. \tag{5.2}$$

The mechanism of blow-up is a pole of second order that reaches the real axis. With the parameter choice that will be used below, namely, $\sigma = 1$ and $\varrho = 1$, the location of this pole is given by

$$z = \frac{\pi}{2} + \frac{5}{6}(1-t)\,i$$

[2]It is interesting to note that when a viscous term νu_{xx} is added to (1.1), then the branch cut is also replaced by a series of poles [15].

[3]A better way to approximate functions with branch point singularities is the quadratic Padé method [1, p. 376], but this will have to wait for future studies.

(plus 2π-periodic copies). The blow-up occurs at $t = 1$, and its profile is qualitatively similar to the blow-up shown in Figure 3.1.

The semidiscretization of the PDE (5.2), which is similar to (2.3), was integrated as described in section 2. The initial condition was obtained by setting $t = 0$ in (5.1). The resulting numerical solution was analyzed by the scheme discussed at the end of section 2, by means of which the singularities were tracked and classified. The output, shown in Table 5.1, confirms that the singularity was located accurately, and there remains little doubt that it is indeed a pole of order 2.

Table 5.1
Location and classification of the pole of (5.1)–(5.2).

Time	Exact location of pole	Computed location of pole	$P(u)$
0.3	$\pi/2 + 7/12\,i$	1.5708e+00 + 5.8333e-01i	2.0000e+00 + 2.3495e-06i
0.6	$\pi/2 + 1/3\,i$	1.5708e+00 + 3.3333e-01i	2.0000e+00 + 1.1298e-08i
0.9	$\pi/2 + 1/12\,i$	1.5708e+00 + 8.3333e-02i	2.0000e+00 − 1.4816e-08i

Having gained a measure of confidence in the proposed strategy, we are now ready to explore some uncharted territory. This is the PDE (1.2), subject to the initial condition

$$u(x, 0) = \cos x. \tag{5.3}$$

Although several explicit solutions to $u_t = u_{xx} + f(u)$ are known (see, e.g., [11] and the references therein), we could not find a solution to (1.2) that is both periodic in x and displays finite time blow-up. Hence (5.3) was chosen somewhat arbitrarily as an initial condition. Note that it is an entire function, so any singularities will have to start at infinity.

There exists a vast literature on blow-up phenomena in diffusion equations such as this; for a survey, see [3]. The majority of these studies is concerned, however, with questions of if and where blow-up occurs and what the profile of the solution is near blow-up. They do not address the dynamics of complex singularities at intermediate times, as is our intention here.

We started by integrating the semidiscrete system (2.3), and its solution, which exhibits a blow-up singularity near $(x, t) = (0, 3.17)$, is shown in Figure 5.1. Initially the mechanism of diffusion dominates, as can be seen in the fact that the amplitude of the cosine curve is damped. At the same time, however, the mean of the solution profile increases, and at some point it becomes so large that the mechanism of nonlinearity becomes dominant. Shortly after this, blow-up occurs.

The corresponding analytical continuation is shown in Figure 5.2, on the domain $z = x+iy$, $x \in [-\pi, \pi]$, $y \in [0, 5]$. (Again, only the upper half-plane is shown, owing to symmetry.) The plot suggests a conjugate pair of pole-like singularities that start at $z = \pm\infty\, i$ and speed toward the real axis, where they meet at the time of blow-up. At a superficial glance, the singularities appear to be isolated poles, but by replotting the figure on a log-scale it becomes evident that the singularity structure is more complicated. Based on the asymptotic estimates cited in [3], [9], we conjecture that the singularity structure is similar to that of $|\log z|/z^2$ near the point of blow-up. Our methods cannot yet classify such singularities.

To examine this behavior more closely, we computed approximate locations of the singularities as described in section 2. The computations suggest that these singularities are

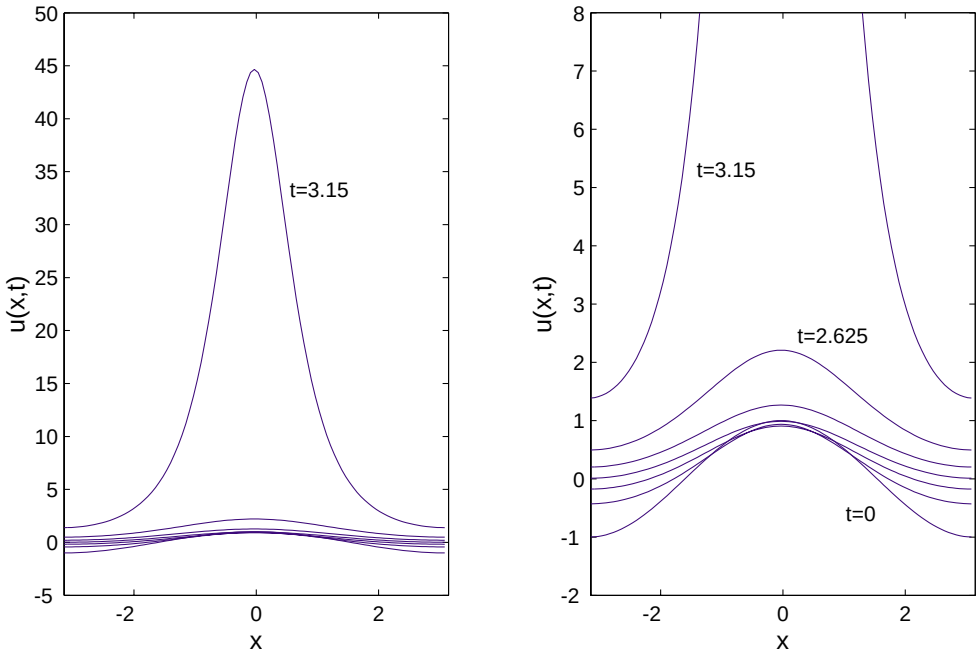


Figure 5.1. *Computed solution of* (1.2)*, subject to the initial condition* (5.3)*, on* $x \in [-\pi, \pi]$ *for values of* $t = 0(0.525)3.15$*. (The figure on the right was clipped at* $u = 8$ *to reveal more detail.)*
[Animation (2.7MB)]

located on the imaginary axis, $z = \pm\, y\, i$. Figure 5.3 shows y as a function of t. (The thin red curve in the figure represents an asymptotic estimate that will be discussed below.)

Unlike the case of the nonlinear heat equation (5.1)–(5.2), the singularities here do not move at a constant speed but instead follow the S-shaped curve seen in Figure 5.3. The singularities start at $z = \pm\, i\, \infty$ and approach the real axis rapidly. They decelerate as they do so, however, and near $t = 0.5$ they (surprisingly, in our opinion) reverse direction. Near $t = 1$ the acceleration toward the real axis turns positive again, which causes the singularities to reverse direction once more—this time roughly at $t = 2$. From this point, the singularities zoom onto the real axis rapidly, and blow-up occurs near $t = 3.17$. We suspect the phases of negative and positive acceleration toward the real axis may have to do, respectively, with the smoothing effect of the diffusion operator vs. the focusing effect of the nonlinearity.

Even though our principal aim was not the computation of the solution very close to blow-up, it is nevertheless interesting to check these computations against known asymptotic results. We quote one typical result, due to Galaktionov and coworkers, that appears to be applicable to the present situation.

Near the blow-up point $x = 0$, the solution to (1.2) is locally given by

$$u(x,t) \sim \left((T-t) + \frac{x^2}{8|\log(T-t)|} \right)^{-1}, \qquad t \longrightarrow T, \tag{5.4}$$

where T is the blow-up time [3], [5], [9].

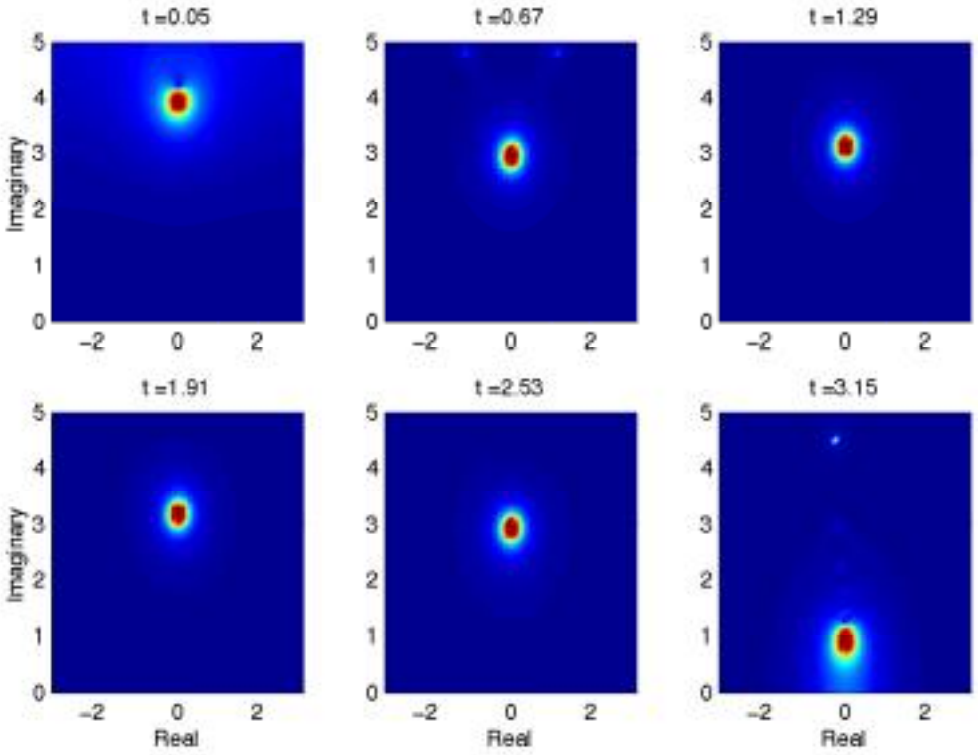


Figure 5.2. *The solution of* (1.2)*, subject to the initial condition* (5.3)*. Values of* $|u(z,t)|$ *are shown in the upper half-plane, as computed numerically with the procedure suggested in this paper.*
[Animation (2.7MB)]

Note. This estimate is usually derived for the Cauchy problem or Dirichlet boundary conditions, not periodic conditions, and it is always assumed that the initial condition is positive-valued, which (5.3) is not. Numerical investigations have led us to believe, however, that the estimate (5.4) may be applicable to the present situation as well.

The estimate (5.4) implies that the solution has pole-like singularities near

$$z = \pm y\, i, \qquad \text{where} \qquad y = \sqrt{8(T-t)|\log(T-t)|}. \tag{5.5}$$

In order to compare this estimate with our numerical approximations, a good estimate of the blow-up time T is required. We computed this by tracking the singularity with the "event" facility available in MATLAB's ODE suite. This yielded the approximation $T = 3.17395$. With this value of T, we plotted the theoretical asymptotic curve $y = y(t)$ defined by (5.5) as the thin red curve in Figure 5.3. For a quantitative comparison, we also list in Table 5.2 the computed and asymptotic values of y for a few values of t near T.

It should be understood that comparisons such as these represent a challenge to the Fourier method, as any spectral method loses accuracy rapidly as the singularities approach the real axis. Nevertheless, we regard the agreement between numerics and asymptotics shown here as satisfactory.

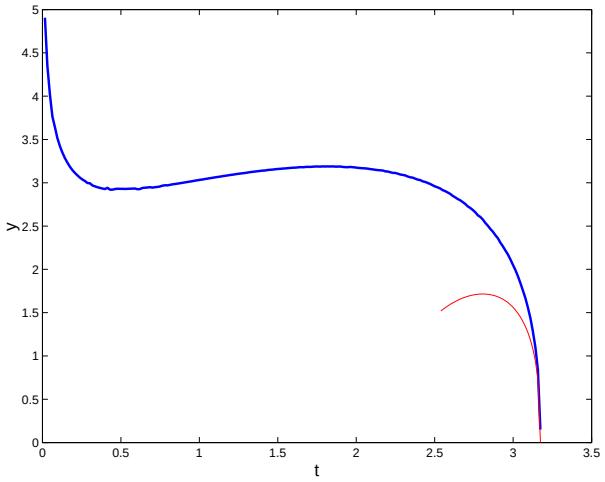


Figure 5.3. *Imaginary part of the singularity, $z = y\,i$, shown in Figure 5.2, as a function of t. The thin red curve represents the asymptotic estimate (5.5), with $T = 3.17395$.*

Table 5.2

Comparison of the computed and asymptotic locations of the singularity of (1.2), subject to (5.3). The asymptotic estimate is defined by (5.5), with $T = 3.17395$.

Time t	Computed y	Asymptotic y
3.171	0.43	0.37
3.172	0.34	0.31
3.173	0.24	0.23

6. Conclusions. Based on the numerical evidence presented here, we feel the proposed strategy should be considered successful. In the three PDEs where explicit solutions were available, the singularity dynamics were tracked accurately (provided one realizes that a series of poles on a straight line is an attempt to approximate a branch cut, as was the case with the Burgers equation). Even for the nonlinear heat equation (1.2), where no closed form solution was available, we believe the comparisons with known asymptotic results were acceptable.

There is, of course, room for improvement. We have used a basic Fourier spectral method combined with the standard adaptive Runge–Kutta method as coded in MATLAB's `ode45` function. This proved adequate in our computations where we made no attempt to follow the singularity to $\max u = 10^{12}$ or so. For better singularity tracking close to blow-up, the Fourier method should be replaced by one of the adaptive methods discussed in [3], [5], and [9]. The details of how to combine these methods with analytic continuation have yet to be worked out, however.

A referee also pointed out that the local error control of `ode45` may be inadequate in the proximity of blow-up, and, in addition, the semidiscrete system may be stiff, so it might be advantageous to switch to a stiff solver. In response to this, we tried MATLAB's `ode15s` function, which executed a little faster but otherwise did not change any of our results appreciably.

Another avenue for improvement is to switch from linear to quadratic Padé approximations. This should enable one to track and classify branch point singularities better; recall the

footnote on page 181.

Related to this point is the fact that in all our model problems, the singularities were well isolated—and only one or two of them near the real axis. If the singularity structure is more complicated, nobody knows what Padé approximants and the epsilon algorithm will do. (For an example of how complicated singularity structure can get, we refer to [8].) In addition, the epsilon algorithm and Padé approximations are beset with risks, such as spurious poles, defects, ill-conditioning, and numerical instabilities. We certainly do not wish to minimize the difficulties associated with numerical analytic continuation, and users of these methods are advised to interpret all results judiciously.

Notwithstanding these words of caution, we are sufficiently pleased with the results obtained here to continue the investigation.

Acknowledgments. The author would like to thank Ben Herbst, Willy Hereman, Dave Sloan, David Trubatch, and anonymous referees for many fruitful suggestions. Milton Maritz provided the explicit solution (5.1)–(5.2). Neil Muller assisted with the Web animations.

REFERENCES

[1] G. A. Baker and P. Graves-Morris, *Padé Approximants,* 2nd ed., Cambridge University Press, Cambridge, UK, 1996.
[2] G. R. Baker, X. Li, and A. C. Morlet, *Analytic structure of two 1D-transport equations with nonlocal fluxes*, Phys. D, 91 (1996), pp. 349–375.
[3] C. Bandle and H. Brunner, *Blowup in diffusion equations: A survey*, J. Comput. Appl. Math., 97 (1998), pp. 3–22.
[4] E. R. Benton and G. W. Platzman, *A table of solutions of the one-dimensional Burgers equation*, Quart. Appl. Math., 30 (1972), pp. 195–212.
[5] M. Berger and R. V. Kohn, *A rescaling algorithm for the numerical calculation of blowing-up solutions*, Comm. Pure Appl. Math., 41 (1988), pp. 841–863.
[6] D. Bessis and J. D. Fournier, *Pole condensation and the Riemann surface associated with a shock in Burgers' equation*, J. Physique Lett., 45 (1984), pp. L-833–L-841.
[7] J. M. Borwein and R. M. Corless, *Emerging tools for experimental mathematics*, Amer. Math. Monthly, 106 (1999), pp. 889–909.
[8] J. P. Boyd, *The Blasius function in the complex plane*, Experiment. Math., 8 (1999), pp. 381–394.
[9] C. J. Budd, W. Huang, and R. D. Russell, *Moving mesh methods for problems with blow-up*, SIAM J. Sci. Comput., 17 (1996), pp. 305–327.
[10] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.
[11] P. A. Clarkson and E. L. Mansfield, *Symmetry reductions and exact solutions of a class of nonlinear heat equations*, Phys. D, 70 (1993), pp. 250–280.
[12] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, *On the Lambert W function*, Adv. Comput. Math., 5 (1996), pp. 329–359.
[13] P. Henrici, *Applied and Computational Complex Analysis, Vol.* I, Wiley-Interscience, New York, 1974.
[14] R. H. Morf, S. A. Orszag, and U. Frisch, *Spontaneous singularity in three-dimensional, inviscid, incompressible flow*, Phys. Rev. Lett., 44 (1980), pp. 572–575.
[15] D. Senouf, R. Caflisch, and N. Ercolani, *Pole dynamics and oscillations for the complex Burgers equation in the small-dispersion limit*, Nonlinearity, 9 (1996) pp. 1671–1702.
[16] C. Sulem, P. L. Sulem, and H. Frisch, *Tracing complex singularities with spectral methods*, J. Comput. Phys., 50 (1983), pp. 138–161.
[17] Y. Tourigny and M. Grinfeld, *Deciphering singularities by discrete methods*, Math. Comp., 62 (1994), pp. 155–169.
[18] J. A. C. Weideman, *Numerical integration of periodic functions: A few examples*, Amer. Math. Monthly, 109 (2002), pp. 21–36.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 2, pp. 187–218

A Simply Stabilized Running Model*

R. M. Ghigliazza†, R. Altendorfer‡, P. Holmes§, and D. Koditschek‡

Abstract. The spring-loaded inverted pendulum (SLIP), or monopedal hopper, is an archetypal model for running in numerous animal species. Although locomotion is generally considered a complex task requiring sophisticated control strategies to account for coordination and stability, we show that stable gaits can be found in the SLIP with both linear and "air" springs, controlled by a simple fixed-leg reset policy. We first derive touchdown-to-touchdown Poincaré maps under the common assumption of negligible gravitational effects during the stance phase. We subsequently include and assess these effects and briefly consider coupling to pitching motions. We investigate the domains of attraction of symmetric periodic gaits and bifurcations from the branches of stable gaits in terms of nondimensional parameters.

Key words. legged locomotion, spring-loaded inverted pendulum, periodic gaits, bifurcation, stability

AMS subject classifications. 34C23, 37J20, 37J25, 37J60, 70Hxx, 70K42, 70K50

PII. S1111111102408311

1. Introduction.

Locomotion, "moving the body's locus," is among the most fundamental of animal behaviors. A large motor science literature addresses gait pattern selection [1], energy expenditure [2], underlying neurophysiology [3], and coordination in animals and machines [4]. In this paper, we explore the stabilizing effect of a very simple control policy on a very simple running model.

Legged locomotion is generally considered a complex task [5] involving the coordination of many limbs and redundant degrees of freedom [6]. In [7], Full and Koditschek note that "locomotion results from complex, high-dimensional, non-linear, dynamically coupled interactions between an organism and its environment." They distinguish locomotion models simplified for the purpose of task specification (templates) from more kinematically and dynamically accurate representations of the true body morphology (anchors). A *template* is a formal reductive model that (1) encodes parsimoniously the dynamics of the body and its payload transport capability, using the minimum number of variables and parameters, and (2) advances an intrinsic hypothesis concerning the control strategy underlying the achievement of this task. *Anchors* are not only more elaborate dynamical systems grounded in the morphology and physiology

*Received by the editors May 23, 2002; accepted for publication (in revised form) by M. Golubitsky January 29, 2003; published electronically May 8, 2003. This work was supported by DoE grant DE-FG02-95ER25238 and DARPA/ONR grant N00014-98-1-0747.
http://www.siam.org/journals/siads/2-2/40831.html
†Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 (rghiglia@princeton.edu). This author was partially supported by Burroughs-Wellcome Training Grant in Biological Dynamics 1001782.
‡Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 (altendor@eecs.umich.edu, kod@umich.edu).
§Department of Mechanical and Aerospace Engineering, Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 (pholmes@math.princeton.edu).

of an animal, but they must also admit the imposition of control policies that result in the realization of the lower dimensional template dynamics. In this context, Full and Koditschek suggest that the spring-loaded inverted pendulum (SLIP) model might reasonably provide a template for sagittal plane motions of the center of mass (COM) of such diverse species as six-legged trotters (cockroaches), four-legged trotters (dogs), two-legged runners (humans and birds), and hoppers (kangaroos). The validation of the SLIP template is based on similarities of ground reaction forces and kinetic and potential energies between these animals running at steady state and the SLIP model with suitably adjusted parameters (see [8]; for a review, see [9]). Details of the anchor system such as pitching motion or multiple leg impacts lead to small deviations from the SLIP predictions, which can be quantified by a more detailed error analysis (see [10] and the references therein).

In related work, McGeer [11] and, more recently, Ruina and colleagues [12, 13] have designed, analyzed, and built passive walking machines that are entirely uncontrolled yet produce stable gaits. These differ from SLIP-type machines in that their rigid legs incur impacts at touchdown, and stable gaits emerge from a balance between energy supplied by motion down an inclined plane and energy losses due to impacts. Their mathematical models are significantly more complicated than the SLIP, and only limited analyses are possible. Similarly, a recent study of Mombaur et al. [14] relies on numerical optimization methods to find the "most stable" periodic gaits of a four-degree-of-freedom hopper endowed with a massy leg and a circular foot. They apply feedforward actuation via programmable leg length and hip torque and note that damper forces and impact losses "may promote stability." In contrast, the SLIP machines investigated in this paper are conservative and operate at constant energy; no friction forces are present, and no impact occurs at touchdown (see section 2 for details of the model).

Models as simple and (relatively) analytically tractable as the SLIP can address two key questions: how much energy and how much information are needed to sustain a gait? With regard to the second question, many researchers (e.g., [4, 15, 6]) implicitly assume that even if "passive dynamic" periodic gaits exist, they are (highly [4]) unstable. A surprising answer to both questions, motivated by hypotheses proposed in [16], was found by Schmitt and Holmes [17, 18, 19, 20] for the mechanics of a lateral leg spring (LLS) model (essentially a SLIP without gravity or flight phases), which describes horizontal plane motions of a rigid body equipped with a pair of massless springy legs that are lifted when leg force drops to zero, are swung forward, and are set down at fixed angles relative to the body. They showed that, even without energy dissipation, the LLS model can exhibit stable periodic gaits. Liftoff events alone trigger the swing phases: continuous (neural) sensing is not required, and stability derives from angular momentum trading from step to step. Moreover, recent experiments [21] have suggested that rapidly running insects do employ such mechanical reaction forces to make heading corrections.

In this paper, we demonstrate and, under simplifying assumptions, prove that stable periodic gaits exist in very simply controlled SLIP models over a physically useful range of parameter values. Specifically, we show that a liftoff-event-triggered reset of the leg angle during flight to achieve a touchdown angle fixed at the same value for each stance phase (hereafter, *fixed-leg reset*) suffices for stability. Such self-stabilized SLIP gaits have already appeared in the literature [10, Figure 2], where periodic SLIP trajectories were compared to

experimental data, although their stability properties were not discussed. Our present work also complements a recent paper of Seyfarth et al. [22], in which parameter ranges for stable, symmetric, periodic SLIP gaits are found by numerical simulation and are compared with data from human running. Here we derive analytical results, perform detailed bifurcation and parameter studies (including a second, nonlinear spring model), explain mechanisms responsible for stable gaits, and elucidate limits to fixed-leg reset stability. We relate our results to [22] where appropriate and summarize the relationship between that and the present work in section 5.

Our main results may be summarized as follows. Using conservation laws and simple geometric relations, we produce closed form approximations (explicit up to the evaluation of a quadrature integral) for the touchdown-to-touchdown Poincaré map and the "stability eigenvalue" of its fixed point for a simplified version of the model; see (2.20), (2.24), and Figures 7 and 11. These allow us to plot branches of stable and unstable periodic gaits (Figures 8 and 13) and to understand how the domains of attraction of the stable gaits depend upon parameters. Particular spring laws appear only in the quadrature. We believe that such explicit approximations have not previously appeared; moreover, exact Poincaré maps, requiring only numerical evaluation of the leg sweep angle during stance, are implicit in our derivation. An appropriate notion of stability for such piecewise-holonomic systems [23] is that of *partial asymptotic stability*. Due to energy conservation and rotational invariance (in the case of coincident "hip joint" and mass center), one or three of the eigenvalues of the linearized Poincaré map are necessarily unity, leaving a single "stability eigenvalue" that may lie within or outside the unit circle. Thus, at best, the orbits are only Liapunov or neutrally stable.

The paper is organized as follows. In section 2, we set up the general rigid body model and then focus on an integrable case, in which pitching motions decouple and gravity is neglected during the stance phase (2.1). This allows us to derive explicit stride-to-stride (Poincaré) maps and obtain expressions characterizing periodic gaits, their stability, and bifurcations. Apart from illustrations, this is all done for general leg-spring laws. We then give convincing numerical evidence that stable gaits persist under the inclusion of gravity during stance (2.2) and under coupling to pitching motions (2.3). In section 3, we illustrate our results using the classical Hooke's law spring (3.1) and a progressively hardening compressed air spring (3.2). In section 4, we reformulate the equations of motion in nondimensional variables and include gravity during stance, thereby clarifying the effects of parameter variations and the resulting range of behaviors exhibited by the model. Finally, section 5 summarizes the work and notes possible extensions.

This work has two main goals: to better understand animal locomotion and to stimulate and enable the creation of "bio-inspired" robots. A significant part of locomotion research is driven by the desire to exploit the advantages of legged robots as opposed to wheeled and tracked vehicles. Nature suggests (and engineers are increasingly concerned to demonstrate) that legged robots can operate over a greater range of environmental and surface conditions, combining dexterity with mobility and efficiency; cf. [24, 10]. In addition, machines that use ballistic flight phases do not require continuous support paths [5].

A natural extension of the work presented here is the design of control algorithms that enlarge the rather small basin of attraction of the SLIP with a constant leg touchdown angle

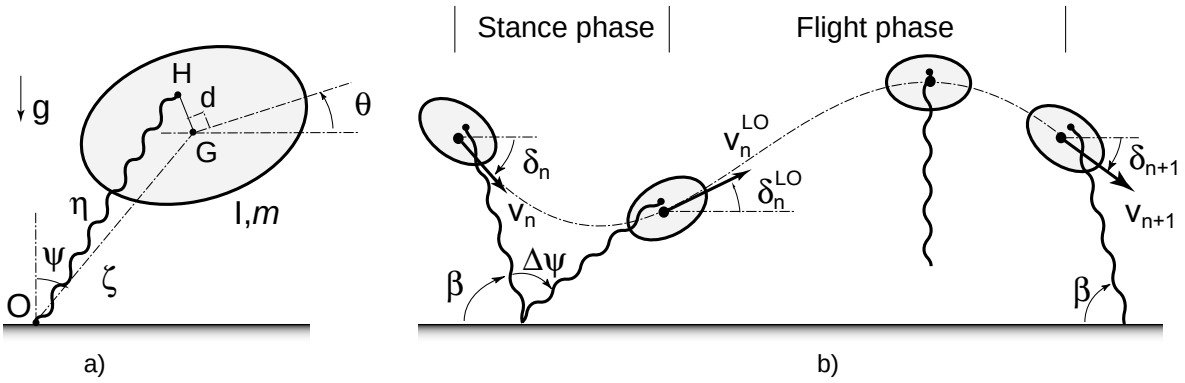


Figure 1. *The hopping rigid body* (a)*, and the stance and flight phases comprising a full stride* (b).

control [25]. A globally attracting "low attention" feedback controller was proposed in [26] by making the leg angle trajectory time-dependent during flight. However, this controller was based on a numerically precomputed leg trajectory and required velocity sensing. This raises the question of how much sensing is required to obtain "large" basins of attraction. A first step in this direction was undertaken in [27], where a necessary condition for the stability of fixed points for arbitrary leg angle trajectories was formulated in terms of the sensor requirements at liftoff. We will further explore this issue in a forthcoming paper [28]. Control enters the present paper only as the fixed feedforward leg placement strategy used at touchdown to define the hybrid switching condition.

2. The model: Equations of motion. Figure 1(a) illustrates our parametrization of the SLIP model as a schematic representation for the stance phase of a running (or hopping) biped with at most one foot on the ground at any time. This model incorporates a rigid body of mass m and moment of inertia I, possessing a massless sprung leg attached at a hip joint, H, a distance d from the COM, G. The figure depicts the attitude or pitch angle θ, the angle ψ formed between the line joining foothold O to the COM and the vertical (gravity) axis, and the distance ζ from foothold to the COM. The quantity

$$\eta = \sqrt{d^2 + \zeta^2 + 2d\zeta\cos(\psi + \theta)} \tag{2.1}$$

measures the (compressed) leg-spring length: the distance between O and the hip pivot H. We take frictionless pin joints at O and H. The body is assumed to remain in the vertical (sagittal) plane, and its state at any point in time is defined by the position of G, (x_G, y_G) referred to a Cartesian inertial frame, and the pitch angle θ; during stance we will also use the generalized polar coordinates ζ, ψ, based at the foothold O, and θ. (Note that ψ increases clockwise, while θ increases counterclockwise.) Unlike many earlier studies of the SLIP, we consider a rigid body with distributed mass and allow pitching motions,[1] although in the present paper we focus our attention upon the uncoupled case $d = 0$ and assume $\theta \equiv 0$, thus largely restricting ourselves to the point mass case.

[1] A bipedal walker with the above-described leg and body geometry with arbitrary radial force in the leg and arbitrary hip torque was considered in [29] in the context of feedback control. However, the investigation did not include gaits with flight phases.

A full stride divides into a stance phase, with foothold O fixed, the leg under compression, and the body swinging forward (ψ increasing), and a flight phase in which the body describes a ballistic trajectory under the sole influence of gravity. The stance phase ends when the spring unloads; the flight phase then begins, continuing until touchdown, which occurs when the landing leg, uncompressed and set at a predetermined angle β, next contacts the ground. See Figure 1. This defines a *hybrid* system in which touchdown and liftoff conditions mark transitions between two dynamical régimes.

Recalling previous robotics research [30] and looking ahead to control studies [28], β could be adjusted from stride to stride (necessitating at least intermittent active neural feedback), but here it will be taken as a fixed parameter. The "fixed leg reset angle" policy of stated interest might be implemented with respect either to the body or to the inertial frame. In the first case, touchdown occurs when the hip reaches the height $\eta_0 \sin(\beta - \theta)$ and in the second case when the hip reaches the height $\eta_0 \sin\beta$. Liftoff occurs automatically when the spring force drops to zero, requiring no sensing, but in any physical implementation, even a fixed-leg reset policy requires some state information to initiate the swing phase (e.g., a contact sensor in the foot or force sensor in the spring).

The kinetic energy of the body is

$$T = \frac{1}{2}m(\dot{\zeta}^2 + \zeta^2\dot{\psi}^2) + \frac{1}{2}I\dot{\theta}^2, \tag{2.2}$$

and its potential energy is

$$V_{tot} = mg\zeta\cos\psi + V\left(\eta\left(\zeta,\psi,\theta\right)\right), \tag{2.3}$$

where $V = V_{spr}$ denotes the spring potential. Forming the Lagrangian $L = T - V$ and writing $\partial V/\partial\eta = V_\eta$, we obtain the equations of motion for the stance phase:

$$\begin{aligned}
\ddot{\zeta} &= \zeta\dot{\psi}^2 - g\cos\psi - \frac{V_\eta\left(\eta\right)}{m\eta}\left(\zeta + d\cos\left(\psi+\theta\right)\right),\\
\zeta\ddot{\psi} &= -2\dot{\zeta}\dot{\psi} + g\sin\psi + d\frac{V_\eta\left(\eta\right)}{m\eta}\left(\sin\left(\psi+\theta\right)\right),\\
\ddot{\theta} &= d\zeta\frac{V_\eta(\eta)}{\eta I}\sin\left(\psi+\theta\right).
\end{aligned} \tag{2.4}$$

The equations of motion during the flight phase are simply the ballistic COM translation and torque-free rotation equations, which may be integrated to yield

$$x_G\left(t\right) = x^{LO} + \dot{x}^{LO}t, \quad y_G\left(t\right) = y^{LO} + \dot{y}^{LO}t - \frac{1}{2}gt^2, \quad \theta\left(t\right) = \theta^{LO} + \dot{\theta}^{LO}t, \tag{2.5}$$

where (x_G, y_G) denotes the COM position and θ the pitch angle, and the superscripts LO refer to the system state at liftoff.

2.1. The case $d = 0$ neglecting gravitational effects in stance. If the leg is attached at the COM ($H \equiv G$), then $d = 0, \zeta \equiv \eta$, the stance phase dynamics simplifies to the "classical"

SLIP, and the pitching equation decouples:

$$\ddot{\zeta} = \zeta\dot{\psi}^2 - g\cos\psi - \frac{V_\zeta(\zeta)}{m}\ , \quad \zeta\ddot{\psi} = -2\dot{\zeta}\dot{\psi} + g\sin\psi,$$
$$\ddot{\theta} = 0 \Rightarrow \theta(t) = \theta(0) + \dot{\theta}(0)\,t. \tag{2.6}$$

The third equation describes the conservation of angular momentum of the body about its COM: $I\dot{\theta} \triangleq p_\theta = \text{const}$.

Neglect of gravity in stance yields an integrable system [31]. A detailed analysis of the validity of this approximation for different spring potentials was performed in [32] using Hamiltonian instead of Lagrangian formalism. This simplification was shown to be too crude over a large range of running gaits, and several closed form approximations to the stance phase dynamics were proposed, although existence and stability of periodic solutions that can arise from concatenation of stance and flight phases were not investigated. Despite the limited accuracy of the gravity-free approximation, we adopt it here in order to gain an analytical understanding of periodic gaits. We will subsequently compare these results to numerical simulations of the full stance dynamics with gravity and show that analogous bifurcation structures persist in the physically more accurate model.

Neglecting gravity, the first two equations of (2.6) simplify to

$$\ddot{\zeta} = \zeta\dot{\psi}^2 - \frac{V_\zeta(\zeta)}{m}\ ,\ \zeta\ddot{\psi} = -2\dot{\zeta}\dot{\psi}. \tag{2.7}$$

The second of these equations expresses the conservation of the moment of linear momentum of the COM about the foot: $m\dot{\psi}\zeta^2 \triangleq p_\psi = \text{const}$. The first equation is, therefore, integrable:

$$\ddot{\zeta} = \frac{p_\psi^2}{m^2\zeta^3} - \frac{V_\zeta(\zeta)}{m} \ \Rightarrow\ m\ddot{\zeta}\dot{\zeta} = \frac{p_\psi^2}{m\zeta^3}\dot{\zeta} - V_\zeta(\zeta)\,\dot{\zeta} \ \Rightarrow$$
$$H \triangleq \left(\frac{m\dot{\zeta}^2}{2} + \frac{p_\psi^2}{2m\zeta^2} + V(\zeta)\right) = \text{const}. \tag{2.8}$$

Indeed, in the absence of dissipative forces, the total energy, which coincides here with the Hamiltonian $H = T+V = E$, is conserved. The original three degrees of freedom reduce to one due to the conservation of moment of linear momentum p_ψ and body angular momentum p_θ individually. The phase portrait during stance is then given by the level sets of H in the region $\zeta \le \eta_0$; Figure 2 illustrates this for a linear spring and also includes comparisons to solutions of the full system (2.6), including gravity. Three cases are shown, with different stiffness/gravity ratios characterized by the nondimensional parameter $\gamma = \frac{k\eta_0}{mg} \in [10, 100]$. As expected [33, 34], the integrable portraits are perturbed by the inclusion of gravity, but orbits retain the same qualitative characteristics. Leg stiffnesses estimated for human running, for example, give $\gamma \in (10, 21)$ [35, 36] (although Seyfarth et al. propose significantly higher values ($\gamma \in (25, 70)$; cf. [22, Fig. 2A]). Errors approach 20% at the lower end of this range at midstance (near $\dot{\zeta} = 0$ in Figure 2(c)) but are smaller at liftoff. Extensive simulation experience confirms that errors decrease with increasing γ (or k) for initial conditions away from the extremes of the physically interesting operating regimes; see [32] for a careful discussion of such modeling errors; also see

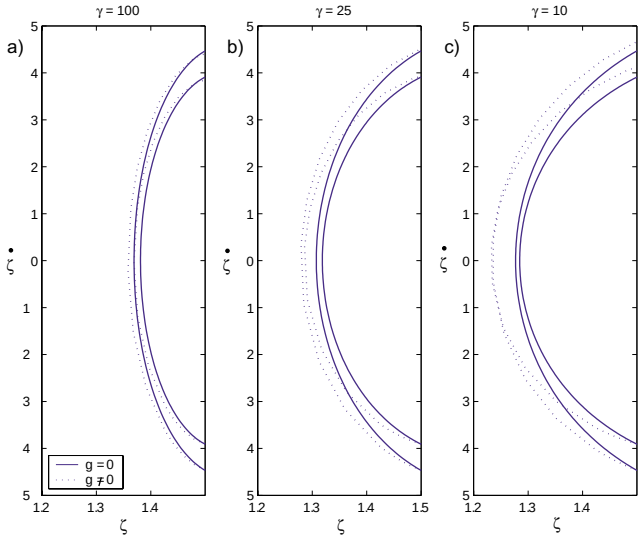


Figure 2. *Solutions of the integrable system (2.7) (solid) and the full ($d = 0$) system (dashed) in the stance phase: $m = 1, \eta_0 = 1.5, \beta = \pi/4$ with linear spring stiffnesses $k = 654$ ($\gamma = 100$) (a), $k = 163.5$ ($\gamma = 25$) (b), and $k = 65.4$ ($\gamma = 10$) (c). Dimensional units, unless otherwise stated, are MKS.*

section 2.2 and Figure 4 below. Moreover, for orbits reflection-symmetric about midstance for which $\psi(t)$ is an odd function, such as the periodic gaits to be found below, the net angular impulse delivered during each stance phase is zero so that, while p_ψ is not conserved, it does regain its touchdown value at liftoff. This also tends to minimize errors.

In principle, we can integrate (2.8), first solving for time in terms of ζ and then inverting and solving for $\zeta(t)$ and $\psi(t)$. In particular, the quadrature determining the angle swept by the leg may be written as $\Delta\psi(v_n, \delta_n) = \int_0^\tau \frac{p_\psi}{m\zeta^2} dt$, where v_n and δ_n denote the COM velocity magnitude and direction relative to horizontal at the nth touchdown instant. Hence the moment of linear momentum for the nth stance phase may be computed as $p_\psi = m\eta_0 v_n \sin(\beta - \delta_n)$. Then, from conservation of energy (2.8), we have

$$\dot{\zeta} = \sqrt{\frac{2}{m}\left(E - V(\zeta)\right) - \frac{p_\psi^2}{m^2\zeta^2}} \;\Rightarrow\; dt = \frac{m\, d\zeta}{\sqrt{2m\left(E - V\left(\zeta\right)\right) - \frac{p_\psi^2}{\zeta^2}}} \tag{2.9}$$

so that the sweep angle may then be expressed as the quadrature

$$\Delta\psi\left(v_n, \delta_n\right) = 2\int_{\zeta_b}^{\eta_0} \frac{\eta_0 v_n \sin(\beta - \delta_n)\, d\zeta}{\zeta^2\sqrt{v_n^2 - \frac{2V(\zeta)}{m} - \frac{\eta_0^2 v_n^2 sin^2(\beta - \delta_n)}{\zeta^2}}}. \tag{2.10}$$

Here we have set $E = \frac{1}{2}mv_n^2$, corresponding to the initial energy at touchdown, and $\zeta_b \le \eta_0$ denotes the midstride (compressed) leg length.

Computations of $\Delta\psi$ in specific cases of a linear spring and an "air spring" with potential $V(\eta) = \frac{c}{2}\left(\frac{1}{\eta} - \frac{1}{\eta_0}\right)^2$ are given in [17]. Schwind and Koditschek [32] develop an approximate

expression for this quadrature and compare it with both the exact integral and the analogous stance sweep angle including gravitational effects. In the present paper, we illustrate the general model again with a linear Hooke's law spring but adopt a different version of the nonlinear air spring model—specifically, that used in [37, 32]. While this potential, $V(\eta) = \frac{c}{2}(\frac{1}{\eta^2} - \frac{1}{\eta_0^2})$, results in an inverse cubic force law of the form $-\frac{c}{\eta^3}$ that is nonzero at touchdown and liftoff, energy is conserved since the leg lengths are the same ($\eta = \eta_0$), and velocities are continuous. The explicit sweep angle expression for this law is given in Appendix A. For the linear spring, $V(\zeta) = \frac{k}{2}(\eta - \eta_0)^2 = \frac{k}{2}(\zeta - \eta_0)^2$, and we have $\Delta\psi = \frac{2\sin(\beta-\delta)}{\sqrt{\tilde{k}}} D(\tilde{k}; \beta - \delta)$, where $\tilde{k} = \frac{k\eta_0^2}{mv_n^2}$ and the function $D(\tilde{k}; \beta - \delta)$ involves elliptic integrals [17, Appendix A.1.2].

The stance phase dynamics described above must be composed with the ballistic dynamics of the flight phase of (2.5), and the overall dynamics and the stability of this piecewise-holonomic system [31] are best described via Poincaré or return maps [34]. It is convenient to choose as generalized coordinates to describe the map the magnitude of touchdown and liftoff velocities v_n^{TD} and v_n^{LO}, respectively, and the relative angles δ_n^{TD} and δ_n^{LO} between the velocity vectors and the horizontal datum; see Figure 1b. The full map is obtained by composition of the stance phase map

$$P_{st} : \begin{bmatrix} v_n^{TD} \\ \delta_n^{TD} \end{bmatrix} \mapsto \begin{bmatrix} v_n^{LO} \\ \delta_n^{LO} \end{bmatrix} \tag{2.11}$$

and the flight map

$$P_{fl} : \begin{bmatrix} v_n^{LO} \\ \delta_n^{LO} \end{bmatrix} \mapsto \begin{bmatrix} v_{n+1}^{TD} \\ \delta_{n+1}^{TD} \end{bmatrix} \tag{2.12}$$

as

$$P = P_{fl} \circ P_{st} : \begin{bmatrix} v_n^{TD} \\ \delta_n^{TD} \end{bmatrix} \mapsto \begin{bmatrix} v_{n+1}^{TD} \\ \delta_{n+1}^{TD} \end{bmatrix}. \tag{2.13}$$

Since $I\dot{\theta} = I\dot{\theta}_0 = \text{const}$ implies that $\theta(t) = \theta_0 + \dot{\theta}_0 t$, and at touchdown in the first protocol the leg is placed at a fixed angle relative to the body, to obtain "sensible" periodic gaits we henceforth assume $\dot{\theta} = \dot{\theta}_0 = 0$. In this case, since $d = 0$ and $\theta \equiv 0$, there is no distinction between the two leg placement protocols.

We now describe the maps in detail, deriving explicit formulae. We shall frequently drop the superscript TD and write $v_n^{TD} = v_n$ and $\delta_n^{TD} = \delta_n$, it being understood that $(v_n, \delta_n) \mapsto P(v_n, \delta_n)$ denotes the touchdown-to-touchdown map.

2.1.1. Stance phase map. The spring is fully extended and stores no potential energy at the beginning or the end of each stance phase. Choosing the reference height for zero gravitational energy at $y = \eta_0 \sin\beta$, the energy at touchdown is therefore purely kinetic, $E_n^{TD} = \frac{1}{2}m(v_n^{TD})^2$, while at liftoff the energy has in general a gravitational component, $E_n^{LO} = \frac{1}{2}m(v_n^{LO})^2 + mg\eta_0(\sin(\beta + \Delta\psi) - \sin\beta)$, the last term being positive, zero, or negative. Appealing to overall energy conservation $E_n^{LO} = E_n^{TD}$, the liftoff velocity is therefore

$$v_n^{LO} = \sqrt{v_n^2 + 2g\eta_0(\sin\beta - \sin(\beta + \Delta\psi))}. \tag{2.14}$$

As noted earlier, if the spring is sufficiently stiff so that gravity is negligible, the moment of linear momentum p_ψ is conserved throughout stance in what is effectively a central force

problem [31]: $p_\psi = m\,\underline{r}_n \times \underline{v}_n = m\,\underline{r}_n^{LO} \times \underline{v}_n^{LO}$. Since $|\underline{r}_n \times \underline{v}_n| = \eta_0 v_n \sin(\delta_n - \beta)$ and $|\underline{r}_n^{LO} \times \underline{v}_n^{LO}| = \eta_0 v_n^{LO} \sin(\delta_n^{LO} - \pi + \Delta\psi + \beta)$, we obtain

$$\delta_n^{LO} = \pi - \Delta\psi - \beta + \sin^{-1}\left(\frac{v_n}{v_n^{LO}}\sin(\delta_n - \beta)\right). \tag{2.15}$$

However, since gravity is ignored in the sweep angle computation of (2.10), for consistency we must also ignore it in assigning a liftoff velocity magnitude in (2.15) and set $v_n^{LO} = v_n$ so that (2.15) simplifies to

$$\delta_n^{LO} = \delta_n + \pi - \Delta\psi(v_n, \delta_n) - 2\beta, \tag{2.16}$$

as in the LLS computations of [17]. Thus the effects of gravity are included in computing liftoff velocity magnitude (2.14) but *not* in approximating liftoff velocity direction (2.16). This "mixed approximation" has the advantage of retaining global energy conservation. Equations (2.14)–(2.16), with (2.10), specify P_{st}. Note that (2.14), along with a (numerical) calculation of the leg sweep angle $\Delta\psi$ and the change in p_ψ due to gravitational moment, defines the exact stance phase map including gravity. We use this in section 4.

We note that p_ψ is reset on each touchdown and that this "trading" of angular momentum from stride to stride will be responsible for asymptotic stability; cf. [17].

2.1.2. Flight phase and overall Poincaré map P. Using similar arguments based on conservation of energy,

$$E_n^{LO} = \frac{1}{2}m(v_n^{LO})^2 + mg\eta_0\left(\sin(\beta + \Delta\psi) - \sin(\beta)\right) = E_{n+1}^{TD} = \frac{1}{2}m(v_{n+1}^{TD})^2,$$

and on conservation of linear momentum in the horizontal direction,

$$v_n^{LO}\cos(\delta_n^{LO}) = v_{n+1}^{TD}\cos(\delta_{n+1}^{TD}), \tag{2.17}$$

we find the flight phase map. For convenience, both maps are specified here:

$$P_{st}: \begin{bmatrix} v_n^{LO} \\ \delta_n^{LO} \end{bmatrix} = \begin{bmatrix} \sqrt{v_n^2 + 2g\eta_0\left(\sin\beta - \sin(\beta + \Delta\psi)\right)} \\ \delta_n + \pi - \Delta\psi - 2\beta \end{bmatrix}, \tag{2.18}$$

$$P_{fl}: \begin{bmatrix} v_{n+1} \\ \cos(\delta_{n+1}) \end{bmatrix} = \begin{bmatrix} \sqrt{(v_n^{LO})^2 + 2g\eta_0\left(\sin(\beta + \Delta\psi) - \sin\beta\right)} \\ \frac{v_n^{LO}}{v_n}\cos(\delta_n^{LO}) \end{bmatrix}. \tag{2.19}$$

The last equation should more properly read $v_{n+1}^{TD}\cos(\delta_{n+1}^{TD}) = v_n^{LO}\cos(\delta_n^{LO})$, but provided β and η_0 remain constant, conservation of energy enforces, without approximations, that $v_{n+1} = v_n$ because the energy at the beginning and the end of each full stance + flight stride is entirely kinetic. The flight map is only implicitly defined, and it is not evident that one can find an expression in terms of v_n^{LO}, δ_n^{LO} above, especially because $\Delta\psi = \Delta\psi(v_n, \delta_n)$ is a complicated function of the touchdown conditions; see (2.10). Nonetheless, using $v_{n+1} = v_n$, the full map simplifies considerably:

$$P: \begin{bmatrix} v_{n+1} \\ \cos(\delta_{n+1}) \end{bmatrix} = \begin{bmatrix} v_n \\ \sqrt{1 - \frac{2g\eta_0}{v_n^2}\left(\sin(\beta + \Delta\psi) - \sin\beta\right)}\cos(\delta_n + \pi - \Delta\psi - 2\beta) \end{bmatrix}. \tag{2.20}$$

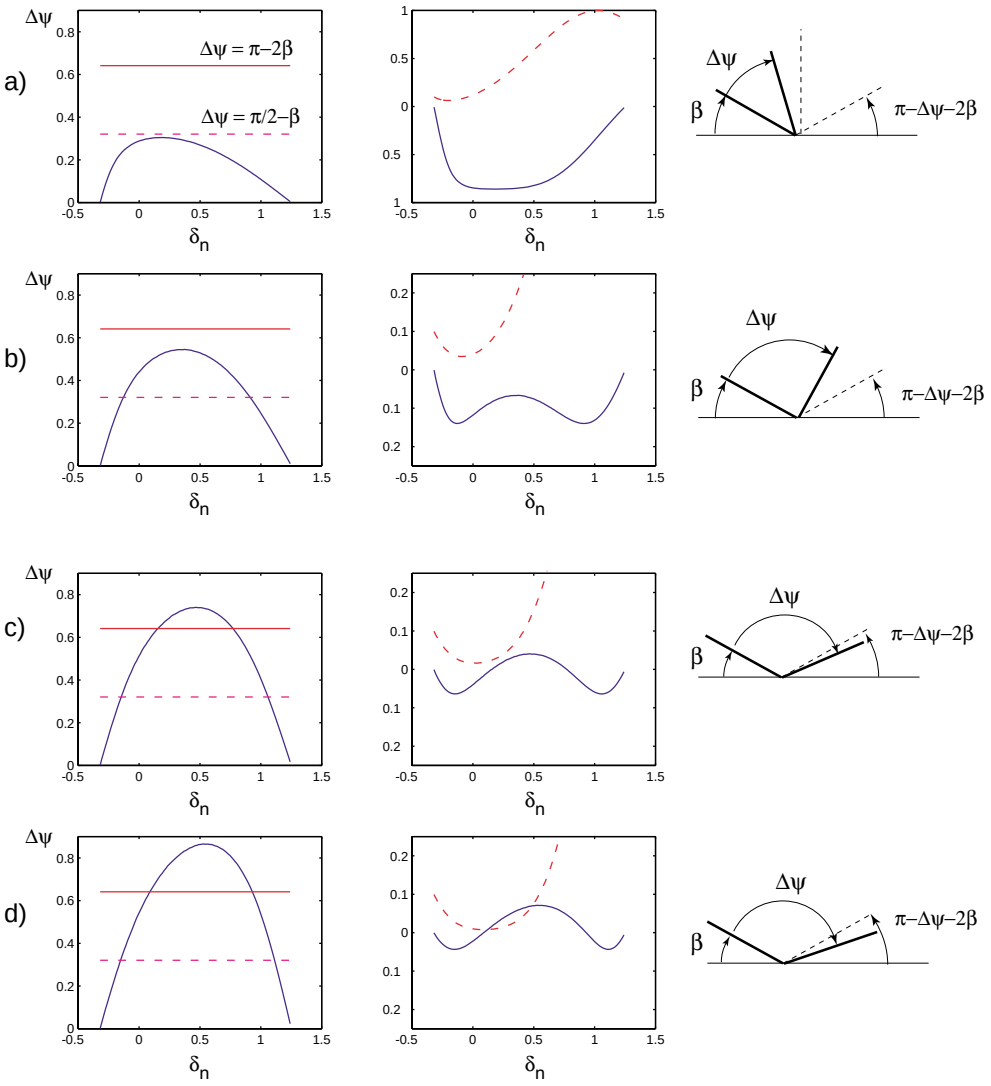


Figure 3. *The first column shows the function $\Delta\psi$ computed for a linear spring with $k = 100, m = 1, \eta_0 = 1.5$, and $\beta = 1.25$. For cases* (a) *through* (d)*, we set $\bar{v} = 1.75, 3.5, 5$, and 6, respectively. The conditions $\Delta\psi = \pi - 2\beta$ (dotted) and $\Delta\psi = \frac{\pi}{2} - \beta$ (dashed) are also shown. The second column shows the left-hand (dotted) and right-hand (solid) sides of inequality* (2.26)*: see sections* 2.1.4–2.1.5*. The third column shows β and $\Delta\psi$ in physical space: solid lines indicate angles at touchdown (β) and liftoff ($\beta + \Delta\psi$). When $\beta + \Delta\psi < \pi - \Delta\psi - 2\beta$ (cases* (a) *and* (b)*), the body leaves the ground at an angle closer to vertical than at touchdown. Note that* (2.26) *is violated for part of the domain in* (d)*.*

This expression is explicit apart from the sweep angle $\Delta\psi(v_n, \delta_n)$, which must be computed from the quadrature of (2.10). Only here does the specific spring potential enter; the rest of the expression for P is derived purely from conservation laws and stance and flight path geometry. Note that, although we have approximated $\Delta\psi$ and hence P_{st} by neglecting the

effect of gravitational torque in changing the COM angular momentum about the foot, the overall composed map P conserves energy, as would the exact solutions of (2.6).

We postpone quantitative analyses of specific spring potentials to sections 3–4; however, we note that analysis of special cases and numerical evidence indicates that for linear and stiffening springs, $\Delta\psi$ has a single maximum. This will suffice for the analysis of the present section. In particular, it is clear that for $\delta_n = \beta - \pi/2$ (glancing contact), $\tau = \Delta\psi = 0$, and for $\delta_n = \beta$ (running directly into the leg), $\Delta\psi = 0$. Thus $\Delta\psi = 0$ at both limits of the admissible δ range, while for any $v_n \neq 0$, $\Delta\psi > 0 \ \forall \ \delta_n \in (\beta - \pi/2, \beta)$, so there must be at least one maximum. We suspect that any physically reasonable spring law will give a $\Delta\psi$ with a unique maximum. The left column of Figure 3 shows $\Delta\psi$ for the linear spring evaluated numerically for several $\bar{v}$ values; they are indistinguishable from those obtained via the analytical expressions of [17].

We remark that, in view of energy conservation and the resulting constancy of v_n, (2.20) defines a one-dimensional map for the touchdown angle δ_n. One could specify the system's state in terms of any other convenient variable, such as the COM height at the apex, which was the choice adopted in [22]; cf. Figure 3(A) of that paper. We prefer to use the touchdown angle and retain the velocity as a second state variable so that, when $d \neq 0$, we may more conveniently couple in the attitude dynamics in terms of θ and $\dot{\theta}$, as was done for yawing motions in the LLS models of [17, 19]. Also, as demonstrated below, branches of periodic orbits and their domains of attraction are conveniently presented in terms of δ_n (cf. Figure 8).

2.1.3. Periodic gaits. The simplest sustained forward motions, in which the hopper maintains a constant average forward speed and lands with the same angle between the velocity vector and the horizontal datum on each step, are period one orbits given by $v_{n+1} = v_n$ and $\delta_{n+1} = \delta_n$. As we see from (2.20), the first condition is always satisfied, whereas the second condition holds if and only if $\Delta\psi\,(v_n, \delta_n) = \pi - 2\beta$.

To verify this, we first check sufficiency. Let $\Delta\psi(v_n, \delta_n) = \pi - 2\beta$. Then $\sin(\beta + \Delta\psi) - \sin\beta = \sin(\pi - \beta) - \sin\beta = 0$, and the map (2.20) reduces to

$$P : \begin{bmatrix} v_{n+1} \\ \cos(\delta_{n+1}) \end{bmatrix} = \begin{bmatrix} v_n \\ \cos(\delta_n) \end{bmatrix}. \tag{2.21}$$

At touchdown following a flight phase, $\delta_n \in [0, \pi]$ (for both locomotion directions). In that range, $\cos\delta_n$ is invertible; hence $\delta_{n+1} = \delta_n$.

Now let $\delta_{n+1} = \delta_n$. For sustained forward motion, $\Delta\psi(v_n, \delta_n) \in [\pi/2 - \beta, \pi - \beta]$ and $\delta_n^{LO} \in [0, \pi/2]$. Assume $\Delta\psi(v_n, \delta_n) > \pi - 2\beta$. Then $\delta_n^{LO} = \delta_n + \pi - \Delta\psi - 2\beta < \delta_n$ and $\cos\delta_n^{LO} > \cos\delta_n \ \forall \delta_n, \delta_n^{LO} \in [0, \pi/2]$. Also, $\beta + \Delta\psi > \pi - \beta$ and $\sin(\beta + \Delta\psi) < \sin(\pi - \beta)$ $\forall \beta \in [0, \pi/2]$ and $\forall \Delta\psi \in [\pi/2 - \beta, \pi - \beta] \Rightarrow \sin(\beta + \Delta\psi) - \sin\beta < 0 \ \forall \beta \in [0, \pi/2]$. Hence we conclude that

$$\cos(\delta_{n+1}) > \left[1 - \frac{2g\eta_0}{v_n^2}\left(\sin(\beta + \Delta\psi) - \sin\beta\right)\right]^{\frac{1}{2}} \cos(\delta_n) > \cos(\delta_n), \tag{2.22}$$

which is a contradiction. A similar argument holds for $\Delta\psi(v_n, \delta_n) < \pi - 2\beta$. Therefore, $\Delta\psi(v_n, \delta_n) = \pi - 2\beta$ is also necessary. Hence $\delta_n^{LO} = \delta_n$, and in the gravity-free approximation with $d = 0$, all one-periodic gaits are reflection-symmetric about midstance [38].

Note that, within limits to be determined below, $v_{n+1} = v_n = \bar{v}$ can be chosen arbitrarily, and the expression $\Delta\psi(\bar{v}, \delta_n) = \pi - 2\beta$ can be solved to obtain the fixed point that we denote by $\bar{\delta}$. Here we appeal to the fact that a parabolic segment of the flight trajectory can always be matched to connect reflection-symmetric stance phases (i.e., those having $\delta_n^{LO} = \delta_n$; see Figure 1), yielding a fixed point of P. Thus, there is a one-parameter ($\bar{v}$-) family of steady periodic gaits for each β and all other parameters fixed. Also see [17] and Figure 8.

We may linearize the general expression (2.20) at a fixed point of the map to obtain the Jacobian matrix

$$DP|_{\delta_n=\bar{\delta}} = \left[\begin{array}{cc} 1 & 0 \\ -\left(1 + \frac{g\eta_0 \cos\beta \cot\delta_n}{\bar{v}^2}\right) & 1 - \left(1 + \frac{g\eta_0 \cos\beta \cot\delta_n}{\bar{v}^2}\right)\frac{\partial\Delta\psi}{\partial\delta_n} \end{array} \right]\Bigg|_{\delta_n=\bar{\delta}}, \tag{2.23}$$

the eigenvalues of which are $\lambda_1 = 1$ and

$$\lambda_2 = 1 - \left(1 + \frac{g\eta_0 \cos\beta \cot\delta_n}{\bar{v}^2}\right)\frac{\partial\Delta\psi}{\partial\delta_n}\Bigg|_{\delta_n=\bar{\delta}}. \tag{2.24}$$

The first eigenvalue, λ_1, lies on the unit circle, corresponding to conservation of energy, but $|\lambda_2|$ may take values either greater than or less than 1. We require $\bar{\delta} \in (0, \beta)$ and $\beta \in \left(0, \frac{\pi}{2}\right)$ for physically admissible gaits; thus the quantity in parentheses in (2.24) is strictly positive, and a key factor in determining λ_2 is the sign of $\frac{\partial\Delta\psi}{\partial\delta_n}$. If $\frac{\partial\Delta\psi}{\partial\delta_n} < 0$, then $|\lambda_2| > 1$, and the fixed point is unstable; if $\frac{\partial\Delta\psi}{\partial\delta_n} > 0$, $|\lambda_2|$ may be less than or greater than 1, and stability or instability may ensue [39, 34] (see below).

As in [17], recognizing that energy is conserved, stability can only be *partially asymptotic*, since perturbations in the direction of the eigenvector of λ_1 neither grow nor decay. As noted at the close of section 2.1.1, both here and in [17], the physical mechanism for stabilization appears to be the trading of angular momentum from stride to stride. As noted by Ruina [23] (cf. [40]), such piecewise-holonomic systems can yield asymptotic stability much like nonholonomically constrained conservative systems [41].

2.1.4. Domain of definition of P. We must recall that the map was derived under the tacit assumption that unimpeded leg motion is possible over the entirety of the configuration space of the kinematic model. This is not true in general, as the toe must not be allowed to penetrate the ground. The touchdown angle β is held constant, and since the spring has a fixed length at rest η_0, the hip height at touchdown is also fixed: $y^{TD} = \eta_0 \sin\beta$. For a gait to exist, this height must be reached during the flight phase, i.e., $y_{G\,\max} \geq \eta_0 \sin\beta$; otherwise, the hopper will "stumble." Integrating the ballistic equations (2.5), the time of flight to reach the apex is $t_{\max} = v_n^{LO} \sin\delta_n^{LO}/g$, and the maximum height is given by (2.5):

$$y_{G\,\max} = y_G(t_{\max}) = \eta_0 \sin(\beta + \Delta\psi) + \frac{(v_n^{LO} \sin\delta_n^{LO})^2}{2g}.$$

Hence the map P is defined if and only if

$$\sin^2\delta_n^{LO} \geq \frac{2g\eta_o(\sin\beta - \sin(\beta + \Delta\psi))}{v_n^2 + 2g\eta_0(\sin\beta - \sin(\beta + \Delta\psi))} \tag{2.25}$$

or, using (2.16),

$$\sin^2(\delta_n + \pi - \Delta\psi - 2\beta) \geq \frac{2g\eta_o(\sin\beta - \sin(\beta + \Delta\psi))}{v_n^2 + 2g\eta_0(\sin\beta - \sin(\beta + \Delta\psi))}. \tag{2.26}$$

Inequality (2.26), which may be implicitly written in the form

$$f(\bar{v}, \delta_n; \beta, \eta_0, g, m, V(\cdot)) \geq 0,$$

specifies the domain of definition of P (admissible values of $(\bar{v}, \delta_n)$) for each choice of physical parameters (β, η_0, g, m) and spring potential V. It appears difficult to give explicit bounds, but we observe that, when $\max \Delta\psi(\bar{v}, \delta_n) \geq \pi - 2\beta$ and reflection-symmetric stance paths with $\delta_n = \bar{\delta}$ exist, we have $\sin(\beta + \Delta\psi) = \sin(\pi - \beta)$, and the right-hand side of (2.26) vanishes. For physically relevant gaits, $\bar{\delta} \in [0, \beta]$ and $\beta < \frac{\pi}{2}$; hence the left-hand side is strictly positive at fixed points unless $\bar{\delta} = 0$. However, since the spring remains compressed during stance, providing a positive radial force, we see that $\frac{d^2\zeta}{d\psi^2} > 0$, which implies $\frac{d^2\hat{y}}{d\hat{x}^2} > 0$, where $\hat{x}$ and $\hat{y}$ are the axes of a rotated orthogonal coordinate system that has its $\hat{y}$-axis aligned with the symmetry axis of the COM path. Hence the COM path is convex (cf. Figure 1), and $\bar{\delta} = 0$ cannot be a fixed point. (The COM path need not be convex when gravity is included; indeed, one may find orbits with $\delta_n^{TD} < 0$.) The second column of Figure 3 shows the two sides of inequality (2.26).

We may therefore conclude via continuous dependence on initial data that the domain of definition of P contains open sets around each fixed point, and, if $|\lambda_2| < 1$ (resp., > 1), local asymptotic stability (resp., instability) holds in the usual sense.

2.1.5. Bifurcations and stability of fixed points of P. To introduce the range of dynamical behaviors of P and better understand its domain of definition, we consider four representative cases depending on the maximum sweep angle $\Delta\psi_{\max}$:

(a) $\Delta\psi_{\max} < \frac{\pi}{2} - \beta$;
(b) $\frac{\pi}{2} - \beta < \Delta\psi_{\max} < \pi - 2\beta$;
(c) $\pi - 2\beta < \Delta\psi_{\max}$ and (2.26) is satisfied everywhere;
(d) $\pi - 2\beta < \Delta\psi_{\max}$ and (2.26) is not satisfied everywhere.

When $\Delta\psi_{\max} \leq \pi/2 - \beta$, the leg is vertical or directed forward at liftoff, so $\delta_{n+1}^{LO} > \delta_n$ and the direction of locomotion reverses once $\delta_n^{LO} > \pi/2$, even though the map may be well defined; see Figures 3(a) and 6(a).

For $\pi/2 - \beta < \Delta\psi_{\max} < \pi - 2\beta$, a domain appears in which $\psi_n^{LO} = \psi_n^{TD} + \Delta\psi_{\max} > 0$ and continuing forward motion is possible. However, the hopper still lifts off and touches down "more vertically" on each step until it eventually bounces backward in this case, too; see Figures 3(b) and 6(b). Indeed, from (2.16) we have $\delta_n^{LO} = \delta_n + \pi - \Delta\psi - 2\beta$, and by assumption (b) $\delta_n^{LO} > \delta_n$. From (2.17) we know that $\cos(\delta_{n+1}) = \frac{v_n^{LO}}{v_n}\cos(\delta_n^{LO})$. Now $\delta_n^{LO} \in \left(0, \frac{\pi}{2}\right)$, and the cosine function is monotonically decreasing. Since the hip position at liftoff is higher than at touchdown, the body has gained gravitational energy at the expense of kinetic energy. This means that $v_n^{LO} < v_n$, and therefore $\cos\delta_{n+1} = \frac{v_n^{LO}}{v_n}\cos\delta_n^{LO} < \cos\delta_n^{LO} < \cos\delta_n$. However, this implies that $\delta_{n+1} > \delta_n$. Thus, starting with an initial angle δ_n, after the stance phase, $\delta_n^{LO} > \delta_n$, and after the flight phase, $\delta_{n+1} > \delta_n^{LO} > \delta_n$. Hence succeeding touchdown angles

increase until progress is reversed; the dynamics is globally unstable, and the Poincaré map has no fixed points.

Cases (c) and (d) are of greater physical interest. In (c), inequality (2.26) is satisfied everywhere, so the domain of definition covers the interval $[\beta - \pi/2, \beta]$. Moreover, two fixed points exist, one of which may be stable, while the other (with higher values of $\bar{\delta}$) is unstable. These fixed points appear in a saddle-node bifurcation [34] at a critical speed $v = v_{SN}$. Indeed, for the smaller $\bar{\delta}$ fixed point, $\frac{\partial \Delta\psi}{\partial \delta_n} > 0$ (see Figure 3(c)), $\bar{\delta} > 0 \Rightarrow \cot\bar{\delta} > 0$, and $\lambda_2 = 1 - (1 + \frac{g\eta_0 \cos\beta \cot\bar{\delta}}{\bar{v}^2})\frac{\partial \Delta\psi}{\partial \delta_n} \stackrel{\triangle}{=} 1 - a\frac{\partial \Delta\psi}{\partial \delta_n}$. For the parameter values chosen here, $a \approx 3 > 0$; thus for $\frac{\partial \Delta\psi}{\partial \delta_n} \in (0, \frac{2}{3})$, $-1 < \lambda_2 < 1$, and we have established asymptotic stability. More generally, since the term $\frac{\partial \Delta\psi}{\partial \delta_n} = 0$ when $\Delta\psi_{\max} = \pi - 2\beta$, by continuous dependence on parameters $\frac{\partial \Delta\psi}{\partial \delta_n}$ is necessarily arbitrarily small for nearby parameter values, implying stability of the fixed point with smaller $\bar{\delta}$ in a neighborhood of the saddle-node bifurcation point. See Figures 3(c) and 6(c).

In case (d), $\Delta\psi_{max} \geq \pi - 2\beta$, but the map P is not everywhere defined: Figure 3(d) shows that inequality (2.26) fails in the interior of $[\beta - \pi/2, \beta]$. A *gap* opens between the fixed points and while a (stable) fixed point still exists to the left of the gap, many orbits, including that shown in Figure 6(d), enter the gap and "stumble."

We now summarize key aspects of the behaviors described above. More detailed analyses for specific spring potentials are given in sections 3–4.

Saddle-node bifurcation. As noted above, a saddle-node bifurcation occurs between regimes (b) and (c). Specifically, for parameter values such that

$$\Delta\psi_{\max}\left(\bar{v}_{SN}, \bar{\delta}_{SN}\right) = \pi - 2\beta \text{ and } \left.\frac{\partial \Delta\psi}{\partial \delta_n}\right|_{\left(\bar{v}_{SN}, \bar{\delta}_{SN}\right)} = 0, \tag{2.27}$$

the fixed points coalesce, and $\lambda_2 = 1$. For fixed physical parameters and $\bar{v} < \bar{v}_{SN}$, no fixed points exist, and periodic gaits are impossible; for a (possibly small) range of velocities $v > \bar{v}_{SN}$, a stable fixed point exists, corresponding to symmetric one-periodic gaits. See Figures 7 and 11.

Gaps. Increases in $\bar{v}$ and the consequent increases in the sweep angle $\Delta\psi_{\max}$ lead to a violation of (2.26), giving birth at a second critical speed $\bar{v} = \bar{v}_{GP}$ to a gap—an interior domain in which the map is not defined. With further increases in $\bar{v}$, the gap progressively expands to occupy a larger interval between the fixed points; see, e.g., Figure 8. Gaps may also appear in the range $\delta_n < 0$ for values of $\bar{v}$ small enough that $\Delta\psi_{\max} < \pi - 2\beta$, although these are of less physical importance, since sustained gaits do not exist in this range (below $\bar{v}_{SN}$). See the discussion of section 3.1 and Figure 7.

Period doubling. We recall expression (2.24) for the second eigenvalue of DP:

$$\lambda_2 = 1 - \left(1 + \frac{g\eta_0 \cos\beta \cot\delta_n}{\bar{v}^2}\right) \left.\frac{\partial \Delta\psi}{\partial \delta_n}\right|_{\delta_n \equiv \bar{\delta}}. \tag{2.28}$$

The quantity in parentheses is strictly positive for $\delta_n = \bar{\delta} \in (0, \beta)$ (symmetric periodic gaits), and $\frac{\partial \Delta\psi}{\partial \delta_n}|_{\delta_n = \bar{\delta}}$ is zero at the saddle-node bifurcation at $\bar{v} = \bar{v}_{SN}$ and thereafter positive at the stable fixed point of P. This suggests that, as the magnitude of $\frac{\partial \Delta\psi}{\partial \delta_n}|_{\delta_n = \bar{\delta}}$ increases with

increasing $\bar{v}$, λ_2 may pass through -1. For general (differentiable) maps, the instability arising from crossing the unit circle at $\lambda_2 = -1$ represents a loss of stability via period-doubling and the birth of a period two orbit [34].

Explicit computations are awkward due to the difficulty of evaluating the sweep angle quadrature (2.10), but we may estimate λ_2 and hence obtain a sufficient condition for period-doubling to occur at high velocities $\bar{v}$ by appealing to the limiting behavior of the $\Delta\psi$ as $\bar{v} \to \infty$. In the next section and in Appendix B, we estimate $\Delta\psi(\bar{v}, \delta_n)$ and the fixed-point location $\delta_n = \bar{\delta}$ in terms of the small parameter $\frac{1}{\bar{v}}$. This permits us to calculate $\frac{\partial \Delta\psi}{\partial \delta_n}|_{f.p.}$ in this limit, which in turn yields the estimate

$$\lambda_2 = -1 - \frac{\eta_0 \cos\beta\,[4mg + V'(\eta_0 \sin\beta)]}{\bar{v}\sqrt{2mV(\eta_0 \sin\beta)}} + \mathcal{O}\left(\frac{1}{\bar{v}^2}\right). \tag{2.29}$$

Since $V(\eta)$ is decreasing on the interval $(0, \eta_0)$ for physically reasonable spring laws, the condition $\lambda_2 = -1$ can indeed be met. Indeed, to guarantee it, bearing in mind the fact that for "low" $\bar{v} = \bar{v}_{SN}$, $\lambda_2 = +1$, it suffices to require $[4mg + V'(\eta_0 \sin\beta)] > 0$ so that λ_2 approaches -1 from below as $\bar{v} \to \infty$, having previously passed down through -1. Thus one would expect period-doubling to occur for relatively soft springs or touchdown angles close to 90^o, e.g., for $k\eta_0(1 - \sin\beta) < 4mg$ in the case of the linear spring of section 3.1. However, we recall that the approximate computation of the sweep angle employed in this section is carried out under the assumption that spring forces dominate gravitational effects, whereas (2.29) indicates that they should be comparable for period-doubling. Evidently, the true behavior of λ_2 depends in a subtle manner on the precise spring law and the other physical parameters.

Nonetheless, numerical evidence suggests that period-doubling does occur for reasonable parameter values and, moreover, that it can occur at relatively low speeds. This observation corrects the misleading claim in [22]: "Bistable solutions do not exist as only symmetric contact phases may result in a periodic movement pattern (Schwind, 1998)," and "More recently, Schwind (1998) showed that for a running spring-mass system only symmetric stance phases with respect to the vertical axis might result in cyclic movement trajectories."[2] Figure 9 shows an example of an attracting period two orbit born in such a bifurcation. Also see section 3.2. We remark that we have not found period-doubling for the Hooke spring with gravity in stance, since whenever we observe $\lambda_2 \leq -1$, the gap has already opened, which destroys any attracting behavior (see also Figure 3(A) of [22], where the gap opens at $\alpha_0 = 68.70^\circ$, whereas the slope of the left fixed point becomes -1 at $\alpha_0 = 68.85^\circ$).

We note that this behavior is markedly different from the LLS dynamics discussed in [17, 19], in which no flight phase occurs, and the bound $\frac{\partial \Delta\psi}{\partial \delta_n} < 2$ (see section 2.1.6) implies that period-doubling cannot occur.

2.1.6. The limiting case $\bar{v} \to \infty$. It was noted in [17] that there is a critical value $\bar{v}$ above which the touchdown kinetic energy exceeds the potential energy stored by a linear spring at zero length. When this happens, $\Delta\psi(\bar{v}, \delta_n)$ no longer has a quadratic shape but approaches

[2]The reader should note that the symmetry of orbits associated with period 1 return maps [36] has no bearing on the existence or properties of higher period discrete time behavior.

the straight line: $\Delta\psi = \pi - 2(\beta - \delta)$ as $\bar{v} \to \infty$. The unstable fixed point is lost, and the (previously stable) fixed point $\delta \to 0^+$, as shown in Figures 7 and 8. (As we shall see, this "change of type" does not occur for the air spring model, which has the physically desirable property that the spring force increases without bound as it is compressed to zero length.) However, for sufficiently large $\bar{v}$ and *any* spring law having bounded energy at nonzero length, kinetic energy dominates both gravitational and elastic energy at finite compression, and the COM follows an almost-straight "ballistic" horizontal path.

In this limit, the quadrature integral of (2.10) can be asymptotically estimated, as shown in Appendix B, leading to the following sweep angle expression:

$$\Delta\psi(\bar{v}, \delta_n) = (\pi - 2\beta + 2\delta_n) - \frac{1}{\bar{v}}\sqrt{\frac{2V(\eta_0 \sin(\beta - \delta_n))}{m}} + \mathcal{O}\left(\frac{1}{\bar{v}^2}\right). \tag{2.30}$$

This allows us to determine the limiting trajectory in physical space. Clearly the stance phase limits to a horizontal motion over the distance $2\eta_0 \cos\beta$ (the top of an inverted isosceles triangle). To compute the flight phase, we note that the fixed-point condition specifies $\Delta\psi(\bar{v}, \bar{\delta}) = \pi - 2\beta$. Calculating $\bar{\delta} \sim \frac{1}{\bar{v}}$ from the $\mathcal{O}(\frac{1}{\bar{v}})$ term of (2.30) (see Appendix B), we obtain an $\mathcal{O}(1)$ vertical component of liftoff velocity:

$$v^{LO}_{\text{vert}} = \bar{v}\sin\bar{\delta} \approx \sqrt{\frac{V(\eta_0 \sin\beta)}{2m}}\ ;\ v^{LO}_{\text{horiz}} = \bar{v}\cos\bar{\delta} \approx \bar{v}. \tag{2.31}$$

Hence the flight duration approaches a constant, and the flight distance grows linearly with $\bar{v}$. The limiting behavior is well defined, but resolution of the flight phase requires an $\mathcal{O}(\frac{1}{\bar{v}})$ calculation.

We note that (2.30) also shows that, as $\bar{v} \to \infty$, the sweep angle approaches the straight line $\Delta\psi = (\pi - 2\beta + 2\delta_n)$ from below, within its domain of definition; in fact, the $\mathcal{O}(\frac{1}{\bar{v}})$ correction to $\Delta\psi$ is the square root of the ratio of potential energy at midstance to kinetic energy at touchdown.

2.2. Gravitational effects during stance. We have argued that, for sufficiently stiff leg springs, elastic force dominates gravitational force during the stance phase. In this situation, their inclusion represents a small perturbation of the idealized case studied above. Order of magnitude estimates indicate that, for the mass and leg length chosen here, a relatively stiff spring (e.g., $k = 2000\frac{N}{m}$, $\gamma = 306$) is required to justify the neglect of g. Typical apex heights are one to two orders of magnitude larger than η_0 in this case. However, even with springs as soft as $k = 100\frac{N}{m}$ ($\gamma = 15.3$), chosen so that flight phase displacements are comparable to those in the stance phase, the hopper exhibits asymptotically stable gaits similar to those of the idealized case. Figure 4 shows four examples of COM trajectories in physical space. Also see Figure 2.

2.3. On pitching dynamics: $d \neq 0$. We have found numerical evidence of periodic gaits even when the leg is not attached at the COM so that the (freely pivoted) body pitches in response to the combined moments due to gravity and the leg-spring force, according to the last equation of (2.4). Figure 5 shows examples of symmetric 1:1 motions in which the pitching angle is periodic with the same (least) period as the COM translation dynamics; note that in

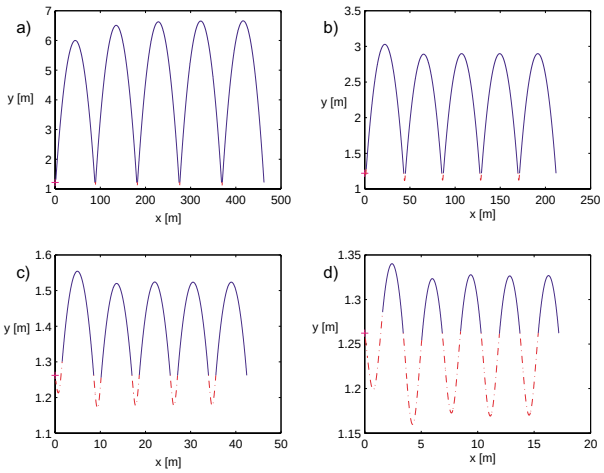


Figure 4. *Stable gaits with gravity included in the stance phase, showing the effect of spring stiffness.* $m = 1, \eta_0 = 1.5$, $k = 2000$ ($\gamma = 306$) (a); $k = 1000$ ($\gamma = 153$) (b); $k = 250$ ($\gamma = 38.2$) (c); *and* $k = 100$ ($\gamma = 15.3$) (d). *Here* $\beta = 0.95$ *for the upper graphs, and* $\beta = 1.0$ *for the lower ones.* v_n *was* $45, 35, 15, 8$, *respectively. Stance phases are shown chain dotted, and flight phases are shown solid; note the differing vertical scales.*

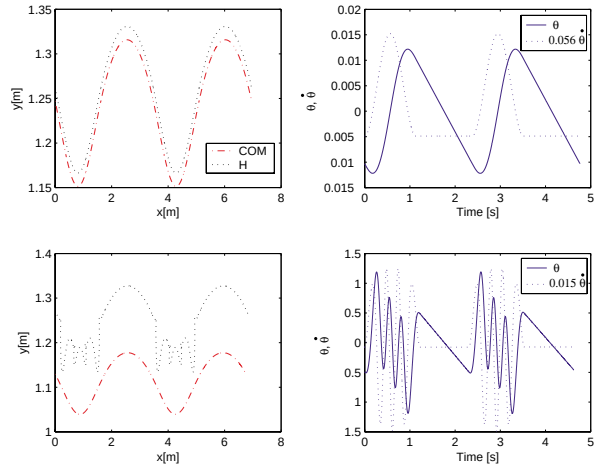


Figure 5. *Periodic gaits of the model with pitching included* ($d \neq 0$). *The left-hand panels show COM and hip paths in physical space, and the right-hand panels show pitch angle and angular velocity. Computed for parameter values* $k = 100, m = 1, \eta_0 = 1.5, \beta = 1$, *and* (a) $d = 0.015, I = 2.25 \times 10^{-2}, \bar{v} = 7.93$; (b) $d = 0.15, I = 2.25 \times 10^{-4}, \bar{v} = 8.26$. *Note that* $\dot{\theta}$ *scales differ in right-hand panels.*

(b) the pitch angle oscillates several times during each stance phase. We have also seen higher order resonances in which the pitching pattern repeats once every n strides and quasi-periodic motions in which the pitch angle remains bounded but is not precisely locked to the stride dynamics. We defer a detailed analysis of these "acrobatic" motions, which appear to include partially asymptotically stable orbits having three eigenvalues of modulus 1 and one inside the unit circle, to a future publication.

3. Two examples. In section 2.1.3, we discussed general conditions for stability, saddle-node bifurcations, period-doubling, and the appearance of a gap, and we classified the solutions in terms of $\Delta\psi_{max}$, the maximum leg angle swept during stance, assuming only that the function $\Delta\psi(v,\delta)$ has a unique maximum but without specifying any particular spring law. In the following section, we consider two specific and commonly used spring models: a linear Hooke's law spring and an air spring that mimics the compressed air strut used in certain hopping robots. Throughout this section, we employ the approximation of section 2.1, ignoring gravity during stance.

3.1. The Hooke's law spring. To further illustrate the four cases discussed in section 2.1 (Figure 3), we numerically evaluate the Poincaré map for a system with spring potential $V(\eta) = \frac{k}{2}(\eta - \eta_0)^2$ and parameters $k = 100, m = 1, \eta_0 = 1.5, \beta = 1.25$. As before, we employ increasing initial speeds $v_0 = 1.75, 3.5, 5, 6$, corresponding to cases (a) to (d), respectively; see Figure 6. A somewhat larger set of touchdown-to-touchdown Poincaré maps P_2 is shown in Figure 7, where we plot the second component $\delta_n \mapsto \delta_{n+1}$ implicit in (2.20).

Note that, as $\bar{v}$ increases, the map first intersects the identity $\delta_{n+1} = \delta_n$ at $\bar{v} = \bar{v}_{SN}$, and fixed points appear in a saddle-node bifurcation. We illustrate this in Figure 8(a) in the form of bifurcation diagrams [34], plotting $\bar{\delta}$ vs. $\bar{v}$. No qualitative changes with β are apparent; this is a general feature that will be discussed in further detail in section 4, where we also assess the effects of gravity in the stance phase. We note that the domain of attraction of the stable fixed point opens and grows following $\bar{v} = \bar{v}_{SN}$ until it is invaded by the gap; thereafter, it shrinks as $\bar{v}$ increases. Also note that the larger δ_n fixed point disappears at a finite speed $\bar{v} \approx 5.9$ due to the change of type of $\Delta\psi$ and the stance map when $\delta \to \beta$, and kinetic energy at touchdown exceeds the potential energy stored in the spring at zero length (cf. [17], and also see Figure 7).

We have also seen gaps in the domain of definition of P_2 for low velocities $\bar{v} < \bar{v}_{SN}$ (before the saddle-node), but these are of less concern since there are no sustained gaits in this range.

As noted in section 2.1.5, period-doubling bifurcations may occur as $\bar{v}$ increases, depending upon the spring potential and other parameters. Figure 9 shows an example of a period two gait born in such a bifurcation for a linear spring system.

3.2. An air spring. The four cases discussed in section 2.1.5 can also be illustrated with an air spring model. As noted above, we adopt the potential $\frac{c}{2}(\frac{1}{\eta^2} - \frac{1}{\eta_0^2})$. We compute orbits and Poincaré maps for a system with the parameters $c = 23, m = 1, \eta_0 = 1.5, \beta = 1.25$ and increasing initial speeds $v_0 = 1.75, 3.5, 5, 6$ shown as cases (a) to (d), respectively, in Figure 10; these should be compared with Figure 6. The corresponding Poincaré maps are shown in Figure 11 for comparison with Figure 7. For small values of speed $\bar{v}$, the map has no fixed points or periodic orbits, and, as for the linear spring, fixed points appear in a saddle-node bifurcation at a critical speed $\bar{v} = \bar{v}_{SN}$. Figure 8b shows a bifurcation diagram for the air spring hopper. For this spring law, which requires infinite energy and force for compression to zero length, no change of type occurs, and the upper, unstable branch of fixed points continues to arbitrarily high velocities.

3.2.1. Period-doubling and chaos. In section 2.1.5, we showed that period-doubling may occur as $\bar{v} \to \infty$. On the other hand, there is also a critical speed $\bar{v}_{GP}$ above which the return

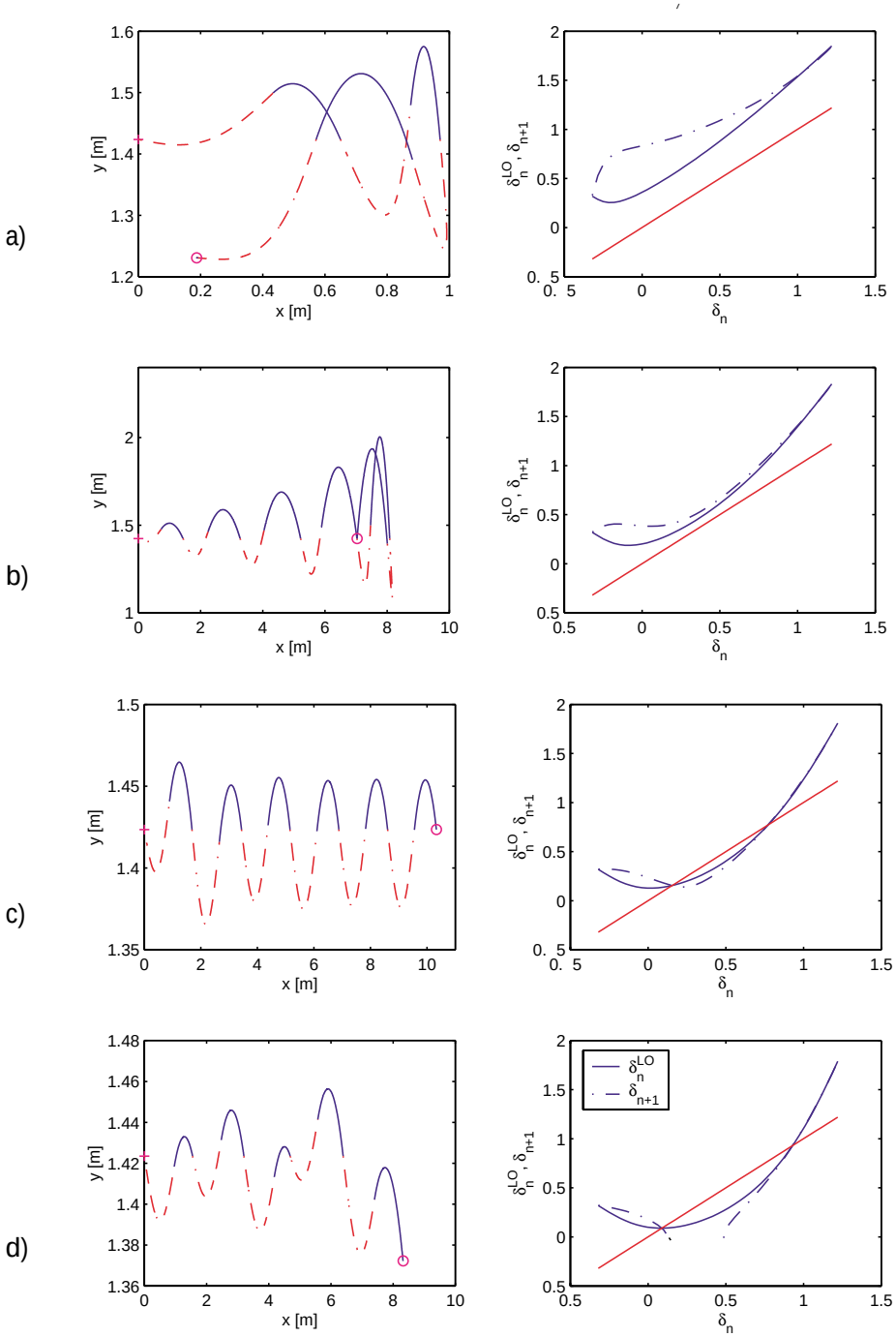


Figure 6. *Motions of the linear spring hopper in physical space (left column) and Poincaré maps (right column). Physical parameters k, m, η_0, β were taken as for Figure 3. Trajectories in physical space were computed from initial condition $\delta_0 = 0.1$, and initial speed was increased from $v_0 = 1.75$ to $3.5, 5$, and 6 for cases* (a) *to* (d), *respectively. Touchdown height is shown by dotted horizontal lines. Poincaré maps were computed for the same speeds. Fixed points occur at intersections of the curves and the line $\delta_n = \delta_{n+1}$. Both the stance map P_{st} (solid) and the full map P (dotted) are shown. In* (d), *note the gap in which the full map is not defined.*

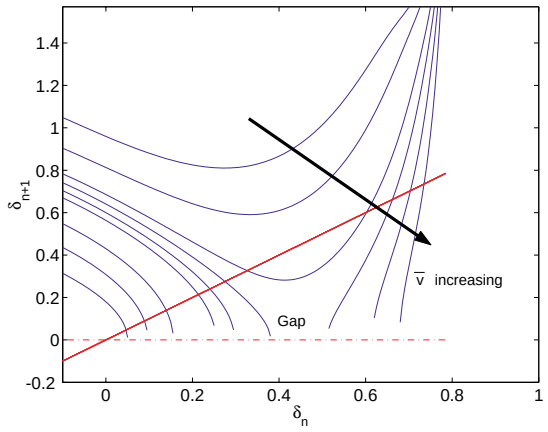


Figure 7. *The Poincaré map P_2 for a linear spring hopper with $k = 10, m = 1, \eta_0 = 1.5, \beta = \pi/4$, and speeds $\bar{v}$ ranging from 3.2 to 8. Note how the two fixed points appear in a saddle-node bifurcation, and a gap then opens as $\bar{v}$ increases. For very high speeds, only one fixed point exists.*

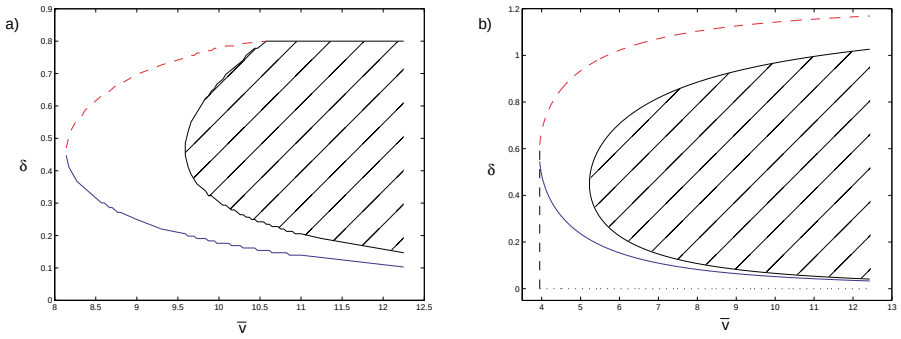


Figure 8. *Bifurcation diagrams for the linear spring hopper with $m = 1, k = 50, \eta_0 = 1.5$, and touchdown angle $\beta = 0.8$ in* (a) *and for the air spring hopper with $m = 1, c = 23, \eta_0 = 1.5$, and $\beta = 1.25$ in* (b)*. Stable branches of fixed points are shown solid, unstable branches are dashed, and cross-hatching identifies the region in which the map is not defined. Saddle-node bifurcations occur at $\bar{v}_{SN} = 8.12$ in* (a) *and $\bar{v}_{SN} = 3.95$ in* (b)*; below these no periodic gaits exist.*

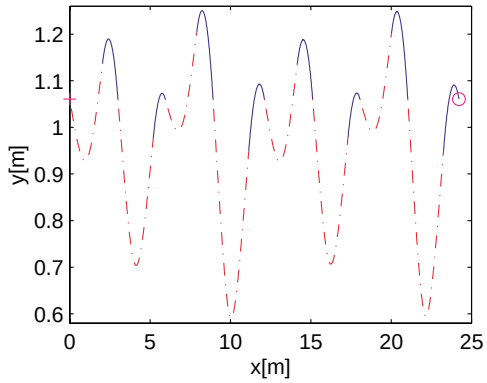


Figure 9. *A period two gait of the linear spring hopper with $k = 10, m = 1, \eta_0 = 1.5, \beta = \pi/4$, and $\bar{v} = 3.95$.*

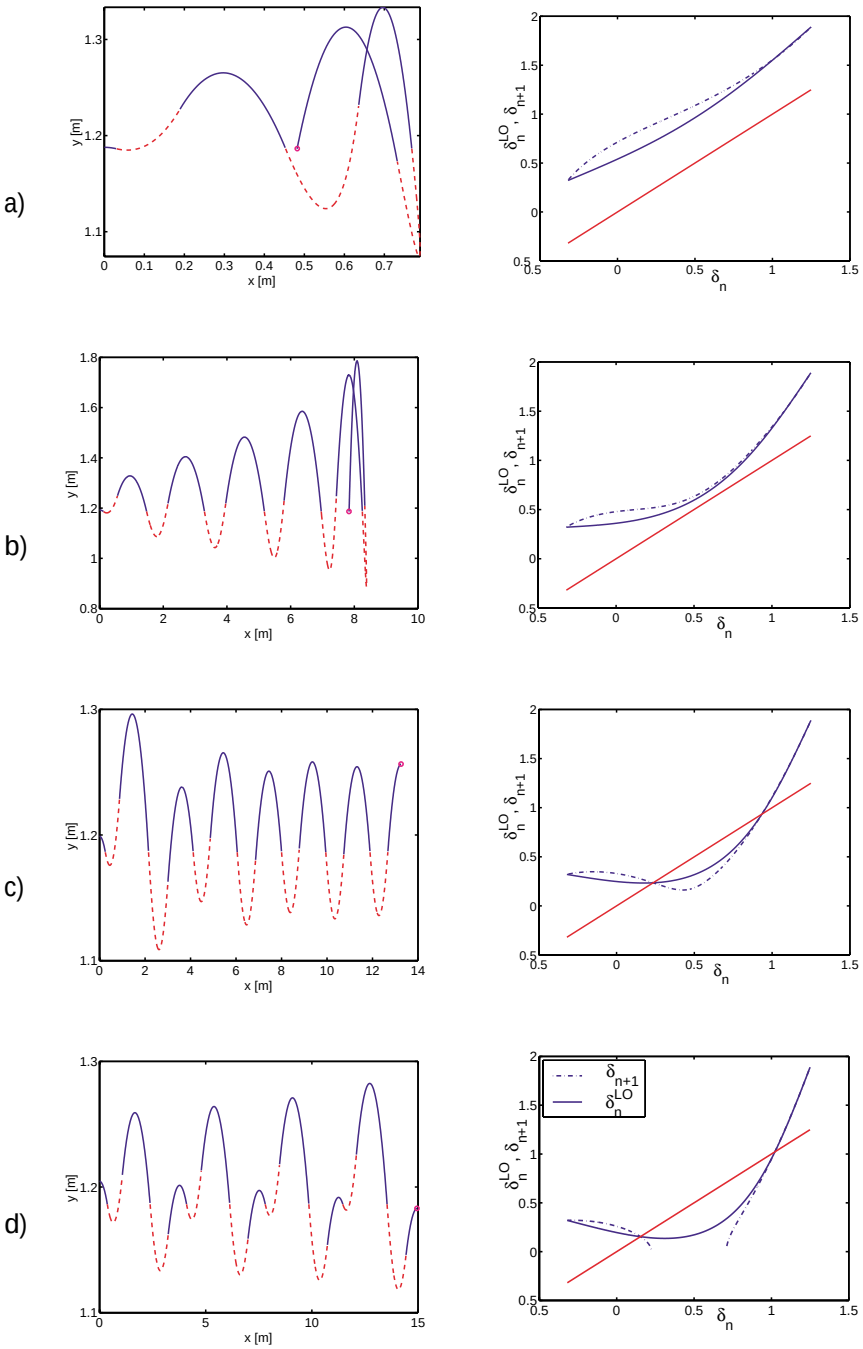


Figure 10. *Physical space motions (left column) and Poincaré maps (right column) for the air spring hopper. Parameter values were chosen as for Figure 3 except for the air spring stiffness $c = 23$ and leg length $\zeta_0 = 1.25$. The physical space trajectories were computed from initial condition $\delta_0 = 0.1$ and initial speeds from $v_0 = 1.75$ to $3.5, 5$, and 6 for cases* (a) *to* (d)*, respectively. Touchdown height is shown by dotted horizontal lines. The maps were computed for the same speeds and angles $\delta \in [0, \beta]$. Fixed points are identified with the intersection of the curves and the line $\delta_n = \delta_{n+1}$. Both the stance map P_{st} (solid) and the full map P (dash-dotted) are shown. The region in which the full map is not defined is apparent in* (d).

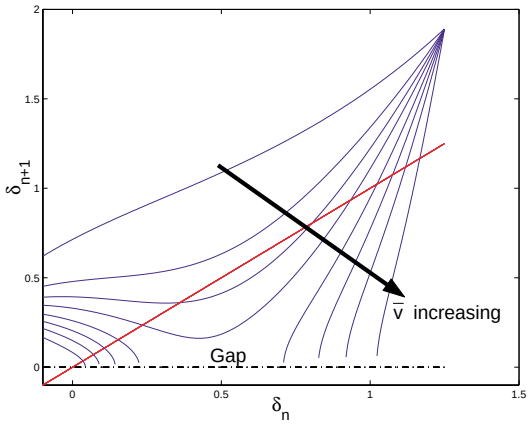
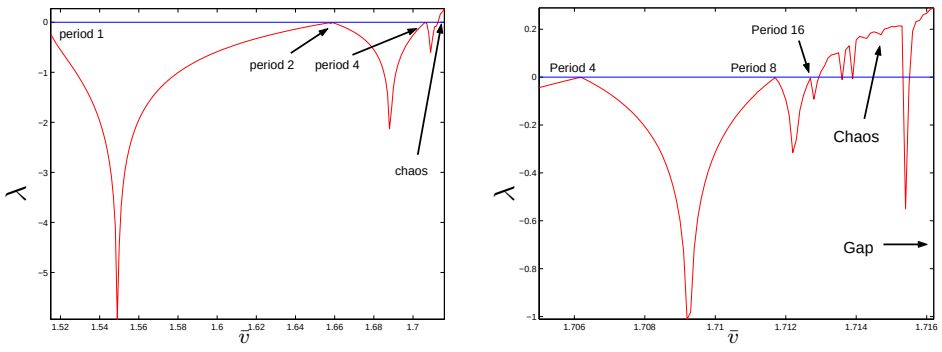


Figure 11. *The Poincaré map P_2 for an air spring hopper with $c = 23, m = 1$, and $\zeta_0 = 1.25$. The speed $\bar{v}$ ranges from 1.75 to 12.2. Note the gap and the fact that the upper (larger $\bar{\delta}$) fixed point persists.*



(a) (b)

Figure 12. *Lyapunov exponents for $\beta = 0.961$, $c/m = 0.01839$, $g = 9.81$, and $\eta_0 = 0.173$. The speed range where the exponent becomes positive is magnified in* (b).

map is not defined over the whole range $\delta_n \in [0, \beta]$. The question then arises whether the gap always opens *before* period-doubling occurs or whether period two and higher period orbits or even chaotic behavior is observed for gap-free return maps. This is not only of theoretical importance; the onset of higher period orbits and chaotic behavior for gap-free return maps would place additional constraints on feedforward control policies that simply keep the leg touchdown angle at $\beta = \text{const}$ [28].

To identify period-2^n orbits and chaos, we numerically approximate the Lyapunov exponent [34] λ of the one-dimensional return map $P_2 : \delta_n \mapsto \delta_{n+1}(\delta_n)$, implicitly defined by the

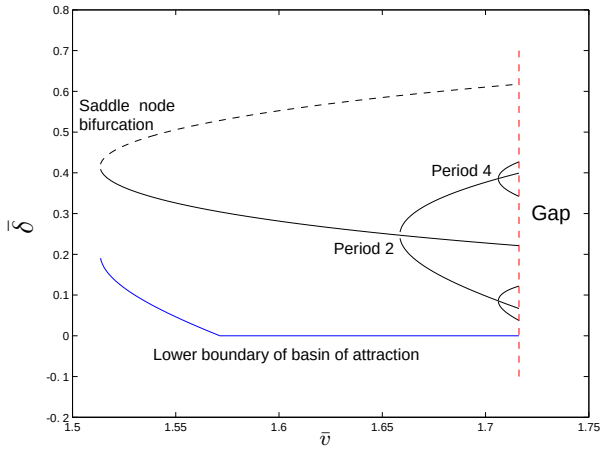


Figure 13. *Bifurcating fixed points for $\beta = 0.961$, $c/m = 0.01839$, $g = 9.81$, and $\eta_0 = 0.173$. The region between the lower blue line and the upper dashed line is the basin of attraction.*

second component of (2.20):

$$\lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \ln \left| \frac{dP_2}{d\delta_n}(\delta_i) \right| =: \lim_{N\to\infty} \lambda_N. \tag{3.1}$$

Specifically, we set

$$\lambda \approx \begin{cases} \lambda_K & \text{if} \quad |\lambda_K - \lambda_{K-1}| < \varepsilon \text{ and } K < \bar{N}, \\ \lambda_{\bar{N}} & \text{else} \quad (\text{with } \bar{N} = 10000 \text{ and } \varepsilon = 10^{-6}) \end{cases} \tag{3.2}$$

and take a range of leg touchdown speeds $\bar{v} \in [1.515, 7.162]$; other parameters are $\beta = 0.961$, $c/m = 0.01839$, $g = 9.81$, and $\eta_0 = 0.173$. In this case, the return map develops a gap at $\bar{v}_{GP} = 1.7162$.

However, it can be seen from Figure 12(a) that λ crosses to positive values at $\bar{v}_{Chaos} \approx 1.713$ (magnified in Figure 12(b)). Similar behavior has been observed for the air spring potential *with* gravity in stance for parameter values $\beta = 0.961$, $c/m = 0.03839$, $g = 9.81$, and $\eta_0 = 0.17$. Here, the Lyapunov exponent becomes positive at $\bar{v}_{Chaos} = 1.6932$, whereas the gap opens at $\bar{v}_{GP} = 1.697$. This is strong numerical evidence for chaotic behavior in the corresponding SLIP. Chaotic behavior has not been observed for the linear spring.

In Figure 13, the corresponding bifurcating branches of fixed points are plotted as a function of the touchdown speed $\bar{v}$ up to the gap at $\bar{v}_{GP}$. Bifurcations of period eight and higher are omitted. The lower boundary of attraction is also shown; this deviates from $\delta_n = 0$ whenever $\delta_{n+1}(0) > \delta_{n+1}(\bar{\delta}_1^{us})$, where $\bar{\delta}_1^{us}$ denotes the unstable fixed point of the period one bifurcation.

4. Nondimensional parameter studies. Parameters intrinsic to the model are the mass m of the body, the moment of inertia I, the gravitational acceleration g, the uncompressed leg length η_0, the leg touchdown angle β, the distance d from hip to the COM, and the stiffnesses k or c. These, together with initial conditions v_0, δ_0, provide a variety of solutions

with different stance and flight phases and basins of attraction. The seven physical parameters can be reduced to a minimal set of *nondimensional* parameters necessary to characterize the model. Rescaling time and lengths by defining $\tilde{t} = \frac{t}{t_0}, \tilde{\zeta} = \frac{\zeta}{\eta_0}, \tilde{d} = \frac{d}{\eta_0}$, and $\tilde{\eta} = \frac{\eta}{\eta_0}$, we can rewrite the equations of motion (2.4) as

$$\begin{aligned}
\tilde{\zeta}'' &= \psi'^2\tilde{\zeta} - \frac{gt_0^2}{\eta_0}\cos\psi - \frac{V_{\tilde{\zeta}}(\eta_0\tilde{\eta})t_0^2}{m\eta_0^2\tilde{\eta}}(\tilde{\zeta} + \tilde{d}\cos(\psi+\theta)),\\
\tilde{\zeta}\psi'' &= -2\psi'\tilde{\zeta}' + \frac{gt_0^2}{\eta_0}\sin\psi + \tilde{d}\frac{V_{\tilde{\zeta}}(\eta_0\tilde{\eta})t_0^2}{m\eta_0^2\tilde{\eta}}\sin(\psi+\theta),\\
\theta'' &= \tilde{d}\frac{V_{\tilde{\zeta}}(\eta_0\tilde{\eta})\,\tilde{\zeta}t_0^2}{I\tilde{\eta}}\sin(\psi+\theta),
\end{aligned} \tag{4.1}$$

where the differentiation $()' \equiv \frac{d}{d\tilde{t}}$ is with respect to the nondimensional time $\tilde{t}$, and $V_{\tilde{\zeta}}(\tilde{\eta}) = V_{\tilde{\eta}}(\tilde{\eta}) = \frac{\partial}{\partial\tilde{\eta}}V(\tilde{\eta})$. It seems physically reasonable to define the characteristic time $t_0 = \frac{\eta_0}{v_0}$, where v_0 is a characteristic speed, such as the COM speed at touchdown, and η_0 is the uncompressed length of the leg spring.

4.1. Hooke's law spring. If we assume a linear spring with $V(\eta_0\tilde{\eta}) = \frac{k\eta_0^2}{2}(\tilde{\eta}-1)^2$ and define the nondimensional parameter groups

$$\tilde{k} \triangleq \frac{kt_0^2}{m} = \frac{k\eta_0^2}{mv_0^2},\ \tilde{g} \triangleq \frac{gt_0^2}{\eta_0} = \frac{g\eta_0}{v_0^2},\ \text{and}\ \tilde{I} \triangleq \frac{I}{m\eta_0^2}, \tag{4.2}$$

the equations of motion, expressed in nondimensional coordinates, become

$$\begin{aligned}
\tilde{\zeta}'' &= \psi'^2\tilde{\zeta} - \tilde{g}\cos\psi - \tilde{k}\left(1-\frac{1}{\tilde{\eta}}\right)(\tilde{\zeta} + \tilde{d}\cos(\psi+\theta)),\\
\tilde{\zeta}\psi'' &= -2\psi'\tilde{\zeta}' + \tilde{g}\sin\psi + \tilde{k}\tilde{d}\left(1-\frac{1}{\tilde{\eta}}\right)\sin(\psi+\theta),\\
\theta'' &= \frac{\tilde{k}\tilde{d}\tilde{\zeta}}{\tilde{I}}\left(1-\frac{1}{\tilde{\eta}}\right)\sin(\psi+\theta).
\end{aligned} \tag{4.3}$$

Here the parameter $\tilde{k} = \frac{k\eta_0^2}{mv_0^2} = \frac{E_{spr}}{E_{kin}}$ expresses the ratio between the potential energy storable by the spring at maximum compression (i.e., to zero length) and the touchdown kinetic energy, whereas $\tilde{g} = \frac{g\eta_0}{v_0^2}$, a Froude number, expresses the ratio of gravitational energy to kinetic energy. Note also that the ratio $\frac{\tilde{k}}{\tilde{g}} = \frac{k\eta_0}{mg} \triangleq \gamma$ is fixed for a given physical system and is independent of initial conditions and, in particular, of the characteristic speed. Seven physical parameters $m, I, d, \eta_0, \beta, g, k$ have been reduced to five: $\tilde{I}, \tilde{d}, \beta, \tilde{g}, \tilde{k}$. In the special case of the hip attached at the COM $\tilde{d} = 0, \theta = \text{const}$ and only three parameters play a role: $\beta, \tilde{g}, \tilde{k}$. This facilitates a parametric analysis of the system. Since β does not appear to change the qualitative behavior of the solutions of (4.3), we represent the "sheets" of periodic solutions in $(\tilde{k}, \tilde{g}, \bar{\delta})$-space.

Since we wish to assess the influence of gravity via $\tilde{g}$, here and for the air spring calculations below, we *include* gravity in the stance phase and make our fixed-point computations numerically.

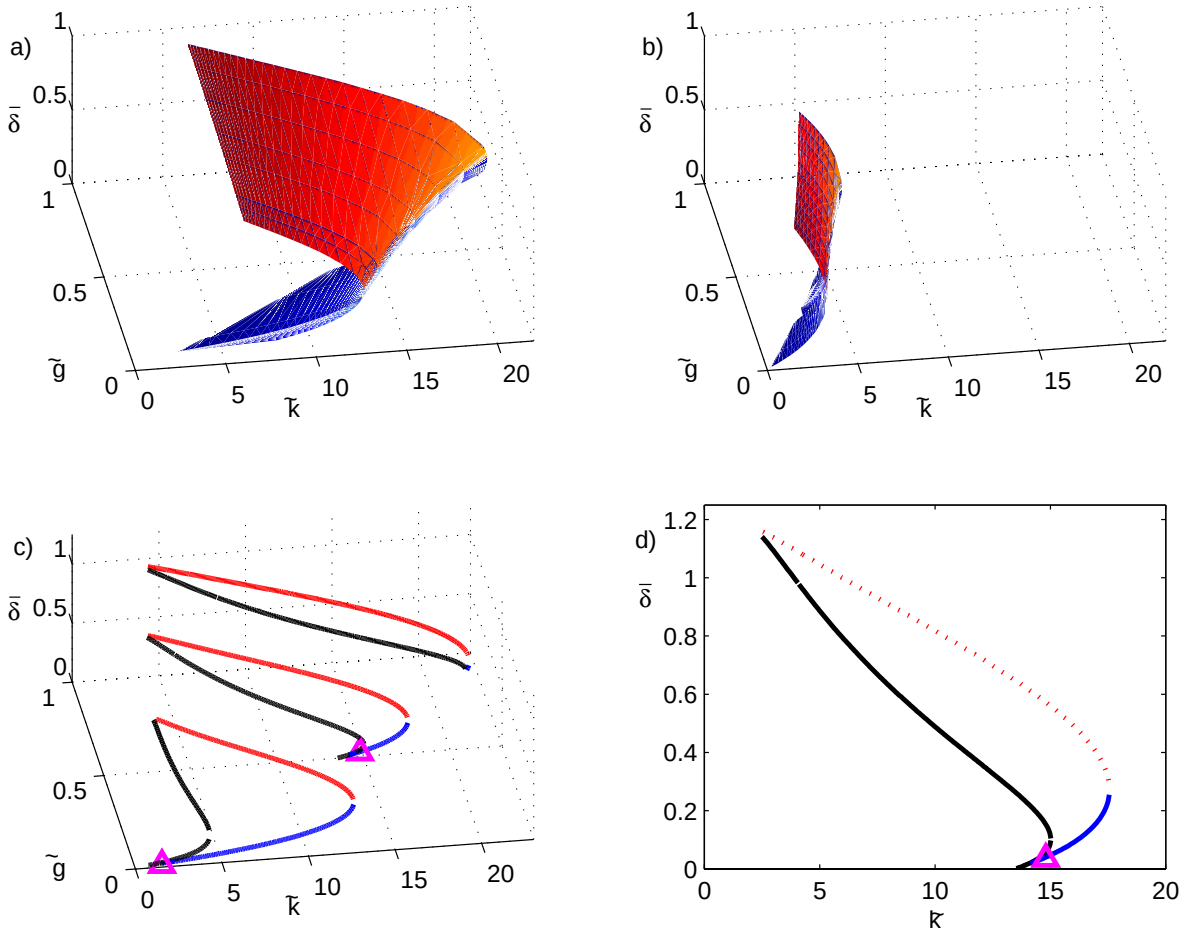


Figure 14. *Bifurcation diagrams for the linear spring in nondimensional $(\tilde{k}, \tilde{g}, \bar{\delta})$-parameter space. Upper panels show the cases $\beta = 1.25$ (a) and $\beta = 1$ (b), respectively; lower panels show three bifurcation diagrams (cross sections of* (a)*) for $\beta = 1.25$ (c) and a single bifurcation diagram for $\beta = 1.25$ and $\tilde{g} = 0.5$ (d). Unstable branches are shown dashed, stable branches are shown solid, period-doubling bifurcation points are indicated by triangles, and boundaries of the gap are indicated by thick black curves.*

Figures 14(a)–(b) show how the stable and unstable branches of the fixed point $\delta_n = \bar{\delta}$ over $(\tilde{k}, \tilde{g})$-space change as β varies. The general shape of the surface of equilibria is preserved, although the influence of $\tilde{g}$ on the saddle-node location $\tilde{k}_{SN}$ lessens as β decreases and $\tilde{k}_{SN}$ itself decreases, corresponding to higher velocities. Also, for fixed β, increases in $\tilde{g}$ cause the lower (stable) branch to shrink until it disappears so that when gravity plays a dominant role (low speed and/or long leg), there is only one *unstable* fixed point; cf. Figure 14(c) with $\tilde{g} \approx 1$. It can also happen, as noted in section 3.1 (Figure 8(a)), that the upper branch terminates and only one (potentially) stable fixed point exists, e.g., near $\beta = 1.25, \tilde{k} \approx 1, \tilde{g} \approx 0.01$ in Figure 14(c). Increasing β has the effect of expanding the domain of attraction both in the $\tilde{k}$ and $\bar{\delta}$ directions. This suggests a choice of high ratios γ (e.g., relatively hard springs) and high values of β in order to maximize the domains of attraction of the stable fixed points.

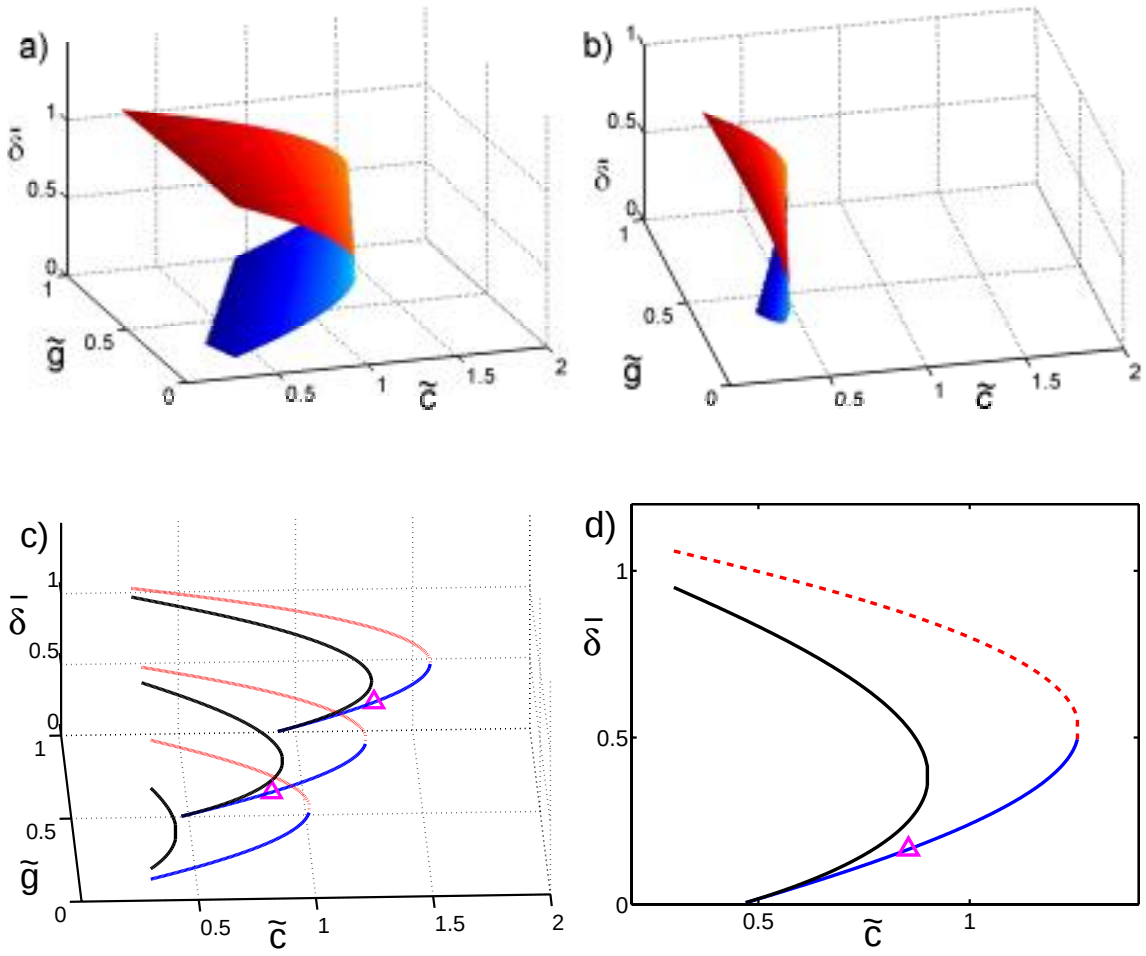


Figure 15. *Bifurcation diagrams for the air spring in nondimensional $(\tilde{c}, \tilde{g}, \bar{\delta})$-parameter space. Upper panels show the cases $\beta = 1.25$ (a) and $\beta = 1$ (b), respectively; lower panels show three bifurcation diagrams (cross sections of* (a)*) for $\beta = 1.25$ (c) and a single bifurcation diagram for $\beta = 1.25$ and $\tilde{g} = 0.5$ (d). Unstable branches are shown dashed, stable branches are shown solid, period-doubling bifurcation points are indicated by triangles, and boundaries of the gap are indicated by thick black curves.*

We note that, for fixed $(\tilde{k}, \tilde{g}, \beta)$, the saddle-node and other bifurcations occur at particular velocities $\bar{v}_{SN}$, etc. It therefore follows from (4.2) that, as k and/or m vary and other physical parameters remain fixed,

$$\bar{v}_{SN} = \eta_0 \sqrt{\frac{k}{m\tilde{k}_{SN}}} \propto \sqrt{\frac{k}{m}}. \tag{4.4}$$

This is presumably the type of scaling mentioned in [22, section 4.1]. See [19] for extensive analyses of this type for the LLS model.

4.2. Air spring. If we assume an air spring with $V(\eta_0\tilde{\eta}) = \frac{c}{2\eta_0^2}(\frac{1}{\tilde{\eta}^2} - 1)$ and define the nondimensional parameter groups

$$\tilde{c} \triangleq \frac{ct_0^2}{m\eta_0^4} = \frac{c}{mv_0^2\eta_0^2}\ ,\ \tilde{g} \triangleq \frac{gt_0^2}{\eta_0} = \frac{g\eta_0}{v_0^2},\ \text{and}\ \tilde{I} \triangleq \frac{I}{m\eta_0^2}, \tag{4.5}$$

the equations of motion, expressed in nondimensional coordinates, become

$$\begin{aligned}
\tilde{\zeta}'' &= \psi'^2\tilde{\zeta} - \tilde{g}\cos\psi + \frac{\tilde{c}}{\tilde{\eta}^4}(\tilde{\zeta} + \tilde{d}\cos(\psi+\theta)),\\
\tilde{\zeta}\psi'' &= -2\psi'\tilde{\zeta}' + \tilde{g}\sin\psi - \frac{\tilde{c}}{\tilde{\eta}^4}\tilde{d}\sin(\psi+\theta)\,,\\
\theta'' &= -\frac{\tilde{c}\tilde{d}\tilde{\zeta}}{\tilde{\eta}^4\tilde{I}}\sin(\psi+\theta)\,.
\end{aligned} \tag{4.6}$$

Note that with these choices, $\tilde{c} = \frac{c}{m\eta_0^2v_0^2} = \frac{E_{spr}}{E_{kin}}$ expresses the ratio between the potential energy stored at infinite spring length and the kinetic energy, whereas $\tilde{g} = \frac{g\eta_0}{v_0^2}$ is again the Froude number. Note also that the ratio $\frac{\tilde{c}}{\tilde{g}} = \frac{c}{mg\eta_0^3} \triangleq \mathring{\gamma}$ is fixed for a given physical system and, like γ above, is independent of initial conditions and of the characteristic speed. Again, seven physical parameters $m, I, d, \eta_0, \beta, g, c$ have been reduced to five: $\tilde{I}, \tilde{d}, \beta, \tilde{g}, \tilde{c}$. In the special case of the hip attached at the COM $\tilde{d} = 0, \theta = \text{const}$ and only three parameters play a role: $\beta, \tilde{g}, \tilde{c}$. The resulting surface plots are generally similar to those of Figure 14 for the linear spring, but they reveal that stable branches persist for large $\tilde{g}$ and that period-doubling occurs "earlier" (for higher $\tilde{c}$ and hence lower $\bar{v}$); see Figure 15.

5. Conclusions. In this paper, we studied the SLIP model of a hopping rigid body in the vertical plane. Exploring suitable limiting cases, we proved the existence of asymptotically stable periodic gaits for a fixed leg-angle (feedforward) touchdown protocol by studying the touchdown-to-touchdown Poincaré map in its exact form and under the approximation that gravity is negligible during stance. Numerical simulations including gravitational effects corroborated this result, revealing regions in the parameter and phase spaces where stable gaits exist. We considered two representative spring laws: a linear spring and a hardening air spring, and we studied bifurcations from the branch of stable gaits, the domains of attraction of those gaits, and the domains of definition of the Poincaré map, picking parameter values appropriate to illustrating key behaviors rather than for comparison with specific animals or machines. Throughout we focused on the classical SLIP, but our formulation includes full rigid body dynamics in the sagittal plane, and we displayed some coupled translation and pitching motions. Future work will include a broader analysis of these aspects.

As noted in the introduction, our work complements the study of [22], which addresses parameter ranges relevant to human running. Using direct numerical solution of the point mass SLIP equations (equivalent to the first two of (2.6)), [22] identifies parameters for which potentially stable period one gaits exist and shows that models with masses, leg lengths, and stiffnesses estimated from human data fall within a narrow range [22, Figure 2(A)]. A set of apex height Poincaré maps and some COM trajectories are also shown [22, Figures 3(A–B)].

It is noted that there is a minimum speed below which periodic gaits do not exist, that "larger variations in leg stiffness and angle of attack are tolerated [for] increasing speed," and that "higher ... velocities require either higher leg stiffness assuming constant angle of attack, or flatter angles of attack for constant leg stiffness" [22, Figures 2(B–C)].

We believe that the present analytical work, with the associated limiting integrable limiting cases, illuminates those observations. Specifically, our bifurcation studies reveal limits to stable parameter ranges bounded by saddle-node and period-doubling bifurcations, the former being responsible for [22]'s minimum speed requirement; our nondimensional analysis shows clear speed/stiffness relations (e.g., (4.4) for the linear spring) and reveals the relative importance of elastic and gravitational effects; and our study of gaps in the domain of definition of P ([22]'s apex/touchdown height constraint) shows that, while stable fixed points or higher period orbits continue to high velocity, their domains of attraction become extremely small. This shows that, with increasing speed, the system is *less* tolerant to dynamical perturbations, even though *parameter* variations are less restricted, as observed in [22] (see also Figure 2(b) in [25]). However, the bifurcation diagrams of Figures 8 and 14 show that, if the nondimensional parameters are maintained in a "good" location (e.g., between the saddle-node and gap of Figure 8) as $\bar{v}$ changes, by suitable tuning of stiffness or touchdown angle, then robust stability can be achieved with simple fixed-leg reset control. This viewpoint is guiding our development of "low attention" feedback controllers, to be described in a future paper [28].

Appendix A. Air spring sweep angle. The sweep angle $\Delta\psi\,(v_n, \delta_n)$ can be obtained explicitly in terms of elementary functions for an air spring model with $V\,(\eta) = \frac{c}{2}(\frac{1}{\eta^2} - \frac{1}{\eta_0^2})$. For $d = 0$ (hip attached to the COM), $\eta \equiv \zeta$. Solving first for the midstance compressed length using conservation of energy and angular momentum, we obtain

$$\frac{1}{2}mv_0^2 = \frac{1}{2}\frac{p_\psi^2}{m\zeta_{\min}^2} + \frac{c}{2}\left(\frac{1}{\zeta_{\min}^2} - \frac{1}{\zeta_0^2}\right) \Rightarrow \zeta_{\min} = \zeta_0\sqrt{\frac{p_\psi^2 + mc}{m^2v_n^2\zeta_0^2 + mc}}. \tag{A.1}$$

Using this in the quadrature (2.10) with $v_n = v_0$ and $p_\psi = m\zeta_0 v_0 \sin(\beta - \delta_n)$, the sweep angle may be computed as

$$\Delta\psi(v_0, \delta_n) = \frac{2p_\psi}{\sqrt{mc + p_\psi^2}} \arccos\left(\sqrt{\frac{p_\psi^2 + mc}{m^2v_0^2\zeta_0^2 + mc}}\right). \tag{A.2}$$

Finally, redefining $\tilde{c} = \frac{c}{m\zeta_0^2v_0^2}$, $z = \sin(\beta - \delta_n)$, $\hat{c} = \frac{\tilde{c}}{1+\tilde{c}}$, and $\hat{z} = \sqrt{\frac{z^2+\tilde{c}}{1+\tilde{c}}}$, $\Delta\psi$ can be rewritten as

$$\Delta\psi(v_0, \delta_n) = 2\frac{z}{\sqrt{1+\tilde{c}}}\frac{\arccos\left(\sqrt{\frac{z^2+\tilde{c}}{1+\tilde{c}}}\right)}{\sqrt{\frac{z^2+\tilde{c}}{1+\tilde{c}}}} = 2\sqrt{1 - \frac{\hat{c}}{\hat{z}^2}}\arccos(\hat{z}). \tag{A.3}$$

Appendix B. Limiting behavior as $\bar{v} \to \infty$. We recall the quadrature integral (2.10)

$$\Delta\psi = 2\eta_0 \sin(\beta - \delta_n)\int_{\zeta_b}^{\eta_0} \frac{d\zeta}{\zeta\sqrt{\zeta^2 - \frac{2\zeta^2V(\zeta)}{m\bar{v}^2} - \eta_0^2\sin^2(\beta - \delta_n)}}, \tag{B.1}$$

where ζ_b is the compressed leg length at midstance, satisfying the energy balance

$$\left[1 - \frac{2V(\zeta_b)}{m\bar{v}^2}\right]\zeta_b^2 = \eta_0^2 \sin^2(\beta - \delta_n). \tag{B.2}$$

(Note that we do not (yet) assume the fixed-point condition $\delta_n = \bar{\delta}$.) Under the standing assumption of analyticity of V in $(0, \eta_0)$, we first estimate the relevant solution of (B.2) and then examine the asymptotics of $\Delta\psi$ as $\bar{v} \to \infty$.

We introduce the small (dimensional) parameter $\epsilon = \frac{1}{\bar{v}^2}$ and expand solutions of (B.2) in the form $\zeta_b \approx \zeta_{b_0} + \epsilon\zeta_{b_1} + \cdots$, obtaining

$$\zeta_{b_0} = \eta_0 \sin(\beta - \delta_n) \ , \quad \zeta_{b_1} = \frac{\zeta_{b_0} V(\zeta_{b_0})}{m}. \tag{B.3}$$

The integral (B.1) may then be written

$$\begin{aligned} \Delta\psi &\approx 2\zeta_{b_0} \int_{\zeta_{b_0}+\epsilon\zeta_{b_1}}^{\eta_0} \frac{d\zeta}{\zeta\sqrt{\zeta^2\left(1 - 2\epsilon\frac{V(\zeta)}{m}\right) - \zeta_{b_0}^2}} \\ &\approx 2\zeta_{b_0} \int_{\zeta_{b_0}+\epsilon\zeta_{b_1}}^{\eta_0} \frac{d\zeta}{\zeta\sqrt{\zeta^2 - \zeta_{b_0}^2}} + \frac{2\epsilon\zeta_{b_0}}{m} \int_{\zeta_{b_0}+\epsilon\zeta_{b_1}}^{\eta_0} \frac{\zeta V(\zeta) d\zeta}{(\zeta^2 - \zeta_{b_0}^2)^{\frac{3}{2}}}. \end{aligned} \tag{B.4}$$

The first integral of (B.4) yields leading terms of $\mathcal{O}(1)$ and $\mathcal{O}(\sqrt{\epsilon}\,)$

$$2\left[\arccos(\sin(\beta - \delta_n)) - \arccos\left(\frac{\zeta_{b_0}}{\zeta_{b_0} + \epsilon\zeta_{b_1}}\right)\right] \approx \pi - 2\beta + 2\delta_n - 2\sqrt{\frac{2\epsilon\zeta_{b_1}}{\zeta_{b_0}}}, \tag{B.5}$$

where the $\mathcal{O}(\sqrt{\epsilon}\,)$ term is computed by setting $\arccos\left(\frac{\zeta_{b_0}}{\zeta_{b_0}+\epsilon\zeta_{b_1}}\right) \approx k_0 + \epsilon^\alpha k_1$ and solving for k_0, k_1, and α.

Integrating the second term of (B.4) twice by parts, we obtain

$$\begin{aligned} &\frac{2\epsilon\zeta_{b_0}}{m}\left(\left[\frac{-V(\zeta)}{\sqrt{\zeta^2 - \zeta_{b_0}^2}} + V'(\zeta)\ln\left(\zeta + \sqrt{\zeta^2 - \zeta_{b_0}^2}\right)\right]\Bigg|_{\zeta_{b_0}+\epsilon\zeta_{b_1}}^{\eta_0}\right. \\ &\qquad \left. - \int_{\zeta_{b_0}+\epsilon\zeta_{b_1}}^{\eta_0} V''(\zeta)\ln\left(\zeta + \sqrt{\zeta^2 - \zeta_{b_0}^2}\right) d\zeta\right). \end{aligned} \tag{B.6}$$

The upper limit of the first term and both logarithmic boundary terms give contributions of $\mathcal{O}(\epsilon)$. Successive integrations by parts of the third term produce a convergent series [42], so it is also of $\mathcal{O}(\epsilon)$. To obtain the dominant $\mathcal{O}(\sqrt{\epsilon}\,)$ contribution, we therefore need only include the lower limit of the first term:

$$\frac{2\epsilon\zeta_{b_0}}{m}\frac{V(\zeta_{b_0} + \epsilon\zeta_{b_1})}{\sqrt{2\epsilon\zeta_{b_0}\zeta_{b_1} + \epsilon^2\zeta_{b_1}^2}} \approx \sqrt{\frac{2\epsilon\zeta_{b_0}}{\zeta_{b_1}}}\,\frac{V(\zeta_{b_0})}{m}. \tag{B.7}$$

Combining (B.5) and (B.7) and using (B.3), we therefore obtain

$$\Delta\psi = (\pi - 2\beta + 2\delta_n) - \sqrt{\frac{2\epsilon\zeta_{b_1}}{\zeta_{b_0}}} + \mathcal{O}(\epsilon). \tag{B.8}$$

We note that $\Delta\psi \to (\pi - 2\beta + 2\delta_n)$ from below as $\epsilon \to 0$. Substituting the expressions (B.3) once more, (B.8) yields (2.30).

The fixed-point condition $\delta_n = \bar{\delta}$ requires that $\Delta\psi\left(\bar{\delta}, v_n\right) = \pi - 2\beta$, implying

$$\bar{\delta} \approx \sqrt{\frac{\epsilon\zeta_{b_1}}{2\zeta_{b_0}}} = \sqrt{\frac{\epsilon V(\zeta_{b_0})}{2m}}. \tag{B.9}$$

Now (B.9) defines $\bar{\delta}$ only implicitly since $\zeta_{b_0} = \eta_0 \sin(\beta - \bar{\delta})$, but setting $\zeta_{b_0} \approx \eta_0 \left(\sin\beta - \bar{\delta}\cos\beta\right)$ for $\bar{\delta} = \mathcal{O}(\sqrt{\epsilon})$ and expanding $V(\zeta_{b_0}) \approx V(\eta_0 \sin\beta) - V'(\eta_0 \sin\beta)\bar{\delta}\cos\beta$, we obtain

$$\bar{\delta} = \sqrt{\frac{V(\eta_0 \sin\beta)}{2m\bar{v}^2}} + \mathcal{O}\left(\frac{1}{\bar{v}^2}\right). \tag{B.10}$$

Thus $\bar{\delta} \to 0^+$ *linearly* with $\frac{1}{\bar{v}}$ as $\bar{v} \to \infty$. We use this in (2.31).

Appendix C. Asymptotic behavior of λ_2. We recall that the second eigenvalue of the stride-to-stride map is given by

$$\lambda_2 = 1 - \left(1 + g\eta_0 \cos\beta \frac{\cot\bar{\delta}}{\bar{v}^2}\right) \left.\frac{\partial\Delta\psi}{\partial\delta_n}\right|_{f.p.}. \tag{C.1}$$

Evaluating the derivative of the sweep angle at the fixed point $(\bar{v}, \bar{\delta})$ using the expressions developed in Appendix B, and noting that $\bar{\delta} \sim \sqrt{\epsilon} = \frac{1}{\bar{v}}$, we obtain

$$\frac{\partial\Delta\psi}{\partial\delta_n} = 2 + \frac{\eta_0 \cos\beta\, V'(\eta_0 \sin\beta)}{\bar{v}\sqrt{2mV(\eta_0 \sin\beta)}} + \mathcal{O}\left(\frac{1}{\bar{v}^2}\right). \tag{C.2}$$

Then using $\cot\bar{\delta} \approx \frac{1}{\bar{\delta}} \propto \bar{v}$ and substituting (C.2) into (C.1) yield the expression (2.29):

$$\lambda_2 = -1 - \frac{\eta_0 \cos\beta\,[4mg + V'(\eta_0 \sin\beta)]}{\bar{v}\sqrt{2mV(\eta_0 \sin\beta)}} + \mathcal{O}\left(\frac{1}{\bar{v}^2}\right). \tag{C.3}$$

REFERENCES

[1] J. Nishii, *Legged insects select the optimal pattern based on the energetic cost*, Biological Cybernetics, 83 (2000), pp. 435–442.

[2] R. Full, D. R. Stokes, A. Ahn, and R. K. Josephson, *Energy absorption during running by leg muscles in a cockroach*, J. Experimental Biology, 201 (1998), pp. 997–1012.

[3] K. G. Pearson, *Central programming and reflex control of walking in the cockroach*, J. Experimental Biology, 56 (1972), pp. 173–193.

[4] M. Ahmadi and M. Buehler, *Stable control of a simulated one-legged running robot with hip and leg compliance*, IEEE Trans. Robotics Automation, 13 (1997), pp. 96–104.

[5] J. K. Hodgkins and M. H. Raibert, *Adjusting step length for rough terrain locomotion*, IEEE Trans. Robotics Automation, 7 (1991), pp. 289–298.
[6] G. M. Nelson and R. D. Quinn, *Posture Control*, IEEE Control Systems Magazine 19 (1999), pp. 9–14.
[7] R. Full and D. E. Koditschek, *Templates and anchors: Neuromechanical hypothesis of legged locomotion on land*, J. Experimental Biology, 83 (1999), pp. 3325–3332.
[8] R. Blickhan and R. Full, *Similarity in multilegged locomotion: Bouncing like a monopode*, J. Comp. Physiol. A, 173 (1993), pp. 509–517.
[9] M. H. Dickinson, C. T. Farley, R. J. Full, M. A. R. Koehl, R. Kram, and S. Lehman, *How animals move: An integrative view*, Science, 288 (2000), pp. 100–106.
[10] R. Altendorfer, N. Moore, H. Komsuoğlu, M. Buehler, H. B. Brown, Jr., D. McMordie, U. Saranli, R. Full, and D. E. Koditschek, *RHex: A biologically inspired hexapod runner*, Autonomous Robots, 11 (2001), pp. 207–213.
[11] T. McGeer, *Passive bipedal running*, Proc. Roy. Soc. London B, 240 (1990), pp. 107–134.
[12] M. Garcia, A. Chatterjee, A. Ruina, and M. J. Coleman, *The simplest walking model: Stability, complexity and scaling*, ASME J. Biomech. Engrg., 120 (1998), pp. 281–288.
[13] M. J. Coleman and A. Ruina, *An uncontrolled walking toy that cannot stand still*, Phys. Rev. Lett., 80 (1998), pp. 3658–3661.
[14] K. D. Mombaur, R. W. Longman, H. G. Bock, and J. P. Schlöder, *Stable one-legged hopping without feedback and with a point foot*, in Proceedings of the 2002 IEEE International Conference on Robotics and Automation, Vol. 4, Washington, DC, 2002, pp. 3978–3983.
[15] E. Celaya and J. M. Porta, *A control structure for the locomotion of a legged robot on difficult terrain*, IEEE Robotics and Automation Magazine, 5 (1998), pp. 43–51.
[16] T. Kubow and R. Full, *The role of the mechanical system in control: A hypothesis of self-stabilization in hexapedal runners*, Phil. Trans. Roy. Soc. London Ser. B Biol. Sci., 354 (1999), pp. 849–861.
[17] J. Schmitt and P. Holmes, *Mechanical models for insect locomotion: Dynamics and stability in the horizontal plane* I. *Theory*, Biological Cybernetics, 83 (2000), pp. 501–515.
[18] J. Schmitt and P. Holmes, *Mechanical models for insect locomotion: Dynamics and stability in the horizontal plane* II. *Application*, Biological Cybernetics, 83 (2000), pp. 517–527.
[19] J. Schmitt and P. Holmes, *Mechanical models for insect locomotion: Stability and parameter studies*, Phys. D, 156 (2001), pp. 139–168.
[20] J. Schmitt, M. Garcia, R. Razo, P. Holmes, and R. J. Full, *Dynamics and stability of legged locomotion in the horizontal plane: A test case using insects*, Biological Cybernetics, 86 (2002), pp. 343–353.
[21] D. L. Jindrich and R. J. Full, *Dynamic stabilization of rapid hexapedal locomotion*, J. Exp. Biol., 205 (2002), pp. 2803–2823.
[22] A. Seyfarth, H. Geyer, M. Günther, and R. Blickhan, *A movement criterion for running*, J. Biomechanics, 35 (2002), pp. 649–655.
[23] A. Ruina, *Non-holonomic stability aspects of piecewise holonomic systems*, Rep. Math. Phys., 42 (1998), pp. 91–100.
[24] U. Saranli, M. Buehler, and D. E. Koditschek, *RHex: A simple and highly mobile hexapod robot*, Internat. J. Robotics Research, 20 (2001), pp. 616–631.
[25] R. Altendorfer, R. M. Ghigliazza, P. Holmes, and D. E. Koditschek, *Exploiting passive stability for hierarchical control*, in Proceedings of the Fifth International Conference on Climbing and Walking Robots (CLAWAR 2002), Professional Engineering Publishing Limited, London, 2002, pp. 177–184.
[26] A. Seyfarth and H. Geyer, *Natural control of spring-like running—optimized self-stabilization*, in Proceedings of the Fifth International Conference on Climbing and Walking Robots (CLAWAR 2002), Professional Engineering Publishing Limited, London, 2002, pp. 81–85.
[27] R. Altendorfer, D. E. Koditschek, and P. Holmes, *Towards a Factored Analysis of Legged Locomotion Models*, Technical report, CSE-TR-467-02, University of Michigan, Ann Arbor, MI, 2002; to appear in Proceedings of the IEEE Automation Conference, Taipei, Taiwan, 2003.
[28] R. Altendorfer, R. M. Ghigliazza, P. Holmes, and D. E. Koditschek, *Hopping on a springy leg: "Low attention" feedback control*, manuscript.
[29] F. Gubina, H. Hemami, and R. B. McGhee, *On the dynamic stability of biped locomotion*, IEEE Trans. Biomedical Engrg., 21 (1974), pp. 102–108.

[30] M. H. Raibert, *Legged Robots that Balance*, MIT Press, Cambridge, MA, 1986.
[31] H. Goldstein, *Classical Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1980.
[32] W. J. Schwind and D. E. Koditschek, *Approximating the stance map of a* 2 *DOF monoped runner*, J. Nonlinear Sci., 10 (2000), pp. 533–568.
[33] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.
[34] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1990.
[35] C. T. Farley and O. González, *Leg stiffness and stride frequency in human running*, J. Biomechanics, 29 (1995), pp. 181–186.
[36] W. J. Schwind, *Spring Loaded Inverted Pendulum Running: A Plant Model*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1998.
[37] U. Saranli, W. J. Schwind, and D. E. Koditschek, *Toward the control of a multi-jointed, monoped runner*, in Proceedings of the IEEE International Conference on Robotics and Automation, Leuven, Belgium, 1998, pp. 2676–2682.
[38] W. J. Schwind and D. E. Koditschek, *Characterization of monoped equilibrium gaits*, in Proceedings of the IEEE International Conference on Robotics and Automation, Albuquerque, NM, 1997, pp. 1986–1992.
[39] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, Benjamin/Cummings, Menlo Park, CA, 1986.
[40] M. J. Coleman and P. Holmes, *Motions and stability of a piecewise holonomic system: The discrete Chaplygin sleigh*, Regul. Chaotic Dyn., 4 (1999), pp. 55–77.
[41] J. I. Neimark and N. A. Fufaev, *Dynamics of Nonholonomic Systems,* Transl. Math. Monogr., 33, AMS, Providence, RI, 1972.
[42] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw–Hill, New York, 1978.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 2, pp. 219–237

Coorbital Periodic Orbits in the Three Body Problem*

Josep M. Cors† and Glen R. Hall‡

Abstract. We consider the dynamics of coorbital motion of two small moons about a large planet which have nearly circular orbits with almost equal radii. These moons avoid collision because they switch orbits during each close encounter. We approach the problem as a perturbation of decoupled Kepler problems as in Poincaré's periodic orbits of the first kind. The perturbation is large but only in a small region in the phase space. We discuss the relationship required among the small quantities (radial separation, mass, and minimum angular separation). Persistence of the orbits is discussed.

Key words. coorbital motion, periodic orbits of the first kind, three body problem

AMS subject classifications. 70F15, 70F07

PII. S1111111102411304

1. Introduction.

In 1981, the Voyager 1 space craft explored and photographed the neighborhood of Saturn (see Aksnes [1]). These photos captured two remarkable small moons subsequently named Janus and Epimetheus. The moons are remarkable because they are of nearly equal mass, and their orbits have nearly equal radii. Because their orbits are so close, their motion is called "coorbital." As implied by Kepler's laws, the inner moon has a smaller period and hence catches up to the outer moon. The moons avoid collision because their mutual gravitation causes them to switch orbits when they approach each other. Figure 1.1 shows a schematic drawing of the paths of the coorbital moons. Figure 1.2 shows computed orbits of the two moons in rotating coordinates. The speed of rotation of the coordinates is that of the orbit of average radius. In this paper, we investigate the dynamics of the motion of a pair of small, comparable mass, coorbital moons as solutions of the Newtonian three body problem.

Several authors have studied the dynamics of coorbital satellites in general and Janus and Epimetheus in particular. One approach is to treat one of the moons as a test particle of zero mass and think of the dynamics as that of the restricted three body problem. Coorbital motion in this situation is very interesting and has been extensively studied both analytically and numerically (see, for example, Salo and Yoder [12] and Llibre and Ollé [4]).

Since the small moons are of comparable mass, it is more natural to treat them equally in the analysis. Two groups have approached the problem from this point of view. One approach is to do a "matching" of two separate dynamics problems—an "outer" problem, where the small moons do not interact, and an "inner" problem, where the interactions dominate the

*Received by the editors July 18, 2002; accepted for publication (in revised form) by J. Meiss January 20, 2003; published electronically May 23, 2003.
http://www.siam.org/journals/siads/2-2/41130.html
†Departament de Matemàtica Aplicada III, Universitat Politècnica de Catalunya, 08240, Manresa, Spain (cors@eupm.upc.es). The work of this author was partially supported by DGICYT grant PB96-1153 and MEC grant PR97-39332370.
‡Department of Mathematics and Statistics, Boston University, Boston, MA 02215 (rockford@math.bu.edu).

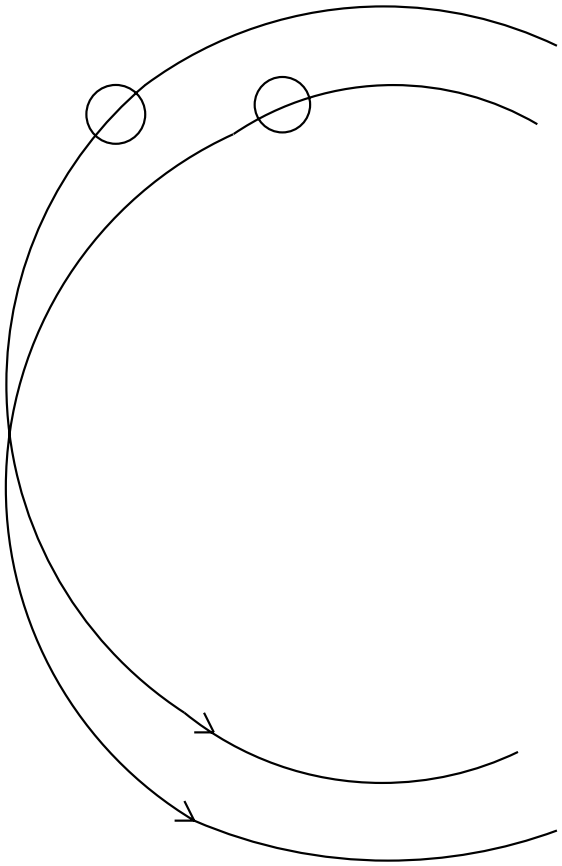


Figure 1.1. *Schematic of coorbital satellites exchanging orbits. The actual exchange takes several revolutions.*

motion. This technique was successfully used by Spirig and Waldvogel [13]. A second approach is that of Murray and Dermott [9]. They consider perturbations of the tangential component of the forces caused by the interactions of the two moons. Their work gives very accurate predictions of the orbits.

In the classical literature, the study of the dynamics of the three body problem, in which two of the bodies are small compared to the third, has a long and noble history. Periodic orbits of this problem in which the interaction between the two small bodies is treated as a perturbation of decoupled Kepler problems have been studied extensively (see, for example,

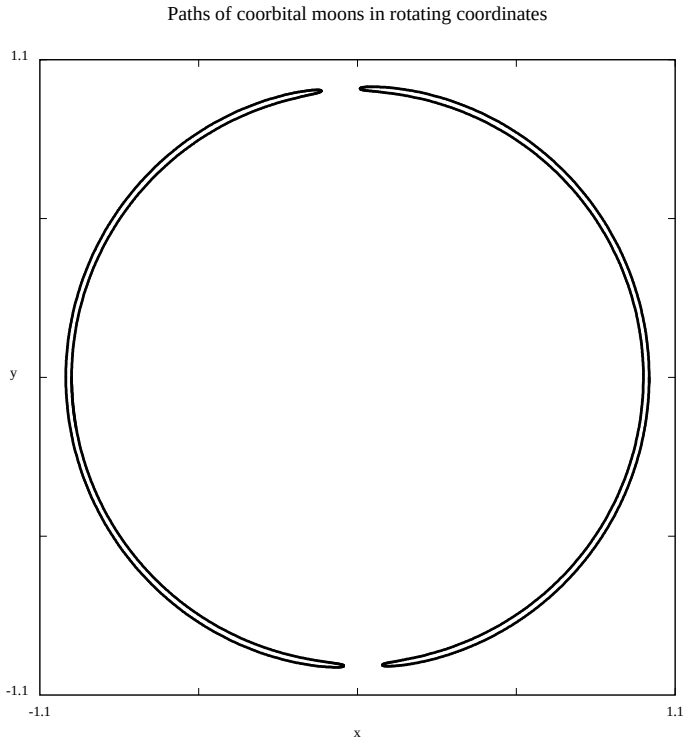


Figure 1.2. *Typical coorbital paths in rotating coordinates. The angle of closest approach is the angle of the gap between the paths of the moons. In this picture, the initial positions are* $(1.01, 0), (0, -0.99)$*; the initial velocities are* $(0.0995037), (0, -1.00504)$*; and the masses are* $0.01^{2.4}$*. (Computed with dstool.)*

Meyer [6], [7]). The situation we study here involves two small masses in nearly circular coplanar orbits about a large mass. Periodic orbits of this type were described by Poincaré as periodic orbits of the "first type" (see Poincaré [11, pp. 75–76] and Goroff's introduction in [11, p. I46]) The problem studied here differs from the classical problem in that we consider solutions for which the distance between the radii of the orbits of the small bodies tends to zero as their masses tend to zero. For this reason, there is no "restricted" (small masses $= 0$) limit, and we must study the problem for small but nonzero masses.

There are three small quantities to consider. The first is the ratio of the difference between the radii of the moons' orbits and the average radius of their orbits. We call this quantity ϵ. The second is the masses of the small moons. We call the masses $\mu_i \epsilon^a$, where μ_1, μ_2 are order one. Finally, we are seeking orbits for which the moons do not pass through syzygy on the same side of the planet. Hence, there is some minimum angle between the rays from the planet to the moons. This minimum angular separation of the moons is the third small quantity; we call it ϵ^b. (This angle is the gap between the orbits of the moons in rotating coordinates; see Figure 1.2.) The main and new result of this paper is the determination of the relative size of these small quantities that admits coorbital motion like that observed for Janus and Epimetheus. In particular, we show the following.

Theorem 1.1. *Let ϵ be the ratio of the difference between the radii of the orbits of the small moons with the average radius, and suppose the ratio of the moons' masses to the mass of the planet is of order ϵ^a and the minimum angular separation of the moons is of order ϵ^b. If $2 < a < 5/2$, then the system admits coorbital motion, and in this case $b = a - 2$.*

The technique we use is to make restrictions on the parameters a and b so as to obtain a low order system in ϵ that we can analyze. Then we show coorbital motion occurs in this truncated system. Restrictions on the parameter a correspond to restricting the choice of masses of the small bodies. Restrictions on the parameter b correspond to limiting our attention to certain regions of the phase space. Hence we can treat a and b as independent parameters. Once we fix the ranges for a and b, we must (a postiori) show that coorbital motion occurs for the masses and in the region of phase space satisfying the restrictions.

The range of values for the parameter a is chosen so that the terms giving the force between the moons have only one summand of low order in epsilon. Care must be taken because terms involving epsilon appear in both the numerators and denominators of terms in the Hamiltonian.

In the final section, we discuss qualitative aspects of coorbital motion. We do not prove stability, but we discuss a lower bound for the persistence of these orbits.

2. Equations of motion and scalings. In this section, we derive our model for the study of motion of small coorbital moons. Since the two small moons have masses of the same order of magnitude, we require that they be treated equally by the equations of motion.

We begin with the Hamiltonian system representing the Newtonian three body problem given by the usual energy function

$$H(q,p) = \frac{p_0^2}{2M} + \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_3} - \frac{GMm_1}{\|q_1 - q_0\|} - \frac{GMm_2}{\|q_2 - q_0\|} - \frac{Gm_1m_2}{\|q_2 - q_1\|},$$

where q_0, p_0 are the positions and momentum in the plane of the large planet with mass M, q_i, p_i, $i = 1, 2$, are the position and momenta in the plane of the small moons with masses m_i, $i = 1, 2$, respectively, and G is the gravitational constant. We choose units of distance so that $G = 1$ (see Meyer [8]) and units of mass so that $M = 1$. This gives the Hamiltonian

$$H(q,p) = \frac{p_0^2}{2} + \frac{p_1^2}{2m_1} + \frac{p_2^2}{2m_2} - \frac{m_1}{\|q_1 - q_0\|} - \frac{m_2}{\|q_2 - q_0\|} - \frac{m_1m_2}{\|q_2 - q_1\|}.$$

Next we isolate the total momentum and center of mass via a change of variables. The usual coordinate choice for this step is the Jacobi coordinates (see Meyer [8]); however, these coordinates treat the moons asymmetrically. We use a coordinate change which treats the moons equivalently. We define new position variables u by

$$\begin{aligned} u_1 &= q_1 - q_0, \\ u_2 &= q_2 - q_0, \\ u_3 &= \frac{1}{1 + m_1 + m_2}(q_0 + m_1q_1 + m_2q_2). \end{aligned}$$

Extending this to a symplectic transformation, we define new momentum variables v by

$$\begin{aligned} v_1 &= -\frac{m_1}{1+m_1+m_2}p_0 + \frac{1+m_2}{1+m_1+m_2}p_1 - \frac{m_1}{1+m_1+m_2}p_2, \\ v_2 &= -\frac{m_2}{1+m_1+m_2}p_0 - \frac{m_2}{1+m_1+m_2}p_1 + \frac{1+m_1}{1+m_1+m_2}p_2, \\ v_3 &= p_0 + p_1 + p_2. \end{aligned}$$

In these variables, the new Hamiltonian function is

$$H(u,v) = \frac{(1+m_1)v_1^2}{2m_1} + \frac{(1+m_2)v_2^2}{2m_2} + v_1v_2 + \frac{v_3^2}{2(1+m_1+m_2)} - \frac{m_1}{\|u_1\|} - \frac{m_2}{\|u_2\|} - \frac{m_1m_2}{\|u_1-u_2\|}.$$

Since the variables v_3, u_3 represent the total momentum and center of mass, respectively, it suffices to consider only u_1, u_2, v_1, v_2.

2.1. The small quantities: Moon masses. As stated above, we take advantage of the fact that the coorbital motion involves three small quantities. We assume that the distance between the semimajor axes of the orbits of the small moons when they are far apart is of order epsilon compared to the average of the two semimajor axes. We assume that the masses are (both) of order ϵ^a (in comparison to the mass of the planet $M = 1$) and that the angle of closest approach of the bodies is of order ϵ^b.

It is convenient to incorporate the small masses first. We let

$$\begin{aligned} m_1 &= \epsilon^a \mu_1, \\ m_2 &= \epsilon^a \mu_2, \end{aligned}$$

where μ_1, μ_2 are of the same order as the mass of the planet $M = 1$. We simultaneously rescale the momentum variables, defining new momentum variables w_1, w_2 by

$$\begin{aligned} w_1 &= \epsilon^{-a} v_1, \\ w_2 &= \epsilon^{-a} v_2. \end{aligned}$$

The change from u, v to u, w variables is a symplectic transformation with multiplier ϵ^{-a} (see Meyer [8]). The resulting system has Hamiltonian function

$$H(u,w) = \frac{w_1^2}{2\mu_1} + \epsilon^a \frac{w_1^2}{2} + \frac{w_2^2}{2\mu_2} + \epsilon^a \frac{w_2^2}{2} + \epsilon^a w_1 w_2 - \frac{\mu_1}{\|u_1\|} - \frac{\mu_2}{\|u_2\|} - \frac{\epsilon^a \mu_1 \mu_2}{\|u_1 - u_2\|},$$

and the corresponding differential equations are

$$\begin{aligned} \dot{u_1} &= \frac{w_1}{\mu_1} + \epsilon^a w_1 + \epsilon^a w_2, \\ \dot{u_2} &= \frac{w_2}{\mu_2} + \epsilon^a w_2 + \epsilon^a w_1, \\ \dot{w_1} &= -\frac{\mu_1}{\|u_1\|^3} u_1 + \frac{\epsilon^a \mu_1 \mu_2}{\|u_2 - u_1\|^3}(u_2 - u_1), \\ \dot{w_2} &= -\frac{\mu_2}{\|u_2\|^3} u_2 + \frac{\epsilon^a \mu_1 \mu_2}{\|u_2 - u_1\|^3}(u_1 - u_2). \end{aligned}$$

2.2. Incorporating the minimum distance between moons. The next step is to group terms in the system above with respect to powers of ϵ. Our goal is to collect terms so that the system corresponding to the lowest order is simple yet reflects the dynamics of coorbital motion. In particular, we cannot just treat all terms involving epsilon as "higher order" since the resulting system is just two decoupled Kepler problems.

To begin, we make the a priori assumption that $a > 0$, that is, that the small masses are small in comparison to the mass of the primary $M = 1$. In addition, since we are seeking coorbital motion, we assume that the moons do not approach the planet, i.e., that $\|u_1\|$ and $\|u_2\|$ are of order one. Finally, while coorbital motion is subtle, it probably does not include times when the velocity is very large. Hence we assume that the (scaled) momentum variables w_1, w_2 are of order one. These a priori assumptions restrict our attention to a region of phase space in which we hope to find solutions corresponding to coorbital motion.

With the a priori assumptions above, we can classify all of the terms with a factor ϵ^a in the Hamiltonian and/or the differential equations as order $O(\epsilon^a)$ *except* the terms having

$$\|u_2 - u_1\|$$

in the denominator. Since we are looking for orbits where the two moons do interact significantly, we must assume that the distance between the moons becomes sufficiently small that the interaction terms are significant. In particular, we assume that

$$\|u_2 - u_1\| \geq O(\epsilon^b)$$

and that the minimum value of $\|u_2 - u_1\|$ along an orbit is of this order. This can be interpreted as allowing the algebra to do the division between the "inner" and the "outer" solutions in the matching method. We set the relative size of a and b below to achieve this division.

Hence, at this stage, we have

$$H(u, w) = \frac{w_1^2}{2\mu_1} + \frac{w_2^2}{2\mu_2} - \frac{\mu_1}{\|u_1\|} - \frac{\mu_2}{\|u_2\|} - \frac{\epsilon^a \mu_1 \mu_2}{\|u_2 - u_1\|} + O(\epsilon^a).$$

2.3. Rotating polar coordinates. We next change to rotating coordinates in the usual way. Let

$$\begin{aligned} x_j &= R(t) u_j, \\ y_j &= R(t) w_j, \end{aligned}$$

where

$$R(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

We focus on orbits of period near 2π and hence radius near 1 without loss of generality. The new Hamiltonian is

$$\begin{aligned} H(x_1, x_2, y_1, y_2) = \frac{\| y_1 \|^2}{2\mu_1} + \frac{\| y_2 \|^2}{2\mu_2} - x_1^{\mathrm{T}} K y_1 - x_2^{\mathrm{T}} K y_2 \\ - \frac{\mu_1}{\| x_1 \|} - \frac{\mu_2}{\| x_2 \|} - \epsilon^a \frac{\mu_1 \mu_2}{\| x_2 - x_1 \|} + O(\epsilon^a), \end{aligned}$$

where

$$K = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

As expected, the next step is to convert to polar coordinates

$$(r_1, r_2, \theta_1, \theta_2, R_1, R_2, \Theta_1, \Theta_2) = (r, \theta, R, \Theta).$$

In these coordinates, the Hamiltonian is given by

$$\begin{aligned} H(r,\theta,R,\Theta) = {} & \frac{1}{2\mu_1}\left(R_1^2 + \left(\frac{\Theta_1}{r_1}\right)^2\right) + \frac{1}{2\mu_2}\left(R_2^2 + \left(\frac{\Theta_2}{r_2}\right)^2\right) \\ & - \Theta_1 - \Theta_2 - \frac{\mu_1}{r_1} - \frac{\mu_2}{r_2} \\ & - \epsilon^a \frac{\mu_1\mu_2}{(r_1^2 + r_2^2 - 2r_1r_2\cos(\theta_2 - \theta_1))^{1/2}} + O(\epsilon^a). \end{aligned}$$

2.4. Incorporating the difference between the semimajor axes. The last of the small quantities in the problem is the difference between the semimajor axes of the moons when they are far from each other and on approximate Kepler orbits. We assume this difference is of order ϵ; i.e., we let

$$\begin{aligned} \epsilon\rho_1 &= r_1 - 1, \\ \epsilon\rho_2 &= r_2 - 1, \\ \epsilon\Phi_1 &= \Theta_1 - \mu_1, \\ \epsilon\Phi_2 &= \Theta_2 - \mu_2. \end{aligned}$$

This change of variables is a symplectic transformation with multiplier ϵ^{-1}. To avoid the singularity at $\epsilon = 0$, we rescale time using

$$dt_{\text{old}}/dt_{\text{new}} = \epsilon, \text{ so } t_{\text{old}} = \epsilon t_{\text{new}}.$$

The new Hamiltonian is

$$\begin{aligned} H(\rho_1,\rho_2,\theta_1,\theta_2,R_1,R_2,\Phi_1,\Phi_2) = {} & \frac{1}{2\mu_1}\left(R_1^2 + \frac{(\mu_1+\epsilon\Phi_1)^2}{(1+\epsilon\rho_1)^2}\right) + \frac{1}{2\mu_2}\left(R_2^2 + \frac{(\mu_2+\epsilon\Phi_2)^2}{(1+\epsilon\rho_2)^2}\right) \\ & - (\mu_1+\mu_2) - \epsilon\Phi_1 - \epsilon\Phi_2 - \frac{\mu_1}{1+\epsilon\rho_1} - \frac{\mu_2}{1+\epsilon\rho_2} \\ & - \frac{\epsilon^a\mu_1\mu_2}{((1+\epsilon\rho_1)^2 + (1+\epsilon\rho_2)^2 - 2(1+\epsilon\rho_1)(1+\epsilon\rho_2)\cos(\theta_2-\theta_1))^{1/2}} + O(\epsilon^a). \end{aligned}$$

Again, we expand and collect similar power terms in ϵ for all but the last summand (which is treated separately below). We obtain

$$H(\rho_1,\rho_2,\theta_1,\theta_2,R_1,R_2,\Phi_1,\Phi_2) = \frac{R_1^2}{2\mu_1} + \epsilon^2\left(\frac{\Phi_1^2}{2\mu_1} - 2\Phi_1\rho_1 + \frac{\mu_1\rho_1^2}{2}\right)$$

$$+\frac{R_2^2}{2\mu_2}+\epsilon^2\left(\frac{\Phi_2^2}{2\mu_2}-2\Phi_2\rho_2+\frac{\mu_2\rho_2^2}{2}\right)$$
$$-\frac{\epsilon^a\mu_1\mu_2}{((1+\epsilon\rho_1)^2+(1+\epsilon\rho_2)^2-2(1+\epsilon\rho_1)(1+\epsilon\rho_2)\cos(\theta_2-\theta_1))^{1/2}}+O(\epsilon^a)+O(\epsilon^3),$$

where we have kept terms of order two in ϵ and the last summand. The constant terms have no effect on the dynamics and have been dropped. The "$O(\epsilon^3)$" refers to those generated by expanding all but the last summand.

2.5. A priori assumption on the masses: Lower bound for parameter a. We can think of the Hamiltonian above as coming from two decoupled Kepler problems for the motion of the moons plus interaction terms. In the rotating coordinates, there is a circle of fixed points corresponding to the circular orbits at radius one. To include the fact that the moons are not on this orbit but rather are moving on orbits nearby (within $\pm\epsilon$), we must include the ϵ^2 order terms in this part of the Hamiltonian.

We are seeking systems for which the terms governing interaction between the two moons are significant only when they are close. Hence we must assume that the terms collected above in "$O(\epsilon^a)$" are actually of higher order than ϵ^2. To do this, we are forced to assume that

$$a>2.$$

This assumption allows the algebra to perform the localization of the interaction terms between the moons.

2.6. Interaction terms between the moons. Finally, we return to the terms governing the interaction between the two moons. We must go carefully since we must keep terms which may become large when the two moons are close.

To deal with the coupling term between the moons, we analyze

$$g(\rho_1,\rho_2,\theta,\epsilon)=\frac{1}{((1+\epsilon\rho_1)^2+(1+\epsilon\rho_2)^2-2(1+\epsilon\rho_1)(1+\epsilon\rho_2)\cos(\theta))^{1/2}}.$$

Here $\theta=\theta_2-\theta_1$, the angular distance between the moons. We must examine the order in epsilon of terms in the partials of g, particularly when θ is small. We extend our a priori assumptions to restrict our attention to orbits for which syzygy does not occur with the two moons on the same side of the planet, i.e., that

$$\theta, 2\pi-\theta\geq O(\epsilon^b),$$

where b is fixed below. We compute $\partial g/\partial\rho_1$ and arrange the terms as follows:

$$\begin{aligned}\frac{\partial g}{\partial\rho_1}&=-\epsilon\frac{(1-\cos\theta)+\epsilon(\rho_1-\rho_2\cos\theta)}{((2-2\cos\theta)(1+\epsilon(\rho_1+\rho_2)+\epsilon^2\rho_1\rho_2)+\epsilon^2(\rho_1-\rho_2)^2)^{3/2}}\\&=-\epsilon\frac{1}{2(2-2\cos\theta)^{1/2}}\left[\frac{1}{(1+\epsilon(\rho_1+\rho_2)+\epsilon^2\rho_1\rho_2+\epsilon^2(\rho_1+\rho_2)^2/(2-2\cos\theta))^{3/2}}\right.\\&\quad\left.+\frac{\epsilon(\rho_1-\rho_2\cos\theta)/(1-\cos\theta)}{(1+\epsilon(\rho_1+\rho_2)+\epsilon^2\rho_1\rho_2+\epsilon^2(\rho_1+\rho_2)^2/(2-2\cos\theta))^{3/2}}\right].\end{aligned}$$

Now, the order in epsilon of these terms is determined by the denominators. We can expand the terms within the square brackets by expanding

$$(1+\delta)^{-3/2}$$

in powers of δ, where

$$\delta = \epsilon(\rho_1+\rho_2) + \epsilon^2\rho_1\rho_2 + \frac{\epsilon^2(\rho_1+\rho_2)^2}{(2-2\cos\theta)}.$$

The last summand of δ is of order ϵ^{2-2b}, so δ^n contains terms of order

$$\epsilon^k\left(\epsilon^2\right)^r\left(\epsilon^{(2-2b)}\right)^{n-(k+r)} = \epsilon^{k+2r+(2-2b)(n-k-r)}.$$

At extreme values of θ,

$$\frac{\epsilon}{2(2-2\cos\theta)^{1/2}} = O(\epsilon^{1-b}).$$

So, expanding $\partial g/\partial\rho_1$ at the extreme values of θ, we obtain terms of order

$$O(\epsilon^{1-b+k+2r+(2-2b)(n-k-r)}) \text{ and } O(\epsilon^{2-3b+k+2r+(2-2b)(n-k-r)}),$$

where $n = 0, 1, 2, \ldots$ and $0 \le k, r$, $k+r \le n$. The terms in $\dot{R}_1$ are ϵ^a times these.

The terms in $\partial g/\partial\rho_2$ (which relates to the $\dot{R}_2 = -\partial H/\partial\rho_2$ term) are of the same order.

For partials with respect to the θ, we must compute more and be considerably more careful. First, we compute $\partial g/\partial\theta$. We could deal with the resulting terms as we did above, but it is more illustrative to examine the power series of $\partial g/\partial\theta$ in epsilon:

$$\begin{aligned}\frac{\partial g}{\partial\theta} &= \frac{(1+\epsilon\rho_1)(1+\epsilon\rho_2)\sin\theta}{((2-2\cos\theta)(1+\epsilon(\rho_1+\rho_2)+\epsilon^2\rho_1\rho_2)+\epsilon^2(\rho_1-\rho_2)^2)^{3/2}}\\ &= \frac{\sin\theta}{(2-2\cos\theta)^{3/2}} - \frac{1}{2}\frac{(\rho_1+\rho_2)\sin\theta}{(2-2\cos\theta)^{3/2}}\epsilon\\ &+ \frac{1}{4}\frac{((3\rho_1^2+3\rho_2^2+2\rho_1\rho_2)\cos\theta+3\rho_1^2+3\rho_2^2-14\rho_1\rho_2)\sin\theta}{(2-2\cos\theta)^{5/2}}\epsilon^2\\ &+ -\frac{1}{8}\frac{((5\rho_1^3+5\rho_2^3+3\rho_1^2\rho_2+3\rho_1\rho_2^2)\cos\theta+13\rho_1^3+13\rho_2^3-21\rho_1^2\rho_2-21\rho_1\rho_2^2)\sin\theta}{(2-2\cos\theta)^{5/2}}\epsilon^3+O(\epsilon^4).\end{aligned}$$

We analyze each summand at extreme values of θ, so $2-2\cos\theta = O(\epsilon^{2b})$ and $\sin\theta = O(\epsilon^b)$.

Evaluating the order of these terms at the extreme values of θ, we see that the first term is of order ϵ^{-2b}. The second term is of order ϵ^{1-2b}, and the third term is at the largest of order ϵ^{2-4b}.

Subsequent terms are of order

$$O(\epsilon^{w-2b+k+2r+(2-2b)(n-k-r)}),$$

where $w = 0, 1, 2$, $n = 1, 2, 3, \ldots$, and $0 \le k, r$, $k+r \le n$. The corresponding terms in the $\dot{\Phi}_i$ are of order ϵ^a times these.

2.7. Bounds for b. We saw above that we must keep terms of order ϵ^2 and lower in order to have any chance of accurately representing the motion of the two moons when they are far apart. In the previous section, we computed the order of terms resulting from the interaction between the moons when they are close to each other. The order of these terms is as follows:

$\dot{R}_i$ contains terms $\epsilon^{a+1-b+k+2r+(2-2b)(n-k-r)}$ and $\epsilon^{a+2-3b+k+2r+(2-2b)(n-k-r)}$, where $n = 0, 1, \ldots$ and $0 \le k, r, \quad k + r \le n$;

$\dot{\Phi}_i$ contains terms $\epsilon^{a-2b}, \quad \epsilon^{a-2b+1}, \epsilon^{a-4b+2}$, $\epsilon^{a+w-2b+k+2r+(2-2b)(n-k-r)}$, where $w = 0, 1, 2, n = 1, 2, 3, \ldots, 0 \le k, r,$ and $k + r \le n$.

In order to obtain a model system in which to search for and study coorbital behavior, we choose an upper bound for a and conditions on b which guarantee that finitely many of these terms are of order less than or equal to 2 in epsilon. We note that the terms

$$\epsilon^{a+1-b+(2-2b)n} \quad \text{and} \quad \epsilon^{a-2b+(2-2b)n}$$

appear in $\dot{R}_i$ and $\dot{\Phi}_i$. In order to guarantee that only finitely many of these terms are of order less than 2 in epsilon, we must assume that $2 - 2b > 0$. Hence we assume $b < 1$. That is, we restrict our attention to the region of the phase space corresponding to positions where the angular separation is greater than or equal to $O(\epsilon^b)$ and $b > 1$.

However, we do not want to eliminate all the terms from the interactions between the moons. To guarantee that at least one interaction term is of order less than 2 in epsilon, we assume that

$$a - 2b < 2$$

or

$$\frac{a-2}{2} < b.$$

The simplest nontrivial system possible would contain just the one interaction term of order ϵ^{a-2b} in $\dot{\Phi}_i$ and none of the higher order terms in $\dot{\Phi}_i$ or $\dot{R}_i$. To obtain this system to order 2 in epsilon, we need to guarantee that

$$a - 2b + 1 > 2 \qquad \text{and} \qquad a - 4b + 2 > 2.$$

All subsequent terms of $\dot{\Phi}_i$ and all terms of $\dot{R}_i$ are of order greater than these powers and/or of order greater than $a > 2$ in epsilon. If $0 < b < 1/2$, then $a - 2b + 1 < a - 4b + 2$, so we must assume that

$$a - 2b + 1 > 2,$$

i.e.,

$$b < \frac{a-1}{2}.$$

If $1/2 \le b < 1$, then $a - 4b + 2 \le a - 2b + 1$, so we must assume that

$$a - 4b + 2 > 2,$$

i.e.,

$$b < \frac{a}{4}.$$

2.8. Summary. To summarize, we have made two sorts of restrictions. First, we have restricted our attention to a specific region of the phase space in which we hope to find coorbital motion. This includes a priori restrictions on the size of some variables (e.g., the distances from the planet to the moons is of order one) and restrictions on the parameter b. Second, we have made restrictions on the possible values of the parameters of the masses of the moons by placing restrictions on the parameter a. The motivation for these restrictions is to isolate a region in phase x parameter space where the lowest order system is simple enough to study (and in which we hope to find coorbital behavior).

We collect the following assumptions:

- The distances of the moons from the planet are bounded below by an order one constant.
- The velocities of the moons are bounded above by an order one constant.
- The difference between the radii of the orbits of the moons is of order ϵ (i.e., the radii of the moons' orbits are close).
- The masses of the moons are both of order ϵ^a (where the mass of the planet is one), and $a > 0$; i.e., the moons are small.
- The angular distance between the moons (call it θ) satisfies

$$\min(\theta) \geq O(\epsilon^b)$$

and

$$\max(2\pi - \theta) \geq O(\epsilon^b),$$

and $b > 0$ (i.e., the moons get close to each other, but syzygy of the moons does not occur with both on the same side of the planet), and at closest approach $\theta = O(\epsilon^b)$.

Recall that the radii of the approximate Kepler orbits of the two moons when they are separated are assumed to differ by order epsilon. In the resulting system, we must keep terms of order two in epsilon in order to distinguish our system from one where the two moons are both on the same circular orbit.

- We assume the masses of the moons are such that $a > 2$. This assumption is equivalent to assuming that the significant interaction between the moons occurs only when they are close.
- We assume that the minimum angular separation of the moons is such that

$$\frac{a-2}{2} < b.$$

This assumption ensures that there is at least one term of order less than two in epsilon resulting from the interaction between the moons.

- We assume that $b < 1$. This assumption is necessary to keep the number of interaction terms between the moons of order less than two in epsilon finite.
- The two assumptions above imply that $a < 4$.

- If $0 < b < 1/2$, we assume that $b < (a-1)/2$. If $1/2 \leq b < 1$, we assume that $b < a/4$. These assumptions yield the simplest nontrivial system in which to search for coorbital behavior.

With these assumptions, we can write the truncated equations (up to order ϵ^2) as

$$\begin{aligned}
\dot{\rho}_1 &= \frac{1}{\mu_1} R_1, \\
\dot{\theta}_1 &= \epsilon^2 \left(\frac{\Phi_1}{\mu_1} - 2\rho_1 \right), \\
\dot{R}_1 &= \epsilon^2 (2\Phi_1 - \mu_1 \rho_1), \\
\dot{\Phi}_1 &= \epsilon^a \frac{\mu_1 \mu_2 \sin(\theta_2 - \theta_1)}{(2 - 2\cos(\theta_2 - \theta_1))^{3/2}}, \\
\dot{\rho}_2 &= \frac{1}{\mu_2} R_2, \\
\dot{\theta}_2 &= \epsilon^2 \left(\frac{\Phi_2}{\mu_2} - 2\rho_2 \right), \\
\dot{R}_2 &= \epsilon^2 (2\Phi_2 - \mu_2 \rho_2), \\
\dot{\Phi}_2 &= -\epsilon^a \frac{\mu_1 \mu_2 \sin(\theta_2 - \theta_1)}{(2 - 2\cos(\theta_2 - \theta_1))^{3/2}}.
\end{aligned}$$

This system is Hamiltonian with

$$\begin{aligned}
H(\rho_1, \rho_2, \theta_1, \theta_2, R_1, R_2, \Phi_1, \Phi_2) = {} & \frac{1}{2\mu_1} R_1^2 + \epsilon^2 \left(\frac{\Phi_1^2}{2\mu_1} - 2\rho_1 \Phi_1 + \frac{\mu_1 \rho_1^2}{2} \right) \\
& + \frac{1}{2\mu_1} R_2^2 + \epsilon^2 \left(\frac{\Phi_2^2}{2\mu_2} - 2\rho_2 \Phi_2 + \frac{\mu_2 \rho_2^2}{2} \right) \\
& - \epsilon^a \frac{\mu_1 \mu_2}{(2 - 2\cos(\theta_2 - \theta_1))^{1/2}}.
\end{aligned}$$

3. Analysis of the reduced system. Two steps remain in the construction of our reduced system. First, we must determine appropriate value(s) for the parameter b, which determines the minimum angular separation between the moons. Second, we must justify (a postiori) our model by finding solutions of our model exhibiting coorbital behavior.

3.1. Projecting to two degrees of freedom. The four degree of freedom system above almost decouples into two systems. We can take advantage of this, projecting the system above onto a two degree of freedom system as follows: Let

$$\begin{aligned}
\rho &= \rho_2 - \rho_1, \\
\theta &= \theta_2 - \theta_1, \\
R &= R_2/\mu_2 - R_1/\mu_1, \\
\Phi &= \Phi_2/\mu_2 - \Phi_1/\mu_1
\end{aligned}$$

to obtain the system

$$\begin{aligned}\dot{\rho} &= R,\\ \dot{\theta} &= \epsilon^2\left(\Phi - 2\rho\right),\\ \dot{R} &= \epsilon^2(2\Phi - \rho),\\ \dot{\Phi} &= -\frac{(\mu_1+\mu_2)\epsilon^a \sin\theta}{(2-2\cos\theta)^{3/2}}.\end{aligned}$$

This is a Hamiltonian system with Hamiltonian

$$H(\rho,\theta,R,\Phi) = \frac{R^2}{2} + \epsilon^2\frac{\rho^2}{2} + \epsilon^2\left(\frac{\Phi^2}{2} - 2\rho\Phi\right) - \frac{(\mu_1+\mu_2)\epsilon^a}{(2-2\cos\theta)^{1/2}}.$$

3.2. Separating fast and slow variables. Next we isolate the slow and fast variables and put the system into a simpler form while maintaining the Hamiltonian structure. Hence all of our changes of variables will be symplectic or symplectic transformations with multipliers.

First, we redistribute the ϵ on the ρ, R variables by letting

$$\begin{aligned}\sqrt{\epsilon}P &= R,\\ \frac{1}{\sqrt{\epsilon}}\gamma &= \rho.\end{aligned}$$

We also rescale time, choosing a new time variable with $dt_{\text{new}}/dt_{\text{old}} = \epsilon$. Note that this change undoes the scaling done at the beginning of section 2.4, so this time variable is "real" time. These changes yield the new system

$$\begin{aligned}\dot{\gamma} &= P,\\ \dot{\theta} &= -2\epsilon^{1/2}\gamma + \epsilon\Phi,\\ \dot{P} &= -\gamma + 2\epsilon^{1/2}\Phi,\\ \dot{\Phi} &= -\frac{(\mu_1+\mu_2)\epsilon^{a-1}\sin\theta}{(2-2\cos\theta)^{3/2}}.\end{aligned}$$

The Hamiltonian is

$$H(\gamma,\theta,P,\Phi) = \frac{P^2}{2} + \frac{\gamma^2}{2} - 2\epsilon^{1/2}\Phi\gamma + \epsilon\frac{\Phi^2}{2} + \epsilon^{a-1}\frac{\mu_1+\mu_2}{(2-2\cos\theta)^{1/2}}.$$

To lowest order, the γ, P equations are those of a harmonic oscillator with equilibrium at $P = 0, \gamma = 2\epsilon^{1/2}\Phi$. We move this equilibrium to the origin by changing variables, setting

$$\tilde{\gamma} = \gamma - 2\epsilon^{1/2}\Phi.$$

However, to keep the change symplectic, we must also alter the other variables. Using the generating function (see Meyer [8])

$$S(\gamma,\theta,\tilde{P},\tilde{\Phi}) = \gamma\tilde{P} + \theta\tilde{\Phi} - 2\epsilon^{1/2}\tilde{\Phi}\tilde{P},$$

we obtain the symplectic change of variables

$$\begin{aligned}\tilde{\gamma} &= \gamma - 2\epsilon^{1/2}\tilde{\Phi},\\ \tilde{\theta} &= \theta - 2\epsilon^{1/2}\tilde{P},\\ \tilde{P} &= P,\\ \tilde{\Phi} &= \Phi.\end{aligned}$$

In these variables, the Hamiltonian is

$$H(\tilde{\gamma},\tilde{\theta},\tilde{P},\tilde{\Phi}) = \frac{\tilde{P}^2}{2} + \frac{\tilde{\gamma}^2}{2} - \epsilon\frac{3}{2}\tilde{\Phi}^2 - \epsilon^{a-1}\frac{\mu_1+\mu_2}{(2-2\cos(\tilde{\theta}+2\epsilon^{1/2}\tilde{P}))^{1/2}}.$$

As above, we expand and keep only terms which contribute terms of order less than or equal to one in epsilon to the differential equation. We must identify these terms in the expansions of

$$\frac{\partial}{\partial\tilde{\theta}}\left(\epsilon^{a-1}\frac{\mu_1+\mu_2}{(2-2\cos(\tilde{\theta}+\epsilon^{1/2}\tilde{P}))^{1/2}}\right)$$

and

$$\frac{\partial}{\partial\tilde{P}}\left(\epsilon^{a-1}\frac{\mu_1+\mu_2}{(2-2\cos(\tilde{\theta}+\epsilon^{1/2}\tilde{P}))^{1/2}}\right).$$

The computations involved are similar to those of sections 2.6 and 2.7,

$$\frac{\partial}{\partial\tilde{\theta}}\left(\frac{1}{(2-2\cos(\tilde{\theta}+\epsilon^{1/2}\tilde{P}))^{1/2}}\right) = -\frac{\sin\tilde{\theta}}{(2-2\cos\tilde{\theta})^{3/2}}$$

$$+\epsilon^{1/2}\tilde{P}\left(-\frac{\cos\tilde{\theta}}{(2-2\cos\tilde{\theta})^{3/2}} + \frac{\sin^2\tilde{\theta}}{(2-2\cos\tilde{\theta})^{5/2}}\right) + \cdots,$$

where " $\cdots$ " refers to terms in higher powers of $\epsilon^{1/2}\tilde{P}$. As above, we examine the higher order terms at the extreme values of $\tilde{\theta}$. In addition, we use the fact that restricting our attention to the region of the phase space where ρ is of order one implies that γ is $O(\epsilon^{1/2})$. Since γ and P behave to lowest order like a harmonic oscillator about $(\epsilon\Phi, 0)$ and Φ is order one, we restrict our attention to the region of phase space where $\tilde{P}$ is $O(\epsilon^{1/2})$. Assuming γ and P are order $O(\epsilon^{1/2})$ is equivalent to assuming that the eccentricity of the moons' orbits is small $(O(\epsilon))$ when they are far apart.

However, we also need to add the assumption that

$$\frac{a-1}{3} > b.$$

Using this and $0 < b < 1$, we have that we need only keep the first summand on the right-hand side above in the equation for $\dot{\Phi}$.

Similarly, we need to assume that

$$\frac{2a-3}{4} > b$$

to guarantee that all terms of

$$\frac{\partial}{\partial \tilde{P}}\left(\epsilon^{a-1}\frac{\mu_1+\mu_2}{(2-2\cos(\tilde{\theta}+\epsilon^{1/2}\tilde{P}))^{1/2}}\right)$$

are of order greater than epsilon in the equation for $\dot{\rho}$.

Hence, to first order in epsilon, we have

$$H(\gamma,\theta,P,\Phi)=\frac{P^2}{2}+\frac{\gamma^2}{2}-\epsilon\frac{3}{2}\Phi^2-\epsilon^{a-1}\frac{\mu_1+\mu_2}{(2-2\cos(\theta))^{1/2}}.$$

We emphasize that this is the reduced system under the assumptions and restrictions of section 2.8 and that

- P is $O(\epsilon^{1/2})$,
- $b<(a-1)/3$, and $b<(2a-3)/4$.

3.3. Fixing the range for a and setting b. We note that in these variables, to lowest order in ϵ, we have the system

$$\begin{aligned}\dot{\gamma}&=P,\\ \dot{\theta}&=-3\epsilon\Phi,\\ \dot{P}&=-\gamma,\\ \dot{\Phi}&=-\frac{(\mu_1+\mu_2)\epsilon^{a-1}\sin\theta}{(2-2\cos\theta)^{3/2}}.\end{aligned}$$

This system decouples. The Hamiltonian for the θ,Φ system is

$$K(\theta,\Phi)=-\frac{3}{2}\epsilon\Phi^2-\frac{(\mu_1+\mu_2)\epsilon^{a-1}}{(2-2\cos\theta)^{1/2}}.$$

Using this expression, we see that solutions of the slow system are closed curves about $\theta=\pi,\Phi=0$. These correspond to the coorbital motions of the two moons. The angle θ, which equals the angle between the moons, oscillates strictly between 0 and 2π, while the angular velocity Φ alternates sign depending on which moon is closer to the planet. The oscillations of γ and P correspond to slight eccentricities of the Kepler orbits followed by the moons when they are far apart.

Hence we have found the coorbital solutions in the truncated system, and this provides a postiori justification of the assumptions necessary to obtain the system.

Moreover, if we consider the level set $K(\theta,\Phi)=k$, then on $\theta=\pi,\Phi=\Phi_0$ we have

$$K(\pi,\Phi_0)=k=-\epsilon\frac{3}{2}\Phi_0^2-\epsilon^{a-1}\frac{\mu_1+\mu_2}{2}\approx-\epsilon\frac{3}{2}\Phi_0^2.$$

The maximum, $\theta_{\max}$, and minimum, $\theta_{\min}$, values of θ on the level curve $K(\theta,\Phi)=k$ occur where $\Phi=0$, and so, to lowest order in ϵ,

$$\theta_{\min}\approx\epsilon^{a-2}\frac{2(\mu_1+\mu_2)}{3\Phi_0^2},$$

and similarly for $\theta_{\max}$.

This implies that the consistent choice for the parameter b is $b = a - 2$. Combining this with the a priori assumptions of section 2.8, we obtain the following: if $0 < b < 1/2$, then $2 < a < 5/2$, while if $1/2 \le b < 1$, then $2 < a < 8/3$. Adding the assumptions of section 3.2, namely, that $b < (a-1)/3$ and $b < (2a-3)/4$, and using $a > 2$, $b = a - 2$, we obtain $2 < a < 5/2$.

This completes the proof of the theorem stated in the introduction.

3.4. Analysis of the moons of Saturn. We next consider examples of coorbital motion observed in the moons of Saturn—first Janus and Epimetheus. The analysis above allows us to relate the three small quantities in coorbital motion, the difference in average radius of the orbit, the mass, and the angle of closest approach. The equation at the end of section 3.3 relates $\theta_{\min}, \mu_1, \mu_2$, and Φ_0 the value of Φ when $\theta = \pi$ to each other. Referring back to the system of section 3.1, we see that if we take the initial value of ρ (the separation of the radii) to be one, then nearly circular orbits will have $\rho \approx 2\Phi$, so we take $\Phi_0 = 1/2$. Substituting gives

$$\theta_{\min} \approx \frac{8}{3}\epsilon^{a-2}(\mu_1 + \mu_2) = \frac{8}{3}\epsilon^{-2}(m_1 + m_2),$$

where we use that the masses satisfy $m_i = \epsilon^a \mu_i$. This agrees to lowest order with the similar relationship found by Murray and Dermott [9].

While careful observation has accurately determined the orbits of Janus and Epimetheus (see Nicholson et al. [10]), the masses cannot be directly measured. In practice, masses of small bodies are determined by the effect these bodies have on other bodies. For Janus and Epimetheus, observation yields that the average radius of the orbits of the moons is 150,432 Km and the difference in the radii when they are far apart is approximately 50Km. This gives an epsilon value of $\epsilon \approx 3.32 \cdot 10^{-4}$. Consistent choices of masses and minimum separation angles which agree with observation are given by $m_1 + m_2 \approx 4.5 \cdot 10^{-9}$ and $\theta_{\min} \approx 6^\circ$. While our theorem specifies only the order of magnitude of the relationship between the masses and epsilon, the value $4.5 \cdot 10^{-9} \approx \epsilon^{2.4}$ is well within the range allowed, i.e., $\epsilon^2 \approx 1 \cdot 10^{-7}$ to $\epsilon^{2.5} \approx 2 \cdot 10^{-9}$.

Other groups of moons of Saturn have very close semimajor axes—for example, Dione and Helene and Thethys, Telesto, and Calypso (see Murray and Dermott [9]). However, the ratios of the masses of the smaller moon to the larger one in these cases is so small that it is more natural to treat these as perturbations of the restricted three body problem. One exception is the pair Prometheus and Pandora, which have approximately equal masses and radii 139,350 Km and 141,700 Km, respectively (data from Murray and Dermott [9, pp. 531–533]). For this pair, the average radius is 140,525 Km and

$$\epsilon = \frac{141,700 - 139,350}{140,525} \approx 0.0167.$$

The minimum mass ratio of these moons with Saturn necessary for them to exhibit coorbital motion is on the order of $0.0167^{5/2} = 3.6 \cdot 10^{-5}$. The actual mass ratio is $2.37 \cdot 10^{-10}$ and, as expected, these moons do not switch orbits. The F ring is between the orbits of these two moons. However, the interaction terms between these moons have a significant effect on their orbits as recent observations and simulations show (see Goldreich and Rapporport [3]).

4. Persistence of coorbital motion. Above we found, under appropriate assumptions on the parameters, orbits in a certain region of the phase space of the three body problem for a truncated approximation of the full system. Ideally we would like to show that there is a large set of such orbits which remains in the region for all time ("stability" of coorbital motion). We can, however, use ideas from KAM theory to at least discuss a lower bound on the minimum time solution's stay in the coorbital region. Here we assume the small moons are of equal mass ($\mu_1 = \mu_2 = 1$).

In section 3.1, we gave a projection of the full four degree of freedom system onto a two degree of freedom system. In the case in which $\mu_1 = \mu_2 = 1$, this projection can easily be altered and extended to form a symplectic transformation as follows:

$$\begin{array}{rclcrcl} \rho &=& \rho_2 - \rho_1, & & \bar{\rho} &=& \rho_2 + \rho_1, \\ \theta &=& \theta_2 - \theta_1, & & \bar{\theta} &=& \theta_2 + \theta_1, \\ R &=& (R_2 - R_1)/2, & & \bar{R} &=& (R_2 + R_1)/2, \\ \Phi &=& (\Phi_2 - \Phi_1)/2, & & \bar{\Phi} &=& (\Phi_2 + \Phi_1)/2. \end{array}$$

The new Hamiltonian is

$$H(\rho, \bar{\rho}, \theta, \bar{\theta}, R, \bar{R}, \Phi, \bar{\Phi}) = R^2 + \bar{R}^2 + \frac{\epsilon^2}{4}(\rho^2 + \bar{\rho}^2) + \epsilon^2(\Phi^2 + \bar{\Phi}^2 - \rho\Phi - \bar{\rho}\bar{\Phi}) - \epsilon^a \frac{1}{(2 - 2\cos(\theta))^{1/2}}.$$

As in section 3.2, we separate the fast and slow variables by setting

$$\begin{array}{rclcrcl} \sqrt{\epsilon}P &=& R, & & \sqrt{\epsilon}\bar{P} &=& \bar{R}, \\ \frac{1}{\sqrt{\epsilon}}\gamma &=& \rho, & & \frac{1}{\sqrt{\epsilon}}\bar{\gamma} &=& \bar{\rho} \end{array}$$

and rescale time to obtain

$$H(\gamma, \bar{\gamma}, \theta, \bar{\theta}, P, \bar{P}, \Phi, \bar{\Phi}) = P^2 + \bar{P}^2 + \frac{1}{4}(\gamma^2 + \bar{\gamma}^2) + \epsilon^{1/2}\gamma\Phi - \epsilon^{1/2}\bar{\gamma}\bar{\Phi} + \epsilon\Phi^2 + \epsilon\bar{\Phi}^2 - \epsilon^{a-1}\frac{1}{(2 - 2\cos(\theta))^{1/2}}.$$

Finally, we let

$$\begin{array}{rclcrcl} \tilde{\gamma} &=& \gamma - 2\epsilon^{1/2}\Phi, & & \tilde{\bar{\gamma}} &=& \bar{\gamma} - 2\epsilon^{1/2}\bar{\Phi}, \\ \tilde{\theta} &=& \theta - 2\epsilon^{1/2}P, & & \tilde{\bar{\theta}} &=& \bar{\theta} - 2\epsilon^{1/2}\bar{P}. \end{array}$$

To reduce excessive decoration, we remove the tildes, obtaining the Hamiltonian to first order in ϵ:

$$H(\gamma, \bar{\gamma}, \theta, \bar{\theta}, P, \bar{P}, \Phi, \bar{\Phi}) = P^2 + \bar{P}^2 + \frac{1}{4}(\gamma^2 + \bar{\gamma}^2) + 2\epsilon\Phi^2 + 2\epsilon\bar{\Phi}^2 - \epsilon^{a-1}\frac{1}{(2 - 2\cos(\theta))^{1/2}}.$$

We note that $\bar{\theta}$ is missing from the Hamiltonian. Hence

$$\dot{\bar{\Phi}} = 0.$$

This corresponds to the conservation of angular momentum. The conjugate variable $\bar{\theta}$ satisfies

$$\dot{\bar{\theta}} = 4\epsilon\bar{\Phi},$$

which corresponds to the constant rotation of the angle at which the closest approach occurs. (The outer moon takes slightly longer to traverse 360° than the inner moon, so the position of closest approach precesses.) We can remove these variables from consideration.

Also, we note that, to order one in epsilon, the unbarred and barred systems decouple. In fact, while there could be terms of order between one and $a-1$ in epsilon which we have not displayed above, the two systems actually decouple up to order $a-1$ in epsilon. (The changes of variables in section 3 do not mix barred with unbarred variables.)

4.1. Period of the slow system and a twist map. We can compute the approximate period of the slow unbarred system as follows: Letting T equal the period of the orbit containing the point $\theta = \pi, \Phi = \Phi_0$, we can use the symmetry of the system across the θ axis and the monotonicity of θ when Φ has constant sign to say that

$$T = 2\int_{\theta_{\min}}^{\theta_{\max}} \frac{dt}{d\theta}\, d\theta.$$

Substituting in the differential equation for $\dot{\theta}$, we obtain

$$T = 2\int_{\theta_{\min}}^{\theta_{\max}} \frac{1}{3\epsilon\Phi}\, d\theta,$$

where the relationship between θ and Φ is given by $K(\theta, \Phi)$ in section 3.3.

For the orbit with initial conditions $\theta = \pi, \Phi = \Phi_0$, we have $1/\Phi \approx 1/\Phi_0$ on most of the interval $(\theta_{\min}, \theta_{\max})$. Also, the values of $\theta_{\min}$ and $\theta_{\max}$ change at higher order in epsilon as Φ is changed, and hence the period T decreases as Φ_0 increases.

This implies that if we put the θ, Φ and γ, P variables into action-angle coordinates (to order one in epsilon) in the four dimensional system (decoupled from $\bar{\gamma}$, $\bar{P}$), then the return map in θ, Φ on a constant energy surface plane given by fixing γ and P satisfies a twist condition. Because the periods in θ, Φ and γ, P are order epsilon apart, the resulting map will be a "small twist." Luckily, both Poincaré's last geometric theorem and the KAM theorem apply in this situation (see Franks [2] and Levy and Moser [5]), providing periodic orbits and invariant circles for the return map and their perturbations.

4.2. Persistence of coorbital motion. Unfortunately, as we noted above, the barred and unbarred systems decouple only to order $a-1$ in epsilon. Hence invariant circles for the return map of the unbarred system correspond to separating invariant sets in the constant energy manifolds of the full system truncated to order $a-1$ in epsilon. We can go slightly beyond this order by looking at the terms we studied at the beginning of section 2.7 since these are the first terms involving both barred and unbarred variables. The term of lowest order is $\epsilon^{a-2b} = \epsilon^{4-a}$.

This implies that if we follow solutions of the full four degree of freedoms system, the value of the integral $K(\theta, \Phi)$ can change no faster than a constant times ϵ^{4-a}. Hence our coorbital solutions remain qualitatively the same for a time period of order $1/\epsilon^{4-a}$.

For Janus and Epimetheus, one unit of time corresponds to one orbit around Saturn or 0.694590 days (see Murray and Dermott [9]). With $\epsilon = 3.32 \cdot 10^{-4}$ and $a = 2.4$, we conclude that the orbits of these moons will remain qualitatively the same for at least on the order of $1/\epsilon^{1.6}$ orbits. This is on the order of 368,000 orbits or 700 years.

This ignores the possibility of large coefficients in the expansion; i.e., it cannot be taken literally—only as an order of magnitude estimate. The precise statement is only that the rate of change of $K(\theta, \Phi)$ is of order ϵ^{4-a} for orbits in the region of the phase space that we are considering.

Acknowledgments. The authors thank Tasso Kaper for his valuable suggestions and patient listening and the referees for useful suggestions.

REFERENCES

[1] K. Aksnes, *The tiny satellites of Jupiter and Saturn and their interactions with the rings*, in Stability of the Solar System and Its Minor Natural and Artificial Bodies, V. G. Szebehely, ed., D. Reidel, Boston, MA, 1985, pp. 3–16.

[2] J. Franks, *A variation on the Poincaré-Birkhoff theorem*, in Hamiltonian Dynamical Systems, Contemp. Math. 81, AMS, Providence, RI, 1988, pp. 111–117.

[3] P. Goldreich and N. Rappaport, *Chaotic Motions of F-Ring Shepherds*, http://www.arxiv.org/abs/astro-ph/0205330.

[4] J. Llibre and M. Ollé, *The motion of Saturn coorbital satellites in the restricted three-body problem*, Astron. Astrophys., 378 (2001), pp. 1087–1099.

[5] M. Levi and J. Moser, *A Lagrangian proof of the invariant curve theorem for twist mappings*, in Smooth Ergodic Theory and Its Applications, Proc. Sympos. Pure Math. 69, A. Katok, R. de la Llave, Y. Pesin, and H. Weiss, eds., AMS, Providence, RI, 2001, pp. 733–746.

[6] K. R. Meyer, *Symmetries and integrals in mechanics*, in Dynamical Systems, M. M. Peixoto, ed., Academic Press, New York, 1973, pp. 259–272

[7] K. R. Meyer, *Periodic Solutions of the N-Body Problem*, Lecture Notes in Math. 1719, Springer-Verlag, Berlin, 1999.

[8] 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.

[9] C. D. Murray and S. F. Dermott, *Solar System Dynamics*, Cambridge University Press, Cambridge, UK, 1999.

[10] P. D. Nicholson, D. P. Hamilton, K. Matthews, and C. Yoder, *New observations of Saturn's coorbital satellites*, Icarus, 100 (1982), pp. 464–484.

[11] H. Poincaré, *New Methods of Celestial Mechanics*, D. Goroff, ed., American Institute of Physics, New York, 1993.

[12] H. Salo and C. F. Yoder, *The dynamics of coorbital satellite systems*, Astron. Astrophys., 205 (1988), pp. 309–327.

[13] F. Spirig and J. Waldvogel, *The three-body problem with two small masses: A singular-perturbation approach to the problem of Saturn's coorbiting satellites*, in Stability of the Solar System and Its Minor Natural and Artificial Bodies, V. G. Szebehely, ed., D. Reidel, Boston, MA, 1985, pp. 223–234.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 2, No. 2, pp. 238–253

Predicting Irregularities in Population Cycles*

Shandelle M. Henson†, James R. Reilly‡, Suzanne L. Robertson§, Matthew C. Schu§, Eric W. D. Rozier§, and J. M. Cushing¶

Abstract. Oscillating population data often exhibit cycle irregularities such as episodes of damped oscillation and abrupt changes of cycle phase. The prediction of such irregularities is of interest in applications ranging from food production to wildlife management. We use concepts from dynamical systems theory to present a model-based method for quantifying the risk of impending cycle irregularity.

Key words. nonlinear population dynamics, stochasticity, periodic solutions, phase switching, basins of attraction

AMS subject classifications. 39A11, 92D25

PII. S1111111102411262

1. Introduction.

Animal numbers can oscillate periodically in many biological populations. Laboratory examples include paramecia [22], blowflies [33], bean weevils [37], moths [3], and flour beetles [6, 19].

Noise, always present in population dynamics, causes cycle irregularities such as outbreaks, switches in oscillation phase [28], and episodes of damped oscillations caused by "saddle fly-bys" [11] or other stochastic visitations of unstable equilibria. Phase switches (when, for example, an "up-down" time series pattern becomes a "down-up" pattern) occur rather frequently in laboratory populations of the flour beetle *Tribolium castaneum*. Figure 1.1 displays the larval numbers from control B of the experiment reported in Desharnais and Costantino [16]. The data oscillate with period two but switch cycle phase at times $t = 3$ and $t = 13$.

Cycle irregularities are important in a variety of applications ranging from food production to forest management to species conservation. In this paper, we use the concepts of dynamical systems theory to devise a method for predicting such irregularities in oscillating data.

In section 2, we use the univariate Ricker map to illustrate the dynamic mechanisms that give rise to cycle irregularities and to develop a method for predicting such irregularities. Section 3 applies the theoretical ideas to simulated data sets generated by two different multivariate models. In section 4 we use the method to predict cycle irregularities in experimental data replicates of laboratory cultures of the flour beetle *Tribolium*.

*Received by the editors July 16, 2002; accepted for publication (in revised form) by G. B. Ermentrout February 20, 2003; published electronically May 23, 2003. This work was supported in part by the National Science Foundation (Cushing, Henson), a Howard Hughes Medical Institute grant to the College of William and Mary (Henson, Reilly, Rozier, Schu), a U. S. Charles Center Summer Research Scholarship (Reilly), and a Verizon Scholarship (Robertson).
http://www.siam.org/journals/siads/2-2/41126,html
†Department of Mathematics, Andrews University, Berrien Springs, MI 49104 (henson@andrews.edu).
‡Department of Entomology, Cornell University, Ithaca, NY 14850 (jrr28@cornell.edu).
§College of William and Mary, Williamsburg, VA 23187 (slrobe@wm.edu, mcschu@wm.edu, ewdavi@wm.edu).
¶Department of Mathematics, University of Arizona, Tucson, AZ 85721 (cushing@math.arizona.edu).

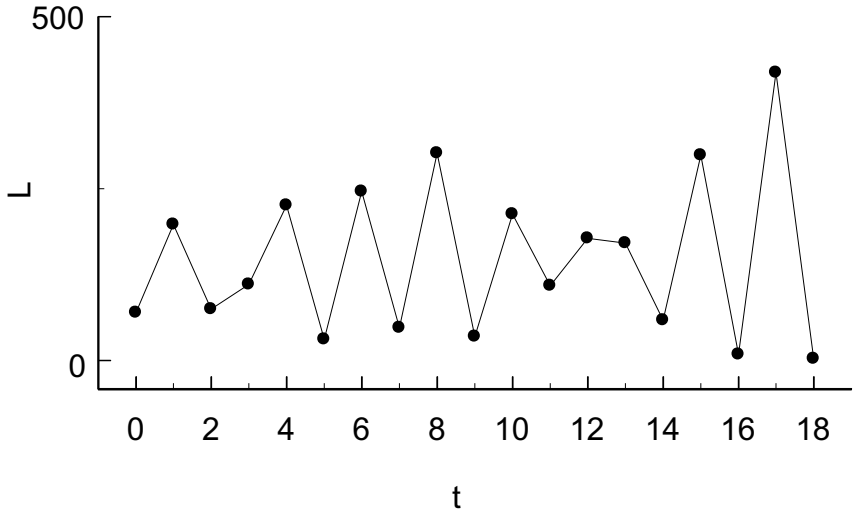


Figure 1.1. *Experimental data from Desharnais and Costantino* [16] *replicate B.* Tribolium castaneum *larval numbers oscillate in a* 2*-cycle. Phase switches occur at times* $t = 3$ *and* $t = 13$*. The time step is two weeks.*

2. Cycle irregularities. Consider a discrete time autonomous population model of the form $x_{t+1} = f(x_t)$ where x is a scalar, a vector of life stage classes, or a vector of interacting populations. To illustrate these ideas, consider specifically the scalar Ricker population model [31, 32, 34]

$$x_{t+1} = bx_t e^{-cx_t}. \tag{2.1}$$

The Ricker map (2.1) has stable periodic solutions at many values of its parameters. For example, when $b = 9$ and $c = 1$, the Ricker map has a stable 2-cycle solution given by $x_0 = 1.099$, $x_1 = 3.296$, $x_2 = 1.099$, $x_3 = 3.296, \ldots$. Since (2.1) is autonomous, the phase shift $x_0 = 3.296$, $x_1 = 1.099$, $x_2 = 3.296$, $x_3 = 1.099$, ... is also a stable 2-cycle solution. Denote the first cycle "Phase 1" and the second "Phase 2." Let B_1 be the set of positive initial conditions $x_0 \in R^+$ that give rise to solutions converging to the Phase 1 cycle, and let B_2 be the set of positive initial conditions that give rise to solutions converging to Phase 2. These "basins of attraction" for the two cycle phases are sets on the real line. Some of the numerically computed boundaries separating these phase basins are shown in Figure 2.1(a). Initial conditions x_0 with $4.840 < x_0 < 7.425$ converge to Phase 1; initial conditions with $2.198 < x_0 < 4.840$ converge to Phase 2; those with $0.345 < x_0 < 2.198$ converge to Phase 1; those with $0.040 < x_0 < 0.345$ converge to Phase 2; and so forth. (Near the origin, the basin structure becomes complicated and cannot be shown in Figure 2.1(a).) Note that the unstable equilibrium $x_u = 2.198$ lies on the basin boundary. The reader should also note that

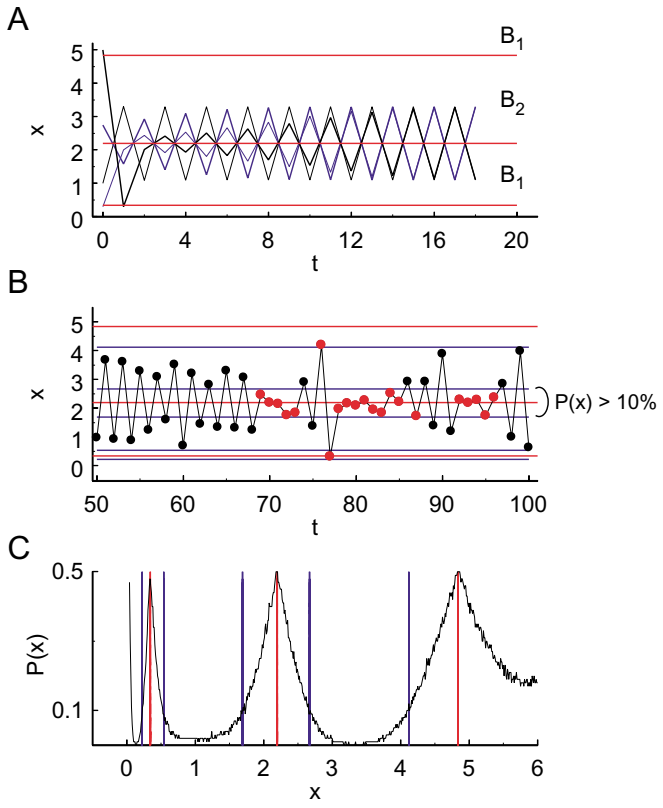


Figure 2.1. *Ricker map simulations with $b = 9$ and $c = 1$.* (a) *Four deterministic Ricker time series with initial conditions $x_0 = 0.3$, $x_0 = 1$, $x_0 = 2.75$, and $x_0 = 5$. Red horizontal lines represent basin boundaries at $x = 0.3445$, $x = 2.198$, and $x = 4.840$. Since $x_0 = 1$ and $x_0 = 5$ are both in basin B_1, their orbits converge to Phase* 1 *(shown in black). Since $x_0 = 0.3$ and $x_0 = 2.75$ are both in basin B_2, their orbits converge to Phase* 2 *(shown in blue).* (b) *Stochastic Ricker simulation with $\sigma^2 = 0.0225$. Red horizontal lines represent basin boundaries as in* (a). *The red lines are bracketed by blue horizontal lines that demarcate the* 10% *zones. Points on the orbit which lie in the* 10% *zones are colored red and have at least a* 10% *probability of switching phase at the next time step.* (c) *The probability $P(x)$ of switching phase at the next time step given the current value of x. Red vertical lines represent values of x at which $P(x) = 50\%$ and correspond to the basin boundaries. Blue vertical lines show values of x at which $P(x) = 10\%$ and correspond to the boundaries of the* 10% *zones.*

the term "basin" is being used here to differentiate two separate basins of attraction for two different solutions lying on the same 2-cycle attractor, while the usual meaning lumps B_1 and B_2 together as a single basin of attraction for the 2-cycle attractor itself. (To view B_1 and B_2 as basins in the traditional sense, we could use the composite map to decompose the 2-cycle attractor into two separate stable equilibria with two separate basins of attraction. However, this complication is unnecessary for our purposes.)

If $x_0 \in B_1$ so that the solution converges to Phase 1, then taking $f(x_0)$ as an initial condition would lead to a solution converging to Phase 2; hence $x_1 = f(x_0) \in B_2$. Similarly,

if $x_0 \in B_2$, then $x_1 \in B_1$. In this way, deterministic orbits must always bounce back and forth between the two basins at each time step as they converge to the appropriate phase of the 2-cycle (Figure 2.1(a)).

Process noise in ecological data is of two basic types: environmental and demographic. Environmental noise is additive on the log scale, while demographic noise is additive on the square root scale [15]. The methods in this paper work equally well for both kinds of noise; however, we will focus on demographic stochasticity. Thus we can incorporate demographic noise by means of

$$x_{t+1} = \left(\sqrt{bx_t e^{-cx_t}} + E_t\right)^2, \tag{2.2}$$

where E_t is a random normal variable with mean zero and variance σ^2 (see Dennis et al. [15]). When $\sigma = 0$, we recover the "deterministic skeleton" (2.1) [36]. As σ increases from zero, the cycles become irregular. In particular, orbits may not always alternate between basins at each time step. If a stochastic orbit remains in the same basin for two consecutive time steps, we say a "phase switch" has occurred. For example, in the stochastic simulation shown in Figure 2.1(b), the orbit is in Phase 2 for $t = 86$ to 91 but in Phase 1 for $t = 97$ to 100. On examination of the basins in which the $t = 92$ to 96 values lie, we see that the orbit switches phase at $t = 97$, since x_{96} and x_{97} are both in B_2.

The conditional probability $P(x)$ of switching phase at the next time step, given a current value of x, is graphed in Figure 2.1(c). $P(x)$ is lowest at the 2-cycle values $x = 1.099$ and $x = 3.296$ and is highest along the basin boundary. $P(x)$ can be considered a kind of measure of how far a point is from the deterministic attractor or, alternately, how close a point is to the basin boundary.

Figures 2.1(b) and (c) show the numerically computed values of x for which $P(x) = 10\%$. These values bracket the basin boundaries; and when x falls inside these "10% zones," there is at least a 10% chance of suffering a phase switch at the next time step. For example, $P(x) > 10\%$ whenever $1.691 < x < 2.669$.

For the stochastic orbit in Figure 2.1(b), all values x_t for which $P(x_t) \geq 10\%$ are colored red. Note how the red tagging provides a fairly good warning of impending cycle irregularities. For example, at $t = 69$, the orbit falls into a 10% zone, and the following values for $t = 70$ to 73 are highly irregular. At $t = 76$ the orbit again enters the 10% zone, and irregularities follow for $t = 77$ to 85. At $t = 92$ the orbit once more lands in the 10% zone, and irregularities, including the aforementioned phase switch at $t = 97$, follow.

Note how the orbit sometimes lingers near the unstable equilibrium $x_u = 2.198$ if stochastically bumped into its vicinity (for example, at times $t = 70, 79, 92$). This phenomenon is similar to a "saddle fly-by," although the unstable equilibrium here is a repellor rather than a saddle. In a "saddle fly-by," the orbit is stochastically bumped near the stable manifold of a saddle and approaches the saddle before moving away. In a "repellor visitation," the orbit is stochastically knocked directly into the neighborhood of a repellor or maps in from a nonlocal stable set and then lingers before moving away. In cycling data, saddle fly-bys and repellor visitations of unstable equilibria typically give rise to episodes of damped oscillation [11].

To summarize this example, we suggest an empirical "rule of thumb":

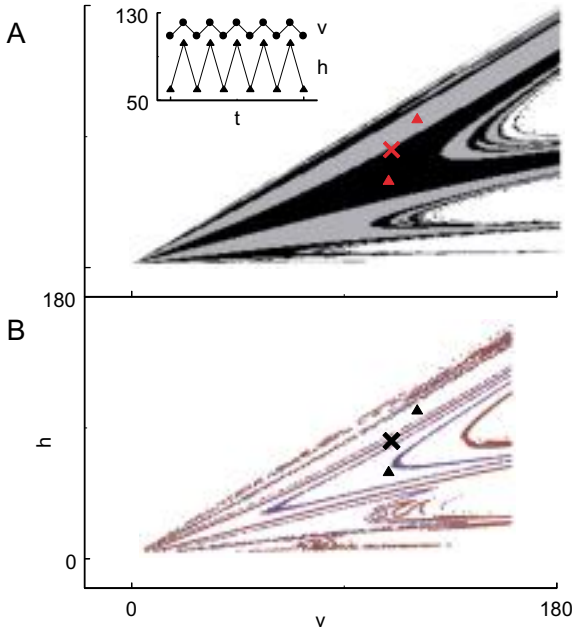


Figure 3.1. *Leaf-herbivore model simulations with* $f = 1.5$, $a = .005$, $r = 3.8$, *and* $\delta = 1.0$. (a) *Basins of attraction for the two phases of the attracting* 2*-cycle. The basins were determined by iterating points in state space (on a grid of mesh* $1/10$*) forward* 200 *time steps with the deterministic model. The red triangles are the* 2*-cycle values* $(v, h) = (108.76, 59.94)$, $(120.9, 102.25)$. *The unstable saddle equilibrium* $(v, h) = (110, 81)$ *is indicated by a red "x." The* 2*-cycle attractor is shown above this plot as a time series.* (b) *Approximation of the* 50% *(red) and* 10% *(blue) contour lines for* $P(v, h)$. *These curves were computed by finding the frequency of phase switching at the next step for points on a grid of mesh* $1/10$ *in state space. Each point tested was iterated one time step forward,* 2000 *times, with* $\sigma_v^2 = .008$ *and* $\sigma_h^2 = .032$. *The set of all initial conditions with* $9.75\% - 10.25\%$ *probability of switching phase on the next step is plotted in blue. The set of all initial conditions with* $49.5\% - 50.5\%$ *probability of switching phase on the next step is plotted in red. The deterministic* 2*-cycle values are indicated by triangles and the unstable equilibrium by an "x."*

Remark 2.1. *The conditional probability* $P(x)$ *of switching phase at the next time step, given a current data point* x*, can be considered a measure of how close the point* x *is to the basin boundary. The set of points for which* $P(x) \geq 10\%$ *brackets the basin boundaries. When the current data point falls within these "*10% *zones," cycle irregularities are likely to follow. The choice of* 10% *is subjective and depends on the acceptable risk level in the particular application. A more conservative approach would use a smaller contour value (and hence a wider zone around the basin boundaries).*

3. Theoretical examples in population models. In this section, we use two examples to illustrate how cycle irregularities occur in multivariate models and in cycles of periods other than 2. The first example is a two-species model with a 2-cycle, and the second is a single species structured model with a 4-cycle.

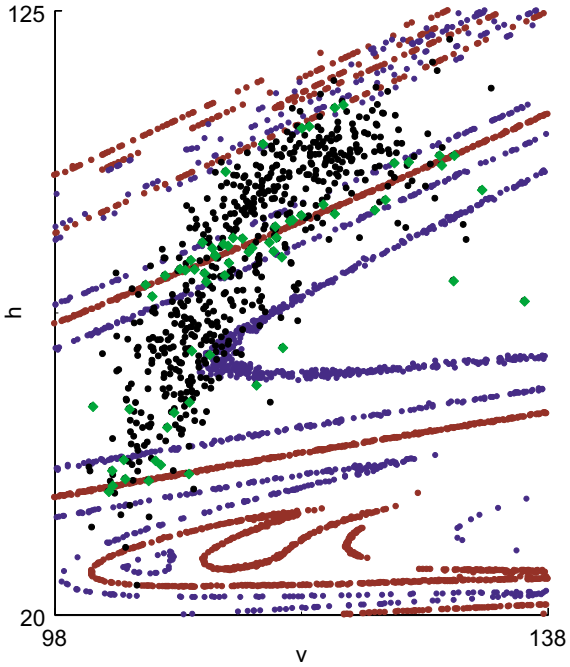


Figure 3.2. *Stochastic orbit of the leaf-herbivore model with* $\sigma_v^2 = .008$ *and* $\sigma_h^2 = .032$. 600 *time steps are plotted in a close-up of state space. Red curves approximate basin boundaries; blue curves bracket the basin boundaries and demarcate the* 10% *zones. Points of the orbit that suffered a phase switch at the next step are colored green; note that most lie within the* 10% *zones, as expected.*

3.1. Leaf-herbivore model. Consider a leaf-herbivore model of Edelstein-Keshet [20],

$$\begin{aligned} v_{t+1} &= f e^{-a h_t} v_t, \\ h_{t+1} &= r\left(\delta - \frac{h_t}{v_t}\right) h_t, \quad v_t \neq 0, \end{aligned} \tag{3.1}$$

where h_t is the number of herbivores at time t on a tree of leaf mass v_t and f, a, r, and δ are positive constants. When $f = 1.5$, $a = 0.005$, $r = 3.8$, and $\delta = 1$, this model has a stable 2-cycle attractor, both phases of which are stable solutions. The basins of attraction for the two phases are shown in Figure 3.1(a), along with the 2-cycle values and an unstable saddle equilibrium lying on the basin boundary.

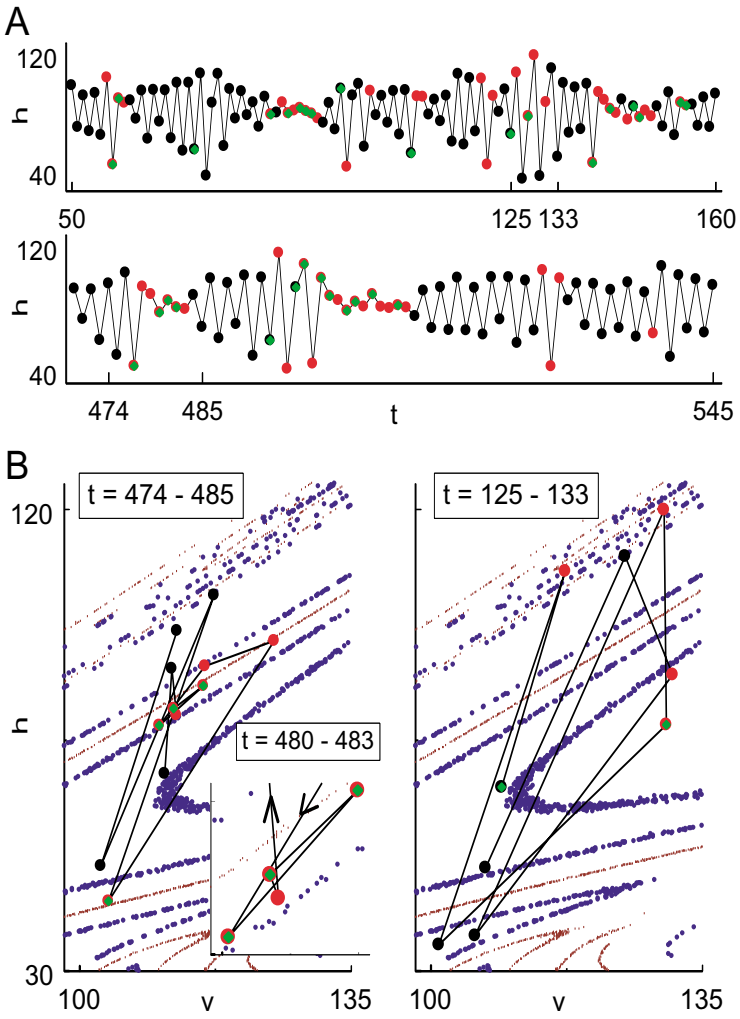


Figure 3.3. *Stochastic orbit of the leaf-herbivore model with* $\sigma_v^2 = .008$ *and* $\sigma_h^2 = .032$. *Points on the orbit that fall within the* 10% *zones are colored red. Points that suffered a phase change at the next time step are indicated by a green diamond.* (a) *h component for two sections of the stochastic time series shown in Figure* 3.2. *Two typical kinds of cycle irregularities are contained in sections* $t = 125$ *to* 133 *and* $t = 474$ *to* 485. *In the first case, the oscillations are irregular (down-up-up) with large amplitude, while in the second case the oscillations are irregular and damped.* (b) *State space close-ups of two sections of the time series in* (a). *For* $t = 474$ *to* 485, *the orbit starts out regularly, following the deterministic* 2*-cycle. Beginning at* $t = 477$, *the orbit lands near the basin boundary, then switches phase, and lands near the basin boundary again. From there the orbit follows the stable manifold toward the unstable equilibrium for a saddle fly-by before returning to the* 2*-cycle. A close-up of the saddle fly-by is shown for* $t = 480$ *to* 483. *The panel for* $t = 125$ *to* 133 *shows how the orbit is knocked far from the* 2*-cycle and visits regions of state space in which the basins are more complicated.*

Consider the stochastic model

$$v_{t+1} = \left(\sqrt{f e^{-a h_t} v_t} + E_{1t}\right)^2, \tag{3.2}$$
$$h_{t+1} = \left(\sqrt{r\left(\delta - \frac{h_t}{v_t}\right) h_t} + E_{2t}\right)^2, \quad v_t \neq 0,$$

where the random vector $\mathbf{E}_t = (E_{1t}, E_{2t})'$ is assumed to have bivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}$ [15].

Given that the stochastic system (3.2) is at (v_t, h_t) in state space, one can numerically compute the probability $P(v_t, h_t)$ that the next stochastic step (v_{t+1}, h_{t+1}) will be a phase switch. Figure 3.1(b) shows two numerically computed contours in state space along which $P(v, h)$ is constant. Along the blue contours, $P(v, h) = 10\%$, and along the red contours, $P(v, h) = 50\%$. Note that the 50% contours follow the basin boundaries, while the 10% contours roughly parallel and "bracket" the boundaries. We will call these bracketing sets "10% zones."

Figure 3.2 shows a stochastic run of 600 time steps in a close-up of the state space plot shown in Figure 3.1(b). The orbit points which suffer a phase switch at the next step are green. Note that most of the green points lie within the 10% zones. Indeed, less than 10% of all points landing outside the 10% zones are expected to give rise to phase switches at the next step. By their definition, the 10% zones give a fairly accurate indication of impending phase switches.

Do the 10% zones also provide a good indicator for the onset of other types of cycle irregularities? In Figure 3.3, we consider two sections of the stochastic time series of Figure 3.2. Orbit points landing within the 10% zones are colored red, while those leading to phase switches are also marked by a green diamond. Two typical kinds of cycle irregularities appear in the sections $t = 125$ to 133 and $t = 474$ to 485 (Figure 3.3(a)). In the first case the time series oscillations are irregular ("down-up-up") with large amplitude, while in the second case the oscillations are irregular and damped. In Figure 3.3(b), we take a closer look in state space at the mechanics of these time series irregularities. The section beginning with $t = 474$ starts off with a fairly regular pattern, following the deterministic 2-cycle. At $t = 477$, the orbit lands near the basin boundary within the 10% zone, then switches phase, and lands near the basin boundary again. From there the orbit follows the stable manifold toward the unstable equilibrium and lingers for a saddle fly-by before returning to the 2-cycle. The section beginning with $t = 125$ starts with a point outside the 10% zone but which in fact leads to an unpredicted phase switch. The next step ($t = 126$) is, however, inside the 10% zone, even though in the time series it appears rather unremarkable (Figure 3.3(a)). Although there is no phase switch at the next step, there is an upcoming irregularity. The orbit is stochastically knocked down across three basin boundaries and then to the right into a 10% zone, after which it switches phase and lands far from the attractor. A similar pattern follows until the orbit lands outside the 10% zone fairly close to the attractor at $t = 133$, after which the pattern becomes more regular.

In both of these cases of irregularity, a 10% zone visitation (at $t = 477$ and $t = 126$) served as a warning of the upcoming cycle disturbance. Furthermore, in both cases, the

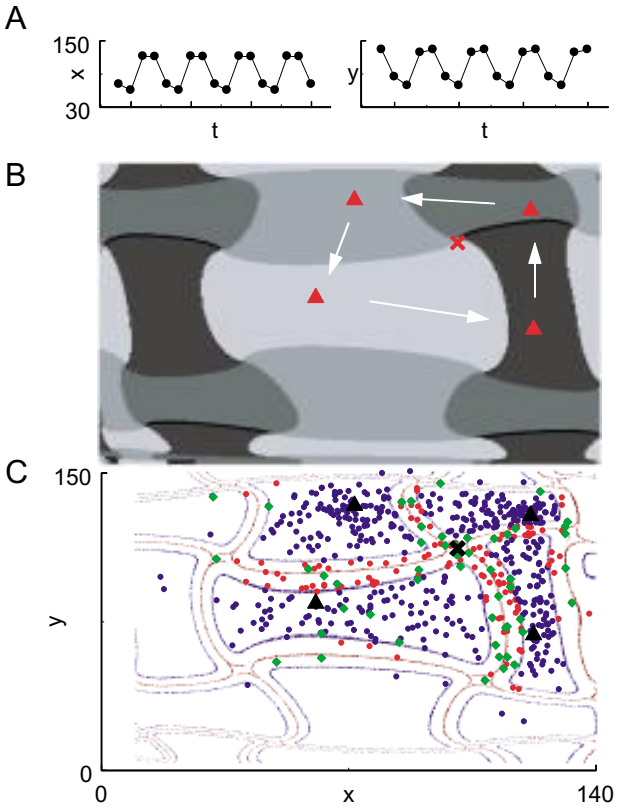


Figure 3.4. *Bartlett model simulations with* $\alpha = 0.02$, $b = 3.14$, *and* $\mu = 0.9$. (a) *Time series for Bartlett model* 4*-cycle.* (b) *Four basins of attraction corresponding to the four phases of the attracting* 4*-cycle. The* 4*-cycle values, indicated by red triangles, are* $(x, y) = (60.67, 85.09)$, $(122.38, 69.18)$, $(121.51, 129.30)$, $(71.64, 133.44)$. *White arrows show the temporal sequence. The red "x" marks the unstable equilibrium at* $(x, y) = (100.80, 112.00)$. (c) 50% *and* 10% *contour lines of* $P(x, y)$ *for the stochastic Bartlett model are represented by red and blue curves, respectively. Here* $\sigma_x^2 = 0.045$ *and* $\sigma_y^2 = 0.025$. *To compute the phase switching frequency,* 2000 *one-step stochastic predictions were made from each point on a grid of mesh* 1/10. *Also shown is a* 600 *step stochastic orbit. Points landing within the* 10% *zone are colored red. Orbit points that switched phase at the next step are indicated by green diamonds. Black triangles mark the values of the deterministic* 4*-cycle, while a black "x" identifies the unstable equilibrium.*

crucial orbit point (at $t = 477$ and $t = 126$) would appear unremarkable in the unfolding time series (Figure 3.3(a)) unless one had knowledge of its proximity in state space to the basin boundary as measured by the 10% zone or some other appropriate measure.

3.2. Bartlett model. In this example, we consider a well-known juvenile-adult model of Bartlett [1]. A stochastic version is

$$x_{t+1} = \left(\sqrt{(b - \alpha y_t)\, y_t} + E_{1t}\right)^2, \tag{3.3}$$

$$y_{t+1} = \left(\sqrt{x_t + (1 - \mu)\, y_t} + E_{2t}\right)^2,$$

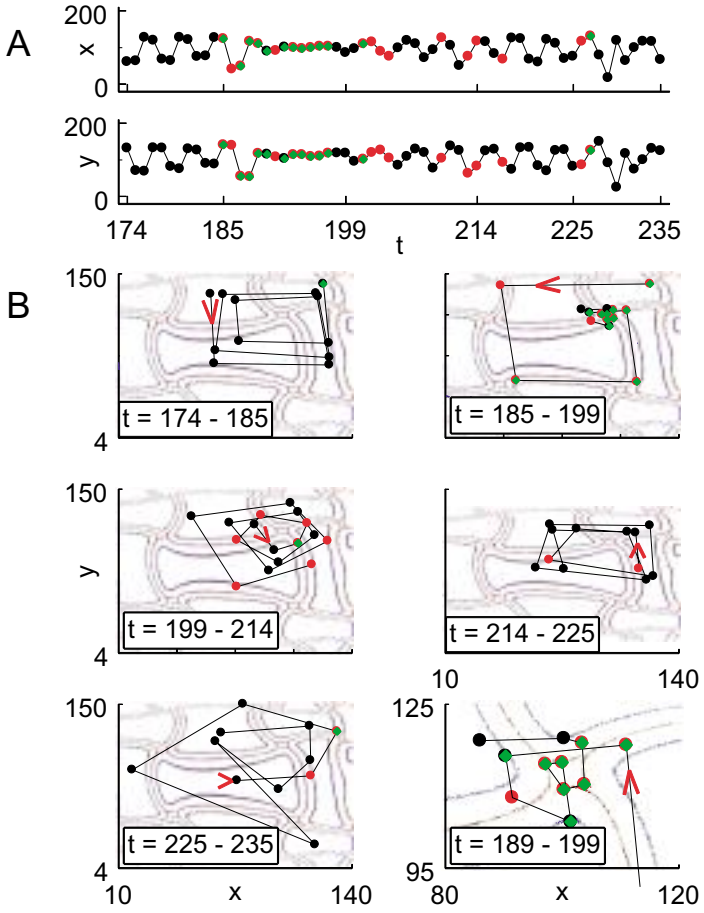


Figure 3.5. *Stochastic simulation of Bartlett model with* $\alpha = 0.02$, $b = 3.14$, $\mu = 0.9$, $\sigma_x^2 = 0.045$, *and* $\sigma_y^2 = 0.025$ *from Figure* 3.4(c). *Orbit points falling within the* 10% *zone are colored red. Green diamonds indicate points that lead to a phase switch at the next time step.* (a) *The time series is divided into five sections of interest (see text).* (b) *State space graphs of the five sections of the stochastic time series in* (a). 50% *and* 10% *contour lines for* $P(x, y)$ *are shown in red and blue, respectively. Orbit points are shown in temporal sequence, with the direction of motion indicated by red arrows.* $t = 174$ *to* 185 *shows a fairly regular section, but we see in state space that the last orbit point is within a* 10% *zone and in fact leads to a phase switch at the next step. A period of irregularity follows, as shown in the second panel for* $t = 185$ *to* 199. *For* $t = 185$ *to* 188, *the time series still appears regular, but we see from state space that the orbit is oscillating near the basin boundary. This leads to a visitation of the unstable equilibrium beginning at* $t = 189$, *and the orbit lingers there before moving away. The move away is shown in the third panel with* $t = 199$ *to* 214 *as the system regains regularity. The fourth panel,* $t = 214$ *to* 225, *is fairly regular, deviating minimally from the* 4*-cycle pattern. The fifth panel of the time series,* $t = 225$ *to* 235, *shows the stochastic orbit landing inside the* 10% *zone at* $t = 226$, *and again at* $t = 227$, *after which a phase change occurs. With the phase change comes a new period of irregularity, as the stochastic orbit is knocked far from the attractor before finally returning to the* 4*-cycle for the last few time steps. The last panel shows a close-up of the equilibrium visitation for* $t = 189$ *to* 199. *The unstable equilibrium is located at* $(v, h) = (100.80, 112.00)$, *within the diamond created by the* 50% *contour lines.*

where x_t and y_t are the numbers of juveniles and adults at time t, respectively, and b, α, and μ are positive constants with $0 < \mu < 1$. Here the random vector $\mathbf{E}_t = (E_{1t}, E_{2t})'$ is assumed to have bivariate normal distribution with mean vector $\mathbf{0}$. The deterministic Bartlett model obtains for $\mathbf{E}_t$ identically $\mathbf{0}$. When $b = 3.14$, $\alpha = 0.02$, and $\mu = 0.9$, the deterministic Bartlett model has a stable 4-cycle attractor [6]. See Figure 3.4(a). The 4-cycle attractor corresponds to four different stable solutions, one for each phase of the cycle. The corresponding four basins of attraction, along with the four values on the stable cycle, are shown in Figure 3.4(b). A repelling equilibrium exists at the vortex of the pinwheel of basin boundaries.

When noise is incorporated into the Bartlett model (3.3), the values of $P(x, y)$ can be computed numerically. Figure 3.4(c) shows 10% contours (blue) and 50% contours (red) in state space. While the 50% contours roughly follow the basin boundaries, there are regions (for example, inside the diamond around the repellor formed by the 50% contour) for which $P(x, y) > 50\%$. A 600 step stochastic simulation also is plotted in Figure 3.4(c). The majority of orbit points that give rise to a phase switch at the next time step (green diamonds) do fall within the 10% zone, as expected.

The 10% zones also warn of other types of cycle irregularities. Figure 3.5 shows a segment of the stochastic simulation in Figure 3.4(c), both as a time series and in state space. Two main periods of cycle irregularity occur during this segment. (A long period of damped oscillations begins at $t = 189$, and an "up-up-down-down" irregularity with large amplitude begins at $t = 227$.) Initially, the segment tends to follow the deterministic 4-cycle fairly closely. However, at $t = 185$ the orbit lands in a 10% zone, and a long episode of irregularity follows. Note that for $t = 185$ to 188, the time series in Figure 3.5(a) still appears regular, but we see from state space in Figure 3.5(b) that the orbit is oscillating near the basin boundary. This leads to a visitation of the unstable equilibrium beginning at $t = 189$, and the orbit lingers there for ten more time steps before moving away for $t = 199$ to 214. For $t = 214$ to 225, the system is fairly regular, deviating minimally from the 4-cycle pattern. The stochastic orbit lands inside the 10% zone at $t = 226$ and again at $t = 227$, after which a phase change occurs. With the phase change comes another period of irregularity, as the stochastic orbit is knocked far from the attractor before finally returning to the 4-cycle for the last few time steps of the segment. Note that the crucial orbit points themselves at $t = 185$ and $t = 226$ would seem unremarkable in the unfolding time series (Figure 3.5(a)) unless one realized they were sufficiently close to the basin boundaries to be in the 10% zones. A recognition, however, that they are in the 10% zones warns of the two impending episodes of cycle disruption.

4. Theory applied to experimental data. In this section, we apply our proposed rule of thumb for predicting cycle irregularities to experimental data.

The discrete stage-structured "LPA" *Tribolium* model has successfully explained and predicted nonlinear phenomena in a variety of contexts, including transitions between dynamic regimes (such as equilibria, 2-cycles, 3-cycles, invariant loops, and chaos), multiple attractors, saddle influences, stable and unstable manifolds, and lattice effects [4, 7, 5, 10, 8, 11, 12, 9, 13, 14, 15, 17, 18, 28, 24, 27, 25, 26, 30]. We now use the LPA model and the methods proposed in this paper to predict the cycle irregularities observed by Desharnais and Costantino [16].

Noise in laboratory cultures of flour beetles is mainly demographic [15]. A version of the

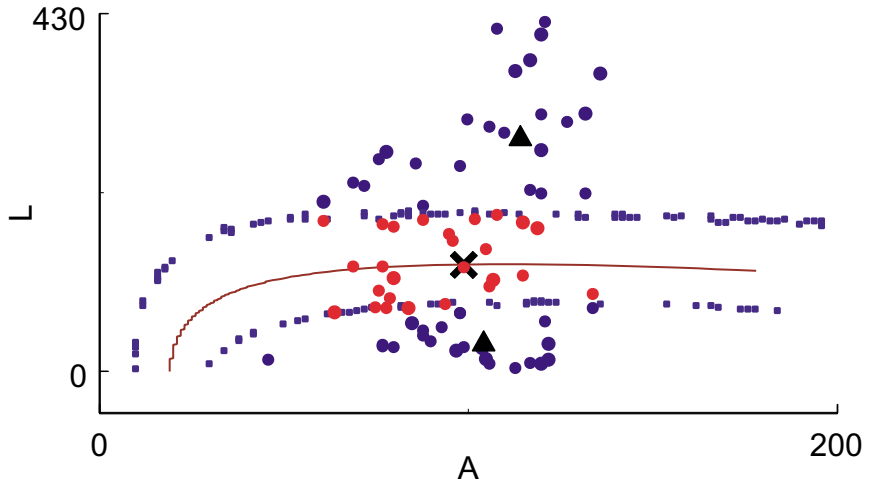


Figure 4.1. *Experimental data shown in cross section of state space at $P = 30$ for LPA model. The red curve shows the location of the basin boundary as calculated by the deterministic model. The blue curves bracket the 10% zones. Here $\sigma_l = 3.848$, $\sigma_p = 2.962$, and $\sigma_a = .2401$. Experimental data control replicates A,B,C, and D from Desharnais and Costantino [16] are projected onto the plane $P = 30$. Data points which fall within the 10% zone are shown in red. Projections of the 2-cycle coordinates $(L, P, A) = (35.26, 136.26, 104.14)$, $(282.24, 17.03, 114.08)$, and the unstable equilibrium $(L, P, A) = (128.03, 61.83, 98.74)$ are indicated by black triangles and a black "x," respectively.*

LPA model incorporating demographic stochasticity is

$$
\begin{aligned}
L_{t+1} &= \left(\sqrt{bA_t \exp\left(-c_{el}L_t - c_{ea}A_t\right)} + E_{1t}\right)^2, \\
P_{t+1} &= \left(\sqrt{(1-\mu_l)\,L_t} + E_{2t}\right)^2, \\
A_{t+1} &= \left(\sqrt{P_t \exp\left(-c_{pa}A_t\right) + (1-\mu_a)\,A_t} + E_{3t}\right)^2,
\end{aligned}
\tag{4.1}
$$

where L denotes the number of (feeding) larvae, P denotes the number of pupae (nonfeeding larvae, pupae, and callow adults), and A denotes the number of (mature) adults. The discrete time interval is two weeks. The coefficient $b > 0$ denotes the average number of larvae recruited per adult per unit time in the absence of cannibalism, μ_l and μ_a are the larval and adult per unit time probabilities of dying from causes other than cannibalism, and the exponentials represent the probabilities that individuals survive cannibalism one unit of time, with "cannibalism coefficients" c_{el}, c_{ea}, $c_{pa} > 0$. The random vector $\mathbf{E}_t = (E_{1t}, E_{2t}, E_{3t})'$ is assumed to have trivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{\Sigma}_{LPA}$. $\mathbf{E}_0, \mathbf{E}_1, \ldots$ are assumed to be uncorrelated. The deterministic skeleton ($\mathbf{\Sigma}_{LPA} = \mathbf{0}$) of model (4.1) is the deterministic LPA model. Local stability results for the LPA model are obtained using standard linearization techniques [8, 23].

The conditioned least squares parameter estimates from the four control replicates reported in Desharnais and Costantino [16] were $b = 8.913$, $c_{el} = 0.008446$, $c_{ea} = 0.008572$,

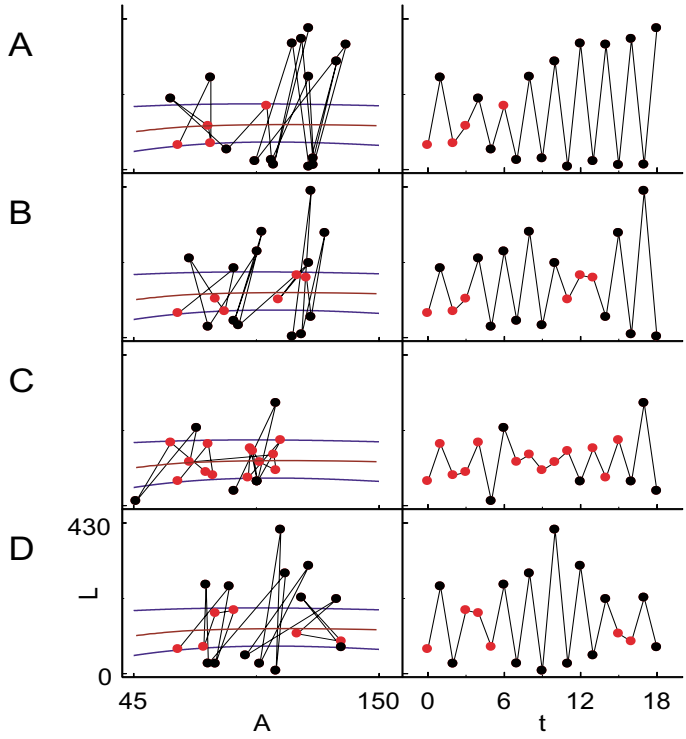


Figure 4.2. *Experimental data control replicates A,B,C, and D from Desharnais and Costantino* [16]. *The panels on the left show the cross-section of LPA state space at* $P = 30$. *Red and blue curves indicate the basin boundary and the* 10% *contours, respectively. Red data points are those within the* 10% *zones. The time series (*L *versus* t*) of the replicates are shown on the right. The time series illustrates how most cycle irregularities can be predicted by the* 10% *zone rule. Note the saddle fly-by in replicate C.*

$\mu_l = 0.5171$, $c_{pa} = 0.01795$, and $\mu_a = 0.1064$, with

$$\mathbf{\Sigma}_{LPA} = \begin{pmatrix} 3.848 & 0.3665 & 0.1440 \\ 0.3665 & 2.962 & -0.5895 \\ 0.1440 & -0.5895 & 0.2401 \end{pmatrix}.$$

At these parameter values, the deterministic LPA model admits an unstable saddle point (rounded to the nearest beetle) of $[128, 62, 99]$. The LPA model also predicts a 2-cycle attractor and hence two stable 2-cycle solutions (one for each phase of the cycle). The basins of attraction for these phases, along with the replicate data, are shown in Figure 4.1. The red curve is the basin boundary, and the blue curves are the 10% contours. The red data points lie within the 10% zone. The two values of the 2-cycle are shown, as well as the saddle, which lies on the basin boundary.

Figure 4.2 presents each replicate both in state space and as a time series. From the time series, it is clear that all cycle irregularities were preceded by a data point falling into the 10% zone. Of course, not all visitations of the 10% zone actually led to cycle disturbances.

5. Summary and discussion. Populations often exhibit temporal cycles, but as a result of noise, the oscillations may display irregularities in phase and amplitude. Understanding and anticipating such irregularities is of great importance in applications from pest control and species conservation to physiology and epidemiology [35]. In this paper, we have proposed a method for understanding and predicting cycle irregularities. The method is model-based. It requires a deterministic model for the population dynamics, together with a stochastic version of the model that describes random deviations from the deterministic predictions. If the deterministic model is autonomous, each phase of the cycle attractor is a stable solution. For each phase there is a basin of attraction, and these basins are separated by basin boundaries which may contain unstable equilibria. In general, the closer to the basin boundary a data point falls, the higher the probability it will stochastically switch phase at the next time step. Amplitude-dampened oscillations, caused by saddle fly-bys or repellor visitations, also occur when the data fall near the basin boundary, since that is where the unstable equilibria lie and exert the most influence. The probability $P(x)$ of phase switching at the next time step, given the current position x, can be computed with the stochastic model. "Warning zones" can be constructed in state space by computing contour lines along which $P(x)$ is constant. These zones follow and contain the basin boundaries. When data fall within the warning zones, the observer is informed that a cycle irregularity is likely to occur and can act accordingly. In this paper, we have defined the warning zones using the 10% contours, i.e., where a greater than 10% chance of phase switching is predicted. A more conservative approach would use a smaller contour value (and hence a wider zone around the basin boundaries). The actual contour value chosen is subjective and depends on the acceptable risk level in the particular application.

Dynamical systems approaches are powerful tools for understanding biological systems. Situation specific applications of model-based methods, including the one presented in this paper, require constructing a model in whose accuracy one has confidence. During the last few decades, rigorous connection of mathematical models with population data has become possible for several laboratory systems (see, for example, Bjørnstad and Grenfell [2], Cushing et al. [9], Fussmann et al. [21], and the references cited in section 4), but quantitatively accurate models for field populations are still rare [29]. Until reliable models become more common and established in ecology, there will be a need for nonparametric methods. An interesting and important open question is whether nonparametric versions of the method presented in this paper can be constructed when good data but no structural models are available.

Acknowledgments. We thank R. F. Costantino, Brian Dennis, and R. A. Desharnais as well as two anonymous referees.

REFERENCES

[1] M. S. Bartlett, *Stochastic Population Models*, Wiley, New York, 1960, p. 50.

[2] O. N. Bjørnstad and B. T. Grenfell, *Noisy clockwork: Time series analysis of population fluctuations in animals*, Science, 293 (2001), pp. 638–643.
[3] C. J. Briggs, S. M. Sait, M. Begon, D. J. Thompson, and H. C. J. Godfray, *What causes generation cycles in populations of stored-product moths?*, J. Animal Ecol., 69 (2000), pp. 352–366.
[4] R. F. Costantino, J. M. Cushing, B. Dennis, and R. A. Desharnais, *Experimentally induced transitions in the dynamic behavior of insect populations*, Nature, 375 (1995), pp. 227–230.
[5] R. F. Costantino, J. M. Cushing, B. Dennis, R. A. Desharnais, and S. M. Henson, *Resonant population cycles in alternating habitats*, Bull. Math. Biol., 60 (1998), pp. 247–273.
[6] R. F. Costantino and R. A. Desharnais, *Population Dynamics and the* Tribolium *Model: Genetics and Demography*, Springer-Verlag, New York, 1991.
[7] R. F. Costantino, R. A. Desharnais, J. M. Cushing, and B. Dennis, *Chaotic dynamics in an insect population*, Science, 275 (1997), pp. 389–391.
[8] J. M. Cushing, R. F. Costantino, B. Dennis, R. A. Desharnais, and S. M. Henson, *Nonlinear population dynamics: Models, experiments, and data*, J. Theor. Biol., 194 (1998), pp. 1–9.
[9] J. M. Cushing, R. F. Costantino, B. Dennis, R. A. Desharnais, and S. M. Henson, *Chaos in Ecology: Experimental Nonlinear Dynamics*, Academic Press, San Diego, 2003.
[10] J. M. Cushing, B. Dennis, R. A. Desharnais, and R. F. Costantino, *An interdisciplinary approach to understanding nonlinear ecological dynamics*, Ecol. Model., 92 (1996), pp. 111–119.
[11] J. M. Cushing, B. Dennis, R. A. Desharnais, and R. F. Costantino, *Moving toward an unstable equilibrium: Saddle nodes in population systems*, J. Animal Ecol., 67 (1998), pp. 298–306.
[12] J. M. Cushing, S. M. Henson, R. A. Desharnais, B. Dennis, R. F. Costantino, and A. A. King, *A chaotic attractor in ecology: Theory and experimental data*, Chaos Solitons Fractals, 12 (2001), pp. 219–234.
[13] B. Dennis, R. A. Desharnais, J. M. Cushing, and R. F. Costantino, *Nonlinear demographic dynamics: Mathematical models, statistical methods, and biological experiments*, Ecol. Monogr., 65 (1995), pp. 261–281.
[14] B. Dennis, R. A. Desharnais, J. M. Cushing, and R. F. Costantino, *Transitions in population dynamics: Equilibria to periodic cycles to aperiodic cycles*, J. Animal Ecol., 66 (1997), pp. 704–729.
[15] B. Dennis, R. A. Desharnais, J. M. Cushing, S. M. Henson, and R. F. Costantino, *Estimating chaos and complex dynamics in an insect population*, Ecol. Monogr., 71 (2001), pp. 277–303.
[16] R. A. Desharnais and R. F. Costantino, *Genetic analysis of a population of Tribolium.* VII. *Stability: Response to genetic and demographic perturbations*, Can. J. Genet. Cytol., 22 (1980), pp. 577–589.
[17] R. A. Desharnais, R. F. Costantino, J. M. Cushing, and B. Dennis, *Estimating chaos in an insect population*, Science, 276 (1997), pp. 1881–1882.
[18] R. A. Desharnais, R. F. Costantino, J. M. Cushing, S. M. Henson, and B. Dennis, *Chaos and population control of insect outbreaks*, Ecol. Let., 4 (2001), pp. 229–235.
[19] R. A. Desharnais and L. Liu, *Stable demographic limit cycles in laboratory populations of* Tribolium castaneum, J. Animal Ecol., 56 (1987), pp. 885–906.
[20] L. Edelstein-Keshet, *Mathematical Models in Biology*, McGraw–Hill, Boston, 1988, p. 105.
[21] G. F. Fussmann, S. P. Ellner, K. W. Shertzer, and N. G. Hairston, *Crossing the Hopf bifurcation in a live predator-prey system*, Science, 290 (2000), pp. 1358–1360.
[22] G. F. Gause, *The Struggle for Existence*, Hafner Publishing Company, New York, 1964, p. 144.
[23] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, Berlin, 1983.
[24] S. M. Henson, R. F. Costantino, 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, R. F. Costantino, R. A. Desharnais, J. M. Cushing, and B. Dennis, *Basins of attraction: Population dynamics with two locally stable* 4*-cycles*, Oikos, 98 (2002), pp. 17–24.
[26] S. M. Henson and J. M. Cushing, *The effect of periodic habitat fluctuations on a nonlinear insect population model*, J. Math. Biol., 36 (1997), pp. 201–226.
[27] S. M. Henson, R. F. Costantino, J. M. Cushing, R. A. Desharnais, B. Dennis, and A. A. King, *Lattice effects observed in chaotic dynamics of experimental populations*, Science, 294 (2001), pp. 602–605.

[28] S. M. Henson, J. M. Cushing, R. F. Costantino, B. Dennis, and R. A. Desharnais, *Phase switching in population cycles*, Proc. Roy. Soc. London Ser. B, 265 (1998), pp. 2229–2234.
[29] S. M. Henson, J. L. Hayward, C. M. Burden, C. J. Logan, and J. G. Galusha, *Predicting the dynamics of aggregate loafing behavior in gulls*, The Auk, submitted, 2003.
[30] A. A. King, R. A. Desharnais, S. M. Henson, R. F. Costantino, J. M. Cushing, and B. Dennis, *Random perturbations and lattice effects in chaotic population dynamics*, Science, 297 (2002), p. 2163a.
[31] R. M. May, *Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos*, Science, 186 (1974), pp. 645–647.
[32] P. A. P. Moran, *Some remarks on animal population dynamics*, Biometrics, 6 (1950), pp. 250–258.
[33] A. J. Nicholson, *The self-adjustment of populations to change*, Cold Spring Harbor Symposia on Quantitative Biology, 22 (1957), pp. 153–173.
[34] W. E. Ricker, *Stock and recruitment*, J. Fish. Res. Bd. Can., 11 (1954), pp. 559–623.
[35] W. M. Schaffer, B. E. Kendall, C. W. Tidd, and L. F. Olsen, *Transient periodicity and episodic predictability in biological dynamics*, IMA J. Math. Appl. Med. Biol., 10 (1993), pp. 227–247.
[36] H. Tong, *Nonlinear Time Series: A Dynamical System Approach*, Oxford University Press, Oxford, UK, 1990.
[37] S. Utida, *Population fluctuations, an experimental and theoretical approach*, Cold Spring Harbor Symposia on Quantitative Biology, 22 (1957), pp. 139–151.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 2, pp. 254–276

A Two-Parameter Study of the Locking Region of a Semiconductor Laser Subject to Phase-Conjugate Feedback*

Kirk Green†, Bernd Krauskopf‡, and Giovanni Samaey§

Abstract. We present a detailed bifurcation analysis of a single-mode semiconductor laser subject to phase-conjugate feedback, a system described by a delay differential equation. Codimension-one bifurcation curves of equilibria and periodic orbits and curves of certain connecting orbits are presented near the laser's locking region in the two-dimensional parameter plane of feedback strength and pump current. We identify several codimension-two bifurcations, including a double-Hopf point, Belyakov points, and a T-point bifurcation, and we show how they organize the dynamics.

This study is the first example of a two-parameter bifurcation study, including bifurcations of periodic and connecting orbits, of a delay system. It was made possible by new numerical continuation tools, implemented in the package DDE-BIFTOOL, and showcases their usefulness for the study of delay systems arising in applications.

Key words. semiconductor lasers, phase-conjugate feedback, delay differential equations, two-parameter continuation, heteroclinic orbits, T-point bifurcation

AMS subject classifications. 37N20, 34K18, 37G10, 37G20

PII. S1111111102416575

1. Introduction.

The majority of lasers in application today are semiconductor lasers. They can be found, for example, in CD-players, laser printers, and optical communication networks. Semiconductor lasers are so-called Class B lasers, in which the polarization of the electric field can be adiabatically eliminated. As a consequence, they can be described well by three-dimensional rate equations, one for the complex electric field $E(t)$ and one for the population inversion $N(t)$ (the number of excited states that can produce a single photon). It turns out that the phase $\phi(t)$ of the electric field follows the two equations for the optical intensity $P(t) = |E(t)|^2$ and the inversion $N(t)$. Therefore, a solitary semiconductor laser is essentially a two-dimensional dynamical system that cannot exhibit chaotic dynamics. The only observable dynamical behavior is a damped periodic exchange between the electric field and the inversion. These oscillations are referred to as *relaxation oscillations* in the laser literature (not to be confused with relaxation oscillations in slow-fast systems); see, for example, [26, 41] for an introduction to the theory of semiconductor lasers.

*Received by the editors October 28, 2002; accepted for publication (in revised form) by M. Golubitsky February 27, 2003; published electronically May 23, 2003.
http://www.siam.org/journals/siads/2-2/41657.html

†Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (kirk.green@bristol.ac.uk). Current address: Department of Computer Science, K. U. Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium.

‡Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (b.krauskopf@bristol.ac.uk).
This author was supported by an EPSRC Advanced Research Fellowship.
§Department of Computer Science, K. U. Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium (giovanni.samaey@cs.kuleuven.ac.be). This author is Research Assistant of the Fund of Scientific Research—Flanders.

The occurrence of interesting dynamics in a semiconductor laser system requires the addition of one or more degrees of freedom to the rate equations. The good news, from a dynamical systems point of view, is that this is easily achievable with the addition of some form of external influence. This may be due to noise, optical injection from another laser, or delayed optical feedback, the subject of this paper.

Optical feedback results when a part of a laser's output light is fed back into the laser after a delay time τ. This feedback can be unwanted; for example, reflections from a CD or optical fiber can seriously interrupt the proper operation of the device. However, more recently, it has been demonstrated that chaotic output from a feedback laser can be used in optical encryption schemes [10, 33, 43].

In an experiment, controllable optical feedback is obtained by adding an external mirror to the laser set-up; see Figure 2.1. If this mirror is a conventional optical reflector, then one speaks of conventional optical feedback (COF), a system that has received much attention due to its relevance for applications and because very complicated dynamics have been found; see, for example, [11, 34]. If the optical feedback comes from a phase-conjugating mirror (PCM), then one speaks of a laser with phase-conjugate feedback (PCF). This is the laser system that we are studying here. It is physically interesting because, unlike in COF, the PCM reverses the phase of the light so that the reflected wave travels back along the same path as the incident wave. This means that the alignment of the laser beam is not so much of an issue. Furthermore, perturbations to the light front on the way to the PCM are undone on the way back. Also, in the PCF laser, many interesting dynamical regimes have been identified; see, for example, [1, 12, 16, 24]; more details can be found in section 2.

Mathematically, optical feedback is described by adding a delay term to the equation describing the electric field. This implies that the system is described by a delay differential equation (DDE) with an infinite-dimensional phase space. Consequently, lasers with optical feedback may exhibit very complicated dynamics, and their analysis is quite hard. Until quite recently, the analysis of the linear stability of steady states and direct simulation of the equations were essentially the only tools to study the dynamics of DDEs arising in applications. However, this is changing with the introduction of advanced tools allowing detailed bifurcation studies of DDEs. These consist of (a) the publicly available Matlab continuation package DDE-BIFTOOL for numerical bifurcation analysis (see [8] and section 4 below) and (b) an algorithm, using DDE-BIFTOOL to obtain the necessary starting data, to compute unstable manifolds of saddle periodic orbits in a suitable Poincaré section [25]. (This algorithm was used in [19] to identify the break-up of a torus in the PCF laser and its subsequent disappearance in a crisis bifurcation.)

Continuation techniques have not yet been widely used to study the dynamics and bifurcations in DDEs arising in applications. The first examples include the series of papers [21, 36, 37], in which connecting bridges of periodic solutions in the COF laser were studied, and similar work on a vertical-cavity surface-emitting laser [40]. Continuation studies of the PCF laser can be found in [17, 18, 19].

The continuation studies mentioned above follow steady states and periodic orbits as a single parameter is changed (usually the strength of the feedback). This is also the case in [18], where the locking range of the PCF laser was studied in detail. This revealed regions of bistability associated with a saddle-node bifurcation, a Hopf bifurcation, and heteroclinic

connections. The next logical step is to construct a full two-dimensional bifurcation diagram.

In this paper, we do just this: we present a consistent two-dimensional bifurcation diagram near the locking region of the PCF laser in the plane of feedback strength versus pump current. (These parameters are physically natural and were also used in an experimental overview of the COF laser in [9].) We follow bifurcations of steady states and connecting orbits in two parameters. We make a first attempt at mapping bifurcations of periodic orbits by detecting a sufficient number of individual bifurcation points at appropriate values of the parameters. We identify several codimension-two bifurcations—most importantly, a double-Hopf point, a Belyakov point, and a bifurcation of heteroclinic orbits known as a T-point. We show how the dynamics of the PCF laser near the locking region are organized around these points.

The paper is organized as follows. In section 2, we introduce the rate equations for the PCF laser. In section 3, we give a brief introduction to the basic theory of DDEs. The capabilities of the continuation package DDE-BIFTOOL, particularly with respect to connecting orbits, are introduced in section 4. In section 5, we present a two-parameter bifurcation analysis of the steady states and a heteroclinic orbit involved in the locking mechanisms of the PCF laser. To allow for a better comparison with previous studies, we also present one-dimensional cross sections through the bifurcation diagram for fixed values of the pump current. In section 6, we look more closely at the bifurcations of steady states and continue the branch of heteroclinic orbits involved in the locking mechanism, all the way to its end in a codimension-two bifurcation of heteroclinic orbits known as a T-point. The heteroclinic orbits along this branch are studied in section 7, where we also identify a codimension-two Belyakov point. In section 8, we map out bifurcations of periodic orbits near the locking region. Finally, in section 9, we draw conclusions and discuss future work.

2. Phase-conjugate feedback.

Our object of study is a semiconductor laser with PCF from a PCM, schematically shown in Figure 2.1. A PCM can be made by utilizing a number of nonlinear optical processes, including stimulated Brillouin scattering, backward stimulated Raman scattering, and three-wave down conversion [12]. Probably the most common way of making a PCM is to use degenerate four-wave mixing, in which three input waves mix to produce a fourth output wave. Two of the input waves are counterpropagating pump waves from, for example, additional semiconductor lasers. The third wave is the incident wave, which may enter at any angle to the pump waves. These three waves couple through a third-order susceptibility $\chi^{(3)}$ and produce a fourth wave, phase-conjugated to the incident wave. This phase-conjugated wave may be more intense than the incident wave due to additional gain provided by the pump waves [41].

In contrast to COF, the laser light is wave-front inverted at the PCM. The return conjugated wave retraces the path of the incident wave, and, therefore, the system is self-aligning. Any distortions of the incident wave between its source and the PCM are undone on the return trip [12]. This produces a highly focused beam [12] that is of considerable advantage when stable output is desired, such as in mode locking [15] and phase locking, where PCF has been shown to reduce the laser noise considerably [1, 16, 42].

Mathematically, a single-mode semiconductor laser subject to weak (instantaneous) PCF can be described by the three-dimensional delay differential system

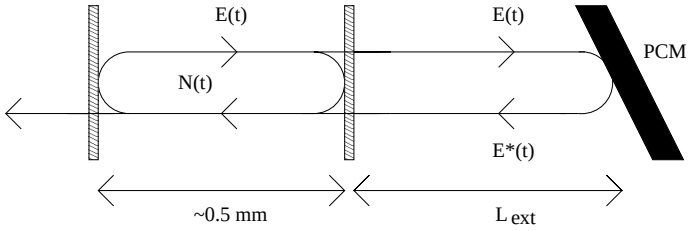


Figure 2.1. *Sketch of a semiconductor laser with PCF.*

$$\begin{aligned}\frac{dE}{dt} &= \frac{1}{2}\left[-i\alpha G_N(N(t)-N_{\text{sol}}) + \left(G(t)-\frac{1}{\tau_p}\right)\right]E(t) \\ &\qquad +\kappa E^*(t-\tau)\exp[2i\delta(t-\tau/2)+i\phi_{\text{PCM}}], \\ \frac{dN}{dt} &= \frac{I}{q} - \frac{N(t)}{\tau_e} - G(t)\,|E(t)|^2\end{aligned} \tag{2.1}$$

for the evolution of the slowly varying complex electric field $E(t) = E_x(t) + iE_y(t)$ and the population inversion $N(t)$ [16, 24]. This type of rate equation of a laser with feedback goes back to Lang and Kobayashi [31], who developed the first delay model for the COF laser. In system (2.1), nonlinear gain is included as $G(t) = G_N(N(t) - N_0)(1 - \epsilon P(t))$, where $\epsilon = 3.57 \times 10^{-8}$ is the nonlinear gain coefficient and $P(t) = |E(t)|^2$ is the intensity. Parameter values are set to realistic values corresponding to a Ga-Al-As semiconductor laser [16, 24], namely, the line-width enhancement factor $\alpha = 3$, the optical gain $G_N = 1190\,\text{s}^{-1}$, the photon lifetime $\tau_p = 1.4\,\text{ps}$, the magnitude of the electron charge $q = 1.6 \times 10^{-19}\text{C}$, the electron lifetime $\tau_e = 2$ ns, and the transparency electron number $N_0 = 1.64 \times 10^8$. Further, $N_{\text{sol}} = N_0 + 1\,/\,(G_N\tau_p)$. The constant phase shift ϕ_{PCM} at the PCM and the detuning parameter δ were both set to zero, as is common in the field [16, 24]. Therefore, the feedback term in system (2.1) reduces to $\kappa E^*(t-\tau)$ and involves the complex conjugated electric field E^*, the feedback rate κ, and the external cavity round-trip time τ. For one-parameter studies of (2.1), we fix τ at the realistic value $\tau = 2/3\,\text{ns}$, corresponding to an external-cavity length of $L_{\text{ext}} \approx 10\,\text{cm}$, and consider the bifurcation parameter $\kappa\tau$; for two-parameter studies, we also free the pump current I.

System (2.1) has $\mathbb{Z}_2$-symmetry under the transformation $(E, N) \to (-E, N)$, where the symmetry group is $\mathbb{Z}_2 = \{1, -1\}$. This corresponds to a rotation over π of the complex E-plane so that any attractor (or other invariant set) either is symmetric or has a symmetric counterpart. This symmetry allows for the possibility of symmetry-breaking and symmetry-restoring bifurcations [24, 28] and also implies restrictions on the types of bifurcations of periodic orbits. For example, symmetric orbits cannot undergo period-doubling bifurcations [29].

It was shown in [24] that the general picture of the dynamics of the PCF laser is that of stable periodic operation interspersed with "bubbles" of more complicated, chaotic dynamics.

In [17], these periodic solutions were shown to be connected to a steady state solution. This steady state solution represents a frequency match between the solitary laser and the pump lasers used in the four-wave mixing [41]. When it is stable, the laser is frequency locked and phase locked. In this region, the laser phase no longer undergoes diffusion, resulting in an extremely narrow line-width, a property that persists even with the addition of noise from spontaneous emission [1, 16, 42]. The region in which this stable locked solution exists is called the *locking region* of the PCF laser.

3. Background on DDEs. System (2.1) is a DDE with an infinite-dimensional phase space. As the reader may be unfamiliar with the theory of DDEs, we now give a brief introduction; see [5, 22, 23] for further details.

Models featuring a delay can be found in many areas of science, such as, for example, biology [35], control theory [14], and, as we have seen, laser physics [26]. They lead to a mathematical description by a DDE which, in its simplest form of a single fixed delay $\tau \in \mathbb{R}$, takes the form

$$\frac{dx(t)}{dt} = F(x(t), x(t-\tau), \eta), \tag{3.1}$$

where

$$F : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$$

is differentiable and $\eta \in \mathbb{R}^p$ is a multiparameter. We consider here only the case of one fixed delay. This is not an oversimplification; for example, both the COF laser and the PCF laser introduced in section 2 are of this important class. We remark, however, that the continuation methods described in section 4 also work for a finite number of fixed delays and even for certain state-dependent delays [32].

The *phase space* of (3.1) is the infinite-dimensional space of continuous functions $\mathcal{C}$ over the delay interval $[-\tau, 0]$ with values in $\mathbb{R}^n$. The space $\mathbb{R}^n$ is called the *physical space*; for example, it is the (E, N)-space for system (2.1). A *point* $q \in \mathcal{C}$ is a continuous function

$$q : [-\tau, 0] \to \mathbb{R}^n.$$

We call $q(0)$ the *head* of q and $q|_{[-\tau,0)} \equiv \{q(t) \,|\, t \in [-\tau, 0)\}$ its *history*. The evolution of a point $q \in \mathcal{C}$ after time $t \geq 0$ is given by the *evolution operator*

$$\Phi^t : \mathcal{C} \to \mathcal{C}.$$

A *solution* of (3.1) is a function

$$x : [-\tau, \infty) \to \mathbb{R}^n, \quad t \mapsto \Phi^t(x_0)$$

for some initial point $x_0 \in \mathcal{C}$.

A *steady state* (or equilibrium) of (3.1) is a point x_0 such that $x_0(t) \equiv \bar{x}_0$ for all $t \in [-\tau, \infty)$ and fixed $\bar{x}_0 \in \mathbb{R}^n$. In other words, $F(\bar{x}_0, \bar{x}_0, \eta^*) = 0$ (for some fixed η^*) and $\Phi^t(x_0) = x_0$ for all $t > 0$. The stability of x_0 is given by the variational equation

$$DF(\bar{x}_0, \eta^*) = A_1(\bar{x}_0, \eta^*)x(t) + A_2(\bar{x}_0, \eta^*)x(t-\tau), \tag{3.2}$$

which is the linearization around x_0, where

$$A_1(\bar{x}_0, \eta^*) = D_1 F(\bar{x}_0, \eta^*) \quad \text{and} \quad A_2(\bar{x}_0, \eta^*) = D_2 F(\bar{x}_0, \eta^*) \tag{3.3}$$

are the derivatives of F with respect to the first and second variable.

When we define the $n \times n$ matrix Δ as

$$\Delta(\bar{x}_0, \eta^*, \lambda) := \lambda I - A_1(\bar{x}_0, \eta^*) - A_2(\bar{x}_0, \eta^*) e^{-\lambda\tau}, \tag{3.4}$$

then the eigenvalues are given as the roots of the *characteristic equation*

$$\det(\Delta(\bar{x}_0, \eta^*, \lambda)) = 0. \tag{3.5}$$

If we also need to compute eigenvectors, we solve

$$\Delta(\bar{x}_0, \eta^*, \lambda)v = 0,\ c^H v - 1 = 0 \quad \text{or} \quad \Delta(\bar{x}_0, \eta^*, \lambda)^H w = 0,\ d^H w - 1 = 0, \tag{3.6}$$

where v and w are right and left eigenvectors, respectively, and H denotes the Hermitian conjugate.

It is a crucial property of DDEs with fixed delays that the eigenvalues are discrete and that there are always finitely many eigenvalues with real part larger than γ for any fixed $\gamma \in \mathbb{R}$; see Theorem 4.4 (i) of [5]. In particular, there are only finitely many unstable eigendirections (associated with eigenvalues with real part greater than zero). As usual, a steady state is called *hyperbolic* if there are no eigenvalues that have zero real part.

A *periodic orbit* is a solution Γ such that $\Phi^{\mathsf{T}}(q) = q$ for some period $\mathsf{T} > 0$ and all $q \in \Gamma$. After choosing a section $\Sigma \in \mathbb{R}^n$ (locally) transverse to Γ, the corresponding Poincaré map P is defined on the space $\mathcal{C}_\Sigma$ of points in $\mathcal{C}$ with headpoints in Σ; that is,

$$P:\ \mathcal{C}_\Sigma \to \mathcal{C}_\Sigma, \quad q \mapsto \Phi^{t_q}(q),$$

where t_q is the return time to Σ. The stability of Γ is given by its *Floquet multipliers*, which are the eigenvalues of the linearization $DP(q)$ of P at the corresponding fixed point $q \in \mathcal{C}_\Sigma$. The linearization $DP(q)$ is found by solving the variational equation, over one period T, along Γ. For any fixed radius $r > 0$, there is only a finite number of Floquet multipliers outside a circle of radius r so that there are always finitely many unstable eigendirections. A periodic point is called *hyperbolic* if there are no Floquet multipliers on the unit circle.

As is the case for ODEs, a local bifurcation occurs when a steady state or a periodic orbit is not hyperbolic. The generic bifurcations are saddle-node and Hopf bifurcations of steady states and saddle-node, period-doubling, and torus (or Neimark–Sacker) bifurcations of periodic orbits.

A solution $x(t)$ of (3.1) at some parameter $\eta = \eta^*$ is called a *connecting orbit* if the limits

$$\lim_{t\to-\infty} x(t) = x^- \quad \text{and} \quad \lim_{t\to+\infty} x(t) = x^+ \tag{3.7}$$

exist, where $x^\pm$ are steady states of (3.1). Connecting orbits are discussed in more detail in section 4.1.

4. Numerical continuation with DDE-BIFTOOL. The continuation package DDE-BIFTOOL [8] has been developed for the numerical bifurcation analysis of DDEs with fixed discrete delays or state-dependent delays [32]. While a number of packages exist for the numerical bifurcation analysis of ODEs, such as AUTO [6] and CONTENT [30], DDE-BIFTOOL is the first publicly available package for the bifurcation analysis of DDEs. Roughly speaking, DDE-BIFTOOL has the same functionality as (the local bifurcation part of) AUTO with new features constantly being added, such as the computation of connecting orbits discussed below in section 4.1.

DDE-BIFTOOL allows the user to find and follow steady states and periodic solutions irrespective of their stability. It also detects the generic codimension-one local bifurcations of steady states and periodic orbits by detecting when an eigenvalue of the linearization has a zero real part or is on the unit circle, respectively. The software is able to switch to the continuation of emanating branches of periodic orbits at bifurcation points (such as a Hopf bifurcation). Furthermore, once a codimension-one bifurcation of a steady state has been detected, it can be followed in two parameters. At present, codimension-one bifurcations of periodic orbits can be detected but cannot be continued in two parameters. A recent addition to DDE-BIFTOOL is the computation of connecting orbits, using projection boundary conditions, and their continuation in two parameters [38]. This algorithm is a natural extension of the method that was implemented for the computation of connecting orbits in ODEs in the HomCont [4] extension of AUTO.

The infinite-dimensional nature of DDEs means that the computation of solutions and their stability is far from trivial. To compute the stability of steady states, DDE-BIFTOOL approximates an appropriate number of the right-most roots of the characteristic equation and corrects them by using Newton iterations. A steady state is represented by the value of the parameter η^*, the steady state position $\bar{x}_0$, and the eigenvalues λ_i of this steady state. A saddle-node bifurcation is detected and represented by a null-vector of $\Delta(\bar{x}_0, \eta^*, 0)$, the matrix defined in (3.4). Similarly, a Hopf bifurcation is represented by the complex null-vector of $\Delta(\bar{x}_0, \eta^*, i\omega)$ and corresponding frequency ω. To represent and follow periodic solutions, orthogonal collocation, based on a piecewise polynomial representation of the solution, is used. A periodic solution is represented by the value of the parameter η^*, the period T, and a time-scaled profile $x^*(t/\mathsf{T})$ on a mesh over the interval $[0, 1]$.

DDE-BIFTOOL can be extended to monitor other quantities that might be of interest to the user. One such example was developed and implemented for the analysis in section 7 to find and follow a neutral saddle (-focus) point. This is a steady state solution where the sum of the real parts of the unstable (complex conjugate) eigenvalues $\lambda_{1,2}$ and the leading stable eigenvalue λ_3 is equal to zero. A neutral saddle is represented by the steady state position $\bar{x}_0$, the values of the parameter η^*, the leading eigenvalues $\lambda_{1,2}$ and λ_3, and their eigenvectors and is subject to the constraint that $\mathrm{Re}(\lambda_{1,2}) + \mathrm{Re}(\lambda_3) = 0$. These fields are used as elements inside the branch structure when computing and continuing a neutral saddle point.

4.1. Computing connecting orbits in DDEs. A recent addition to DDE-BIFTOOL is the continuation of connecting orbits [38]. Both homoclinic orbits ($x^- = x^+$ in (3.7)) and heteroclinic orbits ($x^- \neq x^+$ in (3.7)) can be computed and continued as certain system parameters η are varied. This requires finding a good starting solution for a fixed parameter

value. For a homoclinic orbit, one can start from a nearby periodic orbit with a sufficiently large period. Heteroclinic orbits can be approximated by using time integration or by using an extension of the method of successive continuation [7].

A defining condition for a connecting orbit is that it is contained in both the stable manifold of x^+ and the unstable manifold of x^-. A classical approach in the ODE case is to approximate this condition by truncating the time domain to an interval of length T and to apply (so-called) projection boundary conditions [2]: one end point of the connecting orbit is required to lie in the unstable eigenspace of x^- and the other end point in the stable eigenspace of x^+. The projection boundary conditions therefore replace the stable and unstable manifolds by their linear approximations near the steady states. The error caused by this approximation has been proven to be exponentially decaying with the interval length T [39]. This approach is successfully implemented in standard continuation codes for ODEs such as the HomCont [4] part of AUTO [6].

In the implementation under DDE-BIFTOOL, this approach was extended to the case of DDEs. Because one needs to provide an initial *function segment* (rather than just an initial value as for ODEs), the boundary conditions need to be written in terms of solution segments. Further, x^+ has infinitely many eigenvalues with negative real parts so that it is impossible to write the final function segment as a linear combination of all (infinitely many) stable eigenfunctions. Instead, it is required that the end function segment is in the *orthogonal complement* of all unstable left eigenfunctions.

More specifically, the condition for the initial function segment $x_0(\theta)=x(\theta)$, $\theta \in [-\tau, 0]$, can be written as

$$x_0(\theta) = x^- + \epsilon \sum_{k=1}^{s^-} \alpha_k v_k^- e^{\lambda_k^- \theta} \qquad \left(\sum |\alpha_k|^2 = 1 \right),$$

where s^- is the number of unstable eigenvalues λ^-, with corresponding eigenvectors v^-. The α_k are unknown coefficients, and ϵ is a measure for the desired accuracy. An extra condition is added to ensure continuity at $\theta = 0$. As discussed above, we cannot write the end conditions for the final function segment in a similar way. Instead a special bilinear form [22] is used to express the fact that the final function segment is in the complement of the unstable eigenspace of x^+. This leads to s^+ extra conditions:

$$w_k^{+H}(x(T) - x^+) + \int_{-\tau}^{0} w_k^{+H} e^{-\lambda_k^+(\theta+\tau)} A_2(x^+, \eta) \left(x(T+\theta) - x^+ \right) d\theta = 0\,.$$

Here s^+ is the number of unstable eigenvalues of x^+, w_k^+ are the left eigenvectors corresponding to the eigenvalues λ_k^+, and the matrix A_2 is defined in (3.3). While this integral condition works well in practice, one slight drawback is that it does not control the distance of the end function segment to the steady state.

Connecting orbits arise in one-parameter families: any time-translate is also a connecting orbit. Therefore, a phase condition needs to be added to pick out just one of these orbits. In general, a number s_η of free parameters is required to obtain a generically isolated solution. As a consequence, the equations for x^- and x^+ (steady state equations) and λ_k^- and v_k^- and λ_k^+ and w_k^+ (characteristic equations) have to be added to the defining system. The result

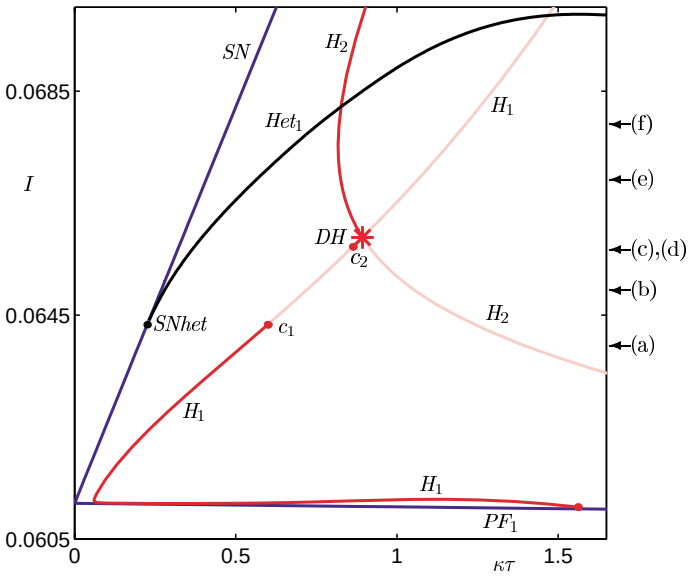


Figure 5.1. *Bifurcations bounding the locking region of the PCF laser in the $(\kappa\tau, I)$-plane. The arrows correspond to the one-parameter continuations shown in Figure 5.2.*

is a system of n differential equations, supplemented with $(s^- + s^+)(n+1) + s^+ + 2$ extra equations, resulting in the need for $s_\eta = s^+ - s^- + 1$ free parameters. We end up with a boundary value problem, which we solve by using a collocation method, where the solution is represented as a piecewise polynomial.

5. The locking region. In this section, we use the continuation package DDE-BIFTOOL to detect and follow the bifurcations involved in the locking mechanism of the PCF laser. Physically, this steady state solution corresponds to a frequency match between the PCM pump lasers and the solitary laser. The resulting bifurcation diagram shown in Figure 5.1 was obtained by starting a two-parameter continuation from bifurcation points identified in one-parameter studies, including the study described in [18]. Plotted in red are curves of Hopf bifurcations $H_{1,2}$, where each point on the curve represents a steady state with a pair of pure imaginary eigenvalues. The Hopf curves $H_{1,2}$ are drawn dark when they are supercritical (the bifurcating periodic orbit is stable), and are drawn in a lighter tone when they are subcritical (the bifurcating periodic orbit is unstable). In blue are plotted curves of saddle-node bifurcations SN and pitch-fork bifurcations PF_1; in both cases, each point represents a steady state with a real eigenvalue equal to zero. Furthermore, we plot a curve of heteroclinic connections Het_1 between two saddle steady states that are each other's symmetric counterparts.

The general picture of the locking region of the PCF laser is that it is bounded by the curve of saddle-node bifurcations SN to the left, the (supercritical parts of the) curves of Hopf bifurcations $H_{1,2}$ to the right, and a curve of pitchfork bifurcations PF_1 below. Passing through the locking region is a curve of heteroclinic orbits Het_1. The area above the curve Het_1 and between the curves SN and H_2 is a region of bistability; that is, the periodic orbit involved in the heteroclinic bifurcation Het_1 and the nonsymmetric steady states born in the

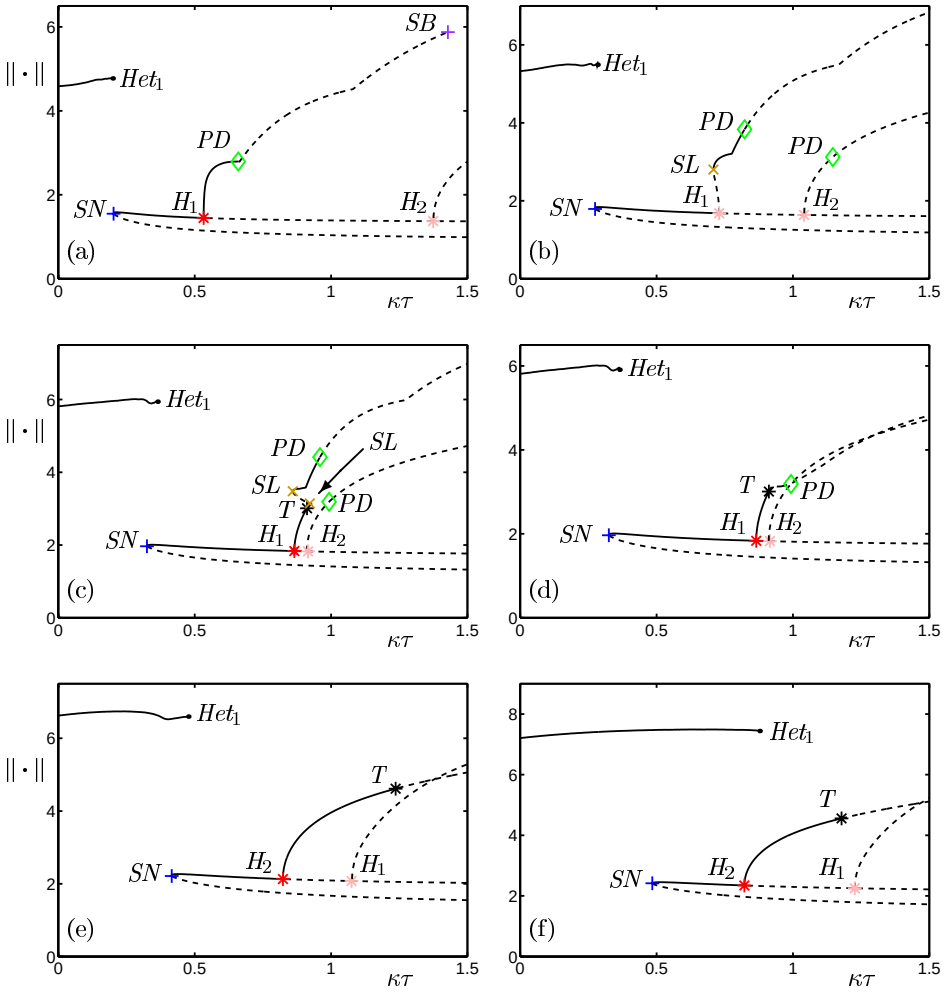


Figure 5.2. *One-parameter continuation of steady states and periodic orbits for fixed pump current I; from* (a) *to* (f) *I takes the values* 0.064, 0.065, 0.0657345, 0.0657347, 0.067, *and* 0.068.

saddle-node bifurcation SN coexist; see Figure 5.2 (b) to (f). Note that, when the heteroclinic curve crosses the curve of Hopf bifurcations H_2, there is a second region of bistability, where two stable periodic solutions coexist.

The nonsymmetric saddle steady states are born in the saddle-node bifurcation SN together with a pair of nonsymmetric stable steady states which correspond to the locked solutions of the PCF laser. These locked solutions are destabilized along the curves of Hopf bifurcations $H_{1,2}$ when they bifurcate with a periodic orbit. Between c_1 and c_2, on H_1, the Hopf bifurcation is subcritical; that is, the bifurcating periodic orbit is unstable. For fixed I, this subcritical Hopf bifurcation leads to a bistability between an attracting steady state and an attracting periodic orbit and a heteroclinic connection between their respective saddles. This bistability is due to a saddle-node bifurcation of limit cycles curve SL connecting c_1 and c_2, and running parallel to the subcritical part of H_1; see Figure 8.1 (a) and [18].

The pitchfork curve PF_1 is a bifurcation of the trivial steady state $(E, N) = (0, \frac{I\tau_e}{q})$. Below PF_1, the trivial steady state is stable. It is destabilized at PF_1 when it bifurcates with a pair of nonsymmetric stable steady states (the locked solutions). Physically, the PCF laser is in its off-state below PF_1. In this system with $\mathbb{Z}_2$-symmetry, the pitchfork bifurcation constitutes the laser threshold; that is, it marks the onset of lasing.

The Hopf curves $H_{1,2}$ intersect at the point DH, at $(\kappa\tau, I) \approx (0.893, 0.06589)$, which is a codimension-two double-Hopf bifurcation point where there are two pairs of complex eigenvalues on the imaginary axis [29]. At this bifurcation point, the center manifold is four-dimensional; in other words, this bifurcation is only possible in a phase-space of dimension greater than or equal to four. Around a double-Hopf point, the system can bifurcate to a number of invariant objects, including two-dimensional tori, which may branch to three-dimensional tori [29]. In fact, we will see below that, near the double-Hopf point, a period-doubling route to chaos [18] becomes a route to chaos via the break-up of a torus.

At $(\kappa\tau, I) \approx (0.225, 0.06433)$, the heteroclinic curve Het_1 ends at the saddle-node curve SN at a saddle-node heteroclinic point $SNhet$. Here the saddle-node bifurcation takes place on a codimension-one heteroclinic connection. If we divide out the symmetry of system (2.1), this is a saddle-node homoclinic bifurcation [29]. Below the point $SNhet$, the saddle-node bifurcation SN takes place on a limit cycle [20].

The lower part of the Hopf curve H_1 is seen to bend off and run roughly parallel to the curve PF_1 before terminating. For low values of fixed pump current I, this means that it is possible for the laser to lock and unlock as the value of $\kappa\tau$ is varied; see Figure 6.2. However, this effect occurs for a very small range of I and would be extremely difficult to observe experimentally.

We now discuss transitions through the two-dimensional bifurcation diagram in Figure 5.1 as we vary the value of $\kappa\tau$ for fixed values of I; this is the approach we took in [18]. The bifurcation diagrams in Figure 5.2 were obtained with DDE-BIFTOOL, and the respective I-values are indicated by the arrows on the right in Figure 5.1. For steady states, we plot $\mathrm{Re}(E)$, and for periodic solutions we plot $|\max(\mathrm{Re}(E)) - \min(\mathrm{Re}(E))|$, offset by the $\mathrm{Re}(E)$-value of the steady state at the Hopf point. Attracting solutions are drawn as solid curves, while unstable solutions are drawn as dashed curves. By studying the eigenvalues of the system, we are able to identify the bifurcations involved. Apart from saddle-node bifurcations SN, Hopf bifurcations $H_{1,2}$, and saddle-focus heteroclinic bifurcations Het_1 already shown in Figure 5.1, we also find saddle-node bifurcations of limit cycles SL, period-doubling bifurcations PD, symmetry-breaking (or restoring) bifurcations SB, and torus (or Neimark–Sacker) bifurcations T. The different bifurcations are color coded throughout; compare these with Figure 5.1, and see Figures 6.1, 6.2, and 8.1.

For low values of I (Figure 5.2 (a)), the saddle-node bifurcation SN occurs very close to the saddle-focus heteroclinic bifurcation Het_1. For $I < 0.06433$, the steady state is destabilized in the supercritical Hopf bifurcation H_1. The ensuing periodic solution then undergoes a period-doubling bifurcation PD, the first along a route to chaos [18, Figure 1]. At $I \approx 0.06433$, the Hopf bifurcation H_1 becomes subcritical, and for larger values of I (Figure 5.2 (b)) we observe the emergence of a saddle-node bifurcation of limit cycles SL. This is the scenario considered in [18]. At $I \approx 0.0657345$, the period-doubling route to chaos is preceded by a torus bifurcation T and two saddle-node bifurcations of limit cycles SL. Consequently, we find a bistability

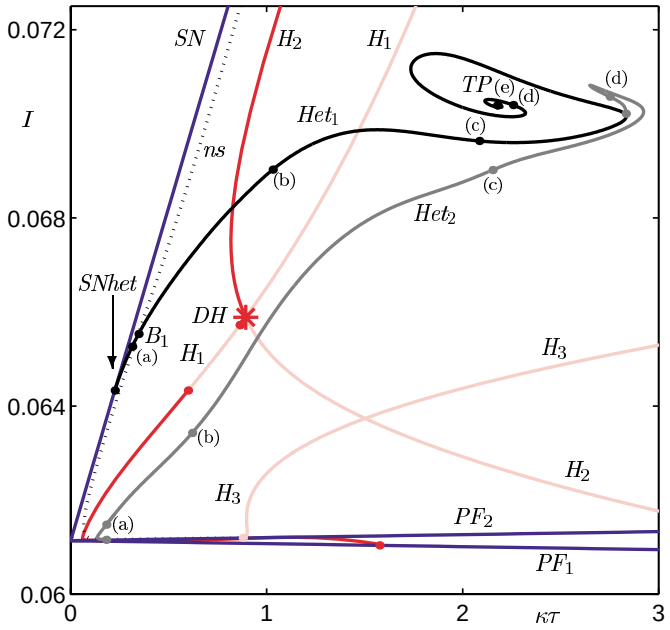


Figure 6.1. *Bifurcations of steady states in the* $(\kappa\tau, I)$*-plane. The labels along the curves* $Het_{1,2}$ *correspond to phase portraits shown in Figures 7.1 and 7.3.*

between two stable periodic solutions (Figure 5.2 (c)). One of the periodic solutions is born in the Hopf bifurcation H_1 at $\kappa\tau \approx 0.8653$ and is destabilized in the torus bifurcation T at $\kappa\tau \approx 0.9118$. The other periodic solution is born in the saddle-node bifurcation of limit cycles SL at $\kappa\tau \approx 0.8598$, first identified in Figure 5.2 (b), and is destabilized in the period-doubling bifurcation PD at $\kappa\tau \approx 0.9593$. For $I > 0.0657346$, the route to chaos via the break-up of a torus persists (Figure 5.2 (d)). This change, from a period-doubling route to chaos to a route to chaos via the break-up of a torus, appears to be due to the presence of the double-Hopf point. At the double-Hopf point, the Hopf curves H_1 and H_2 pass through one another. Figure 5.2 (e) shows the situation for $I = 0.067$. The first Hopf bifurcation that destabilizes the nonsymmetric saddle steady state is now H_2, which leads to a stable periodic orbit that is destabilized in a torus bifurcation T at $\kappa\tau \approx 1.236$. As I is increased further, there is an increasingly larger region of bistability between a stable periodic solution and the nonsymmetric stable steady state at the left boundary of the region of stability. This is a result of the heteroclinic curve Het_1 moving away from the saddle-node bifurcation curve SN, and crossing into the locking region. Finally, for $I > 0.068222$, the periodic orbit involved in the heteroclinic bifurcation exists past the first Hopf bifurcation. This is shown in Figure 5.2 (f), where for $I = 0.068$ we observe a bistability between the periodic orbit involved in the heteroclinic bifurcation and the periodic orbit born in the Hopf bifurcation H_2.

6. Bifurcations of steady states. In Figure 6.1, we increase the area of the $(\kappa\tau, I)$-plane under consideration to take a somewhat more global point of view. We follow the curves of saddle-node bifurcations SN, pitchfork bifurcations PF_1, Hopf bifurcations $H_{1,2}$, and hetero-

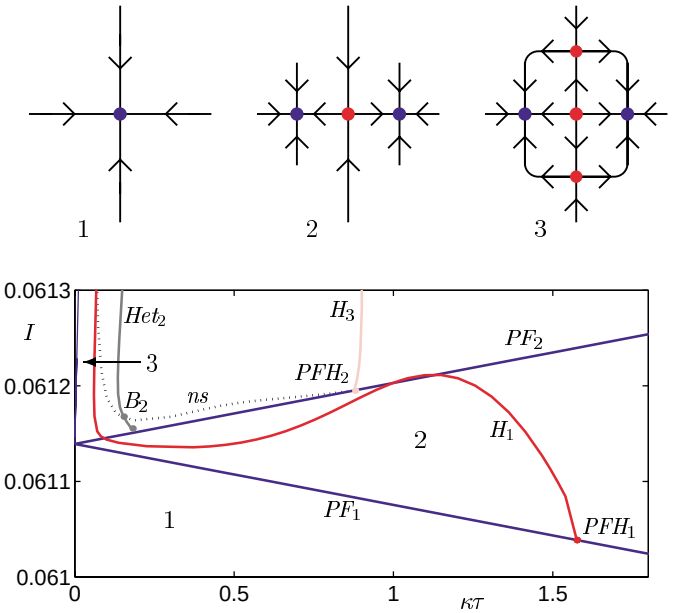


Figure 6.2. *Enlargement of bifurcation diagram in Figure 6.1 near the laser threshold and sketch of stability of steady states in different regions (directions not sketched are attracting).*

clinic bifurcations Het_1 shown further in Figure 5.1. We also follow the additional curves of pitchfork bifurcations PF_2, Hopf bifurcations H_3, and heteroclinic bifurcations Het_2 and show a neutral saddle curve *ns*. This provides a consistent picture of the bifurcations that can be continued, namely, bifurcations of steady states and heteroclinic orbits.

Figure 6.2 shows an enlargement of Figure 6.1 near the laser threshold, illustrating the interaction of the pitchfork curves $PF_{1,2}$ with the Hopf curves $H_{1,3}$. The steady states in three different regions are sketched, where stable steady states are drawn as blue points and saddle (unstable) steady states as red points. As was mentioned earlier, below the pitchfork curve PF_1, the trivial steady state is stable (region 1). Physically, the curve PF_1 marks the onset of lasing of the PCF laser. (Below this curve the laser is off.) At the curve PF_1, the trivial steady state is destabilized, and a pair of nonsymmetric stable steady states emerge (region 2); these are destabilized at the Hopf curve H_1. The curve of pitchfork bifurcations PF_2 represents another bifurcation of the trivial steady state. At this curve the trivial steady state has a zero eigenvalue, which means that a pair of nonsymmetric saddle steady states is born (region 3). These saddles are those involved in the saddle-node bifurcation *SN*, and they appear as the lower branch of saddle steady states identified in Figure 5.2. At the Hopf curve H_3, this nonsymmetric saddle steady state undergoes a Hopf bifurcation, where it bifurcates with a saddle periodic orbit. The codimension-two pitchfork-Hopf bifurcation points $PFH_{1,2}$ represent the ends of the Hopf curves $H_{1,3}$, respectively. At a pitchfork-Hopf point the center manifold is three-dimensional, and one finds a real eigenvalue and two pure imaginary eigenvalues on the imaginary axis [29].

The shape of the Hopf curve H_1 means that the laser can lock and unlock for a fixed value

of the pump current I as the feedback strength $\kappa\tau$ is varied. For example, for $I = 0.0612$ a locked solution is born in a saddle-node bifurcation SN at $\kappa\tau \approx 0.00428$. This solution is unlocked (becomes unstable) at the Hopf curve H_1 at $\kappa\tau \approx 0.0606$, and the ensuing solution is once again locked at the Hopf curve H_1 at $\kappa\tau \approx 0.977$. A final intersection with the Hopf curve H_1 at $\kappa\tau \approx 1.227$ results in an unlocked solution. However, we note that these transitions would be extremely difficult to observe experimentally due to the small ranges in the parameters involved.

The curves of heteroclinic bifurcations $Het_{1,2}$ are explained in the next section. We note that, as is to be expected, they end at the curves SN and PF_2, where the nonsymmetric saddle steady states involved in the heteroclinic orbits are born.

7. Global bifurcations. Figure 6.1 shows that the curve of heteroclinic bifurcations Het_1 curls up near the point marked TP. This indicates that the center point of the spiral is a codimension-two point known as a *T-point* [13]. At TP the heteroclinic connection between the two nonsymmetric steady states is destroyed. This results in the creation of two heteroclinic orbits from the bifurcating heteroclinic orbit. To show that this is indeed the case, we must look at the heteroclinic orbits themselves.

Figure 7.1 shows heteroclinic orbits, calculated with DDE-BIFTOOL, for the parameter values along the curve Het_1 indicated in Figure 6.1. The first and second columns show E_x and N, respectively, as a function of time on the truncation interval that was used by the boundary value solver; the third column shows the heteroclinic orbit projected onto the E-plane; and the fourth column shows the heteroclinic orbit projected onto (E, N)-space.

Near the saddle-node bifurcation SN (Figure 7.1 (a)), the orbit is seen to leave one saddle steady state and spiral into its symmetric counterpart; this was also found in [18]. As one moves along the curve Het_1, the heteroclinic orbit starts to increase in size in (E, N)-space (Figure 7.1 (b4) and (c4)). As the heteroclinic curve Het_1 approaches the T-point TP in the $(\kappa\tau, I)$-plane, the heteroclinic orbit continues to grow in (E, N)-space (Figure 7.1 (d)) until just prior to reaching the T-point TP it is seen to pass very near the origin of the E-plane. At the same time, the value of the inversion N grows (Figures 7.1 (e1) and (e2)) with a final rapid oscillation before ending up at the other nonsymmetric saddle steady state. The maximum value of inversion N reached is very close to the value of N of the trivial steady state (Figure 7.1 (e4)). This is a clear indication that we are very near the forthcoming T-point bifurcation. The evolution of the connecting orbits is also shown in a movie accompanying Figure 7.1.

One can find the new heteroclinic orbits at the T-point TP connecting the trivial steady state to the nonsymmetric steady states, by providing the boundary value solver of DDE-BIFTOOL with the position and stability information of the trivial steady state. As is to be expected at a T-point, there is a codimension-two connection from the nonsymmetric steady state to the trivial steady state and a codimension-zero connection from the trivial steady state to the symmetric counterpart of the nonsymmetric steady state.

Figure 7.2 (a) shows the heteroclinic orbit between the two nonsymmetric steady states. As was detailed in section 4.1, the codimension of a heteroclinic orbit is given by the dimension of the unstable manifold of the end steady state minus the dimension of the unstable manifold of the initial steady state plus one [38]. The nonsymmetric steady states have

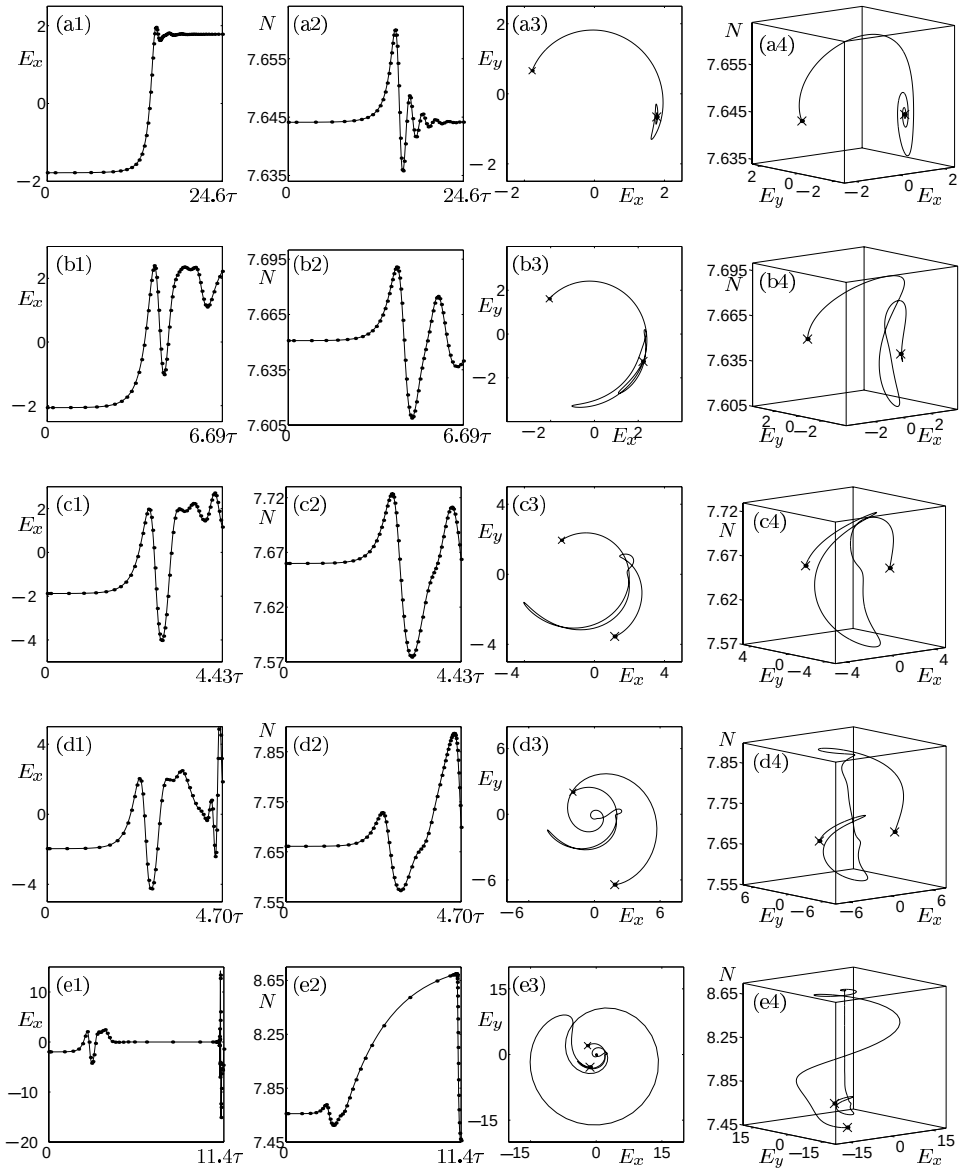


Figure 7.1. *Heteroclinic orbits along the curve Het_1; from* (a) *to* (e) $(\kappa\tau, I)$ *takes the values* $(0.314, 0.065264)$, $(1.303, 0.069026)$, $(2.085, 0.069635)$, $(2.201, 0.070357)$, *and* $(2.177, 0.070394)$. *(The movie file linked to the framed image shows the development of the heteroclinic orbits along the branch Het_1 in four panels. Top left: Position along the curve Het_1 in the $(\kappa\tau, I)$-plane. Top right: The heteroclinic orbit projected onto (E, N)-space. Bottom left: E_x on the truncation interval. Bottom right: N on the truncation interval.)*

one-dimensional unstable manifolds. Consequently, the heteroclinic orbit (and its symmetric counterpart) shown in Figure 7.2 (a) is of codimension one. This orbit is seen to start at one of the nonsymmetric steady states and then spend much time at the trivial steady state (red part of the orbit) before a sudden oscillation back to the end nonsymmetric steady state (blue part of the orbit). These two parts correspond to the two heteroclinic orbits that we find at

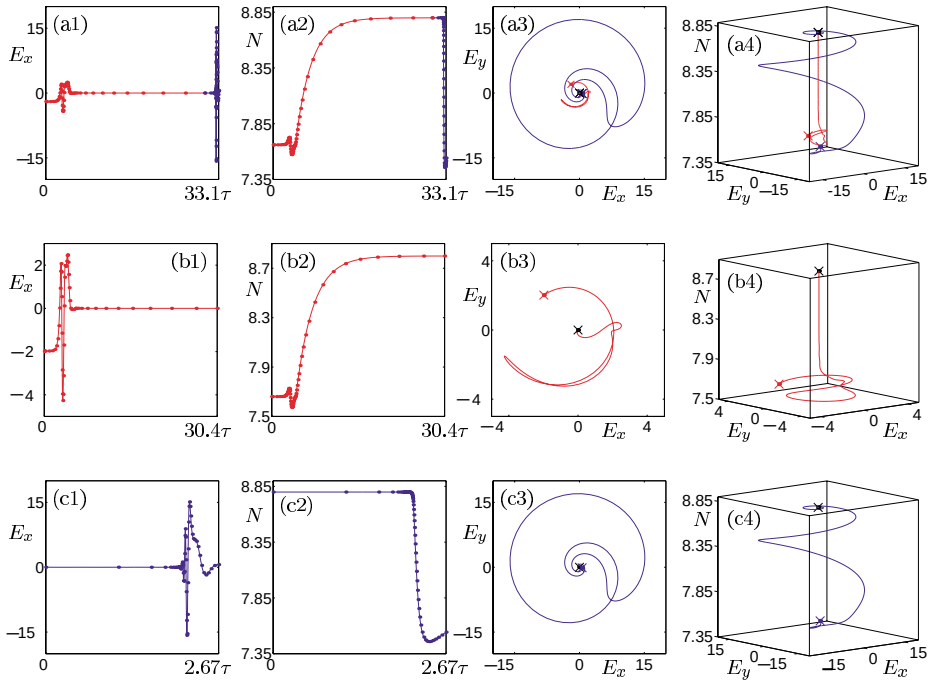


Figure 7.2. *Codimension-one heteroclinic orbit very close to the T-point TP* (a) *and the corresponding codimension-two* (b) *and codimension-zero* (c) *heteroclinic orbits at the T-point TP.*

the T-point. The first of these is shown in Figure 7.2 (b). The nonsymmetric steady states have one-dimensional unstable manifolds, and the trivial steady state has a two-dimensional unstable manifold; therefore, the heteroclinic orbit shown in Figure 7.2 (b) is of codimension two. This heteroclinic connection exists only at the point *TP*, and its continuation would require freeing a third parameter, which is beyond the scope of this study. The second heteroclinic orbit found at the T-point *TP* is shown in Figure 7.2 (c). Because the trivial steady state has a two-dimensional unstable manifold and the nonsymmetric steady states have one-dimensional unstable manifolds, this heteroclinic orbit is of codimension zero. In other words, this heteroclinic orbit exists for all values of $(\kappa\tau, I)$ in a local neighborhood of the T-point *TP*.

In Figures 6.1 and 6.2, the dotted curve *ns* represents a neutral saddle curve, along which the saddle steady states born in the saddle-node bifurcation *SN* have zero saddle quantity; that is, $\sigma = \mathrm{Re}(\lambda_{1,2}) + \lambda_3 \equiv 0$, $\mathrm{Re}(\lambda_{1,2}) > 0$, $\lambda_3 < 0$. We note that the curve *ns* starts at the pitchfork curve PF_2, the curve in which the nonsymmetric saddle steady states associated with the neutral saddle are born; see Figure 6.2. The curve *ns* is seen to intersect the curve of heteroclinic orbits Het_1 at the point B_1, a codimension-two Belyakov point [45]. Along the curve Het_1, the heteroclinic orbits between the saddle-node heteroclinic bifurcation *SNhet* and the Belyakov bifurcation B_1 have a negative saddle quantity corresponding to a bifurcating attracting periodic orbit associated with the heteroclinic bifurcation. Above the Belyakov

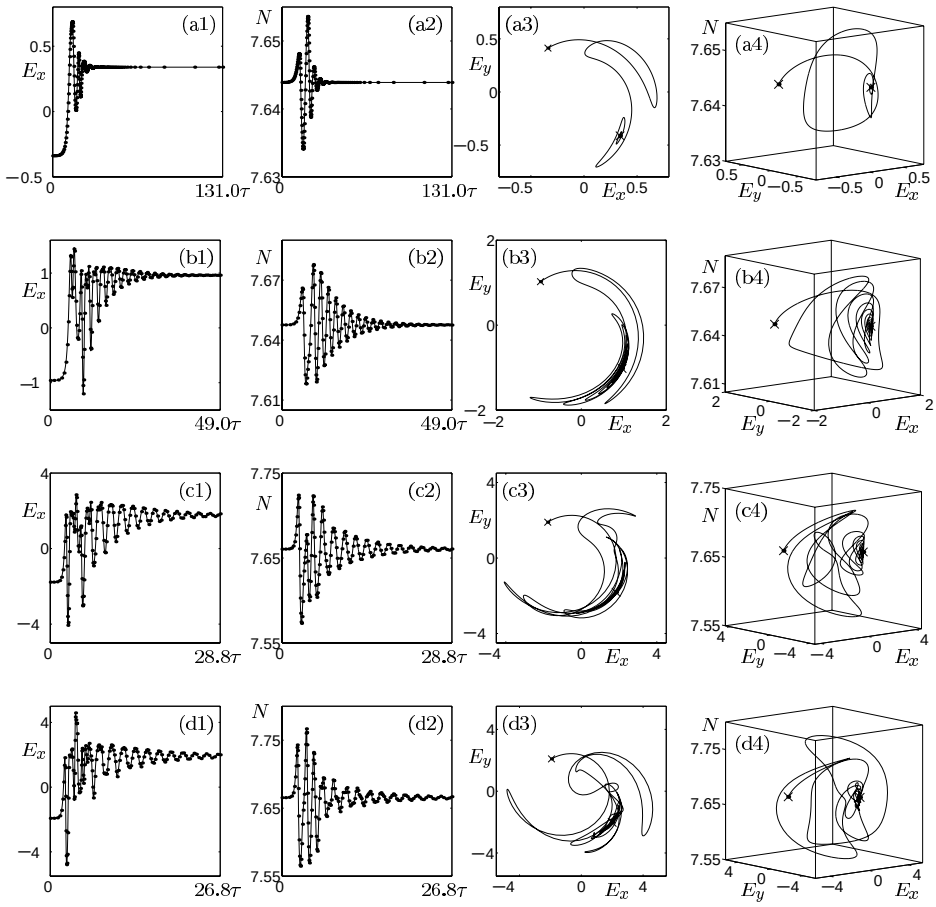


Figure 7.3. *Heteroclinic orbits along the curve Het_2; from* (a) *to* (d) $(\kappa\tau, I)$ *takes the values* $(0.182, 0.061478)$, $(0.620, 0.063430)$, $(2.153, 0.069019)$, *and* $(2.751, 0.070576)$. *(The movie file linked to the framed image shows the development of the heteroclinic orbits along the branch Het_2 in four panels. Top left: Position along the curve Het_2 in the $(\kappa\tau, I)$-plane. Top right: The heteroclinic orbit projected onto (E, N)-space. Bottom left: E_x on the truncation interval. Bottom right: N on the truncation interval.)*

point B_1, the saddle quantity is positive; this is the case of a chaotic Shil'nikov bifurcation and implies the existence of an infinite number of saddle periodic orbits associated with the heteroclinic bifurcation. Near the Belyakov point B_1 the PCF laser is excitable; that is, if a locked steady state solution is perturbed enough, it will produce a large pulse by following the nearby heteroclinic orbit before ending up at the other symmetric locked solution; see also [27]. This may lead to multipulse solutions, as was recently shown for a semiconductor laser with optical injection [45].

The existence of a Belyakov point and a T-point implies that there are infinite sequences of codimension-one homoclinic and heteroclinic orbits in their local vicinities. However, it appears to be very difficult to find and follow these solutions. So far we were unable to find further branches of connecting orbits near the Belyakov point. Near the T-point, there

are codimension-zero heteroclinic orbits, some of which one could try to find and continue. Theory dictates that the regions where they exist are bounded by curves of codimension-one heteroclinic orbits [13]. Also, in this case, we could not locate a starting value which enabled us to follow branches of these orbits.

However, during our investigations, DDE-BIFTOOL detected another branch Het_2 of connecting orbits. This new branch is shown in Figure 6.1. As is the case for Het_1, it also represents heteroclinic orbits, shown in Figure 7.3, between the nonsymmetric saddle steady states. The curve Het_2 starts at the pitchfork curve PF_2, where the nonsymmetric saddle steady states associated with the heteroclinic connection are born; again see Figure 6.2. The heteroclinic orbit shown in Figure 7.3 (a) is very close to the pitchfork curve PF_2; note that the nonsymmetric steady states, which have just been created, are very close together (Figure 7.3 (a3)). As one moves along Het_2, the orbits oscillate more as they spiral into the end steady state (Figure 7.3 (b)). Finally, near the end of the curve Het_2, the damped oscillations become more irregular with the size of the orbit increasing only slightly in (E, N)-space (Figure 7.3 (c) and (d)). One sees the emergence of an extra "arm" of the orbit in Figure 7.3 (d3). This may be an indication of a bifurcating periodic orbit responsible for the destruction of the curve Het_2 at $(\kappa\tau, I) \approx (2.833, 0.0702)$. It is clear that the heteroclinic curves Het_1 and Het_2 cannot cross, as they involve the same branch of the one-dimensional unstable manifold of the nonsymmetric saddle steady states. We note that, unlike the case of the curve Het_1, there is no interaction of the curve Het_2 with the trivial saddle steady state. The evolution of the connecting orbits is also shown in a movie accompanying Figure 7.3.

As is seen in Figure 6.2, the curve Het_2 crosses the curve ns at the point B_2, another codimension-two Belyakov point. Below the Belyakov point B_2, the heteroclinic orbits along the curve Het_2 have a negative saddle quantity corresponding to a bifurcating attracting periodic orbit; above B_2, the saddle quantity is positive, corresponding to an infinite number of bifurcating saddle periodic orbits. Again, we were unable to find further branches of heteroclinic orbits near this Belyakov point.

8. Bifurcations of periodic orbits. In this section, we make a first attempt at providing a full two-dimensional picture, including bifurcations of periodic orbits, of the locking region of the PCF laser. We performed a number of one-parameter studies, detected bifurcations of periodic orbits by studying the Floquet multipliers, and plotted the results in the $(\kappa\tau, I)$-plane.

Figure 8.1 shows colored points indicating bifurcations of periodic orbits. For different values of I, a one-parameter continuation of a periodic orbit was performed. By studying the Floquet multipliers of the system, the following bifurcations were detected: saddle-node bifurcations of limit cycles SL (yellow points) and symmetry-breaking (or restoring) bifurcations SB (purple points), both when a real Floquet multiplier passes through the unit circle at $+1$; period-doubling bifurcations PD (green points) when a real Floquet multiplier passes through the unit circle at -1; and torus (or Neimark–Sacker) bifurcations T (black points) when a complex pair of Floquet multipliers pass through the unit circle. Figure 8.1 (a) shows bifurcations of the periodic orbit originating from the Hopf curve H_1, while Figure 8.1 (b) shows bifurcations of the periodic orbit originating from the Hopf curve H_2.

Below the double-Hopf point DH, Figure 8.1 (a) clearly identifies that a saddle-node bifurcation of limit cycles SL runs close to the curve of subcritical Hopf bifurcations H_1. The

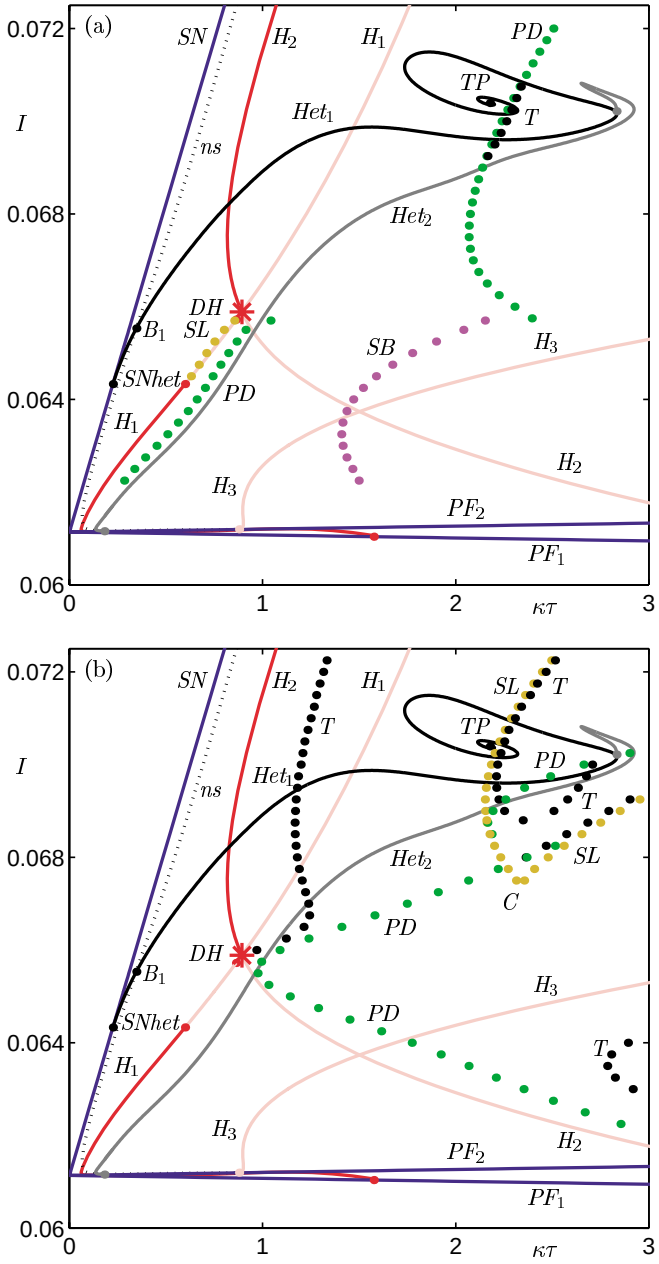


Figure 8.1. *Bifurcation diagrams in the* $(\kappa\tau, I)$*-plane; panel* (a) *shows bifurcations of periodic orbits originating from the Hopf curve* H_1*, and panel* (b) *shows bifurcations of periodic orbits originating from the Hopf curve* H_2*.*

stable periodic orbit emerging from the curve *SL* is destabilized at the curve of period-doubling bifurcations *PD* (the first in a route to chaos); see Figures 5.2 (b) and (c). Near the double-Hopf point *DH*, there is an interaction of several bifurcations [29] which we are unable to

resolve. (In particular, our calculations suggest that the curve *PD* does not end at the double-Hopf point *DH* but turns sharply to the right.) The periodic orbit that is destabilized at *PD* is destroyed in a symmetry-restoring bifurcation at the curve *SB*. Above the double-Hopf point *DH*, the bifurcation scenario is very different. The saddle periodic orbits originating from the Hopf curve H_1 are seen to undergo period-doubling bifurcations *PD* and, in a small region, torus bifurcations *T*. An obvious question is: Why is there a sudden change from a curve of symmetry-breaking bifurcations *SB*, below the double-Hopf point *DH*, to a curve of period-doubling bifurcations *PD*, above the double-Hopf point *DH*? With the present tools we cannot conclusively answer this question. However, initial investigations have revealed a possible heteroclinic connection to a saddle periodic orbit near the end points of these curves (around $(\kappa\tau, I) \approx (2.3, 0.0657)$).

The bifurcation scenario of Figure 8.1 (b) is somewhat clearer. It is now clear that the torus bifurcation *T*, identified in Figures 5.2 (e) and (f), originates from the double-Hopf point *DH*. In the lower half-plane, a large curve of period-doubling bifurcations *PD* and a smaller curve of torus bifurcations *T* are identified. The torus bifurcation *T* originating from the double-Hopf point *DH* leads to interesting dynamics. Theory states that the curve *T* comes with narrow resonance tongues, also known as *Arnold tongues* [29]. Inside these tongues we find phase-locked periodic solutions on the torus which lead to additional curves of bifurcations of periodic orbits. Furthermore, away from the curve *T*, the tongues intersect, and this results in complicated bifurcations to chaotic dynamics [29]. A detailed analysis of these resonances is beyond the scope of this paper.

We also identify a region bounded by curves of saddle-node bifurcations of limit cycles *SL*. On the curve *SL* there appears to be a codimension-two cusp point *C* of saddle-node bifurcations of limit cycles *SL* at $(\kappa\tau, I) \approx (2.3, 0.0675)$. (While it is not possible to confirm this cusp point by a two-parameter continuation, one-parameter investigations give clear evidence of a cusp.) Such regions are common in periodically driven systems [3] and laser models [44]. Inside this region, we identify torus bifurcations *T* and period-doubling bifurcations *PD*. It is known that a curve of torus bifurcations can meet a curve of saddle-node bifurcations of limit cycles at a Bogdanov–Takens bifurcation point [29]. Here we also expect a change from supercritical to subcritical of the saddle-node bifurcation of limit cycles curve. (This is also known as a 1:1 resonance.) Indeed, the lower curve *T* inside the bounded region is seen to run very close to the curve *SL*. A period-doubling curve can also meet a torus curve. Again, we would expect a change from supercritical to subcritical of the period-doubling bifurcation, a codimension-two point known as a 1:2 resonance [29]. However, at present we cannot accurately identify such points, and we leave a detailed study of bifurcations of periodic orbits as an open problem.

9. Conclusions.

We have provided a state-of-the-art two-parameter bifurcation analysis of the locking region of the PCF laser, where we paid particular attention to the computation and continuation of connecting orbits. The general picture is that the locking region is bounded by a saddle-node bifurcation and/or a heteroclinic bifurcation on one side and by Hopf bifurcations on the other; this is consistent with the one-parameter study of [18]. Pitchfork bifurcations were shown to lead to interactions between the trivial steady state and the nonsymmetric steady states of the PCF laser, thus forming the laser threshold. A number

of codimension-two bifurcations were found to be organizing centers for the dynamics of the PCF laser. These include pitchfork-Hopf bifurcations, a double-Hopf point, a Belyakov point, and a T-point bifurcation. Finally, we made a first attempt at mapping out bifurcations of periodic orbits in two parameters. This led to an overall consistent picture, in line with what is to be expected from both theory and studies of other semiconductor laser systems.

Some questions remain about further codimension-two points, particularly a cusp bifurcation and possible 1:1 and 1:2 resonance points. Furthermore, it remains a challenge to find connecting orbits in the vicinities of the Belyakov points and to find the codimension-one connecting orbits bounding the regions near the T-point where codimension-zero heteroclinic orbits exist.

To our knowledge, what was presented here is a first two-parameter bifurcation study of a DDE with continuation techniques. Our study highlights the usefulness of continuation tools for DDEs and, we hope, may encourage readers in other fields to use these new techniques.

REFERENCES

[1] G. P. Agrawal and G. R. Gray, *Effect of phase-conjugate feedback on the noise characteristics of semiconductor lasers*, Phys. Rev. A (3), 46 (1992), pp. 5890–5898.

[2] W. J. Beyn, *The numerical computation of connecting orbits in dynamical systems*, IMA J. Numer. Anal., 9 (1990), pp. 379–405.

[3] H. W. Broer and B. Krauskopf, *Chaos in periodically driven systems*, in Fundamental Issues of Nonlinear Laser Dynamics, AIP Conf. Proc. 548, American Institute of Physics, New York, 2000, pp. 31–53.

[4] 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.

[5] 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.

[6] E. Doedel, T. Fairgrieve, B. Sandstede, A. Champneys, Y. Kuznetsov, and X. Wang, *AUTO* 97: *Continuation and Bifurcation Software for Ordinary Differential Equations*, http://indy.cs.concordia.ca/auto/main.html, 1997.

[7] E. J. Doedel, M. J. Friedman, and B. I. Kunin, *Successive continuation for locating connecting orbits*, Numer. Algorithms, 17 (1997), pp. 103–124.

[8] K. Engelborghs, T. Luzyanina, and G. Samaey, *DDE-BIFTOOL v*2.00*: A Matlab Package for Bifurcation Analysis of Delay Differential Equations*, Tech. rep. TW-330, Department of Computer Science, K. U. Leuven, Belgium, http://www.cs.kuleuven.ac.be/~koen/delay/ddebiftool.shtml, 2000.

[9] I. Fischer, T. Heil, and W. Elsässer, *Emission dynamics of semiconductor lasers subject to delayed optical feedback*, in Fundamental Issues of Nonlinear Laser Dynamics, AIP Conf. Proc. 548, American Institute of Physics, New York, 2000, pp. 218–237.

[10] I. Fischer, Y. Liu, and P. Davis, *Synchronization of chaotic semiconductor laser dynamics on subnanosecond time scales and its potential for chaos communication*, Phys. Rev. A (3), 62 (2000), 011801.

[11] A. Gavrielides, *Nonlinear dynamics of semiconductor lasers: Theory and experiments*, in Fundamental Issues of Nonlinear Laser Dynamics, AIP Conf. Proc. 548, American Institute of Physics, New York, 2000, pp. 191–217.

[12] C. R. Giuliano, *Applications of optical phase conjugation*, Physics Today, 34 (1981), pp. 27–35.

[13] P. Glendinning and C. Sparrow, *T-points: A codimension two heteroclinic bifurcation*, J. Statist. Phys., 43 (1986), pp. 479–488.

[14] H. Glüsing-Lüerssen, *A behavioral approach to delay-differential systems*, SIAM J. Control Optim., 35 (1997), pp. 480–499.

[15] G. R. Gray, D. H. DeTienne, and G. P. Agrawal, *Mode locking in semiconductor lasers by phase-conjugate optical feedback*, Opt. Lett., 20 (1995), pp. 1295–1297.
[16] G. R. Gray, D. Huang, and G. P. Agrawal, *Chaotic dynamics of semiconductor lasers with phase-conjugate feedback*, Phys. Rev. A (3), 49 (1994), pp. 2096–2105.
[17] K. Green and B. Krauskopf, *Bifurcation analysis of frequency locking in a semiconductor laser with phase-conjugate feedback*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., to appear.
[18] 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.
[19] K. Green, B. Krauskopf, and K. Engelborghs, *Bistability and torus break-up in a semiconductor laser with phase-conjugate feedback*, Phys. D, 173 (2002), pp. 114–129.
[20] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1993.
[21] B. Haegeman, K. Engelborghs, D. Roose, D. Pieroux, and T. Erneux, *Stability and rupture of bifurcation bridges in semiconductor lasers subject to optical feedback*, Phys. Rev. E (3), 66 (2002), 046216.
[22] J. K. Hale, *Theory of Functional Differential Equations*, Appl. Math. Sci. 3, Springer-Verlag, New York, 1977.
[23] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.
[24] B. Krauskopf, G. R. Gray, and D. Lenstra, *Semiconductor laser with phase-conjugate feedback: Dynamics and bifurcations*, Phys. Rev. E, 58 (1998), pp. 7190–7196.
[25] B. Krauskopf and K. Green, *Computing unstable manifolds of periodic orbits in delay differential equations*, J. Comput. Phys., 186 (2003), pp. 230–249.
[26] B. Krauskopf and D. Lenstra, eds., *Fundamental Issues of Nonlinear Laser Dynamics*, AIP Conf. Proc. 548, American Institute of Physics, New York, 2000.
[27] B. Krauskopf, K. Schneider, J. Sieber, S. Wieczorek, and M. Wolfrum, *Excitability and self-pulsations near homoclinic bifurcations in semiconductor laser systems*, Opt. Commun., 215 (2003), pp. 367–379.
[28] B. Krauskopf, G. H. M. Van Tartwijk, and G. R. Gray, *Symmetry properties of lasers subject to optical feedback*, Opt. Commun., 177 (2000), pp. 347–353.
[29] Y. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd ed., Springer-Verlag, Berlin, 1998.
[30] Y. A. Kuznetsov and V. V. Levitin, *CONTENT: A multiplatform environment for continuation and bifurcation analysis of dynamical systems*, ftp://ftp.cwi.nl/pub/CONTENT, 1997.
[31] R. Lang and K. Kobayashi, *External Optical Feedback Effects on Semiconductor Injection Laser Properties*, IEEE J. Quantum Elec., 16 (1980), pp. 347–355.
[32] T. Luzyanina, K. Engelborghs, and D. Roose, *Numerical bifurcation analysis of differential equations with state-dependent delays*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 737–754.
[33] C. R. Mirasso, *Applications of semiconductor lasers to secure communications*, in Fundamental Issues of Nonlinear Laser Dynamics, AIP Conf. Proc. 548, American Institute of Physics, New York, 2000, pp. 112–127.
[34] J. Mørk, B. Tromborg, and J. Mark, *Chaos in semiconductor lasers with optical feedback: Theory and experiment*, IEEE J. Quantum Elec., 28 (1992), pp. 93–108.
[35] J. D. Murray, *Mathematical Biology*, vol. 19, Springer-Verlag, Berlin, 1980.
[36] D. Pieroux, T. Erneux, B. Haegeman, K. Engelborghs, and D. Roose, *Bridges of periodic solutions and tori in semiconductor lasers subject to delay*, Phys. Rev. Lett., 87 (2001), 193901.
[37] D. Pieroux, T. Erneux, T. Luzyanina, and K. Engelborghs, *Interacting pairs of periodic solutions lead to tori in lasers subject to delayed feedback*, Phys. Rev. E, 63 (2001), 036211.
[38] G. Samaey, K. Engelborghs, and D. Roose, *Numerical computation of connecting orbits in delay differential equations*, Numer. Algorithms, 30 (2002), pp. 335–352.
[39] B. Sandstede, *Convergence estimates for the numerical approximation of homoclinic solutions*, IMA J. Numer. Anal., 17 (1997), pp. 437–462.
[40] M. Sciamanna, T. Erneux, F. Rogister, O. Deparis, P. Megret, and M. Blondel, *Bifurcation bridges between external-cavity modes lead to polarization self-modulation in vertical-cavity surface-emitting lasers*, Phys. Rev. A (3), 65 (2002), 041801.

[41] G. H. M. Van Tartwijk and G. P. Agrawal, *Laser instabilities: A modern perspective*, Progr. Quantum Electron., 22 (1998), pp. 43–122.
[42] G. H. M. Van Tartwijk, H. J. C. Van der Linden, and D. Lenstra, *Theory of a diode laser with phase-conjugate feedback*, Opt. Lett., 17 (1995), pp. 1590–1592.
[43] G. D. Van Wiggeren and R. Roy, *Communication with chaotic lasers*, Science, 279 (1998), pp. 1198–1200.
[44] S. Wieczorek, B. Krauskopf, and D. Lenstra, *Unnested islands of period-doubling in an injected semiconductor laser*, Phys. Rev. E (3), 64 (2001), 056204.
[45] S. Wieczorek, B. Krauskopf, and D. Lenstra, *Multipulse excitability in a semiconductor laser with optical injection*, Phys. Rev. Lett., 88 (2002), 063901.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 3, pp. 277–296

Synergetic System Analysis for the Delay-Induced Hopf Bifurcation in the Wright Equation*

Michael Schanz† and Axel Pelster‡

Abstract. We apply the synergetic elimination procedure for the stable modes in nonlinear delay systems close to a dynamical instability and derive the normal form for the delay-induced Hopf bifurcation in the Wright equation. The resulting periodic orbit is confirmed by numerical simulations.

Key words. delay-induced bifurcations, normal forms, center manifold theory, synergetics

AMS subject classifications. 34K17, 37G10, 37L10

DOI. 10.1137/S1111111102412802

1. Introduction.

Within the last decades, synergetics has provided powerful concepts and methods to describe self-organization processes in various branches of science [1, 2, 3, 4, 5, 6]. The spontaneous formation of spatial, temporal, or functional patterns in complex systems has been successfully investigated by working out general principles and by mapping them onto universal mathematical structures. The important result is due to the fact that in the vicinity of a dynamical instability the high-dimensional set of nonlinear evolution equations modeling a complex system on a microscopic or a mesoscopic scale can approximately be reduced to a low-dimensional set of order parameter equations describing the evolving pattern formation on a macroscopic scale. To obtain such a simplified, reduced description of self-organization processes, the synergetic system analysis proceeds as follows. A linearization of the evolution equations around a stationary solution shows that a dynamical instability is always accompanied by a time-scale hierarchy between numerous fast modes s and few slow modes u. A rigorous treatment of the full nonlinear evolution equations in the vicinity of the dynamical instability leads to a characteristic interdependence between both hierarchy levels which may be illustrated by a circular causality chain. On the one hand, the slaving principle of synergetics states that the numerous fast modes s quasi-instantaneously take values which are prescribed by the few slow modes u according to $s(t) = h(u(t))$ with the center manifold $h(u)$. On the other hand, an adiabatic elimination of the fast enslaved modes s yields equations for the slow order parameters u which depend, in general, on the center manifold $h(u)$ due to the nonlinear feedback.

In its original formulation, the synergetic system analysis was developed for complex systems which can be modeled by ordinary and partial differential equations as well as their

*Received by the editors August 8, 2002; accepted for publication (in revised form) by R. Murray January 21, 2003; published electronically August 15, 2003.
http://www.siam.org/journals/siads/2-3/41280.html
†Institute of Parallel and Distributed Systems, University of Stuttgart, Breitwiesenstraße 20-22, D-70565 Stuttgart, Germany (Michael.Schanz@informatik.uni-stuttgart.de).
‡Institute of Theoretical Physics, Free University of Berlin, Arnimallee 14, D-14195 Berlin, Germany (pelster@physik.fu-berlin.de).

stochastic generalizations. Some time ago, the general concepts and methods of synergetics were extended to delay differential equations to deal with dynamical instabilities which are induced by the finite propagation time of signals in feedback loops [7]. Taking into account the infinite-dimensional character of a delay system [8, 9], the adiabatic elimination of the stable modes leads to a low-dimensional set of order parameter equations which turn out to be of the form of ordinary differential equations; i.e., they no longer contain memory effects. The predictions of the synergetic system analysis have been quantitatively tested by investigating the delay-induced Hopf bifurcation of the electronic system of a first-order phase-locked loop (PLL) [7]. The periodic orbit which results from the corresponding order parameter equation near the bifurcation point has been confirmed by both a multiple scale procedure and numerical simulations [10, 11]. Although this application exemplarily proves the order parameter concept for delay systems, it does not allow us to draw conclusions about the slaving principle. As the lowest nonlinear term in the scalar delay differential equation of the PLL is a cubic one, the center manifold does not influence the order parameter equation of the Hopf bifurcation in the lowest order. In order to check both ingredients of the circular causality chain, i.e., the order parameter concept and the slaving principle, for delay systems, it is thus indispensable to study a scalar delay differential equation with a quadratic nonlinearity. Such dynamical systems were studied, for instance, in the context of Lotka–Volterra models [12] and machining models [13, 14].

Another candidate is provided by the evolution equation

$$\frac{d}{dt}z(t) = R\left[z(t) - z(t-\tau)^2\right]. \tag{1.1}$$

With vanishing time delay τ, it represents a system which is named after the Belgian mathematician P. F. Verhulst from the 19th century [15]. It is used as a simplified model for the population dynamics of a species in an environment with limited food supply [16]. The synergetic system analysis for the Verhulst system with time delay (1.1) has already been performed in [10]. There also the well-known equation of Wright [17]

$$\frac{d}{dt}z(t) = -Rz(t-\tau)\left[1 + z(t)\right] \tag{1.2}$$

has been treated, where R denotes a system parameter and τ a delay time. This delay differential equation is mentioned by Wright [17] as arising in the application of probability methods to the theory of asymptotic prime number density. Cunningham [18] depicts it as a "growth equation" representing a mathematical description of a fluctuating population of organisms under certain environmental conditions. In addition, it may describe the operation of a control system working with potentially explosive chemical reactions. Performing an appropriate scaling of time

$$t = \tau\, t', \quad z'(t') = z(\tau\, t') \tag{1.3}$$

converts the Wright equation (1.2) to its standard form with the control parameter

$$R' = \tau\, R. \tag{1.4}$$

Thus varying the delay time τ corresponds to changing the control parameter R'. By omitting the prime $'$ for the respective quantities, the standard form of the Wright equation reads

$$\frac{d}{dt}z(t) = -Rz(t-1)\,[1+z(t)]\,. \tag{1.5}$$

In this paper, we restrict ourselves to analyzing this standard form of the Wright equation.

The Wright equation (1.5) shows a delay-induced instability, namely, a Poincaré–Andronov–Hopf bifurcation at the critical value

$$R_c = \frac{\pi}{2} \tag{1.6}$$

of the control parameter R. In [19], it is shown that the oscillatory solution in the vicinity of this instability, i.e., the emerging limit cycle, can be calculated approximately using the method of *averaging*. This approximation reads in the lowest order

$$z(t) = A\sqrt{R-\frac{\pi}{2}}\,\cos\left(\frac{\pi}{2}t\right) + \mathcal{O}\left(R-\frac{\pi}{2}\right), \tag{1.7}$$

where the amplitude A has the value

$$A = \sqrt{\frac{40}{3\pi-2}}. \tag{1.8}$$

In section 2, we start with a linear stability analysis of the Wright equation (1.5) which confirms, of course, the delay-induced Poincaré–Andronov–Hopf bifurcation when the control parameter R approaches the critical value (1.6). Near this instability, we perform a nonlinear synergetic treatment in section 3 and study in detail how the center manifold influences the order parameter equation. In section 4, the resulting order parameter equation is transformed to the normal form of a Hopf bifurcation, where the emerging periodic orbit is determined one order higher than the lowest-order result (1.7) and (1.8). The numerical investigations of section 5 confirm the emerging periodic orbit; furthermore, we discuss the global bifurcation scenario of the Wright equation (1.5).

2. Linear stability analysis. The solution of the delay differential equation (1.5) for times $t \geq 0$ depends on the initial values of the function $z(t)$ in the entire interval $[-1,0]$. In order to properly define such an initial value problem, Hale [8] and Krasovskii [9] proposed to transform the equation of motion (1.5) for a function $z(t)$ in the usual state space Γ to the extended state space $\mathcal{C}$ of continuous complex valued functions z_t, which are defined on the interval $[-1,0]$:

$$\frac{d}{dt}z_t(\Theta) = (\mathcal{G}\,z_t)\,(\Theta) = \begin{cases} \dfrac{d}{d\Theta}z_t(\Theta), & -1 \leq \Theta < 0, \\ \mathcal{F}[z_t], & \Theta = 0. \end{cases} \tag{2.1}$$

Following the notation of [7], we introduced not only the new function $z_t \in \mathcal{C}$, which is connected to the original function $z(t) \in \Gamma$ through

$$z_t(\Theta) = z(t+\Theta)\,, \quad -1 \leq \Theta \leq 0, \tag{2.2}$$

but also the nonlinear functional

$$\mathcal{F}[z_t] = \sum_{k=1}^{2} \int_{-1}^{0} d\Theta_1 \cdots \int_{-1}^{0} d\Theta_k \, \omega^{(k)}(\Theta_1, \ldots, \Theta_k) \prod_{l=1}^{k} z_t(\Theta_l) \tag{2.3}$$

with the two scalar densities

$$\omega^{(1)}(\Theta_1) = -R\boldsymbol{\delta}(\Theta_1 + 1), \tag{2.4}$$

$$\omega^{(2)}(\Theta_1, \Theta_2) = -R\boldsymbol{\delta}(\Theta_1 + 1)\boldsymbol{\delta}(\Theta_2). \tag{2.5}$$

The stationary states of this system

$$z^{\mathrm{I}}_{\mathrm{stat}} = 0\,, \qquad z^{\mathrm{II}}_{\mathrm{stat}} = -1 \tag{2.6}$$

are candidates for the reference state from which we start our further investigations. For the main body of the article, we focus our attention on the stationary state $z^{\mathrm{I}}_{\mathrm{stat}}$ and choose it as the reference state. The other stationary state $z^{\mathrm{II}}_{\mathrm{stat}}$ will be discussed in section 5 together with the global bifurcation scenario of the Wright equation (1.5).

Then we linearize the system (2.1) with respect to the stationary state $z^{\mathrm{I}}_{\mathrm{stat}} = 0$ by using the decomposition

$$z_t(\Theta) = z^{\mathrm{I}}_{\mathrm{stat}} + \zeta_t(\Theta)\,, \quad -1 \le \Theta \le 0. \tag{2.7}$$

This leads to the following linearized equation of motion for the deviation $\zeta_t(\Theta)$ from the stationary state $z^{\mathrm{I}}_{\mathrm{stat}} = 0$:

$$\frac{d}{dt}\zeta_t(\Theta) = (\mathcal{G}_L\,\zeta_t)\,(\Theta) = \begin{cases} \dfrac{d}{d\Theta}\zeta_t(\Theta), & -1 \le \Theta < 0, \\ \mathcal{L}[\zeta_t], & \Theta = 0, \end{cases} \tag{2.8}$$

where the linear functional is given by

$$\mathcal{L}\,[\zeta_t] = \int_{-1}^{0} d\Theta\, \omega(\Theta)\zeta_t(\Theta) \tag{2.9}$$

with the scalar density

$$\omega(\Theta) = \left.\frac{\delta\mathcal{F}[z_t]}{\delta z_t(\Theta)}\right|_{z_t(\Theta)=z^{\mathrm{I}}_{\mathrm{stat}}} = -R\,\boldsymbol{\delta}(\Theta + 1). \tag{2.10}$$

Inserting the solution ansatz

$$\zeta_t(\Theta) = \phi^{\lambda}(\Theta)e^{\lambda t}, \quad -1 \le \Theta \le 0, \tag{2.11}$$

into (2.8) leads to the eigenvalue problem of the infinitesimal generator $\mathcal{G}_L$:

$$\lambda\phi^{\lambda}(\Theta) = \left(\mathcal{G}_L\phi^{\lambda}\right)(\Theta), \quad -1 \le \Theta \le 0. \tag{2.12}$$

Taking into account the definition of $\mathcal{G}_L$ in (2.8), the eigenfunction $\phi^\lambda(\Theta)$ is determined to be

$$\phi^\lambda(\Theta) = N_\lambda e^{\lambda\Theta}, \quad -1 \le \Theta \le 0, \tag{2.13}$$

and the eigenvalue λ follows from

$$\lambda = L(\lambda), \tag{2.14}$$

where $L(\lambda)$ is defined by

$$L(\lambda) = \int_{-1}^{0} d\Theta\, \omega(\Theta) e^{\lambda\Theta}. \tag{2.15}$$

Using the scalar density (2.10), we obtain the following transcendental characteristic equation:

$$-Re^{-\lambda} - \lambda = 0. \tag{2.16}$$

Thus the spectrum of the linear operator $\mathcal{G}_L$ has the following properties [8]:

- It consists of a countable infinite number of eigenvalues which cumulate for $\Re(\lambda) \to -\infty$.
- It is confined by an upper threshold for the real parts of the eigenvalues.
- At the bifurcation point, i.e., the instability, some of the eigenvalues reach the imaginary axes and thus become unstable.

Further properties of the eigenvalues of the characteristic equation (2.16) follow from the Hayes theorem, which can be found in [20]. It states that all solutions of the transcendental equation

$$p + qe^{-\lambda} - \lambda = 0 \tag{2.17}$$

possess a negative real part if and only if (a) $p < 1$ and (b) $p < -q < \sqrt{a_1^2 + p^2}$. Here a_1 represents the solution of the transcendental equation $a_1 = p\tan(a_1)$ which lies in the interval $[0, \pi)$. For the special case $p = 0$, one can show that a_1 is equal to $\pi/2$. The shaded region in Figure 2.1 represents that region of the parameter space q, p where both conditions of the Hayes theorem are fulfilled. The upper boundary line stems from (a) $p < 1$ and (b1) $p < -q$, whereas the lower boundary line follows from (a) $p < 1$ and (b2) $-q < \sqrt{a_1^2 + p^2}$.

Comparing (2.16) with (2.17), we obtain the identification $q = -R$ and $p = 0$. Changing the control parameter R from 0 to $\pi/2$, the corresponding point in the parameter space q, p moves along the q-axis from the point $q = 0$ to $q = -\pi/2$ (see the arrow in Figure 2.1). At this critical value, it reaches the boundary of the shaded stability region; i.e., no longer do all solutions of the characteristic equation (2.16) have a negative real part. Therefore, an instability occurs at $R_c = \pi/2$.

Figure 2.2 confirms this result by illustrating the movement of the ten solutions of the characteristic equation (2.16) with the largest real part when the control parameter is increased from 0 to $\pi/2$. The eigenvalues were obtained with a Newton algorithm, and the control parameter R was increased in equidistant steps. For $R = 0$, there exists only one

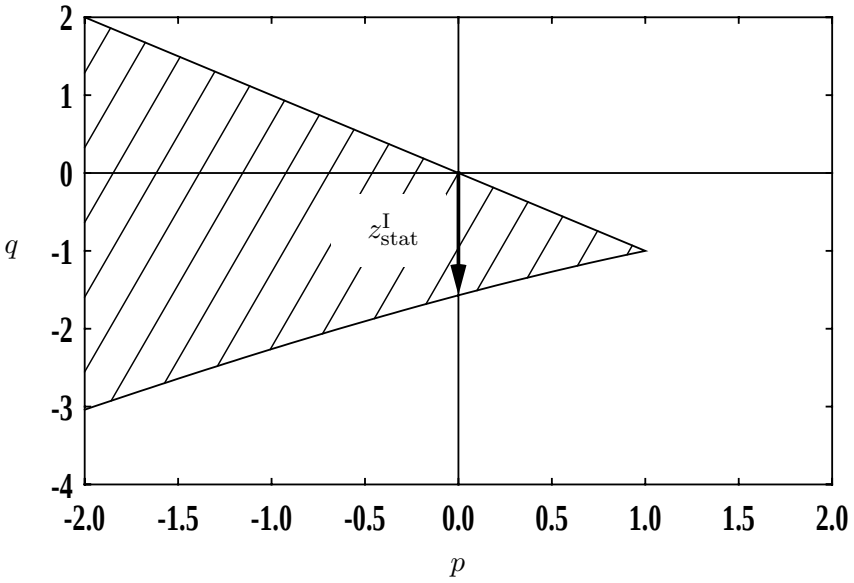


Figure 2.1. *The Hayes theorem is fulfilled within the shaded region.*

real eigenvalue 0 as the linearized delay differential equation (2.8) degenerates to an ordinary differential equation. For $R > 0$, this eigenvalue remains real and becomes negative. Furthermore, a countable infinite number of conjugate complex eigenvalues and another real eigenvalue emerge from an infinite negative real part. At the value $R = 1/e$, both real eigenvalues meet at the point $(-1/0)$. They are converted to a pair of conjugated complex eigenvalues for $1/e < R < \pi/2$. These two complex conjugated eigenvalues have zero real part at the instability $R_c = \pi/2$, which thus represents a Hopf bifurcation. We can further analyze this instability by introducing the smallness parameter

$$\varepsilon = \frac{R - R_c}{R_c} \quad \Longleftrightarrow \quad R = R_c(1 + \varepsilon) \tag{2.18}$$

for the deviation from the critical control parameter $R_c = \pi/2$. In particular, we can determine both eigenvalues $\lambda_u^{\pm}(\varepsilon)$ with nearly vanishing real part at $\varepsilon \approx 0$ from the characteristic equation (2.16):

$$\lambda_u^{\pm}(\varepsilon) = \frac{R_c^2}{1 + R_c^2}\,\varepsilon \pm i R_c \left(1 + \frac{1}{1 + R_c^2}\,\varepsilon\right) + \mathcal{O}\left(\varepsilon^2\right). \tag{2.19}$$

In the vicinity of the instability $\varepsilon \approx 0$, we read off from Figure 2.2 that only the two eigenvalues (2.19) have nearly vanishing real part; all other eigenvalues have a large negative real part:

$$\Re[\lambda_u^{\pm}(\varepsilon \approx 0)] \approx 0; \quad \Re[\lambda_s^{j}(\varepsilon \approx 0)] < 0, \quad j = 1, \ldots, \infty. \tag{2.20}$$

This characteristic property of the linearized system (2.8) leads to the time-scale hierarchy

$$T_u^{\pm} = \frac{1}{\Re[\lambda_u^{\pm}(\varepsilon \approx 0)]} \gg T_s^{j} = \frac{1}{\Re[\lambda_s^{j}(\varepsilon \approx 0)]}, \quad j = 1, \ldots, \infty. \tag{2.21}$$

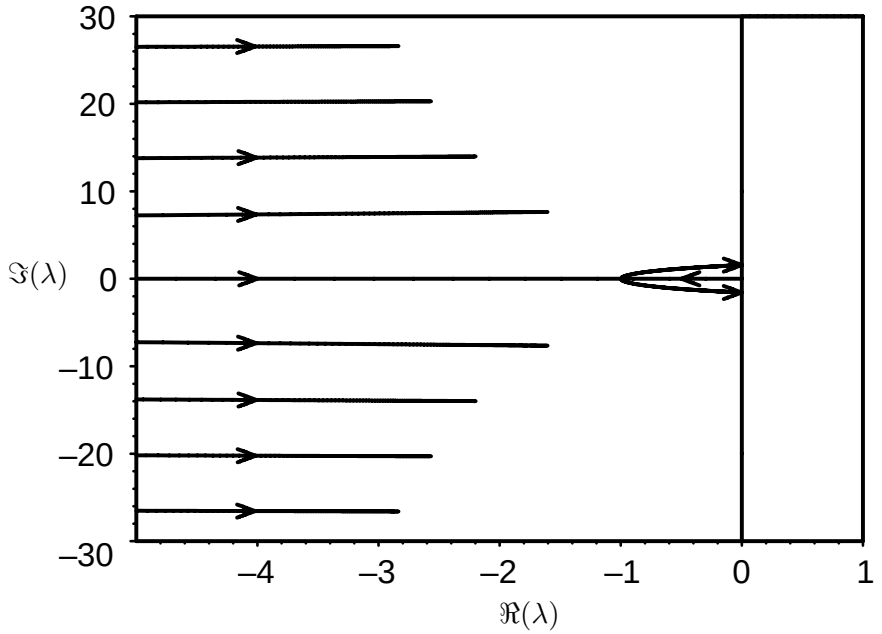


Figure 2.2. *Movement of the ten solutions of the characteristic equation* (2.16) *with the largest real part.*

Thus the infinite-dimensional extended state space $\mathcal{C}$ decomposes in a two-dimensional subspace $\mathcal{U}$ of the linear unstable modes and a remaining infinite-dimensional subspace $\mathcal{S}$ of the linear stable modes [7]. As a consequence, the extended state function z_t can be decomposed near the instability according to

$$z_t(\Theta) = z_{\text{stat}}^{\text{I}} + u_t(\Theta) + s_t(\Theta) = u_t(\Theta) + s_t(\Theta), \quad -1 \leq \Theta \leq 0, \tag{2.22}$$

as we have $z_{\text{stat}}^{\text{I}} = 0$. Here u_t and s_t denote the respective contributions of z_t in the subspaces $\mathcal{U}$ and $\mathcal{S}$. In order to project into these subspaces, we need the linear unstable modes

$$\phi^{\lambda_u^\pm}(\Theta) = N_{\lambda_u^\pm} e^{\lambda_u^\pm \Theta}, \quad -1 \leq \Theta \leq 0, \tag{2.23}$$

of the system (2.8) which have already been determined in (2.13). However, this knowledge is not sufficient, as the infinitesimal generator $\mathcal{G}_L$ is not self-adjoint. Therefore, we also need the linear unstable modes

$$\psi^{\dagger\lambda_u^\pm}(s) = N_{\lambda_u^\pm} e^{-\lambda_u^\pm s}, \quad 0 \leq s \leq 1, \tag{2.24}$$

of the adjoint system

$$\frac{d}{dt}\zeta_t^\dagger(s) = -\left(\mathcal{G}_L^\dagger\, \zeta_t^\dagger\right)(s) = \begin{cases} \dfrac{d}{ds}\zeta_t^\dagger(s), & 0 < s \leq 1, \\ -\mathcal{L}^\dagger[\zeta_t^\dagger], & s = 0, \end{cases} \tag{2.25}$$

where the linear functional

$$\mathcal{L}^\dagger\left[\zeta_t^\dagger\right] = \int_0^1 ds\, \omega(-s)\zeta_t^\dagger(s) \tag{2.26}$$

also contains the scalar density (2.10). Indeed, the solution ansatz

$$\zeta_t^\dagger(s) = \psi^{\dagger\lambda}(s)e^{-\lambda s}, \quad 0 \le s \le 1, \tag{2.27}$$

converts (2.25) to the eigenvalue problem

$$\lambda\psi^{\dagger\lambda}(s) = \left(\mathcal{G}_L^\dagger \psi^{\dagger\lambda}\right)(s), \quad 0 \le s \le 1, \tag{2.28}$$

which is adjoint to (2.12). Note that $\psi^{\dagger\lambda}$ and $\zeta_t^\dagger$ are elements of the dual extended state space $\mathcal{C}^\dagger$, which consists of continuous complex valued functions on the interval $[0, 1]$. The relation between both extended state spaces $\mathcal{C}$ and $\mathcal{C}^\dagger$ is defined by the bilinear form [7]

$$\left(\psi^\dagger|\phi\right) = \psi^\dagger(0)\phi(0) - \int_{-1}^{0} d\Theta \int_0^\Theta ds\, \psi^\dagger(s-\Theta)\omega(\Theta)\phi(s). \tag{2.29}$$

Using this bilinear form, one can show that the eigenfunctions (2.23) and (2.24) are biorthonormal:

$$\left(\psi^{\dagger\lambda_i}|\phi^{\lambda_j}\right) = \delta_{ij}, \quad i, j = \pm. \tag{2.30}$$

This determines the yet unknown normalization constants to be

$$N_{\lambda_u^\pm} = \frac{1}{\sqrt{1+\lambda_u^\pm}}, \tag{2.31}$$

so they reduce near the instability because of (2.19) to

$$N_{\lambda_u^\pm} = \frac{1}{\sqrt{1 \pm iR_c}} + \mathcal{O}(\varepsilon). \tag{2.32}$$

Furthermore, the bilinear form (2.29) allows us to define the projector into the two-dimensional subspace $\mathcal{U}$ of the unstable modes:

$$(\mathcal{P}_u \bullet)(\Theta) = \sum_{i=\pm} \phi^{\lambda_u^i}(\Theta)\left(\psi^{\dagger\lambda_u^i}\middle|\bullet\right). \tag{2.33}$$

Correspondingly, the projector into the remaining infinite-dimensional subspace $\mathcal{S}$ of the stable modes reads

$$\mathcal{P}_s \bullet = (\mathcal{I} - \mathcal{P}_u)\bullet. \tag{2.34}$$

Applying the projector $\mathcal{P}_u$ to $z_t \in \mathcal{C}$ leads to $u_t \in \mathcal{U}$ according to

$$u_t(\Theta) = (\mathcal{P}_u z_t)(\Theta) = \sum_{i=\pm} u^i(t)\phi^{\lambda_u^i}(\Theta), \quad -1 \le \Theta \le 0, \tag{2.35}$$

where the amplitudes of the linear unstable modes $\phi^{\lambda_u^\pm}(\Theta)$ are defined by

$$u^\pm(t) = \left(\psi^{\dagger\lambda_u^\pm}\middle|z_t\right). \tag{2.36}$$

Later on, these amplitudes represent the order parameters which indicate the emergence of an instability. Analogously, the projector (2.34) leads to the stable modes

$$s_t(\Theta) = (\mathcal{P}_s z_t)(\Theta), \quad -1 \le \Theta \le 0. \tag{2.37}$$

3. Nonlinear synergetic analysis. After having performed a linear stability analysis around the reference state $z^{\mathrm{I}}_{\mathrm{stat}} = 0$ in the vicinity of the instability $R_c = \pi/2$, we now return to our original nonlinear evolution equation (2.1) in the extended state space $\mathcal{C}$. We proceed by decomposing the generator $\mathcal{G}$ into its linear part $\mathcal{G}_L$ and a remaining effective nonlinear part:

$$\frac{d}{dt} z_t(\Theta) = \left(\mathcal{G}_L\, z_t\right)(\Theta) + X_0(\Theta)\mathcal{F}^{\mathrm{eff}}[z_t], \quad -1 \le \Theta \le 0. \tag{3.1}$$

Here we introduce the scalar function

$$X_0(\Theta) = \begin{cases} 0, & -1 \le \Theta < 0, \\ 1, & \Theta = 0, \end{cases} \tag{3.2}$$

and the effective nonlinear functional

$$\mathcal{F}^{\mathrm{eff}}[z_t] = \int_{-1}^{0} d\Theta_1 \int_{-1}^{0} d\Theta_2\, \omega^{(2)}(\Theta_1, \Theta_2)\, z_t(\Theta_1) z_t(\Theta_2) \tag{3.3}$$

with the scalar density (2.5). Using the projectors (2.33) and (2.34) and their properties (2.35)–(2.37), we can investigate the respective contributions of the order parameters $u^{\pm}(t)$ and the linear stable modes $s_t \in \mathcal{S}$ to the nonlinear dynamics (3.1). Thus we obtain the following system of coupled nonlinear mode equations:

$$\frac{d}{dt} u^{\pm}(t) = \lambda^{\pm}_u u^{\pm}(t) + \psi^{\dagger \lambda^{\pm}_u}(0)\, \mathcal{F}^{\mathrm{eff}} \left[\sum_{j=\pm} \phi^{\lambda^j_u} u^j(t) + s_t \right], \tag{3.4}$$

$$\frac{d}{dt} s_t(\Theta) = \left(\mathcal{G}_L s_t\right)(\Theta) + \left(\left(\mathcal{I} - \mathcal{P}_u\right) X_0\right)(\Theta)\, \mathcal{F}^{\mathrm{eff}} \left[\sum_{j=\pm} \phi^{\lambda^j_u} u^j(t) + s_t \right]. \tag{3.5}$$

It is still exact and describes completely the nonlinear dynamics. However, a solution to these equations can only be found by means of an approximation method. Such a well-established approximative solution is provided by the slaving principle of synergetics [1, 2, 3, 4, 5].

To this end, we start with the time-scale hierarchy (2.21) near the instability which leads to the fact that the dynamics of the stable modes $s_t \in \mathcal{S}$ evolves much faster than the order parameters $u^{\pm}(t)$. In [7] it has been shown for a quite general class of delay differential equations that such a time-scale hierarchy leads to a slaving of the stable modes; i.e., the numerous fast modes $s_t \in \mathcal{S}$ quasi-instantaneously take values which are prescribed by the few slow order parameters $u^{\pm}(t)$. In our context, the slaving principle states mathematically that the dynamics of the stables modes $s_t \in \mathcal{S}$ is determined by the center manifold $h(\Theta, u^+, u^-)$ according to

$$s_t(\Theta) = h\left(\Theta, u^+(t), u^-(t)\right). \tag{3.6}$$

Inserting this ansatz into (3.5) leads to an implicit equation for the center manifold $h(\Theta, u^+, u^-)$:

$$\begin{aligned} \sum_{i=\pm} \frac{\partial h\left(\Theta, u^+(t), u^-(t)\right)}{\partial u^i(t)} \left(\lambda^i_u u^i(t) + \psi^{\dagger \lambda^i_u}(0)\; \mathcal{F}^{\mathrm{eff}} \left[\sum_{j=\pm} \phi^{\lambda^j_u} u^j(t) + h \right] \right) \\ = \left(\mathcal{G}_L h\right)(\Theta) + \left(\left(\mathcal{I} - \mathcal{P}_u\right) X_0\right)(\Theta)\, \mathcal{F}^{\mathrm{eff}} \left[\sum_{j=\pm} \phi^{\lambda^j_u} u^j(t) + h \right]. \end{aligned} \tag{3.7}$$

It can be approximately solved in the vicinity of the instability as follows. We assume that the order parameters $u^{\pm}(t)$ possess a certain dependence on the smallness parameter (2.18) which is typical for a Hopf bifurcation:

$$u^{\pm}(t) = \mathcal{O}(\varepsilon^{1/2}). \tag{3.8}$$

Furthermore, we perform for the center manifold $h(\Theta, u^+, u^-)$ the lowest-order ansatz

$$h(\Theta, u^+, u^-) = \sum_{j_1=\pm} \sum_{j_2=\pm} H_{j_1 j_2}(\Theta) u^{j_1}(t) u^{j_2}(t), \tag{3.9}$$

as $r = 2$ is the order of the effective nonlinear functional (3.3). From (3.8) and (3.9), it follows then in lowest order of ε that the effective nonlinear functional $\mathcal{F}^{\text{eff}}$ in (3.7) can be approximated by

$$\mathcal{F}^{\text{eff}}\left[\sum_{j=\pm} \phi^{\lambda_u^j} u^j(t) + h\right] \approx \sum_{j_1=\pm} \sum_{j_2=\pm} F^{\text{eff}}_{j_1 j_2} u^{j_1}(t) u^{j_2}(t), \tag{3.10}$$

where the coefficients $F^{\text{eff}}_{j_1 j_2}$ read

$$F^{\text{eff}}_{j_1 j_2} = \int_{-1}^{0} d\Theta_1 \int_{-1}^{0} d\Theta_2 \, \omega^{(2)}(\Theta_1, \Theta_2) \, \phi^{\lambda_u^{j_1}}(\Theta_1) \phi^{\lambda_u^{j_2}}(\Theta_2). \tag{3.11}$$

Taking into account (2.5) and (2.23), these coefficients turn out to be

$$F^{\text{eff}}_{++} = F^{\text{eff}\,*}_{--} = -RN^2_{\lambda_u^+} e^{-\lambda_u^+}, \quad F^{\text{eff}}_{+-} = F^{\text{eff}}_{-+} = -RN_{\lambda_u^+} N_{\lambda_u^-} e^{-\lambda_u^+}. \tag{3.12}$$

As a consequence, we conclude from (3.7) in lowest order of ε that the coefficients $H_{j_1 j_2}(\Theta)$ of the center manifold (3.9) are given by

$$H_{j_1 j_2}(\Theta) = F^{\text{eff}}_{j_1 j_2} K_{j_1 j_2}(\Theta), \tag{3.13}$$

where the coefficients $K_{j_1 j_2}(\Theta)$ follow from

$$K_{j_1 j_2}(\Theta) = \left([\mathcal{G}_L - \Lambda]^{-1} (\mathcal{P}_u X_0 - X_0) \right)(\Theta) \tag{3.14}$$

with the abbreviation

$$\Lambda = \sum_{k=1}^{2} \lambda_u^{j_k}. \tag{3.15}$$

In [7] it is shown that the operator $[\mathcal{G}_L - \Lambda]^{-1}$ has the explicit representation

$$\begin{aligned} \left([\mathcal{G}_L - \Lambda]^{-1} \chi \right)(\Theta) &= \int_0^{\Theta} ds \, e^{\Lambda(\Theta - s)} \chi(s) \\ &+ [L(\Lambda) - \Lambda]^{-1} \left(\chi(0) - \int_{-1}^{0} d\Theta \int_0^{\Theta} ds \, e^{\Lambda(\Theta - s)} \omega(\Theta) \chi(s) \right) e^{\Lambda \Theta}, \end{aligned} \tag{3.16}$$

where $L(\lambda)$ is already defined in (2.15). After some calculation, which also involves (2.10), (2.23), (2.24), (2.33), and (3.2), it thus follows that the coefficients (3.14) are given by

$$K_{j_1 j_2}(\Theta) = \sum_{j=\pm} \frac{N^2_{\lambda_u^j}}{\lambda_u^j - \Lambda} e^{\lambda_u^j \Theta} - \frac{e^{\Lambda\Theta}}{L(\Lambda) - \Lambda}. \tag{3.17}$$

Thus, together with (2.10), (2.15), and (3.15), we obtain

$$K_{++}(\Theta) = K^*_{--}(\Theta) = -\frac{N^2_{\lambda_u^+} e^{\lambda_u^+ \Theta}}{\lambda_u^+} + \frac{N^2_{\lambda_u^-} e^{\lambda_u^- \Theta}}{\lambda_u^- - 2\lambda_u^+} + \frac{e^{2\lambda_u^+ \Theta}}{Re^{-2\lambda_u^+} + 2\lambda_u^+}, \tag{3.18}$$

$$K_{+-}(\Theta) = K_{-+}(\Theta) = -\frac{N^2_{\lambda_u^+} e^{\lambda_u^+ \Theta}}{\lambda_u^-} - \frac{N^2_{\lambda_u^-} e^{\lambda_u^- \Theta}}{\lambda_u^+} + \frac{e^{(\lambda_u^+ + \lambda_u^-)\Theta}}{Re^{-(\lambda_u^+ + \lambda_u^-)} + \lambda_u^+ + \lambda_u^-}. \tag{3.19}$$

This completes the lowest-order result for the center manifold $h(\Theta, u^+, u^-)$, which is given by (3.9), (3.12), (3.13), (3.18), and (3.19).

Thus we can now consider the order parameter equation (3.4). In lowest order in ε we take into account (2.5), (2.25), (3.3), (3.6), and (3.9) so that it reduces to

$$\begin{aligned} \frac{d}{dt} u^\pm(t) &= \lambda_u^\pm u^\pm(t) \\ &- RN_{\lambda_u^\pm} \prod_{l=1}^{2} \left[\sum_{j=\pm} \phi^{\lambda_u^j}(\vartheta_l) u^j(t) + \sum_{j_1=\pm} \sum_{j_2=\pm} H_{j_1 j_2}(\vartheta_l) u^{j_1}(t) u^{j_2}(t) \right], \end{aligned} \tag{3.20}$$

where we set

$$\vartheta_l = \begin{cases} -1, & l = 1, \\ 0, & l = 2. \end{cases} \tag{3.21}$$

Note that the order parameter equation (3.20) turns out to be an ordinary differential equation; i.e., it no longer contains memory effects. Furthermore, we observe that the center manifold explicitly enters the order parameter equation (3.20) as a direct consequence of the quadratic nonlinearity of the Wright equation (1.5). We remark that this effect, which is essential for the present synergetic analysis, was neglected in the neurophysiological study in [23]. In the subsequent section we show how the order parameter equation (3.20) is converted to the normal form of a Hopf bifurcation.

4. Normal form. Now we perform a nonlinear transformation of the order parameters which eliminates those terms which are irrelevant for the normal form of a Hopf bifurcation. As far as the so-called *near identity transformation* and the theory of normal forms in general is concerned, we refer to the [19, 21, 22]. The terms in (3.20) which are relevant for the normal form of a Hopf bifurcation read

$$\frac{d}{dt} u^\pm(t) = \lambda_u^\pm u^\pm(t) + q_0^\pm u^\pm(t)^2 + q_1^\pm u^\pm(t) u^\mp(t) + q_2^\pm u^\mp(t)^2 + k_1^\pm u^\pm(t)^2 u^\mp(t) \tag{4.1}$$

as we can neglect quartic terms and nonresonant cubic terms due to the rotating wave approximation. The respective coefficients in (4.1) are given by

$$q_0^{\pm} = -RN_{\lambda_u^{\pm}}\phi^{\lambda_u^{\pm}}(-1)\phi^{\lambda_u^{\pm}}(0) = -RN_{\lambda_u^{\pm}}{}^3 e^{-\lambda_u^{\pm}}, \tag{4.2}$$

$$\begin{aligned} q_1^{\pm} &= -RN_{\lambda_u^{\pm}}\left[\phi^{\lambda_u^{\pm}}(-1)\phi^{\lambda_u^{\mp}}(0) + \phi^{\lambda_u^{\pm}}(0)\phi^{\lambda_u^{\mp}}(-1)\right] \\ &= -RN_{\lambda_u^{\pm}}{}^2 N_{\lambda_u^{\mp}}\left(e^{-\lambda_u^{\pm}} + e^{-\lambda_u^{\mp}}\right), \end{aligned} \tag{4.3}$$

$$q_2^{\pm} = -RN_{\lambda_u^{\pm}}\phi^{\lambda_u^{\mp}}(-1)\phi^{\lambda_u^{\mp}}(0) = -RN_{\lambda_u^{\pm}}N_{\lambda_u^{\mp}}{}^2 e^{-\lambda_u^{\mp}}, \tag{4.4}$$

$$\begin{aligned} k_1^{\pm} = -RN_{\lambda_u^{\pm}}\Big\{&\phi^{\lambda_u^{\pm}}(-1)\left[H_{+-}(0) + H_{-+}(0)\right] + \phi^{\lambda_u^{\mp}}(-1)H_{++}(0) \\ &+ \phi^{\lambda_u^{\pm}}(0)\left[H_{+-}(-1) + H_{-+}(-1)\right] + \phi^{\lambda_u^{\mp}}(0)H_{++}(-1)\Big\}, \end{aligned} \tag{4.5}$$

where we did not write down the explicit form of $k_1^{\pm}$ for simplicity. Then the previous order parameters $u^{\pm}(t)$ are transformed to new order parameters $v^{\pm}(t)$ by the near identity transformation

$$u^{\pm}(t) = v^{\pm}(t) + \alpha_0^{\pm}v^{\pm}(t)^2 + \alpha_1^{\pm}v^{\pm}(t)v^{\mp}(t) + \alpha_2^{\pm}v^{\mp}(t)^2, \tag{4.6}$$

with the yet-unknown coefficients $\alpha_0^{\pm}$, $\alpha_1^{\pm}$, and $\alpha_2^{\pm}$. As the $u^{\pm}(t)$ are small quantities in the vicinity of the instability, the same holds for the $v^{\pm}(t)$. Inserting (4.6) in (4.1), we obtain a system of ordinary differential equations of the form

$$M(t)\frac{d}{dt}\begin{pmatrix} v^+(t) \\ v^-(t) \end{pmatrix} = \begin{pmatrix} w^+(t) \\ w^-(t) \end{pmatrix}, \tag{4.7}$$

where the matrix $M(t)$ is defined by

$$M(t) = \begin{pmatrix} 1 + 2\alpha_0^+ v^+(t) + \alpha_1^+ v^-(t) & \alpha_1^+ v^+(t) + 2\alpha_2^+ v^-(t) \\ \alpha_1^- v^-(t) + 2\alpha_2^- v^+(t) & 1 + 2\alpha_0^- v^-(t) + \alpha_1^- v^+(t) \end{pmatrix}. \tag{4.8}$$

For simplicity, we do not write the explicit form of $w^+(t)$ and $w^-(t)$, but we note that they contain $v^+(t)$ and $v^-(t)$ at least in first order. Thus we obtain from (4.7)

$$\frac{d}{dt}\begin{pmatrix} v^+(t) \\ v^-(t) \end{pmatrix} = M^{-1}(t)\begin{pmatrix} w^+(t) \\ w^-(t) \end{pmatrix}, \tag{4.9}$$

with the inverse matrix

$$M(t)^{-1} = \frac{1}{\operatorname{Det} M(t)}\begin{pmatrix} M_{22}(t) & -M_{12}(t) \\ -M_{21}(t) & M_{11}(t) \end{pmatrix}, \tag{4.10}$$

where the determinant has the form

$$\begin{aligned} \operatorname{Det} M(t) = 1 + v^+(t)\left(2\alpha_0^+ + \alpha_1^-\right) \\ + 2v^+(t)v^-(t)\left(\alpha_0^+\alpha_0^- - \alpha_2^+\alpha_2^-\right) + 2v^+(t)^2\left(\alpha_0^+\alpha_1^- - \alpha_1^+\alpha_2^-\right) + \text{c.c.} \end{aligned} \tag{4.11}$$

Expanding the right-hand side of (4.9) in powers of $v^+(t)$ and $v^-(t)$ up to the third order, we yield

$$\begin{aligned}\frac{d}{dt}v^{\pm}(t) &= \lambda^{\pm}v^{\pm}(t) \\ &+ \left(q_1^{\pm} - \alpha_0^{\pm}\lambda^{\pm}\right)v^{\pm}(t)^2 + \left(q_0^{\pm} - \alpha_1^{\pm}\lambda^{\mp}\right)v^{+}(t)v^{-}(t) + \left[q_2^{\pm} + \alpha_2^{\pm}(\lambda^{\pm} - 2\lambda^{\mp})\right]v^{\mp}(t)^2 \\ &+ \Big[k_1^{\pm} + q_0^{\pm}(\alpha_1^{\mp} - \alpha_0^{\pm}) - q_0^{\mp}\alpha_1^{\pm} + q_1^{\pm}\alpha_1^{\pm} + 2q_2^{\pm}\alpha_2^{\mp} - 2q_2^{\mp}\alpha_2^{\pm} + \alpha_1^{\pm}\alpha_1^{\mp}\lambda^{\pm} \\ &+ 2\alpha_2^{\pm}\alpha_2^{\mp}(2\lambda^{\pm} - \lambda^{\mp}) + \alpha_0^{\pm}\alpha_1^{\pm}(\lambda^{\pm} + 2\lambda^{\mp})\Big]v^{\pm}(t)^2v^{\mp}(t).\end{aligned} \tag{4.12}$$

Now we can fix the yet-unknown coefficients $\alpha_0^{\pm}$, $\alpha_1^{\pm}$, and $\alpha_2^{\pm}$ of the near identity transformation (4.6) in such a way that all quadratic terms vanish. This leads to the conditions

$$\alpha_0^{\pm} = \frac{q_0^{\pm}}{\lambda_u^{\pm}}, \quad \alpha_1^{\pm} = \frac{q_1^{\pm}}{\lambda_u^{\mp}}, \quad \alpha_2^{\pm} = \frac{q_2^{\pm}}{2\lambda_u^{\mp} - \lambda_u^{\pm}}. \tag{4.13}$$

Thus (4.12) reduces to the normal form of a Hopf bifurcation

$$\frac{d}{dt}v^{\pm}(t) = \lambda_u^{\pm}v^{\pm}(t) + b^{\pm}v^{\pm}(t)^2v^{\mp}(t), \tag{4.14}$$

where the Hopf parameter $b^{\pm}$ is given by

$$b^{\pm} = k_1^{\pm} + \frac{q_0^{\pm}q_1^{\pm}(4\lambda_u^{\pm 2} - \lambda_u^{\mp 2}) + q_1^{\pm}q_1^{\mp}(2\lambda_u^{\pm}\lambda_u^{\mp} - \lambda_u^{\mp 2}) + 2q_2^{\pm}q_2^{\mp}\lambda_u^{\pm}\lambda_u^{\mp}}{\lambda_u^{\pm}\lambda_u^{\mp}(2\lambda_u^{\pm} - \lambda_u^{\mp})}. \tag{4.15}$$

Taking into account (2.18), (2.19), (2.23), and (2.32) as well as (3.12), (3.13), (3.18), and (3.19) together with (4.2)–(4.6), this Hopf parameter $b^{\pm}$ reads in the vicinity of the instability as

$$b^{\pm} = -\frac{R_c}{5(1+R_c^2)^{\frac{3}{2}}}\left[(3R_c - 1) \pm i(R_c + 3)\right] + \mathcal{O}(\varepsilon). \tag{4.16}$$

Performing the ansatz

$$v^{\pm}(t) = r(t)e^{\pm i\varphi(t)}, \tag{4.17}$$

the normal form (4.14) is transformed to polar coordinates

$$\frac{d}{dt}r(t) = r(t)\left[\Re\left(\lambda_u^{\pm}\right) + \Re\left(b^{\pm}\right)r(t)^2\right], \tag{4.18}$$

$$\frac{d}{dt}\varphi(t) = \pm\left[\Im\left(\lambda_u^{\pm}\right) + \Im\left(b^{\pm}\right)r(t)^2\right]. \tag{4.19}$$

Thus, taking into account (2.19) and (4.16) near the instability, the oscillatory solution results in

$$r_{\text{stat}} = \sqrt{-\frac{\Re\left(\lambda_u^{\pm}\right)}{\Re\left(b^{\pm}\right)}} = \sqrt{\frac{5R_c}{3R_c - 1}}\sqrt[4]{1+R_c^2}\sqrt{\varepsilon} + \mathcal{O}(\varepsilon), \tag{4.20}$$

$$\frac{d}{dt}\varphi(t) = \pm\left[\Im\left(\lambda_u^{\pm}\right) + \Im\left(b^{\pm}\right)r_{\text{stat}}^2\right] = R_c - \frac{R_c}{3R_c - 1}\varepsilon + \mathcal{O}\left(\varepsilon^2\right). \tag{4.21}$$

In order to compare this result with numerical simulations, we have to convert this oscillatory solution back to the original state space Γ. At first we observe that we obtain, for $z_t \in \mathcal{C}$ from (2.22), (2.35), (3.6), and (3.9) near the instability,

$$z_t(\Theta) = \sum_{j=\pm} \phi^{\lambda_u^j}(\Theta) u^j(t) + \sum_{j_1=\pm} \sum_{j_2=\pm} H_{j_1 j_2}(\Theta) u^{j_1}(t) u^{j_2}(t). \tag{4.22}$$

Taking into account the near identity transformation (4.6) together with (2.23), this yields up to the first order in ε

$$\begin{aligned} z_t(\Theta) = {} & N_{\lambda_u^+} e^{\lambda_u^+ \Theta} v^+(t) + N_{\lambda_u^-} e^{\lambda_u^- \Theta} v^-(t) \\ & + a_0(\Theta) v^+(t)^2 + a_1(\Theta) v^+(t) v^-(t) + a_2(\Theta) v^-(t)^2, \end{aligned} \tag{4.23}$$

where the coefficients $a_0(\Theta)$, $a_1(\Theta)$, $a_2(\Theta)$ read as

$$a_0(\Theta) = \left[N_{\lambda_u^+} e^{\lambda_u^+ \Theta} \alpha_0^+ + N_{\lambda_u^-} e^{\lambda_u^- \Theta} \alpha_2^- + H_{++}(\Theta) \right], \tag{4.24}$$

$$a_1(\Theta) = \left[N_{\lambda_u^+} e^{\lambda_u^+ \Theta} \alpha_1^+ + N_{\lambda_u^-} e^{\lambda_u^- \Theta} \alpha_1^- + H_{+-}(\Theta) + H_{-+}(\Theta) \right], \tag{4.25}$$

$$a_2(\Theta) = \left[N_{\lambda_u^+} e^{\lambda_u^+ \Theta} \alpha_2^+ + N_{\lambda_u^-} e^{\lambda_u^- \Theta} \alpha_0^- + H_{--}(\Theta) \right]. \tag{4.26}$$

Due to the relation (2.2) between $z(t) \in \Gamma$ and $z_t \in \mathcal{C}$, we conclude from (4.23) that

$$\begin{aligned} z(t) = {} & N_{\lambda_u^+} v^+(t) + N_{\lambda_u^-} v^-(t) \\ & + a_0(0) v^+(t)^2 + a_1(0) v^+(t) v^-(t) + a_2(0) v^-(t)^2. \end{aligned} \tag{4.27}$$

Near the instability, we obtain from (2.32)

$$N_{\lambda_u^\pm} = \frac{1}{\sqrt[4]{1+R_c^2}} e^{\pm i \psi_1} + \mathcal{O}(\varepsilon) \tag{4.28}$$

with some phase ψ_1, whereas (4.17)–(4.19) lead to

$$v^\pm(t) = r_{\text{stat}} e^{\pm i \varphi(t)} \tag{4.29}$$

with the radius (4.20) and the phase

$$\varphi(t) = \Omega(\varepsilon) t + \varphi_0. \tag{4.30}$$

Here the frequency turns out to be

$$\Omega(\varepsilon) = R_c - \frac{R_c}{3R_c - 1} \varepsilon. \tag{4.31}$$

Furthermore, we yield from (4.24)–(4.26), by taking into account (3.12), (3.13), (3.18), (3.19), and (4.28) in the lowest order of ε,

$$a_0(0) = \frac{1}{\sqrt{5(1+R_c^2)}} e^{i\psi_2}, \quad a_1(0) = 0, \quad a_2(0) = \frac{1}{\sqrt{5(1+R_c^2)}} e^{-i\psi_2}, \tag{4.32}$$

where ψ_2 denotes some phase. Thus we obtain the following result for $z(t) \in \Gamma$ near the instability:

$$z(t) = c_0(\varepsilon) + c_1(\varepsilon)\cos\left[\varphi(t) + \psi_1\right] + c_2(\varepsilon)\cos\left[2\varphi(t) + \psi_2\right] + \mathcal{O}\left(\varepsilon^{\frac{3}{2}}\right), \tag{4.33}$$

where the respective coefficients read as

$$c_0(\varepsilon) = 0, \quad c_1(\varepsilon) = 2\sqrt{\frac{5R_c}{3R_c - 1}}\sqrt{\varepsilon}, \quad c_2(\varepsilon) = 2\,\frac{\sqrt{5}R_c}{3R_c - 1}\varepsilon. \tag{4.34}$$

Now we compare the oscillatory solution (1.7), (1.8), which was obtained by using the method of averaging, with ours, (4.30), (4.31), (4.33), (4.34), by taking into account the critical value (1.6) of the control parameter. We conclude that both results coincide in the lowest order $\varepsilon^{1/2}$, but our result is even correct up to the order ε.

From the near identity transformation (4.6) as well as from (4.20) and (4.29), we conclude that the order parameters $u^{\pm}(t)$ turn out to be of the order $\varepsilon^{1/2}$. This result is consistent with our original assumption (3.8), which was the basis of our approximate solution of the implicit equation for the center manifold (3.7) in the vicinity of the instability. Thus our synergetic system analysis is justified a posteriori by self-consistency.

Note that the same perturbative result (4.30), (4.31), (4.33), (4.34) for the oscillatory solution above the Hopf bifurcation can be derived with the multiple scale method [10]. It represents a technical procedure to deduce the normal form, once the bifurcation type is known, by using the knowledge of how the respective quantities depend on the smallness parameter $\varepsilon = (R - R_c)/R_c$. Although the multiple scale method was originally developed for ordinary differential equations [24, 25, 26], it can be also applied to delay differential equations (see, for instance, the treatment in [27]).

5. Numerical investigation. In order to numerically verify our analytical result, we integrated the underlying delay differential equation of Wright (1.5). By doing so, we varied the control parameter R in the vicinity of the instability $R_c = \pi/2$ in such a way that the smallness parameter $\varepsilon = (R - R_c)/R_c$ took 200 equidistant values between 10^{-5} and 10^{-1}. We used a Runge–Kutta–Verner method of the IMSL library as an integration routine with a step-size of 10^{-3} and performed a linear interpolation between the respective values in the memory interval. In particular, in the immediate vicinity of the instability, the phenomenon of critical slowing down led to a transient behavior. To exclude this, we iterated the discretized delay differential equation for each value of the control parameter at least 10^6 times. Afterward, we calculated the power spectrum with a complex fast Fourier transform (FFT) so that the basic frequency Ω of the oscillatory solution could be determined with high resolution. Then we performed a real FFT with the period $T = 2\pi/\Omega$ of the simulated periodic signal $z(t) = z(t + T)$:

$$z(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty}\left[a_k\cos\left(k\Omega t\right) + b_k\sin\left(k\Omega t\right)\right]. \tag{5.1}$$

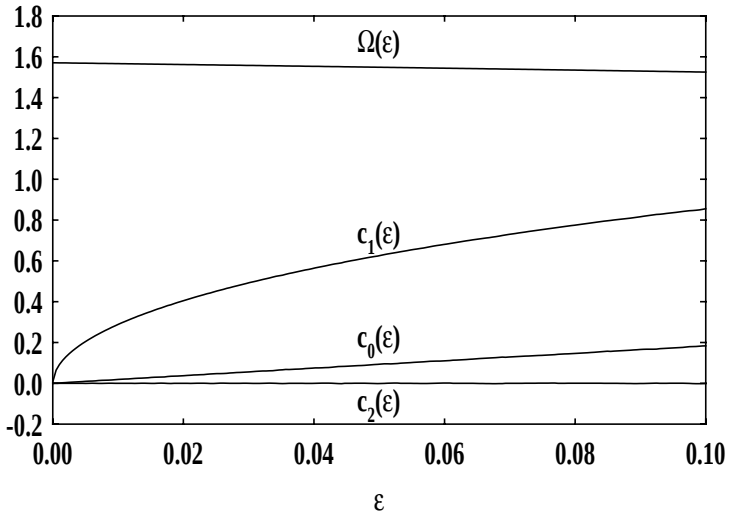


Figure 5.1. *Frequency Ω and Fourier coefficients c_0, c_1, c_2 of the oscillatory solution of the Wright equation after the Hopf bifurcation versus the smallness parameter $\varepsilon = (R - R_c)/R_c$.*

The Fourier coefficients follow from integrations with respect to one period $T = 2\pi/\Omega$:

$$a_k = \frac{2}{T}\int_0^T f(t)\cos(k\Omega t)\,dt, \quad k = 0, 1, \ldots, \infty; \tag{5.2}$$

$$b_k = \frac{2}{T}\int_0^T f(t)\sin(k\Omega t)\,dt, \quad k = 1, \ldots, \infty. \tag{5.3}$$

From (5.1) follows then the spectral representation

$$z(t) = c_0 + \sum_{k=1}^{\infty} c_k \cos(k\Omega t + \phi_k) \tag{5.4}$$

with the quantities

$$c_0 = \frac{a_0}{2}, \quad c_k = \sqrt{a_k^2 + b_k^2}, \quad \phi_k = -\arctan\frac{b_k}{a_k}, \quad k = 1, \ldots, \infty. \tag{5.5}$$

Thus our analytical result (4.30), (4.33) can be interpreted as the first terms within a spectral representation (5.4), where the frequency $\Omega = 2\pi/T$ and the Fourier coefficients c_0, c_1, c_2 are given by (4.31) and (4.34). Numerically analyzing the Hopf bifurcation with the FFT, the results for Ω, c_0, c_1, c_2 are plotted in Figure 5.1 versus the smallness parameter ε. Comparing the respective numerical and analytical results, we observe some deviations for small and for large values of the smallness parameter ε. The former are due to the phenomenon of critical slowing down (i.e., the system stays longer in the transient state when the instability is approached), and the latter arise from the neglected higher-order corrections in the analytical approach. Therefore, we restricted our numerical analysis to the intermediate interval

Table 5.1

Plotting the analytical and numerical values for the frequency $\Omega(\varepsilon)$ and the Fourier coefficients $c_0(\varepsilon)$, $\ln c_1(\varepsilon)$, $c_2(\varepsilon)$ of the oscillatory solution of the Wright equation after the Hopf bifurcation versus ε, we obtain straight lines whose axes intercept and whose slopes are determined.

Quantity	Analytical expression		Analytical value		Numerical value	
	Intercept	Slope	Intercept	Slope	Intercept	Slope
$\Omega(\varepsilon)$	R_c	$\dfrac{R_c}{3R_c-1}$	1.5708	−0.4231	1.5707	−0.4024
$c_0(\varepsilon)$	0	0	0.0	0.0	$-2\cdot10^{-4}$	$4\cdot10^{-2}$
$\ln c_1(\varepsilon)$	$\dfrac{1}{2}\ln\dfrac{20R_c}{3R_c-1}$	$\dfrac{1}{2}$	1.06781	0.5	1.06126	0.4999
$c_2(\varepsilon)$	0	$2\,\dfrac{\sqrt{5}R_c}{3R_c-1}$	0.0	1.8923	$2\cdot10^{-4}$	1.832

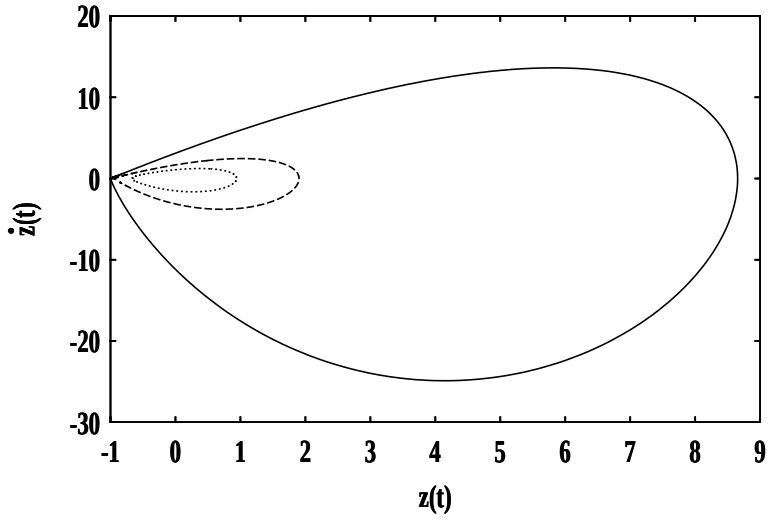


Figure 5.2. *Oscillatory solutions of the Wright equation* (1.5) *for three values of the control parameter R:* 1.7 *(dotted line),* 2.0 *(dashed line), and* 3.247 *(solid line).*

$[10^{-5}, 10^{-1}]$ of the smallness parameter ε. In Table 5.1, we see that the analytically and numerically determined quantities agree quantitatively very well. Thus our synergetic system analysis for the delay-induced Hopf bifurcation in the Wright equation is numerically verified.

For the sake of completeness, we have also investigated oscillatory solutions for values of the control parameter R which are larger than the critical one $R_c = \pi/2$. Figure 5.2 shows that all these periodic solutions oscillate around the stationary state $z^{\mathrm{I}}_{\mathrm{stat}} = 0$, which becomes unstable at $R_c = \pi/2$. It turns out that a global bifurcation occurs for $R^g_c = 3.247$ as then the

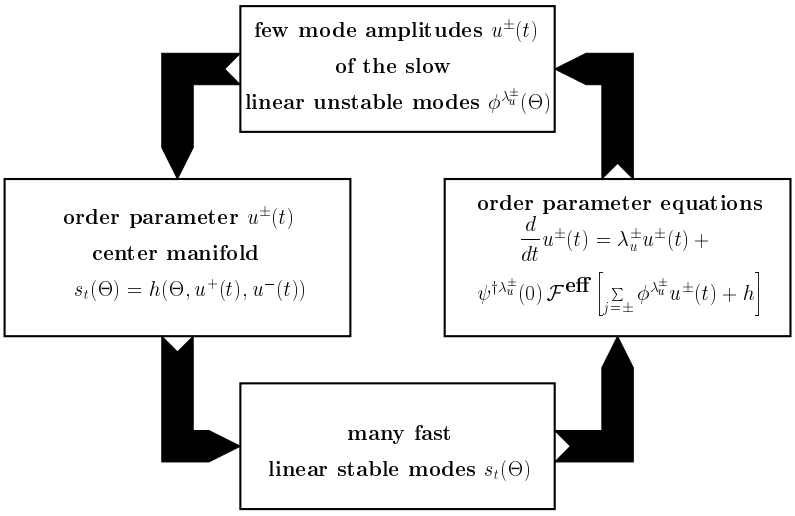


Figure 6.1. *Circular causality chain of synergetics for the Hopf bifurcation of a delay differential equation. On the one hand, the center manifold of the slaving principle guarantees that many fast linear stable modes $s_t(\Theta)$ quasi-instantaneously take values which are prescribed by the few slow linear unstable modes $u^{\pm}(t)$. On the other hand, the adiabatic elimination of the fast enslaved modes $s_t(\Theta)$ influences the resulting order parameter equation.*

oscillatory solution comes close to the other stationary state $z_{\text{stat}}^{\text{II}} = -1$, which turns out to be linear unstable for all values of the control parameter $R > 0$. Indeed, performing a linear stability analysis according to section 2 around the stationary state $z_{\text{stat}}^{\text{II}} = -1$ leads to the characteristic equation

$$R - \lambda = 0, \tag{5.6}$$

so we have from (2.16) the identification $p = R$ and $q = 0$. (Compare this with the shaded stability region in Figure 2.1.)

6. Summary and outlook. In this article, a linear stability analysis of the Wright equation (1.5) around the stationary state $z_{\text{stat}}^{\text{I}} = 0$ showed that a delay-induced Hopf bifurcation occurs at the critical value $R_c = \pi/2$ of the control parameter R. Within a subsequent nonlinear synergetic analysis, we adiabatically eliminated the stable modes and derived the normal form of this Hopf bifurcation. It is explicitly influenced by the center manifold in the lowest order, as the Wright equation (1.5) has a quadratic nonlinearity. Solving the normal form, we obtained a periodic solution above the Hopf bifurcation which was numerically verified.

In contrast to the corresponding analysis of the electronic system of a first-order PLL with time delay [11], this paper not only confirms the order parameter concept for delay systems but also represents a successful test for the slaving principle of synergetics, i.e., for the influence of the center manifold on the order parameter equations. Thus the validity of the circular causality chain of synergetics (see Figure 6.1) has been demonstrated for the Hopf bifurcation of a delay differential equation.

It remains to investigate the circular causality chain for other bifurcations. For instance,

it may be interesting to revisit a Hopf bifurcation of codimension two in delayed systems as it occurs in some robotics applications [28]. Furthermore, the Floquet theory for delay differential equations, and thus the linear stability analysis for a periodic reference state, was already established in [29, 30, 31]. However, a corresponding synergetic system analysis which derives the order parameter equations and the normal forms for bifurcations of oscillatory solutions is still missing [10, 11, 32].

Acknowledgments. We are thankful to Michael Bestehorn and Rudolf Friedrich for contributing various useful comments at an initial stage of this work. Furthermore, we thank Hermann Haken and Arne Wunderlin for teaching us synergetics for many years. Finally, Axel Pelster is grateful for the hospitality of Günter Wunner at the I. Institute of Theoretical Physics at the University of Stuttgart as this article was finished there.

REFERENCES

[1] H. Haken, *Synergetics: An Introduction*, 3rd ed., Springer-Verlag, Berlin, 1983.
[2] H. Haken, *Advanced Synergetics*, corrected second printing, Springer-Verlag, Berlin, 1987.
[3] H. Haken, *Information and Self-Organization*, Springer-Verlag, Berlin, 1988.
[4] H. Haken, *Synergetic Computers and Cognition*, Springer-Verlag, Berlin, 1991.
[5] H. Haken, *Principles of Brain Functioning*, Springer-Verlag, Berlin, 1996.
[6] I. Grabec and W. Sachse, *Synergetics of Measurements, Prediction, and Control*, Springer Ser. Synergetics 68, Springer-Verlag, Berlin, 1997.
[7] W. Wischert, A. Wunderlin, A. Pelster, M. Olivier, and J. Groslambert, *Delay-induced instabilities in nonlinear feedback systems*, Phys. Rev. E (3), 49 (1994), pp. 203–219.
[8] J. K. Hale, *Theory of Functional Differential Equations*, 2nd ed., Springer-Verlag, New York, 1977.
[9] N. Krasovskii, *Stability of Motion*, Stanford University Press, Stanford, CA, 1963.
[10] M. Schanz, *Zur Analytik und Numerik zeitlich verzögerter synergetischer Systeme*, Dissertation, Universität Stuttgart, Stuttgart, Germany, 1997.
[11] M. Schanz and A. Pelster, *Analytical and numerical investigations of the phase-locked loop with time delay*, Phys. Rev. E, 67 (2003), 056205.
[12] G. Stépán, *Great delay in a predator-prey model*, Nonlinear Anal., 10 (1986), pp. 913–929.
[13] G. Stépán and T. Kalmár-Nagy, *Nonlinear regenerative machine tool vibrations*, in Proceedings of the 16th ASME Biennial Conference on Mechanical Vibration and Noise, S. C. Sinha, ed., Sacramento, CA, 1997, pp. 1–11.
[14] T. Kalmár-Nagy, G. Stépán, and F. C. Moon, *Subcritical Hopf bifurcation in the delay equation model for machine tool vibrations*, Nonlinear Dynam., 26 (2001), pp. 121–142.
[15] P. F. Verhulst, *Notice sur la loi que la population suit dans son accroissement*, Corr. Math. Phys., 10 (1838), pp. 113–121.
[16] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1990.
[17] E. M. Wright, *A non-linear difference-differential equation*, J. Reine Angew. Math., 194 (1955), pp. 66–87.
[18] W. J. Cunningham, *A nonlinear differential-difference equation of growth*, Proc. Nat. Acad. Sci. U.S.A., 40 (1954), pp. 708–713.
[19] J. E. Marsden and M. McCracken, *The Hopf Bifurcation and Its Applications*, Appl. Math. Sci. 19, Springer-Verlag, New York, 1976.
[20] N. D. Hayes, *Roots of the transcendental equation associated with a certain difference-differential equation*, J. London Math. Soc., 25 (1950), pp. 226–232.
[21] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1992.

[22] R. H. Rand and D. Armbruster, *Perturbation Methods, Bifurcation Theory and Computer Algebra*, Appl. Math. Sci. 65, Springer-Verlag, New York, 1987.

[23] P. Tass, A. Wunderlin, and M. Schanz, *A theoretical model of sinusoidal forearm tracking with delayed visual feedback*, J. Biol. Phys., 21 (1995), pp. 83–112.

[24] J. Kevorkian, *The two-variable expansion procedure of the approximate solution of certain nonlinear differential equations*, in Space Mathematics (Proc. Summer Seminar, Ithaca, NY, 1963), Part 3, AMS, Providence, RI, 1966, pp. 206–275.

[25] W. Lick, *Two-variable expansions and singular perturbation problems,* SIAM J. Appl. Math., 17 (1969), pp. 815–825.

[26] A. Wunderlin and H. Haken, *Scaling theory for nonequilibrium systems*, Z. Phys. B, 21 (1975), pp. 393–401.

[27] E. Grigorieva, H. Haken, S. A. Kashchenko, and A. Pelster, *Travelling waves dynamics in nonlinear interferometer with spatial field transformer in feedback*, Phys. D, 125 (1999), pp. 123–141.

[28] G. Stépán and G. Haller, *Quasiperiodic oscillations in robot dynamics*, Nonlinear Dynam., 8 (1995), pp. 513–528.

[29] C. Simmendinger, O. Hess, and A. Wunderlin, *Analytical treatment of delayed feedback control*, Phys. Lett. A, 245 (1998), pp. 253–258.

[30] C. Simmendinger, A. Wunderlin, and A. Pelster, *Analytical approach for the Floquet theory of delay differential equations*, Phys. Rev. E (3), 59 (1999), pp. 5344–5353.

[31] T. Insperger and G. Stépán, *Stability chart for the delayed Mathieu equation*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 1989–1998.

[32] M. Schanz and A. Pelster, *On the period-doubling scenario in dynamical systems with time delay*, in Proceedings of the 15th IMACS World Congress on Scientific Computation, Modeling and Applied Mathematics, Berlin, 1997, Wissenschaft und Technik Verlag, 1 (1997), pp. 215–220.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 3, pp. 297–322

Learning about Reality from Observation*

William Ott† and James A. Yorke†

Abstract. Takens, Ruelle, Eckmann, Sano, and Sawada launched an investigation of images of attractors of dynamical systems. Let A be a compact invariant set for a map f on $\mathbb{R}^n$ and let $\phi : \mathbb{R}^n \to \mathbb{R}^m$ be a "typical" smooth map, where $n > m$. When can we say that A and $\phi(A)$ are similar, based only on knowledge of the images in $\mathbb{R}^m$ of trajectories in A? For example, under what conditions on $\phi(A)$ (and the induced dynamics thereon) are A and $\phi(A)$ homeomorphic? Are their Lyapunov exponents the same? Or, more precisely, which of their Lyapunov exponents are the same? This paper addresses these questions with respect to both the general class of smooth mappings ϕ and the subclass of delay coordinate mappings.

In answering these questions, a fundamental problem arises about an arbitrary compact set A in $\mathbb{R}^n$. For $x \in A$, what is the smallest integer d such that there is a C^1 manifold of dimension d that contains all points of A that lie in some neighborhood of x? We define a tangent space T_xA in a natural way and show that the answer is $d = \dim(T_xA)$. As a consequence, we obtain a Platonic version of the Whitney embedding theorem.

Key words. prevalence, attractor, embedding, enveloping manifold, Lyapunov exponent, delay coordinate map

AMS subject classifications. 37C70, 37H15, 37M25, 28C20, 60B11

DOI. 10.1137/S1111111102407421

1. Introduction.

In *The Republic*, Plato writes of people who are chained in a cave for all of their lives, unable to observe life directly. Behind these people a fire burns and real objects cast shadows on the cave wall for them to see. Forced to base their knowledge of reality on inferences made from the shadows, they equate the shadows with reality. While philosophers may vigorously debate epistemological theory, it is certainly true that experimentalists are limited to observations that may not encode the full complexity of their systems.

As Ruelle and Takens have observed, it is very difficult to directly observe all aspects of the evolution of a high dimensional dynamical system such as a turbulent flow. Out of necessity, it is frequently the case that experimentalists study such systems by measuring a relatively low number of different quantities. We assume that all measurements have infinite precision in what follows. A central experimental question is the following.

Question 1.1. Is the measured data sufficient for us to understand the evolution of the dynamical system? In particular, does the measured data contain enough information to reconstruct dynamical objects of interest and recover coordinate independent dynamical properties such as attractor dimension and Lyapunov exponents? How many exponents can be determined?

*Received by the editors May 13, 2002; accepted for publication (in revised form) by M. Golubitsky March 11, 2003; published electronically August 15, 2003. This research was partially supported by the National Science Foundation under grants DMS0104087 and DMS0072700.
http://www.siam.org/journals/siads/2-3/40742.html
†University of Maryland College Park, College Park, Maryland 20742 (ott@math.umd.edu, yorke@ipst.umd.edu).

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a map and suppose $A \subset \mathbb{R}^n$ is a compact invariant set. Let $\phi : \mathbb{R}^n \to \mathbb{R}^m$ be a smooth map. We always assume $m > 0$. We think of ϕ as a measurement function measuring m physical quantities, and for each point x in the state space $\mathbb{R}^n$ we say that $\phi(x)$ is the *measurement* associated with x. Motivated by an experimental point of view, we say that *observations are deterministic* if there exists an induced map $\bar{f}$ on $\phi(A)$ such that the following diagram commutes:

$$\begin{array}{ccc} A & \xrightarrow{f} & A \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\phi} \\ \phi(A) & \xrightarrow{\bar{f}} & \phi(A) \end{array}$$

The dynamics generated by $\bar{f}$ may be thought of as the shadows that traverse Plato's hypothetical cave wall. The global goal is to infer as much as possible about the dynamical system f from knowledge of the induced dynamics. In the absence of induced dynamics, experimenters increase m by either making more measurements or using delay coordinate maps. Assuming $\bar{f}$ exists, there is a considerable literature on how to compute the Lyapunov exponents associated with the induced system. Do these values correspond to those of the full system? What do we need to check to see this? We would like to state theorems of the following type.

Prototypical Theorem 1.2. *For a typical measurement map ϕ, if the induced map $\bar{f}$ exists and has certain properties, then the measurement map ϕ preserves dynamical objects of interest, and dynamical invariants of the full system may be computed from the induced dynamics.*

Under what conditions do our observations allow us to make predictions? James Clerk Maxwell wrote of the fundamental importance of continuous dependence on initial data [2, 9]:

> "It is a metaphysical doctrine that from the same antecedents follow the same consequents. No one can gainsay this. But it is not of much use in a world like this, in which the same antecedents never again concur, and nothing ever happens twice .... The physical axiom which has a somewhat similar aspect is 'That from like antecedents follow like consequents'."

We ask what we can conclude if observations are deterministic and if the induced map $\bar{f}$ is continuous. Using a translation invariant concept of "almost every" on infinite dimensional vector spaces described in section 2, we obtain the main C^0 conclusion.

Notation 1.3. For a map ψ we denote the restriction of ψ to a subset S of the domain of ψ by $\psi[S]$. Notice that this notation is not standard.

Let $\mathrm{Fix}(\bar{f})$ and $\mathrm{Per}_2(\bar{f})$ denote the collection of fixed points and period two points, respectively, of $\bar{f}$.

C^0 Theorem 1.4. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a map and let A be a compact invariant set. For almost every map $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, there is an induced map $\bar{f}$ satisfying*

1. *$\bar{f}$ is continuous and invertible, and*
2. *$\mathrm{Fix}(\bar{f})$ and $\mathrm{Per}_2(\bar{f})$ are countable*

if and only if the following hold.

1. *The measurement map ϕ is one-to-one on A.*

2. *The sets* $\mathrm{Fix}(f[A])$ *and* $\mathrm{Per}_2(f[A])$ *are countable.*
3. *The map* $f[A]$ *is continuous and invertible.*

Remark 1.5. If one can infer a property of A from a corresponding property of $\phi(A)$, we say that the property is *observable*. The boundedness of A is observable in the sense that if A is unbounded, then $\phi(A)$ is unbounded for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$. This applies to each of the embedding theorems in this paper.

Remark 1.6. Our goal is to obtain results with few or preferably no assumptions on f and A. Hypotheses should instead be placed on the observed objects, $\phi(A)$ and $\bar{f}$. This point of view motivates the definition of a Platonic result.

Definition 1.7. A result is said to be *Platonic* if it contains no hypotheses on the dynamical system f aside from the assumption of a finite dimensional Euclidean phase space.

Does a typical measurement function preserve differential structure? If f is a diffeomorphism, A is a smooth submanifold of $\mathbb{R}^n$, and $\dim(A)$ is known a priori, one may appeal to the Whitney embedding theorem [6]. This theorem states that if A is a compact C^r k-dimensional manifold, where $r \geq 1$, then there is a C^r embedding of A into $\mathbb{R}^m$, where $m \geq 2k+1$. This situation is generic in the sense that the set of embeddings of A is open and dense in $C^r(A, \mathbb{R}^m)$. However, the experimentalist lacking a priori knowledge of the structure of A cannot rely on embedding theorems of Whitney type.

In section 3 we define a notion of tangent space, denoted $T_x A$, suitable for a general compact subset A of $\mathbb{R}^n$, and we prove a manifold extension theorem. This result allows us to prove a Platonic version of the Whitney embedding theorem and to formulate a notion of diffeomorphism on A equivalent to the notion of injective immersion on A. We formulate our C^1 embedding theorems using this notion of diffeomorphism. Our Platonic C^1 Theorem states that for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, the existence of an invertible quasidifferentiable (see section 6) induced map $\bar{f}$ on $\phi(A)$ satisfying mild assumptions implies that ϕ is a diffeomorphism on A.

Platonic C^1 Theorem 1.8. *Suppose* $f : \mathbb{R}^n \to \mathbb{R}^n$ *is a map. For almost every* $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, *if there exists an invertible quasidifferentiable (see section 6) induced map* $\bar{f}$ *on* $\phi(A)$ *satisfying*

1. $\mathrm{Fix}(\bar{f})$ *and* $\mathrm{Per}_2(\bar{f})$ *are countable,*
2. $\dim T_y(\phi(A)) < m \ \forall y \in \phi(A)$, *and*
3. $D\bar{f}(y)[T_y\phi(A)]$ *is invertible* $\forall y \in \phi(A)$,

then the measurement mapping ϕ *is a diffeomorphism on* A.

It is difficult for a scientist to measure a large number of independent quantities simultaneously. For this reason one introduces the class of delay coordinate mappings. This mapping class was introduced into the literature by Takens [23].

Definition 1.9. Let $g \in C^1(\mathbb{R}^n, \mathbb{R})$. The *delay coordinate map* $\phi(f, g) : \mathbb{R}^n \to \mathbb{R}^m$ is given by

$$\phi(f, g)(x) = (g(x), g(f(x)), \ldots, g(f^{m-1}(x)))^{\mathrm{T}}. \tag{1.1}$$

Analogues of several of our embedding results hold for the class of delay coordinate mappings. Since the delay coordinate mappings form a subspace of $C^1(\mathbb{R}^n, \mathbb{R}^m)$, it should be stressed that the delay coordinate results do not follow from the corresponding results about almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$. The following result addresses the observation of differentiable dynamics.

Delay Coordinate Map Theorem 1.10. *Let f be a diffeomorphism on $\mathbb{R}^n$ and let A be a compact invariant set. For almost every $g \in C^1(\mathbb{R}^n, \mathbb{R})$, if there is a quasidifferentiable induced map $\bar{f}$ satisfying*

1. *$\bigcup_{i=1}^{2m} \mathrm{Per}_i(\bar{f})$ is countable and*
2. *for each $p \in \{1, \ldots, m\}$ and $y \in \mathrm{Per}_p(\bar{f})$ we have*

$$D\bar{f}^p(y)[T_y\phi(f,g)(A)] \neq \gamma \cdot I \text{ for every } \gamma \in \mathbb{R}, \tag{1.2}$$

then the delay map $\phi(f,g)$ is a diffeomorphism on A.

Assume that f and $\bar{f}$ are quasidifferentiable and invertible on A and $\phi(A)$, respectively, with invertible quasiderivatives at each point $x \in A$ and $y \in \phi(A)$. Suppose that ϕ is a diffeomorphism on A. We say that a Lyapunov exponent $\lambda(y,v)$ of $\bar{f}$ at $y \in \phi(A)$ is *true* if it does not depend on the choice of quasiderivative $D\bar{f}$ and if it is also a Lyapunov exponent of f at $\phi^{-1}(y) \in A$. The works of Eckmann, Ruelle, Sano, and Sawada provide heuristic computational procedures for obtaining m Lyapunov exponents for a trajectory (y_k) of $\bar{f}$. They use the subset of measurement mappings generated by so-called delay coordinate mappings, the mapping class considered in the famous fundamental paper of Takens [23]. In particular, the Eckmann and Ruelle algorithm (ERA) [3, 4] uses a linear fitting of the tangent map and has proven to be computationally efficient in giving the complete Lyapunov spectrum of many dynamical systems. Mera and Morán [14] find conditions ensuring the convergence of this algorithm for a smooth dynamical system on a $C^{1+\alpha}$ submanifold supporting an ergodic invariant Borel probability measure. Our exponent characterization theorem establishes a rigorous connection between the observed Lyapunov exponents and the Lyapunov exponents of $f[A]$. Under our assumptions, an observed Lyapunov exponent $\lambda(y,v)$ is a true Lyapunov exponent if and only if $v \in T_y\phi(A)$.

Suppose A is a manifold of dimension d. Implementation of the full ERA yields m observed Lyapunov exponents, d of which are true. The remaining $m-d$ exponents are spurious artifacts of the embedding process. In order to identify the d true exponents, one must either devise a method to identify the spurious exponents a fortiori or modify the ERA to completely avoid the computation of spurious exponents. Several authors propose a modified ERA in which the tangent maps are computed only on the tangent spaces and not on the ambient space $\mathbb{R}^m$. Mera and Morán [15] discuss the convergence of the modified ERA. This technique eliminates the computation of spurious exponents but requires that tangent spaces be computed along orbits. We propose a new technique based on the exponent characterization theorem that allows for the a fortiori determination of the spurious exponents without requiring the computation of tangent spaces along orbits. We describe this algorithm in section 7 following the statement of the exponent characterization theorem.

1.1. The case of linear f and ϕ. We illustrate our ansatz with the case where f and ϕ are linear.

Proposition 1.11. *Let f be linear on $\mathbb{R}^n$, and let A be an invariant subspace on which f is an isomorphism. If the restriction of f to A is not a scalar multiple of the identity, then for almost every $\phi \in \mathrm{Lin}(\mathbb{R}^n, \mathbb{R}^m)$ in the sense of Lebesgue measure, there is an induced map on $\phi(A)$ if and only if ϕ is an isomorphism on A.*

Key issues are raised by this proposition. Notice that if there exists $c \in \mathbb{R}$ for which $f(x) = cx$ for all $x \in A$, then $y \mapsto cy$ is the induced map on $\phi(A)$ even if ϕ is not one-to-one on A. Since this is a theory of observation, when possible the assumptions should be verifiable from observation. The following alternative version of the proposition transfers the assumption onto the induced dynamics in a manner that will be followed throughout this paper.

Proposition 1.12. *Let f be linear on $\mathbb{R}^n$, and let A be an invariant subspace on which f is an isomorphism. For almost every $\phi \in \mathrm{Lin}(\mathbb{R}^n, \mathbb{R}^m)$, there is an induced map on $\phi(A)$, and this induced map is not identically a scalar multiple of the identity if and only if ϕ is an isomorphism on A and the restriction of f to A is not a scalar multiple of the identity.*

Remark 1.13. The hypothesis that f is an isomorphism on A is observable in the sense mentioned earlier. The key point is that if $f[A]$ is not one-to-one, then for almost every $\phi \in \mathrm{Lin}(\mathbb{R}^n, \mathbb{R}^m)$ there does not exist an injective induced map $\bar{f}$ on $\phi(A)$.

1.2. What does "typical" mean? The conclusions of the linear propositions hold for almost every linear ϕ with respect to Lebesgue measure. In the general situation we will consider the space of C^1 measurement mappings. In order to prove versions of our Prototypical Theorem, we must first clarify what we mean by a "typical" measurement mapping ϕ. The notion of typicality may be cast in topological terms. In this setting, "typical" would be used to refer to an open and dense subset or a residual subset of mappings. For example, consider the topological Kupka–Smale theorem.

Definition 1.14. Let M be a smooth, compact manifold. A diffeomorphism $f \in \mathrm{Diff}^r(M)$ is said to be *Kupka–Smale* if the following hold:

1. The periodic points of f are hyperbolic.
2. If p and q are periodic points of f, then $W^s(p)$ is transverse to $W^u(q)$.

Theorem 1.15 (Kupka–Smale [17]). *The set of Kupka–Smale diffeomorphisms is residual in* $\mathrm{Diff}^r(M)$.

The topological notion of typicality is not the appropriate conceptualization for the experimentalist interested in a probabilistic result on the likelihood of a given property in a function space. Any Cantor set of positive measure illustrates the difference between the topological and measure theoretic notions of a small set. The discord between topological typicality and probabilistic typicality is also evident in the following dynamical examples.

Example 1.16. Arnold [1] studied the family of circle diffeomorphisms

$$f_{\omega,\epsilon}(x) = x + \omega + \epsilon \sin(x) \pmod{2\pi}, \tag{1.3}$$

where $0 \le \omega \le 2\pi$ and $0 \le \epsilon < 1$ are parameters. For each ϵ, we define the set

$$S_\epsilon = \{\omega \in [0, 2\pi] : f_{\omega,\epsilon} \text{ has a stable periodic orbit}\}. \tag{1.4}$$

For $0 < \epsilon < 1$, the set S_ϵ is a countable union of disjoint open intervals (one for each rational rotation number) and is an open dense subset of $[0, 2\pi]$. However, the Lebesgue measure of S_ϵ converges to 0 as $\epsilon \to 0$.

There are even more striking examples where the Baire categorical and measure theoretic notions of typicality yield diametrically opposite conclusions about the size of a set.

Example 1.17. Misiurewicz [16] proved that the mapping $z \mapsto e^z$ on the complex plane is topologically transitive, implying that a residual set of initial points yields dense trajectories. On the other hand, Lyubich [13] and Rees [18] proved that Lebesgue almost every initial point has a trajectory whose limit set is a subset of the real axis.

Finally, we consider Lyapunov exponents. This example is particularly relevant because the work of Eckmann, Ruelle, Sano, and Sawada on the computation of these exponents motivated this paper.

Example 1.18 (Lyapunov exponents). Let $f : M \to M$ be a C^1 diffeomorphism on a compact finite dimensional Riemannian manifold M. For $(x, v) \in TM$, $\|v\| \neq 0$, the number

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \|Df^n(x)v\|, \tag{1.5}$$

should the limit exist, is called the Lyapunov exponent of f at (x, v), denoted $\lambda(x, v)$. We say that $x \in M$ is a *regular point* for f if there are Lyapunov exponents

$$\lambda_1(x) > \cdots > \lambda_l(x) \tag{1.6}$$

and a splitting

$$T_xM = \bigoplus_{i=1}^{l} E_i(x) \tag{1.7}$$

of the tangent space to M at x such that

$$\lim_{n\to\pm\infty} \frac{1}{n} \log \|Df^n(x)u\| = \lambda_j(x) \quad (u \in E_j(x) \setminus \{0\} \text{ and } 1 \leq j \leq l).$$

While the periodic points of f are always regular points, frequently the set of regular points is a topologically small subset of M. Quite often this set is Baire first category, and it may even be finite [24]. From a measure theoretic point of view the situation is completely different.

Theorem 1.19 (Oseledec multiplicative ergodic theorem [24, 11]). *The set of regular points for f has full measure with respect to any f-invariant Borel probability measure on M.*

The Oseledec theorem holds in the more general context of measurable cocycles over invertible measure-preserving transformations of a Lebesgue space (X, μ) [11]. Let $f : X \to X$ be an invertible measure-preserving transformation, and let $L : X \to GL(n, \mathbb{R})$ be a measurable cocycle over X. If

$$\log^+ \|L^{\pm 1}(x)\| \in L^1(X, \mu), \tag{1.8}$$

then almost every $x \in X$ is a regular point for (f, L).

The following example illustrates that Lyapunov exponents may not exist for a residual set of points. Let $p > 1$ and $q > 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$ and $p \neq q$. Consider the Markov map $f : [0, 1] \to [0, 1]$ defined by

$$f(x) = \begin{cases} px & \text{if } 0 \leqslant x < \frac{1}{p}, \\ qx - \frac{q}{p} & \text{if } \frac{1}{p} \leqslant x \leqslant 1. \end{cases} \tag{1.9}$$

This transformation represents the full shift on two symbols with probabilities $1/p$ and $1/q$. Lebesgue measure is invariant under f and ergodic; thus the Lyapunov exponent at Lebesgue almost every $x \in [0,1]$ exists and is equal to

$$\frac{\log(p)}{p} + \frac{\log(q)}{q} \tag{1.10}$$

by virtue of the Birkhoff ergodic theorem. On the other hand, we claim that no Lyapunov exponent exists for a residual set of points. For $n \in \mathbb{N}$, set

$$V_{p,n}(x) = \frac{1}{n}(|\{0 \leqslant i \leqslant n-1 : f^i(x) \in [0,1/p)\}|). \tag{1.11}$$

Fix $\alpha > 1/p$ and $\beta < 1/p$. Define for each $N \in \mathbb{N}$ the sets $C_N = \{x : \exists\, n \geqslant N$ for which $V_{p,n}(x) \geqslant \alpha\}$ and $D_N = \{x : \exists\, n \geqslant N$ for which $V_{p,n}(x) \leqslant \beta\}$. The set C_N contains an open interval to the right of each preimage of $1/p$, and thus C_N contains an open and dense subset of $[0,1]$. Similarly, D_N contains an open interval to the left of each preimage of $1/p$, and thus D_N also contains an open and dense subset of $[0,1]$. No Lyapunov exponent exists for points in the residual set

$$\bigcap_{N=1}^{\infty} C_N \cap D_N \tag{1.12}$$

because $V_{p,n}(x)$ does not converge for such points.

Motivated by the probabilistic interpretation of typicality, we will use the notion of prevalence developed in [7, 8]. See the references given in [8] for closely related concepts. The notion of prevalence generalizes the translation invariant concept of Lebesgue full measure to infinite dimensional Banach spaces.

1.3. Overview of this paper. Section 2 develops the relevant prevalence theory and demonstrates that cardinality and boundedness are observable properties. In section 3 we define a notion of tangent space suitable for general compact subsets of $\mathbb{R}^n$, and we prove the manifold extension theorem. The manifold extension theorem is used in section 4 to derive a Platonic version of the Whitney embedding theorem. We present our embedding theorems in sections 5 and 6 and our results on delay coordinate mappings and Lyapunov exponents in section 7.

1.4. The transference method. Schematically our embedding theorems are developed in the following way. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a dynamical system, and let A be a compact invariant set. We want to require no regularity assumptions about f, nor do we wish to assume that f is invertible. For a map g, a subset D of the domain of g, and any property L, write $(g, L; D)$ to indicate that the restriction of g to D has property L. Let $\mathcal{S}$ denote a collection of properties of a dynamical system. Let $\mathcal{Q}$ denote a collection of properties of maps in the measurement function space $C^1(\mathbb{R}^n, \mathbb{R}^m)$. For example, $\mathcal{Q}$ might consist of the assertion that $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ is a homeomorphism on A. We are interested in the ability of the observer to make inferences, that is, in results of the form

$$(\bar{f}, \mathcal{L}; \phi(A)) \Rightarrow (\phi, \mathcal{Q}) \text{ for almost every } \phi, \tag{1.13}$$

where $\mathfrak{L}$ is a collection of properties of $\bar{f}$. In other words, the existence of an induced map $\bar{f}$ satisfying properties $\mathfrak{L}$ implies that ϕ satisfies properties $\mathfrak{Q}$. We first prove

$$(f, \mathfrak{S}; A) \Rightarrow ((\bar{f}, \mathfrak{L}_1; \phi(A)) \Leftrightarrow (\phi, \mathfrak{Q})) \text{ for a.e. } \phi. \tag{1.14}$$

The Platonic version of the theorem is obtained by replacing each assumption on f with one on $\bar{f}$. For $P \in \mathfrak{S}$, we replace the assumption

$$(f, P; A) \tag{1.15}$$

with one on $\bar{f}$, giving

$$(\bar{f}, \mathfrak{L}_1 \cup \mathfrak{S}; \phi(A)) \Leftrightarrow ((\phi, \mathfrak{Q}) \text{ and } (f, \mathfrak{S}; A)) \text{ for a.e. } \phi. \tag{1.16}$$

In particular, (1.13) holds with $\mathfrak{L} = \mathfrak{L}_1 \cup \mathfrak{S}$. In essence, the Platonic version has been obtained by transferring the hypotheses $(f, P; A)$ for $P \in \mathfrak{S}$ onto the induced dynamics. Prevalence statements allow for these transfers. Properties for which this program may be implemented are said to be *observable*.

2. Prevalence (measure-theoretic transversality).

Let V be a complete metric linear space.

Definition 2.1. A Borel measure μ on V is said to be *transverse* to a Borel set $S \subset V$ if the following hold:

1. There exists a compact set $U \subset V$ for which $0 < \mu(U) < \infty$, and
2. for every $v \in V$ we have $\mu(S + v) = 0$.

For example, μ might be Lebesgue measure supported on a finite dimensional subspace of V.

Definition 2.2. A Borel set $S \subset V$ is called *shy* if there exists a measure transverse to S. More generally, a subset of V is called shy if it is contained in a shy Borel set. The complement of a shy set is called a *prevalent* set.

A subset of $\mathbb{R}^n$ is shy if and only if it has Lebesgue measure zero. For a map ϕ contained in a prevalent subset S of a linear function space V, we say that ϕ is typical. Employing the language of the finite dimensional case, we say that *almost every* element of V lies in S (in the sense of prevalence).

Using the notion of prevalence, researchers have reformulated several topological and dynamical theorems. Sauer, Yorke, and Casdagli prove in [21] a prevalence version of the Whitney embedding theorem.

Theorem 2.3 (prevalence Whitney embedding theorem [21]). *Let A be a compact subset of $\mathbb{R}^n$ of box dimension d, and let m be an integer greater than $2d$. For almost every smooth map $\phi : \mathbb{R}^n \to \mathbb{R}^m$,*

1. *ϕ is one-to-one on A, and*
2. *ϕ is an immersion on each compact subset C of a smooth manifold contained in A.*

This theorem is not Platonic because the dimension assumption is on A. In section 4 we prove a Platonic Whitney embedding theorem as a corollary of the manifold extension theorem.

The reformulation of a genericity theorem of Kupka–Smale type requires a notion of prevalence for nonlinear function spaces such as the space of diffeomorphisms of a compact smooth manifold. Kaloshin in [10] develops such a notion and proves a prevalence version of the Kupka–Smale theorem for diffeomorphisms.

2.1. Cardinality preservation. In sections 5, 6, and 7, we will need to know how a typical smooth projection affects the cardinality of a set. We show that for a set $A \subset \mathbb{R}^n$, A and $\phi(A)$ have the same cardinality for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$. We begin by assuming that A is a countable set.

Proposition 2.4. *Let $A \subset \mathbb{R}^n$ be a countable set. Almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ is one-to-one on A. In particular, if A is countably infinite, then $\phi(A)$ is also countably infinite for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$.*

Proof. We write $A = \{x_i : i \in \mathbb{N}\}$. For $i \neq j$ let $C_{ij} = \{\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m) : \phi(x_i) = \phi(x_j)\}$. We first show that C_{ij} is shy. Let $B(x_i, r_i)$ be a metric ball such that $x_j \notin B(x_i, r_i)$. Let $\beta : \mathbb{R}^n \to \mathbb{R}$ be a C^∞ map such that

1. $\beta > 0$ on $B(x_i, r_i)$, and
2. $\operatorname{supp}(\beta) = \overline{B(x_i, r_i)}$.

Let $v \in \mathbb{R}^m$ be a nonzero vector, and let μ be the Lebesgue measure supported on the one dimensional subspace

$$\{tv\beta : t \in \mathbb{R}\}. \tag{2.1}$$

For any $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, it is evident that $\phi + tv\beta \in C_{ij}$ for at most one $t \in \mathbb{R}$. Thus C_{ij} is a shy subset of $C^1(\mathbb{R}^n, \mathbb{R}^m)$ because μ is transverse to it. The set

$$\bigcap_{\substack{i,j \in \mathbb{N} \\ i \neq j}} C^1(\mathbb{R}^n, \mathbb{R}^m) \setminus C_{ij} \tag{2.2}$$

consists of functions that map A injectively into $\mathbb{R}^m$. This set is prevalent because the countable intersection of prevalent sets is prevalent (see [7]). ■

Plato would have us consider the prisoner's question, where the cardinality of A is not known a priori. For a typical ϕ, does the countability of $\phi(A)$ imply the countability of A? The next proposition answers this question affirmatively with the help of the following lemma.

Lemma 2.5. *Let $A_0 \subset \mathbb{R}^n$ be an uncountable set. Lebesgue almost every function $\phi \in \operatorname{Lin}(\mathbb{R}^n, \mathbb{R}^m)$ maps A_0 to an uncountable set.*

Proof. It suffices to consider the scalar case $m = 1$. For each $\phi \in \operatorname{Lin}(\mathbb{R}^n, \mathbb{R})$ there exists a unique vector $v \in \mathbb{R}^n$ such that $\phi(x) = (x, v)$ for all $x \in \mathbb{R}^n$. Suppose by way of contradiction that the set

$$\{\phi \in \operatorname{Lin}(\mathbb{R}^n, \mathbb{R}) : \phi(A_0) \text{ is countable}\} \tag{2.3}$$

has positive measure. This implies that there exist n linearly independent vectors $\{v_i : i = 1, \dots, n\}$ such that the functions ϕ_{v_i} given by $x \mapsto (x, v_i)$ map A_0 to a countable set. Let A_1 be an uncountable subset of A_0 such that $\phi_{v_1}(A_1) = \{y_1\}$. Inductively construct a collection of sets $\{A_i : i = 1, \dots, n\}$ satisfying the following:

1. A_i is uncountable for each i,
2. $A_i \subset A_{i-1}$ for each i, and
3. $\phi_{v_i}(A_i) = \{y_i\}$.

We have $\phi_{v_i}(A_n) = \{y_i\}$ for each i, so A_n consists of one point. This contradiction establishes the lemma. ■

Proposition 2.6. *Let A_0 be an uncountable set. For almost every*

$$\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m), \tag{2.4}$$

$\phi(A_0)$ is uncountable.

Proof. Once again it suffices to consider the scalar case $m = 1$. We show that the set

$$S = \{\phi \in C^1(\mathbb{R}^n, \mathbb{R}) : \phi(A_0) \text{ is countable}\} \tag{2.5}$$

is shy. Let $\{\phi_{e_i}\}$ be a basis for $\mathrm{Lin}(\mathbb{R}^n, \mathbb{R})$, and let μ be the Lebesgue measure on $\mathbb{R}^n$. Write $\alpha = (\alpha_i)$ for a vector in $\mathbb{R}^n$ and for $\phi \in C^1(\mathbb{R}^n, \mathbb{R})$ set

$$\phi_\alpha := \phi + \sum_{i=1}^{n} \alpha_i \phi_{e_i}. \tag{2.6}$$

If S is not shy, there exists some $g \in S$ such that

$$\mu\{\alpha : g_\alpha(A_0) \text{ is countable}\} > 0, \tag{2.7}$$

where μ denotes n-dimensional Lebesgue measure. Without loss of generality, assume that $g(A_0)$ is countable. There is at least one point y such that $g^{-1}(y) \cap A_0$ is uncountable. Shrinking A_0 if necessary, without loss of generality we may assume that g maps A_0 to a single point; that is, g is constant on A_0. There exist n linearly independent vectors $\{v_i\}$ such that the functions $\phi_{v_i} + g$ map A_0 to a countable set. As in the proof of Lemma 2.5, we inductively construct a collection of sets $\{A_i : i = 1, \ldots, n\}$ satisfying the following:

1. A_i is uncountable for each i,
2. $A_i \subset A_{i-1}$ for each i, and
3. $(\phi_{v_i} + g)(A_i) = \{y_i\}$.

We have $(\phi_{v_i} + g)(A_n) = \{y_i\}$ for each i, so A_n consists of one point. This contradiction establishes the proposition. ∎

2.2. Preservation of unboundedness. We now consider the question of how a typical smooth projection affects the boundedness of a set. For a typical ϕ, does the boundedness of $\phi(A)$ imply that A is bounded?

Proposition 2.7 (unboundedness preservation). *Assume $A \subset \mathbb{R}^n$ is unbounded. Then $\phi(A)$ is unbounded for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$.*

Proof. It suffices to assume $m = 1$. We show that the set

$$V = \{\phi \in C^1(\mathbb{R}^n, \mathbb{R}) : \phi(A) \text{ is bounded}\} \tag{2.8}$$

is shy. As above, let μ be the Lebesgue measure on $\mathbb{R}^n$, and for $\phi \in C^1(\mathbb{R}^n, \mathbb{R})$ and $(\alpha_i) \in \mathbb{R}^n$ write

$$\phi_\alpha := \phi + \sum_{i=1}^{n} \alpha_i \phi_{e_i}. \tag{2.9}$$

If V is not shy, there exists some $g \in V$ such that

$$\mu\{\alpha : g_\alpha(A) \text{ is bounded}\} > 0. \tag{2.10}$$

Without loss of generality, assume that $g(A) \subset [-d, d]$ for some $d > 0$. There exist n linearly independent vectors $\{v_i\}$ and scalars $c_i > 0$ such that the functions $g + \phi_{v_i}$ map A into $[-c_i, c_i]$. Thus A is contained in the set

$$\bigcap_{i=1}^{n} \phi_{v_i}^{-1}([-c_i - d, c_i + d]), \tag{2.11}$$

a bounded solid polygon. This contradiction establishes the proposition. ■

Remark 2.8. We conclude that for a typical $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, the boundedness of $\phi(A)$ implies that A is bounded. That is, the boundedness of A is an observable property.

3. Enveloping manifolds. Let A be a compact subset of $\mathbb{R}^n$, and let $x \in A$. We say that a C^1 manifold M is an enveloping manifold for A at x if there exists a neighborhood $N(x)$ of x such that $M \supset N(x) \cap A$ and if the dimension of M is minimal with respect to this property. We demonstrate the existence of a C^1 enveloping manifold M for each $x \in A$.

Definition 3.1. Let $D_x A$ be the set of all directions v for which there exist sequences (y_i) and (z_i) in A such that $y_i \to x$, $z_i \to x$, and $\frac{z_i - y_i}{\|z_i - y_i\|} \to v$. The tangent space at x relative to A, denoted $T_x A$, is the smallest linear space containing $D_x A$.

We note that this is one of the two obvious ways to define the tangent space at a point in an arbitrary compact subset of $\mathbb{R}^n$. The other would be to fix $y_i = x$ in the above definition, but the resulting tangent space would be too small for our purposes. In general neither the tangent space itself nor its dimension will vary continuously with $x \in A$. Nevertheless, the tangent space varies upper semicontinuously with $x \in A$. More precisely, we have the following lemma.

Lemma 3.2. *The function $x \mapsto \dim(T_x A)$ is upper semicontinuous on A. In fact, $T_x A$ depends upper semicontinuously on $x \in A$ in the sense that if $x_i \to x$, where $x_i \in A$, and $v_i \to v$, where $v_i \in T_{x_i} A$, then $v \in T_x A$. In other words, $\{(x, v) : x \in A, v \in T_x A\}$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^n$. If $T_x A$ has constant dimension on a set $A_0 \subset A$, then $T_x A$ is continuous on A_0 in the same sense.*

Definition 3.3. The *tangent dimension* of A, denoted $\dim_T(A)$, is given by

$$\dim_T(A) = \max_{x \in A} (\dim T_x A). \tag{3.1}$$

Example 3.4. In Figure 3.1 the tangent space $T_p A$ is two dimensional, while $T_x A$ is one dimensional for all other points $x \in A$. Choosing $(y_i) \subset A$ and $(z_i) \subset A$ such that $y_i \to p$, $z_i \to p$, and y_i and z_i lie on a vertical line for each i, we obtain the tangent vector $v \in T_p A$. Thus $\dim_T(A) = 2$.

We are now in a position to state a surprising theorem.

Theorem 3.5 (manifold extension theorem). *For each $x \in A$ there exists an enveloping manifold M for A at x with $T_x M = T_x A$.*

Conjecture 3.6. We believe that integrability is an intrinsic feature of the definition of the tangent space. We therefore conjecture that a global version of the manifold extension theorem holds. Namely, there exists a manifold M such that $\dim(M) = \dim_T(A)$ and $A \subset M$.

Proof. Recall that for a map ψ we denote the restriction of ψ to a subset S of the domain of ψ by $\psi[S]$. Let $m = \dim(T_x A)$. There exists a compact neighborhood N of x such that $\dim(T_y A) \leq m$ for all $y \in N \cap A$. Let π denote the orthogonal projection of $\mathbb{R}^n$ onto $T_x A$.

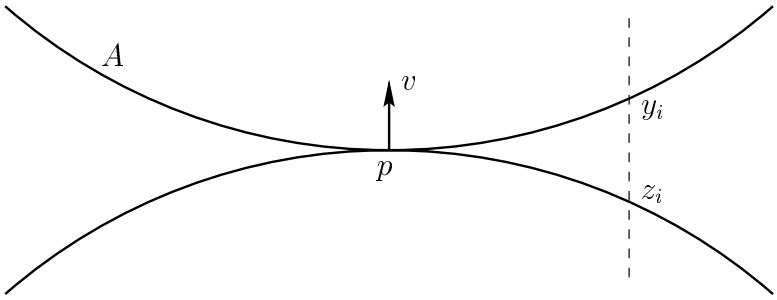


Figure 3.1. *A cusp.*

The projection map π induces the splitting $\mathbb{R}^n = T_xA \oplus E_x$. Using this splitting, write (p, q) for points in $\mathbb{R}^n$. If $((p_i, q_i))$ is a sequence such that $(p_i, q_i) \in N \cap A$ for each i and $(p_i, q_i) \to x$, then $\frac{\|q_{i+1}-q_i\|}{\|p_{i+1}-p_i\|} \to 0$. We may assume N has been chosen sufficiently small so that π maps T_yA injectively into T_xA for each $y \in N \cap A$ and that $\pi[N \cap A]$ is one-to-one. Hence we may define ψ on $\pi(N \cap A)$ by $\psi(p) := q$ for $(p, q) \in N \cap A$. Repeated use of our main technical tool, the Whitney extension theorem, will allow us to extend ψ to a C^1 function defined on a neighborhood in T_xA of $\pi(A \cap N)$. We first state a C^1 version of the Whitney extension theorem for compact domains.

Definition 3.7. Let $Q \subset \mathbb{R}^m$ be a compact set, and assume $f : Q \to \mathbb{R}^k$ and $L : Q \to \mathrm{Lin}(\mathbb{R}^m, \mathbb{R}^k)$ are given functions.

Notation 3.8.

1. $R(y, z) := \frac{f(z)-f(y)-L(y)\cdot(z-y)}{\|z-y\|}$ $(\forall y, z \in Q,\ y \neq z)$.
2. For $\delta > 0$, set

$$\rho(\delta) := \sup_{\substack{y,z\in Q \\ 0<\|z-y\|\le\delta}} \|R(y, z)\|. \tag{3.2}$$

The pair (f, L) is said to be a Whitney C^1 function pair on Q if f and L are continuous and if ρ satisfies

$$\rho(\delta) \to 0 \text{ as } \delta \to 0. \tag{3.3}$$

Notice that (3.3) is equivalent to the following uniformity condition stated by Whitney in [25]: Given any $w \in Q$ and $\epsilon > 0$, there exists $\delta > 0$ such that if $y \in Q$ and $z \in Q$ satisfy $\|y - w\| < \delta$ and $\|z - w\| < \delta$, then $\|R(y, z)\| \le \epsilon$.

Theorem 3.9 (Whitney extension theorem [5, 12, 22, 25]). *Given a Whitney C^1 function pair (f, L) defined on a compact subset Q of $\mathbb{R}^m$, there exists a C^1 function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^k$ such that $\tilde{f} = f$ and $D\tilde{f} = L$ on Q.*

We now continue the proof of our manifold extension theorem. Let

$$d(y) = \dim(T_yA) \tag{3.4}$$

for $y \in A \cap N$. For $k \le m$, let $X_k = \{y \in N \cap A : d(y) = k\}$. We first find a function whose graph is a C^1 manifold which envelops X_m. For each $y \in N \cap A$, the tangent space

T_yA may be viewed as a subspace of $T_xA \oplus E_x = \mathbb{R}^n$. For $y \in X_m$, define the linear operator $L_m(y) : T_xA \to E_x$ as follows. For $(v, w) \in D_yA$, let $L_m(y)v = w$. By linearity $L_m(y)$ is determined on T_yA. The linear operator $L_m(y)$ depends continuously on $y \in X_m$ since T_yA depends continuously on $y \in X_m$ by Lemma 3.2. The function pair (ψ, L_m) is Whitney C^1 on $\pi(X_m)$ because the uniformity condition of Whitney is implied by Definition 3.1. Notice that the Whitney extension theorem can now be used only to extend $\psi[\pi(X_m)]$ because no obvious candidate exists for $L(y)$ for $y \notin X_m$. By applying the Whitney extension theorem, extend ψ to a function $\tilde{\psi}_1$ defined on $\pi(N)$. Notice that if $X_m = N \cap A$, the result is proved since the graph of $\tilde{\psi}_1$ constitutes an enveloping manifold for A at x.

The general case is handled inductively. Construct $\tilde{\psi}_1$ as above, and make the nonlinear change of variable $(p, q) \to (p, q - \tilde{\psi}_1(p)) := (p, \psi_2(p))$. Consider the map $\psi_2[\pi(X_m) \cup \pi(X_{m-1})]$, and let $y \in \text{graph}(\psi_2[\pi(X_m) \cup \pi(X_{m-1})])$. The tangent space $T_y(\text{graph}(\psi_2[\pi(A)]))$ may be viewed as a subspace of $T_xA \oplus E_x = \mathbb{R}^n$. Define the linear map $L_{m-1}(y) : T_xA \to E_x$ as follows. If $y \in \text{graph}(\psi_2[\pi(X_m)])$, set $L_{m-1}(y) \equiv 0$. If $y \in \text{graph}(\psi_2[\pi(X_{m-1})])$, enlarge $T_y(\text{graph}(\psi_2[\pi(A)]))$ to a linear space $\tilde{T}_y$ of dimension m by adjoining one vector in T_xA orthogonal to $T_y(\text{graph}(\psi_2[\pi(A)]))$. For $(v, w) \in \tilde{T}_y$, let $L_{m-1}(y)v = w$. The linear operator $L_{m-1}(y)$ depends continuously on $y \in \text{graph}(\psi_2[\pi(X_m) \cup \pi(X_{m-1})])$ by Lemma 3.2. The function pair (ψ_2, L_{m-1}) is Whitney C^1 on $\pi(X_m) \cup \pi(X_{m-1})$ because the uniformity condition of Whitney is implied by Definition 3.1. By applying the Whitney extension theorem, extend $\psi_2[\pi(X_m) \cup \pi(X_{m-1})]$ to a function $\tilde{\psi}_2$ defined on $\pi(N)$. Make the nonlinear change of variables $(p, q) \to (p, q - \tilde{\psi}_2(p)) = (p, \psi_3(p))$.

Assume now that the functions $\tilde{\psi}_1, \tilde{\psi}_2, \ldots, \tilde{\psi}_{k-1}$ and ψ_k have been constructed. Consider the map

$$\psi_k \left[\bigcup_{i=m-k+1}^{m} \pi(X_i) \right]. \tag{3.5}$$

For each point y in the set

$$\text{graph} \left(\psi_k \left[\bigcup_{i=m-k+1}^{m} \pi(X_i) \right] \right), \tag{3.6}$$

the tangent space $T_y(\text{graph}(\psi_k[\pi(A)]))$ may be viewed as a subspace of $T_xA \oplus E_x = \mathbb{R}^n$. Define the linear map $L_{m-k+1}(y) : T_xA \to E_x$ as follows. If $y \in \text{graph}(\psi_k[\pi(X_m) \cup \cdots \cup \pi(X_{m-k+2})])$, set $L_{m-k+1}(y) \equiv 0$. If $y \in \text{graph}(\psi_k[\pi(X_{m-k+1})])$, enlarge $T_y(\text{graph}(\psi_k[\pi(A)]))$ to a linear space $\tilde{T}_y$ of dimension m by adjoining $k-1$ vectors in T_xA orthogonal to

$$T_y(\text{graph}(\psi_k[\pi(A)])). \tag{3.7}$$

For $(v, w) \in \tilde{T}_y$, let $L_{m-k+1}(y)v = w$. By Definition 3.1 and Lemma 3.2 the function pair

$$(\psi_k, L_{m-k+1}) \tag{3.8}$$

is Whitney C^1 on the set

$$\bigcup_{i=m-k+1}^{m} \pi(X_i). \tag{3.9}$$

By applying the Whitney extension theorem, extend the function

$$\psi_k\left[\bigcup_{i=m-k+1}^{m}\pi(X_i)\right] \tag{3.10}$$

to a function $\tilde{\psi}_k$ defined on $\pi(N)$. Make the change of variables $(p,q)\to(p,q-\tilde{\psi}_k(p)) := (p,\psi_{k+1}(p))$. After $m+1$ steps we obtain a map

$$\Psi := \sum_{i=1}^{m+1}\tilde{\psi}_i \tag{3.11}$$

defined on $\pi(N)$. The graph of Ψ constitutes an enveloping manifold M for A at x. ■

Remark 3.10. Although our inductive procedure is canonical, observe that the Whitney extension theorem makes no claim of uniqueness. Assume that (f,L_1) and (f,L_2) are Whitney C^1 function pairs defined on a compact subset Q of $\mathbb{R}^m$ as in Theorem 3.9. Let $y\in\text{graph}(f)$, and let π denote the orthogonal projection of $\mathbb{R}^m\times\mathbb{R}^k$ onto $\mathbb{R}^m$. The tangent space $T_y(\text{graph}(f))$ may be viewed as a subspace of $\mathbb{R}^m\times\mathbb{R}^k$. The linear operators $L_1(y)$ and $L_2(y)$ must satisfy $L_1(y)v=L_2(y)v=w$ for all $(v,w)\in T_y(\text{graph}(f))$. However, $L_1(y)$ and $L_2(y)$ are determined only for $(v,w)\in T_y(\text{graph}(f))$. If $v\notin\pi(T_y(\text{graph}(f)))$, then $L_1(y)$ and $L_2(y)$ may be such that $L_1(y)v\neq L_2(y)v$.

4. Platonic embedology.

Recall the prevalence version of the Whitney embedding theorem.

Theorem 4.1 (prevalence Whitney embedding theorem [21]). *Let A be a compact subset of $\mathbb{R}^n$ of box dimension d, and let m be an integer greater than $2d$. For almost every smooth map $\phi:\mathbb{R}^n\to\mathbb{R}^m$,*

1. *ϕ is one-to-one on A, and*
2. *ϕ is an immersion on each compact subset C of a smooth manifold contained in A.*

The manifold extension theorem implies a Platonic version of this result. We need a notion of diffeomorphism appropriate for a general compact subset A of $\mathbb{R}^n$.

Definition 4.2. We say that a measurement map $\phi\in C^1(\mathbb{R}^n,\mathbb{R}^m)$ is a *diffeomorphism* on A if ϕ is injective on A and if for each $x\in A$ there exists an enveloping manifold M for A at x that is mapped diffeomorphically onto an enveloping manifold for $\phi(A)$ at $\phi(x)$.

We are now in a position to formulate the Platonic Whitney embedding theorem.

Theorem 4.3 (Platonic Whitney embedding theorem). *Let $A\subset\mathbb{R}^n$ be compact. For almost every $\phi\in C^1(\mathbb{R}^n,\mathbb{R}^m)$, if $\phi(A)$ satisfies $\dim_T\phi(A)<\frac{m}{2}$, then ϕ is a diffeomorphism on A.*

Conjecture 4.4. The Platonic Whitney embedding theorem holds under the weaker assumption that $\dim_T\phi(A)<m$.

The proof of this result requires an understanding of the relationship between the box dimension of A and the dimension of the tangent spaces T_xA for $x\in A$. Working only with the definitions, the relationship is unclear. Illumination is provided by the manifold extension theorem.

Lemma 4.5. *Let $A\subset\mathbb{R}^n$ be compact. For each $x\in A$, there exists a neighborhood N of x such that $\dim(T_xA)\geqslant\dim_B(A\cap N)$.*

Proof. Fix $x \in A$. By the manifold extension theorem, there exists an enveloping manifold M for A at x and a neighborhood N of x such that $M \supset N \cap A$. The set $N \cap A$ is contained in a C^1 manifold of dimension $\dim(T_x A)$, and therefore $\dim(T_x A) \geqslant \dim_B(A \cap N)$. ■

We now commence with the proof of the Platonic Whitney embedding theorem. Suppose there exists $x \in A$ such that $\dim(T_x A) \geqslant \frac{m}{2}$. In this case, we would have that $\dim(T_{\phi(x)}\phi(A)) \geqslant \frac{m}{2}$ for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ as a consequence of the fact that almost every linear transformation has full rank. Therefore, we may assume that $\dim(T_x A) < \frac{m}{2}$ for all $x \in A$. By the manifold extension theorem and the compactness of A, A is contained in a finite union $\bigcup_{i=1}^k M_i$ of enveloping manifolds such that $\dim(M_i) < \frac{m}{2}$ for each i. Box dimension is finitely stable, so one has

$$\dim_B(A) \leqslant \dim_B\left(\bigcup_{i=1}^{k} M_i\right) = \max_i \dim_B(M_i) < \frac{m}{2}. \tag{4.1}$$

The prevalence version of the Whitney embedding theorem, Theorem 2.3, implies that almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ is a diffeomorphism on A.

Remark 4.6. Suppose one knows only that $\dim_B(\phi(A)) < \frac{m}{2}$ for a typical ϕ. It is difficult to draw any conclusions in this case. Sauer and Yorke [20] exhibit a compact subset A of $\mathbb{R}^{10}$ with $\dim_B(A) = 3.5$ such that $\dim_B(\phi(A)) < 3$ for every $\phi \in C^1(\mathbb{R}^{10}, \mathbb{R}^6)$.

5. Observing a continuous dynamical system. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a dynamical system, and let A be a compact invariant set. We make no a priori regularity assumptions about f. Let $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, and let $B \subset \mathbb{R}^n$ be an open metric ball. Recall that if there exists a map $\bar{f} : \phi(A) \to \phi(A)$ such that for $x \in A$ the diagram

$$\begin{array}{ccc} A & \xrightarrow{f} & A \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\phi} \\ \phi(A) & \xrightarrow{\bar{f}} & \phi(A) \end{array}$$

commutes, then we say that $\bar{f}$ is the *induced map* associated with f.

Remark 5.1. If f is continuous, then the existence of $\bar{f}$ implies the continuity of $\bar{f}$.

Definition 5.2. The pair $(x_1, x_2) \in A \times A$ is *coincident* if $\phi(x_1) = \phi(x_2)$. The pair $(x_1, x_2) \in A \times A$ is said to be *dynamically separated* by B if

1. (x_1, x_2) is coincident and
2. $x_1 \notin B$, $x_2 \notin B$, $f(x_1) \in B$, and $f(x_2) \notin B$.

Definition 5.3. Let S_B be the set of maps ϕ in $C^1(\mathbb{R}^n, \mathbb{R}^m)$ for which the following hold:

1. There exists some pair (x_1, x_2) dynamically separated by B, and
2. for all such pairs we have $\phi(f(x_1)) = \phi(f(x_2))$.

Lemma 5.4. *The set S_B is a shy subset of $C^1(\mathbb{R}^n, \mathbb{R}^m)$.*

Proof. We construct a measure transverse to S_B. Let $\beta : \mathbb{R}^n \to \mathbb{R}$ be a C^∞ map such that $\beta > 0$ on B and $\operatorname{supp}(\beta) = \bar{B}$. Let $v \in \mathbb{R}^m$ be a nonzero vector. Let μ be the Lebesgue measure supported on the one dimensional subspace

$$\{tv\beta : t \in \mathbb{R}\}. \tag{5.1}$$

For any $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, it is evident that $\phi + tv\beta \in S_B$ for at most one $t \in \mathbb{R}$. Thus S_B is shy because μ is transverse to it. ■

Definition 5.5. Let $\mathrm{Fix}(f)$ denote the set of fixed points of f. Let $\mathrm{Per}_2(f)$ denote the set of periodic points of f of period 2.

Proposition 5.6. *Suppose $f[A]$ is continuous and invertible. Assume that $\mathrm{Fix}(f[A])$ and $\mathrm{Per}_2(f[A])$ are countable sets. For almost every map $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, the following are equivalent:*

1. *The map ϕ is one-to-one on A.*
2. *The induced map $\bar{f}$ exists (and is therefore continuous).*

Proof. ((1) ⇒ (2)) Define $\bar{f} := \phi \circ f \circ \phi^{-1}$.

((2) ⇒ (1)) Let $\{B_i\}$ be a countable collection of open metric balls such that if $x, y \in A$ satisfy $x \neq y$, then there exists some B_i such that $x \in B_i$ and $y \notin B_i$. Consider the following three sets:

$$\begin{cases} G_1 = \{\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m) : \phi \text{ is one-to-one on } \mathrm{Fix}(f[A])\}, \\ G_2 = \{\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m) : \phi \text{ is one-to-one on } \mathrm{Per}_2(f[A])\}, \\ G_3 = \bigcap_{i=1}^{\infty} (S_{B_i})^C. \end{cases} \tag{5.2}$$

The set G_1 is a prevalent subset of $C^1(\mathbb{R}^n, \mathbb{R}^m)$ by Proposition 2.4 because the fixed points of $f[A]$ are countable. Similarly, G_2 is prevalent. The set G_3 is a prevalent subset of $C^1(\mathbb{R}^n, \mathbb{R}^m)$ because $(S_{B_i})^C$ is prevalent for each i by Lemma 5.4 and because the countable intersection of prevalent sets is prevalent (see [7]). Thus $G_1 \cap G_2 \cap G_3$ is a prevalent subset of $C^1(\mathbb{R}^n, \mathbb{R}^m)$. Let $\phi \in G_1 \cap G_2 \cap G_3$, and assume that ϕ is not one-to-one on A. It follows that no induced map $\bar{f}$ exists. Since $\phi \notin S_{B_i}$ for all i, there exist a metric ball B_i and a coincident pair (x_1, x_2) dynamically separated by B_i such that $\phi(f(x_1)) \neq \phi(f(x_2))$. ■

Proposition 2.6 allows us to improve this result by transferring the dynamical hypotheses onto the induced dynamics. We need a lemma indicating that the existence of a point of discontinuity of $f[A]$ precludes the existence of a continuous induced map for a typical measurement function.

Lemma 5.7. *Suppose $f[A]$ is discontinuous at some point $x \in A$. Then for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, no continuous induced map exists.*

Theorem 5.8. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a map. For almost every map $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, there is an induced map $\bar{f}$ satisfying*

1. *$\bar{f}$ is continuous and invertible, and*
2. *$\mathrm{Fix}(\bar{f})$ and $\mathrm{Per}_2(\bar{f})$ are countable*

if and only if the following hold.

1. *The measurement map ϕ is one-to-one on A.*
2. *The sets $\mathrm{Fix}(f[A])$ and $\mathrm{Per}_2(f[A])$ are countable.*
3. *The map $f[A]$ is continuous and invertible.*

Proof. We employ the transference method. If $f[A]$ is continuous and invertible and $\mathrm{Fix}(f[A])$ and $\mathrm{Per}_2(f[A])$ are countable sets, then Proposition 5.6 implies the result. If $\mathrm{Fix}(f[A])$ or $\mathrm{Per}_2(f[A])$ is uncountable, then Proposition 2.6 implies that the statement of the theorem holds for almost every ϕ. Lemma 5.7 implies the result if $f[A]$ is discontinuous

at some point. If $f[A]$ is not invertible, then for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ no invertible induced map exists. ■

We now consider the possibility of recovering differential information.

6. Observing differentiable dynamics. Assume that f is a diffeomorphism on $\mathbb{R}^n$. The concept of a measurement function ϕ being an immersion on A usually requires A to be a manifold, but there is now an obvious extension.

Definition 6.1. We say the map $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ is an *immersion* on A if $D\phi(x)[T_xA] : T_xA \to T_{\phi(x)}\phi(A)$ is one-to-one for each $x \in A$.

Motivated by the theory of infinite dimensional dynamical systems, we formulate our C^1 results using the notion of quasidifferentiability.

Definition 6.2. The function f is said to be quasidifferentiable on the set A if $f[A]$ is continuous and if for each $x \in A$ there exists a linear map $Df(x) : \mathbb{R}^n \to \mathbb{R}^n$, the quasiderivative of f at x, such that

$$\frac{f(x_i) - f(y_i) - Df(x)(x_i - y_i)}{\|x_i - y_i\|} \to 0 \tag{6.1}$$

for all sequences $(x_i) \subset A$ and $(y_i) \subset A$ such that $x_i \to x$ and $y_i \to x$.

Remark 6.3. The function f is Whitney C^1 if and only if f is quasidifferentiable and the quasiderivative varies continuously. Since continuity is observable, the C^1 embedding results to follow may be formulated with "Whitney C^1" in place of "quasidifferentiable."

We would like to prove under the assumptions of Proposition 5.6 that for almost every ϕ, the existence of a quasidifferentiable induced map $\bar{f}$ implies that ϕ is an injective immersion on A. However, one extra hypothesis on f is needed; namely, for each $x \in \mathrm{Fix}(f[A])$, we have

$$Df(x)[T_xA] \neq \gamma \cdot I \text{ for every } \gamma \in \mathbb{R}. \tag{6.2}$$

To see the need for this hypothesis, suppose that f is the identity map, A is countable, and there exists $x \in A$ such that $\dim(T_xA) = n > m$. In this case, the identity map on $\phi(A)$ is the induced map for every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, yet every ϕ fails to be immersive at x.

Consider a countable set $\{B_i = B(y_i, r_i)\}$ of open metric balls in $\mathbb{R}^n$ that separates points. Let $T(A) = \{(x, v) : x \in A,\ v \in T_xA\}$.

Definition 6.4. Let W_{B_i} be the set of measurement maps in $C^1(\mathbb{R}^n, \mathbb{R}^m)$ with the following properties:

1. There exists some point $(x, v) \in T(A)$ such that $v \neq 0$, $x \notin B(y_i, 2r_i)$, $f(x) \in B(y_i, r_i)$, $D\phi(x)v = 0$, and
2. for all such points we have $D\phi(f(x)) \circ Df(x)v = 0$.

Lemma 6.5. *The set W_{B_i} is shy.*

Proof. Let $F_1, \ldots, F_t$ be a basis for the nm-dimensional space of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let $\beta : \mathbb{R}^n \to \mathbb{R}$ be a C^∞ map with the following properties:

$$\begin{cases} (1) & \beta(x) = 1 \text{ for } x \in B(y_i, \frac{5}{4}r_i), \\ (2) & \mathrm{supp}(\beta) = \overline{B(y_i, \frac{3}{2}r_i)}, \\ (3) & 0 < \beta \leq 1 \text{ on } B(y_i, \frac{3}{2}r_i). \end{cases} \tag{6.3}$$

Let P be the subspace of $C^1(\mathbb{R}^n, \mathbb{R}^m)$ spanned by the collection $\{\beta F_i : i = 1, \dots, t\}$, and endow P with Lebesgue measure. For any ϕ, the set of vectors (α_i) for which

$$\phi + \beta \sum_{i=1}^{t} \alpha_i F_i \in W_{B_i} \tag{6.4}$$

is a subset of P of measure zero. ■

Lemma 6.6. *Let $x \in \mathrm{Fix}(f[A])$, and assume that $Df(x)[T_xA] \neq \gamma \cdot I$ for all $\gamma \in \mathbb{R}$. The set Z_x of measurement mappings satisfying*

1. $\ker(D\phi(x)) \cap T_xA \neq \{0\}$ *and*
2. $Df(x)(\ker(D\phi(x)) \cap T_xA) \subset \ker(D\phi(x))$

is a shy subset of $C^1(\mathbb{R}^n, \mathbb{R}^m)$.

Proof. Consider the orthogonal decomposition $\mathbb{R}^n = T_xA \oplus E_x$. Let L be the subset of $\mathrm{Lin}(\mathbb{R}^n, \mathbb{R}^m)$ consisting of maps that vanish on E_x and have norm at most one. Endow L with the normalized Lebesgue probability measure μ. For any $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, we claim that

$$\mu(\{F \in L : \phi + F \in Z_x\}) = 0. \tag{6.5}$$

If $\dim T_xA \leq m$, then (6.5) follows from the fact that almost every linear transformation has full rank. If $\dim T_xA > m$, then it suffices to consider the scalar case $m = 1$. Let $d = \dim(T_xA)$, and let $\{\phi_{e_i}\}$ be an orthonormal basis for $\mathrm{Lin}(T_xA, \mathbb{R})$, the unit ball of which we identify with L. Let ϕ_w represent $D\phi(x)[T_xA]$ with respect to the basis $\{\phi_{e_i}\}$. For a map $\phi_v \in \mathrm{Lin}(T_xA, \mathbb{R})$ such that $v + w \neq 0$, it is necessary that $v + w$ be an eigenvector of $Df(x)[T_xA]^T$ in order to have

$$Df(x)(\ker(\phi_{v+w}) \cap T_xA) \subset \ker(\phi_{v+w}). \tag{6.6}$$

If $Df(x)[T_xA]^T$ does not have an eigenvalue of multiplicity d, then (6.5) holds. Finally, notice that $Df(x)[T_xA]^T$ has an eigenvalue of multiplicity d if and only if $Df(x)[T_xA]$ is a scalar multiple of the identity. ■

Proposition 6.7. *Suppose f is a diffeomorphism on $\mathbb{R}^n$. Assume that*

$$\mathrm{Fix}(f[A]) \text{ and } \mathrm{Per}_2(f[A]) \tag{6.7}$$

are countable sets. Assume that for each $x \in \mathrm{Fix}(f[A])$ we have

$$Df(x)[T_xA] \neq \gamma \cdot I \text{ for every } \gamma \in \mathbb{R}. \tag{6.8}$$

Then for almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, if there is a quasidifferentiable induced map $\bar{f}$, then the measurement map ϕ is an injective immersion on A.

Proof. Consider the following sets:

$$\begin{cases} G_4 = \bigcap_{i=1}^{\infty} (W_{B_i})^C, \\ G_5 = \bigcap_{x \in \mathrm{Fix}(f[A])} (Z_x)^C. \end{cases} \tag{6.9}$$

The sets G_4 and G_5 are prevalent by Lemmas 6.5 and 6.6, respectively. For ϕ in the prevalent set

$$\bigcap_{j=1}^{5} G_j, \tag{6.10}$$

the existence of a quasidifferentiable induced map $\bar{f}$ implies that ϕ is an injective immersion on A. ■

Once again, Proposition 2.6 allows us to transfer some of the hypotheses of this theorem onto the induced dynamics.

Theorem 6.8. *Suppose f is a diffeomorphism on $\mathbb{R}^n$. For almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, if there is a quasidifferentiable induced map satisfying*

1. $\mathrm{Fix}(\bar{f})$ *and* $\mathrm{Per}_2(\bar{f})$ *are countable and*
2. *For each $y \in \mathrm{Fix}(\bar{f})$, $D\bar{f}(y)[T_y\phi(A)] \neq \gamma \cdot I$ for every $\gamma \in \mathbb{R}$,*

then the following hold.

1. *The measurement map ϕ is an injective immersion on A.*
2. $\mathrm{Fix}(f[A])$ *and* $\mathrm{Per}_2(f[A])$ *are countable.*
3. *For each $x \in \mathrm{Fix}(f[A])$, $Df(x)[T_xA] \neq \gamma \cdot I$ for every $\gamma \in \mathbb{R}$.*

Proof. It suffices to consider the cases in which the hypotheses of Proposition 6.7 fail to hold. If $\mathrm{Fix}(f[A]) \cup \mathrm{Per}_2(f[A])$ is uncountable, then for almost every ϕ there cannot exist an induced map satisfying "$\mathrm{Fix}(\bar{f})$ and $\mathrm{Per}_2(\bar{f})$ are countable" by Proposition 2.6. Suppose there exist $x \in \mathrm{Fix}(f[A])$ and $\gamma \in \mathbb{R}$ such that

$$Df(x)[T_xA] = \gamma \cdot I. \tag{6.11}$$

For almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, $D\phi(x)[T_xA]$ has full rank. If $\dim(T_xA) \geqslant m$, then the full rank of $D\phi(x)[T_xA]$ implies that $D\phi(x)$ maps T_xA onto $T_{\phi(x)}\phi(A)$, and therefore the existence of a quasidifferentiable induced map would imply

$$D\bar{f}(\phi(x))[T_{\phi(x)}\phi(A)] = \gamma \cdot I. \tag{6.12}$$

If $\dim(T_xA) < m$, then the full rank of $D\phi(x)[T_xA]$ implies that $D\phi(x)$ maps T_xA injectively into $T_{\phi(x)}\phi(A)$ and therefore surjectively onto $T_{\phi(x)}\phi(A)$. In this case, the existence of a quasidifferentiable induced map would imply

$$D\bar{f}(\phi(x))[T_{\phi(x)}\phi(A)] = \gamma \cdot I. \tag{6.13}$$ ■

Using the manifold extension theorem, we strengthen this theorem by utilizing the previously introduced notion of a diffeomorphism on A. We recall that definition here.

Definition 6.9. We say that a measurement map $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ is a *diffeomorphism* on A if ϕ is injective on A and if for each $x \in A$ there exists an enveloping manifold M for A at x that is mapped diffeomorphically onto an enveloping manifold for $\phi(A)$ at $\phi(x)$.

Theorem 6.10. *Suppose f is a diffeomorphism on $\mathbb{R}^n$. For almost every $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, if there is a quasidifferentiable induced map $\bar{f}$ satisfying*

1. $\mathrm{Fix}(\bar{f})$ *and* $\mathrm{Per}_2(\bar{f})$ *are countable and*

2. *For each* $y \in \mathrm{Fix}(\bar{f})$, $D\bar{f}(y)[T_y\phi(A)] \neq \gamma \cdot I$ *for every* $\gamma \in \mathbb{R}$,

then the following hold.

1. *The measurement map* ϕ *is a diffeomorphism on* A.
2. $\mathrm{Fix}(f[A])$ *and* $\mathrm{Per}_2(f[A])$ *are countable.*
3. *For each* $x \in \mathrm{Fix}(f[A])$, $Df(x)[T_xA] \neq \gamma \cdot I$ *for every* $\gamma \in \mathbb{R}$.

Remark 6.11. Mera and Morán [14] provide a test for determining whether or not observed trajectories of $\bar{f}$ are consistent with the assumption that $\bar{f}$ belongs to a certain regularity class.

The C^1 theorem, Theorem 6.10, is not Platonic because we assume that f is a diffeomorphism on $\mathbb{R}^n$. We formulate a Platonic version of the C^1 theorem by selecting new hypotheses on the induced map $\bar{f}$. The key modification is the replacement of the dynamical assumption on the nature of $D\bar{f}(y)[T_y\phi(A)]$ for $y \in \mathrm{Fix}(\bar{f})$ with the structural assumption that $\dim T_y(\phi(A)) < m$ for all $y \in \phi(A)$. The smoothness of f becomes observable in this new setting. After presenting several technical preliminaries, we state and prove the main result. We assume only that f is a map throughout this section.

Lemma 6.12. *If* $\dim T_x(A) \geqslant m$ *for some* $x \in A$, *then for almost every* $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ *one has* $\dim T_{\phi(x)}\phi(A) \geqslant m$.

Proof. The result follows from the fact that almost every linear transformation from one finite dimensional vector space to another has full rank. ■

Lemma 6.13. *Suppose there exist sequences* $(x_i) \subset A$, $(y_i) \subset A$, *and* $x \in A$ *such that* $x_i \to x$, $y_i \to x$, *and* $\frac{x_i - y_i}{\|x_i - y_i\|} \to v \in T_xA$, *but*

$$\left(\frac{f(x_i) - f(y_i)}{\|x_i - y_i\|}\right) \tag{6.14}$$

does not converge to a vector in $\mathbb{R}^n$. *For almost every* $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, *there does not exist a quasidifferentiable induced map* $\bar{f}$ *on* $\phi(A)$ *with* $\dim T_y\phi(A) < m$ *for all* $y \in \phi(A)$.

Proof. We need to consider two cases. Assume that the sequence

$$\left(\frac{f(x_i) - f(y_i)}{\|x_i - y_i\|}\right) \tag{6.15}$$

has two limit points v_1 and v_2. There cannot exist a quasidifferentiable induced map $\bar{f}$ on $\phi(A)$ if $v \notin \ker(D\phi(x)[T_xA])$ and $v_1 - v_2 \notin \ker(D\phi(f(x))[T_{f(x)}A])$. This condition is prevalent, and therefore the lemma holds in the first case. Now suppose that the sequence (6.15) tends to infinity. If either $\dim(T_xA) \geqslant m$ or $\dim(T_{f(x)}A) \geqslant m$, then Lemma 6.12 implies that for almost every ϕ one does not have $\dim T_y\phi(A) < m$ for all $y \in \phi(A)$. If both $\dim(T_xA) < m$ and $\dim(T_{f(x)}A) < m$, then for almost every ϕ it follows that $D\phi(x)[T_xA]$ and $D\phi(f(x))[T_{f(x)}A]$ are injective. For such a ϕ, the existence of a quasidifferentiable induced map $\bar{f}$ on $\phi(A)$ would imply

$$\frac{\bar{f} \circ \phi(x_i) - \bar{f} \circ \phi(y_i)}{\|\phi(x_i) - \phi(y_i)\|} = \frac{\phi \circ f(x_i) - \phi \circ f(y_i)}{\|\phi(x_i) - \phi(y_i)\|} \to \infty, \tag{6.16}$$

which is a contradiction. ■

Theorem 6.14 (Platonic C^1 Theorem). *Suppose* $f : \mathbb{R}^n \to \mathbb{R}^n$ *is a map. For almost every* $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, *if there exists an invertible quasidifferentiable induced map* $\bar{f}$ *on* $\phi(A)$ *satisfying*

1. $\mathrm{Fix}(\bar{f})$ *and* $\mathrm{Per}_2(\bar{f})$ *are countable,*
2. $\dim T_y(\phi(A)) < m\ \forall y \in \phi(A)$*, and*
3. $D\bar{f}(y)[T_y\phi(A)]$ *is invertible* $\forall y \in \phi(A)$*,*

then the following hold.

1. *The measurement mapping* ϕ *is a diffeomorphism on* A*.*
2. *The mapping* $f[A]$ *is invertible.*
3. *The sets* $\mathrm{Fix}(f[A])$ *and* $\mathrm{Per}_2(f[A])$ *are countable.*
4. *The dynamical system* f *is quasidifferentiable on* A *and* $Df(x)[T_xA]$ *is invertible for all* $x \in A$*.*
5. *For each* $x \in A$*,* $\dim(T_xA) < m$*.*

Proof. See sections 5 and 6 for the definitions of the sets G_1, G_2, G_3, and G_4. Let

$$G_6 = \{\phi \in C^1(\mathbb{R}^n, \mathbb{R}^m) : D\phi(x)[T_xA] \text{ is injective for each } x \in \mathrm{Fix}(f[A])\}. \tag{6.17}$$

If $\mathrm{Fix}(f[A])$ is countable and $\dim(T_xA) < m$ for each $x \in A$, then G_6 is prevalent. We employ the transference method to prove the Platonic C^1 Theorem.

If f satisfies conclusions (2), (3), (4), and (5), then for ϕ in the prevalent set

$$\left(\bigcap_{j=1}^{4} G_j\right) \bigcap G_6, \tag{6.18}$$

the existence of a quasidifferentiable induced map $\bar{f}$ on $\phi(A)$ implies that ϕ is an injective immersion on A. If $f[A]$ is not invertible, then for almost every ϕ, no invertible induced map exists. If $\mathrm{Fix}(f[A]) \cup \mathrm{Per}_2(f[A])$ is uncountable, then Proposition 2.6 implies that no induced map satisfying hypothesis 1 exists for almost every ϕ. If there exists $x \in A$ for which $\dim(T_xA) \geqslant m$, then Lemma 6.12 implies that $\dim T_{\phi(x)}\phi(A) \geqslant m$ for almost every ϕ, and for such ϕ hypothesis 2 is not satisfied.

Suppose f is not quasidifferentiable on A. If $f[A]$ is not continuous, then Lemma 5.7 implies that for almost every ϕ there does not exist a quasidifferentiable induced map $\bar{f}$ on $\phi(A)$. If f fails to be quasidifferentiable on A because the hypotheses of Lemma 6.13 are satisfied, then this lemma implies that for almost every ϕ there does not exist a quasidifferentiable induced map $\bar{f}$ on $\phi(A)$ with $\dim T_y\phi(A) < m$ for all $y \in \phi(A)$. The remaining possibility is that for some $x \in A$ there exists a nonlinear map taking T_xA into $T_{f(x)}A$. For almost every ϕ, this precludes the existence of a quasidifferentiable induced map $\bar{f}$. Finally, suppose f is quasidifferentiable on A but $Df(x)[T_xA]$ is not invertible for some $x \in A$. In this case, for almost every ϕ, there does not exist a quasidifferentiable induced map $\bar{f}$ on $\phi(A)$ satisfying hypothesis 3. ■

7. Delay coordinate embeddings and Lyapunov exponents.

We state delay coordinate embedding versions of our results and prove the exponent characterization theorem.

7.1. Delay coordinate maps.

The following theorems do not follow from the previously established corresponding theorems for the general class of smooth measurement mappings because the delay coordinate mappings form a subspace of $C^1(\mathbb{R}^n, \mathbb{R}^m)$. Nevertheless, their veracity is established using essentially the same reasoning.

Theorem 7.1. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a map. For almost every $g \in C^1(\mathbb{R}^n, \mathbb{R})$, there is an induced map $\bar{f}$ satisfying*

1. *$\bar{f}$ is continuous and invertible, and*
2. *$\bigcup_{i=1}^{2m} \mathrm{Per}_i(\bar{f})$ is countable*

if and only if the following hold.

1. *The delay coordinate map $\phi(f, g)$ is one-to-one on A.*
2. *The set $\bigcup_{i=1}^{2m} \mathrm{Per}_i(f[A])$ is countable.*
3. *The map $f[A]$ is continuous and invertible.*

Theorem 7.2. *Let f be a diffeomorphism on $\mathbb{R}^n$. For almost every $g \in C^1(\mathbb{R}^n, \mathbb{R})$, if there is a quasidifferentiable induced map $\bar{f}$ satisfying*

1. *$\bigcup_{i=1}^{2m} \mathrm{Per}_i(\bar{f})$ is countable and*
2. *for each $p \in \{1, \dots, m\}$ and $y \in \mathrm{Per}_p(\bar{f})$ we have*

$$D\bar{f}^p(y)[T_y\phi(f, g)(A)] \neq \gamma \cdot I \text{ for every } \gamma \in \mathbb{R}, \tag{7.1}$$

then the following hold.

1. *The delay coordinate map $\phi(f, g)$ is a diffeomorphism on A.*
2. *The set $\bigcup_{i=1}^{2m} \mathrm{Per}_i(f[A])$ is countable.*
3. *For each $p \in \{1, \dots, m\}$ and each $x \in \mathrm{Per}_i(f[A])$, we have*

$$Df^p(x)[T_xA] \neq \gamma \cdot I \text{ for every } \gamma \in \mathbb{R}. \tag{7.2}$$

7.2. Lyapunov exponents. We conclude section 7 with a discussion of Lyapunov exponents. Assume f and $\bar{f}$ are quasidifferentiable and invertible on A and $\phi(A)$, respectively, with invertible quasiderivatives at each point $x \in A$ and $y \in \phi(A)$. Suppose ϕ is a diffeomorphism on A. Assume $y \in \phi(A)$ is a regular point for $\bar{f}$, and recall that this implies the existence of a decomposition

$$\mathbb{R}^m = \bigoplus_{i=1}^{l} E_i(y) \tag{7.3}$$

such that

$$\lim_{k \to \pm\infty} \frac{1}{k} \log \|D\bar{f}^k(y)v\| = \lambda_j(y) \quad (v \in E_j(y) \setminus \{0\} \text{ and } 1 \leq j \leq l).$$

Since the set of regular points $R(\bar{f})$ is invariant in the sense that

1. $y \in R(\bar{f}) \Rightarrow \bar{f}^k(y) \in R(\bar{f})$ for all $k \in \mathbb{Z}$ and
2. $D\bar{f}^{\pm 1}(E_i(y)) = E_i(\bar{f}^{\pm 1}(y))$ for $i = 1, \dots, l$,

we associate the Lyapunov exponents $\lambda_1 > \cdots > \lambda_l$ with the trajectory (y_k). Counting multiplicities, there are m Lyapunov exponents associated with (y_k), and we label them $\chi_1, \dots, \chi_m$ such that

$$\chi_1 \geqslant \chi_2 \geqslant \cdots \geqslant \chi_m. \tag{7.4}$$

In light of Remark 3.10 following the manifold extension theorem, we state the following definitions.

Definition 7.3. We say that a Lyapunov exponent $\lambda(y, v)$ of $\bar{f}$ is a *tangent* Lyapunov exponent if $v \in T_y\phi(A)$. A Lyapunov exponent $\lambda(y, v)$ of $\bar{f}$ is said to be a *transverse* Lyapunov exponent if it is not a tangent exponent.

Definition 7.4. A Lyapunov exponent $\lambda(y, v)$ of $\bar{f}$ is said to be a *true* Lyapunov exponent if it does not depend on the choice of quasiderivative $D\bar{f}$ and if it is also a Lyapunov exponent of f at $\phi^{-1}(y)$. We say that a Lyapunov exponent $\lambda(y, v)$ of $\bar{f}$ is *spurious* if there exists a quasiderivative $D\bar{f}$ for which

$$\lim_{k\to\pm\infty} \frac{1}{k} \log \|D\bar{f}^k(y)v\| \tag{7.5}$$

either does not exist or is not a Lyapunov exponent of f at $\phi^{-1}(y)$.

Theorem 7.5 (exponent characterization theorem). *Assume f and $\bar{f}$ are quasidifferentiable and invertible on A and $\phi(A)$, respectively, with invertible quasiderivatives at each point $x \in A$ and $y \in \phi(A)$. Suppose ϕ is a diffeomorphism on A. Assume that $y \in \phi(A)$ is a regular point for $\bar{f}$ such that $\dim T_z\phi(A) = \dim T_y\phi(A)$ for all $z \in \overline{(y_k)}$. The following characterizations hold for a Lyapunov exponent $\lambda(y, v)$ of $\bar{f}$.*

1. *If the exponent $\lambda(y, v)$ is tangent, then it is a true exponent.*
2. *If the exponent $\lambda(y, v)$ is transverse, then it is a spurious exponent.*

The tangent exponents of $\bar{f}$ correspond to the tangent exponents of f.

Remark 7.6. The tangent space $T_y\phi(A)$ admits the decomposition

$$T_y\phi(A) = \bigoplus_{i=1}^{l} V_i(y), \tag{7.6}$$

where $V_i(y)$ is a subspace of $E_i(y)$ for $i = 1, \ldots, l$.

Remark 7.7. From a computational point of view, one is interested in constructing algorithms to efficiently and accurately compute the Lyapunov spectrum and identify the true exponents. The existing technique [3, 19, 15] requires that one modify the ERA by computing the tangent maps only on the tangent spaces and not on the ambient space $\mathbb{R}^m$. Assuming A is a smooth submanifold, Mera and Morán [15] state conditions under which this modified ERA converges. Clearly this technique eliminates the computation of spurious exponents. However, one has to compute the tangent spaces along the entire orbit. In light of the exponent characterization theorem, we propose a new algorithm that eliminates the need to compute these tangent spaces.

Definition 7.8. A *forward filtration* of $\mathbb{R}^m$ is a nested collection of subspaces

$$\emptyset = F_0(y) \subset F_1(y) \subset F_2(y) \subset \cdots \subset F_m(y) = \mathbb{R}^m \tag{7.7}$$

such that

$$\lim_{k\to+\infty} \frac{1}{k} \log \|D\bar{f}^k(y)v\| = \chi_{m-j+1} \tag{7.8}$$

for $v \in F_j(y) \setminus F_{j-1}(y)$.

Definition 7.9. A *backward filtration* of $\mathbb{R}^m$ is a nested collection of subspaces

$$\emptyset = B_0(y) \subset B_1(y) \subset B_2(y) \subset \cdots \subset B_m(y) = \mathbb{R}^m \tag{7.9}$$

such that

$$\lim_{k \to -\infty} \frac{1}{k} \log \|D\bar{f}^k(y)v\| = \chi_j \tag{7.10}$$

for $v \in B_j(y) \setminus B_{j-1}(y)$.

Suppose that forward and backward filtrations have been computed. Assume that one may determine computationally if a given $(m-1)$-dimensional subspace of $\mathbb{R}^m$ contains $T_y\phi(A)$. For $j = 1, \ldots, m$, compute the Lyapunov vector

$$v_j \in B_j \cap F_{m-j+1}. \tag{7.11}$$

We now fix j and determine if $v_j \in T_y\phi(A)$. If $\mathrm{Span}\{v_i : i \neq j\} \supset T_y\phi(A)$, then $v_j \notin T_y\phi(A)$. If $\mathrm{Span}\{v_i : i \neq j\} \not\supseteq T_y\phi(A)$, then $v_j \in T_y\phi(A)$, and χ_j is a true Lyapunov exponent. The true Lyapunov exponents and $T_y\phi(A)$ have been determined. It would be interesting to compare the performance of this algorithm to that of existing ERA techniques.

Proof. Statement 1 follows from the fact that ϕ is a diffeomorphism on A. We establish 2 with a perturbation argument. Let $\alpha > 1$, and let $d = \dim T_y\phi(A)$. For each $z \in \overline{(y_k)}$ there exists an enveloping manifold M_z for $\phi(A)$ at z with $T_zM_z = T_z\phi(A)$ and $\dim(M_z) = d$. Let

$$\{B(z, r_z) : z \in \overline{(y_k)}\} \tag{7.12}$$

be a collection of metric balls such that

$$B(z, r_z) \cap \phi(A) \subset \mathrm{Int}(M_z). \tag{7.13}$$

By compactness there exists a finite subcover

$$\left\{ B\left(z_i, \frac{r_{z_i}}{2}\right) : i = 1, \ldots, N \right\} \tag{7.14}$$

of $\overline{(y_k)}$. We inductively construct a sequence $\{D\bar{f}_k : k = 1, \ldots, N\}$ of perturbations of $D\bar{f}$. Let $\beta : \mathbb{R}^m \to \mathbb{R}$ be a C^∞ map such that

$$\begin{cases} (1) & 1 \leq \beta \leq \alpha, \\ (2) & \beta(z) = \alpha \text{ for } z \in B(z_1, \frac{r_{z_1}}{2}), \text{ and} \\ (3) & \beta(z) = 1 \text{ on } \mathbb{R}^m \setminus B(z_1, r_{z_1}). \end{cases} \tag{7.15}$$

For each $z \in B(z_1, r_{z_1}) \cap M_{z_1}$, $\mathbb{R}^m$ admits the orthogonal decomposition

$$\mathbb{R}^m = T_z(M_{z_1}) \oplus E_z. \tag{7.16}$$

Using this decomposition, we define $D\bar{f}_1$ as follows.

1. $D\bar{f}_1[\phi(A) \cap \mathbb{R}^m \setminus B(z_1, r_{z_1})] = D\bar{f}[\phi(A) \cap \mathbb{R}^m \setminus B(z_1, r_{z_1})]$.
2. For $z \in \phi(A) \cap B(z_1, r_{z_1})$, define $D\bar{f}_1(z)$ by

$$D\bar{f}_1(z)v = \begin{cases} D\bar{f}(z)v & \text{if } v \in T_z(M_{z_1}), \\ \beta(z)D\bar{f}(z)v & \text{if } v \in E_z. \end{cases} \tag{7.17}$$

In this fashion, we inductively construct the family of perturbations $\{D\bar{f}_k : k = 1, \ldots, N\}$. For $v \in (T_y\phi(A))^{\perp}$, we have

$$\underline{\lim}_{k\to\infty} \frac{1}{k} \log \|D\bar{f}_N^k(y)v\| \geq \lambda(y, v) + \log(\alpha). \tag{7.18}$$

Since $\alpha > 1$ was arbitrary, it follows that if $\lambda(y, v)$ is transverse, then it is spurious. ■

Acknowledgments. The authors would like to thank Dr. Michael Brin for his numerous insightful comments on a draft of this paper. We thank Jaroslav Stark for directing us to the Whitney extension theorem and David Broomhead for a discussion concerning the nature of tangent spaces associated with arbitrary sets.

REFERENCES

[1] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, 1983.
[2] L. Campbell and W. Garnett, *The Life of James Clerk Maxwell*, Macmillan, London, 1882; reprinted by Johnson Reprint, New York, 1969.
[3] J.-P. Eckmann, S. O. Kamphorst, D. Ruelle, and S. Ciliberto, *Lyapunov exponents from time series*, Phys. Rev. A (3), 34 (1986), pp. 4971–4979.
[4] J.-P. Eckmann and D. Ruelle, *Ergodic theory of chaos and strange attractors*, Rev. Modern Phys., 57 (1985), pp. 617–656.
[5] L. Evans and R. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[6] M. Hirsch, *Differential Topology*, Springer-Verlag, New York, 1976.
[7] B. Hunt, T. Sauer, and J. Yorke, *Prevalence: A translation-invariant "almost every" on infinite-dimensional spaces*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 217–238.
[8] B. Hunt, T. Sauer, and J. Yorke, *Prevalence: An addendum*, Bull. Amer. Math. Soc. (N.S.), 28 (1993), pp. 306–307.
[9] B. R. Hunt and J. A. Yorke, *Maxwell on chaos*, Nonlinear Sci. Today, 3 (1993), pp. 2–4.
[10] V. Kaloshin, *Some prevalent properties of smooth dynamical systems*, Proc. Steklov Inst. Math., 213 (1996), pp. 115–140.
[11] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, Cambridge, UK, 1995.
[12] S. Krantz and H. Parks, *The Geometry of Domains in Space*, Birkhäuser Boston, Boston, 1999.
[13] M. Lyubich, *On the typical behavior of the trajectories of the exponential*, Russian Math. Surveys, 41 (1986), pp. 207–208.
[14] M. Mera and M. Morán, *Convergence of the Eckmann and Ruelle algorithm for the estimation of Lyapunov exponents*, Ergodic Theory Dynam. Systems, 20 (2000), pp. 531–546.
[15] M. Mera and M. Morán, *Degrees of freedom of a time series*, J. Statist. Phys., 106 (2002), pp. 125–145.
[16] M. Misiurewicz, *On iterates of e^z*, Ergodic Theory Dynam. Systems, 1 (1981), pp. 103–106.
[17] J. Palis, Jr., and W. de Melo, *Geometric Theory of Dynamical Systems*, Springer-Verlag, New York, 1982.
[18] M. Rees, *The exponential map is not recurrent*, Math. Z., 191 (1986), pp. 593–598.

[19] M. Sano and Y. Sawada, *Measurement of the Lyapunov spectrum from a chaotic time series*, Phys. Rev. Lett., 55 (1985), pp. 1082–1085.
[20] T. Sauer and J. Yorke, *Are the dimensions of a set and its image equal under typical smooth functions?*, Ergodic Theory Dynam. Systems, 17 (1997), pp. 941–956.
[21] T. Sauer, J. Yorke, and M. Casdagli, *Embedology*, J. Statist. Phys., 65 (1991), pp. 579–616.
[22] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, 1983.
[23] F. Takens, *Detecting strange attractors in turbulence*, in Dynamical Systems and Turbulence, Lecture Notes in Math. 898, Springer-Verlag, Berlin, 1981, pp. 366–381.
[24] M. Viana, *Teorema Ergódico de Oseledets*, http://w3.impa.br/~viana/.
[25] H. Whitney, *Analytic extensions of differentiable functions defined in closed sets*, Trans. Amer. Math. Soc., 36 (1934), pp. 63–89.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 2, No. 3, pp. 323–380

Wave Structure and Nonlinear Balances in a Family of Evolutionary PDEs*

Darryl D. Holm† and Martin F. Staley‡

Abstract. We investigate the following family of evolutionary 1+1 PDEs that describes the balance between convection and stretching for small viscosity in the dynamics of one-dimensional nonlinear waves in fluids:

$$m_t + \underbrace{um_x}_{\text{convection}} + \underbrace{b\,u_x m}_{\text{stretching}} = \underbrace{\nu\, m_{xx}}_{\text{viscosity}} \quad \text{with} \quad u = g * m.$$

Here $u = g * m$ denotes $u(x) = \int_{-\infty}^{\infty} g(x-y)m(y)\,dy$. This convolution (or filtering) relates velocity u to momentum density m by integration against the kernel $g(x)$. We shall choose $g(x)$ to be an even function so that u and m have the same parity under spatial reflection. When $\nu = 0$, this equation is both reversible in time and parity invariant. We shall study the effects of the balance parameter b and the kernel $g(x)$ on the solitary wave structures and investigate their interactions analytically for $\nu = 0$ and numerically for small or zero viscosity.

This family of equations admits the classic Burgers "ramps and cliffs" solutions, which are stable for $-1 < b < 1$ with small viscosity.

For $b < -1$, the Burgers ramps and cliffs are unstable. The stable solution for $b < -1$ moves leftward instead of rightward and tends to a stationary profile. When $m = u - \alpha^2 u_{xx}$ and $\nu = 0$, this profile is given by $u(x) \simeq \operatorname{sech}^2(x/(2\alpha))$ for $b = -2$ and by $u(x) \simeq \operatorname{sech}(x/\alpha)$ for $b = -3$.

For $b > 1$, the Burgers ramps and cliffs are again unstable. The stable solitary traveling wave for $b > 1$ and $\nu = 0$ is the "pulson" $u(x,t) = cg(x-ct)$, which restricts to the "peakon" in the special case $g(x) = e^{-|x|/\alpha}$ when $m = u - \alpha^2 u_{xx}$. Nonlinear interactions among these pulsons or peakons are governed by the superposition of solutions for $b > 1$ and $\nu = 0$,

$$m(x,t) = \sum_{i=1}^{N} p_i(t)\,\delta(x - q_i(t)), \quad u(x,t) = \sum_{i=1}^{N} p_i(t)\,g(x - q_i(t)).$$

These pulson solutions obey a finite-dimensional dynamical system for the time-dependent speeds $p_i(t)$ and positions $q_i(t)$. We study the pulson and peakon interactions analytically, and we determine their fate numerically under adding viscosity.

Finally, as outlook, we propose an n-dimensional vector version of this evolutionary equation with convection and stretching, namely,

$$\frac{\partial}{\partial t}\mathbf{m} + \underbrace{\mathbf{u}\cdot\nabla\mathbf{m}}_{\text{convection}} + \underbrace{\nabla\mathbf{u}^T\cdot\mathbf{m} + (b-1)\,\mathbf{m}(\operatorname{div}\mathbf{u})}_{\text{stretching}} = 0$$

for a defining relation, $\mathbf{u} = G * \mathbf{m}$. These solutions show quasi-one-dimensional behavior for $n, k = 2, 1$ that we find numerically to be stable for $b = 2$. The corresponding superposed solutions of

*Received by the editors July 10, 2002; accepted for publication (in revised form) by M. Golubitsky January 30, 2003; published electronically August 23, 2003.

http://www.siam.org/journals/siads/2-3/41094.html

†Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, MS B284, Los Alamos, NM 87545. Mathematics Department, Imperial College of Science, Technology and Medicine, London SW7 2AZ, UK (dholm@lanl.gov).

‡Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, MS B284, Los Alamos, NM 87545 (mstaley@lanl.gov).

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 2, No. 3, pp. 323–380

the vector b-equation in n dimensions exist, with coordinates $\mathbf{x} \in R^n$, $s \in R^k$, $n-k>0$, and $2N$ parameters $\mathbf{P}_i(s,t), \mathbf{Q}_i(s,t) \in R^n$,

$$\mathbf{m}(\mathbf{x},t) = \sum_{i=1}^{N} \int \mathbf{P}_i(s,t)\, \delta\big(\mathbf{x} - \mathbf{Q}_i(s,t)\big) ds, \quad \mathbf{m} \in R^n,$$

$$\mathbf{u}(\mathbf{x},t) = \sum_{i=1}^{N} \int \mathbf{P}_i(s,t)\, G\big(\mathbf{x} - \mathbf{Q}_i(s,t)\big) ds, \quad \mathbf{u} \in R^n.$$

These are momentum surfaces (or filaments for $k=1$), defined on surfaces (or curves) $\mathbf{x} = \mathbf{Q}_i(s,t)$, $i = 1, 2, \ldots, N$. For $b=2$, the $\mathbf{P}_i(s,t), \mathbf{Q}_i(s,t) \in R^n$ satisfy canonical Hamiltonian equations for geodesic motion on the space of n-vector valued k-surfaces with cometric G.

Key words. symmetries and conservation laws, reverse symmetries, invariant manifolds and their bifurcations, reduction; Hamilton's principle; Hamiltonian structures, symmetries, variational principles, conservation laws; Hamiltonian systems on groups of diffeomorphisms and on manifolds of mappings and metrics

AMS subject classifications. 70H33, 70H25, 37K05, 37K65

DOI. 10.1137/S1111111102410943

1. Introduction.

1.1. The b-family of fluid transport equations.

We shall investigate a one-dimensional version of fluid convection and stretching that is described by the family of 1+1 evolutionary equations

$$m_t + \underbrace{u m_x}_{\text{convection}} + \underbrace{b\, u_x m}_{\text{stretching}} = 0 \quad \text{with} \quad u = g * m \tag{1.1}$$

in independent variables time t and one spatial coordinate x.

We shall seek solutions for the fluid velocity $u(x,t)$ that are defined either on the real line and vanishing at spatial infinity or on a periodic one-dimensional domain. Here $u = g * m$ denotes the convolution (or filtering)

$$u(x) = \int_{-\infty}^{\infty} g(x-y) m(y)\, dy, \tag{1.2}$$

which relates velocity u to momentum density m by integration against kernel $g(x)$ over the real line. We shall choose $g(x)$ in the defining relation (1.2) to be an even function so that u and m have the same parity.

The family of equations (1.1) is characterized by the kernel g and the real dimensionless constant b, which is the ratio of stretching to convection. The function $g(x)$ will determine the traveling wave shape and length scale for (1.1), while the constant b will provide a balance or bifurcation parameter for the nonlinear solution behavior. Special values of b will include the first few positive and negative integers.

The quadratic terms in (1.1) represent the competition, or balance, in fluid convection between nonlinear steepening and amplification due to b-dimensional stretching. Equation

(1.1) with $b \neq -1$ arises in the nonlinear dynamics of shallow water waves, as shown in [7, 17, 18]. Equation (1.1) with $b = 2$ and $b = 3$ appears in the theory of integrable PDEs [7, 17, 15]. The three-dimensional analogue of (1.1) with $b = 2$ was introduced in a larger variational context in [26, 27]. Applying the proper viscosity to this three-dimensional analogue with $b = 2$ and enforcing incompressibility produces the Navier–Stokes-alpha model of turbulence [9]. The one-dimensional version of this turbulence model is

$$m_t + \underbrace{u m_x}_{\text{convection}} + \underbrace{b\, u_x m}_{\text{stretching}} = \underbrace{\nu\, m_{xx}}_{\text{viscosity}} \quad \text{with} \quad u = g * m. \tag{1.3}$$

We shall compare our analysis of (1.1) with numerical simulations of (1.3) for small viscosity.

1.2. Outline of the paper. After summarizing previous investigations of particular cases in the b-family of convection equations (1.1), section 2 discusses its symmetries and other general properties such as parity and reversibility. Section 3 discusses the derivation of the b-equation (1.1) among a family of asymptotically equivalent equations for unidirectional shallow water waves. Section 4 discusses the traveling waves for (1.1) and derives their pulson solutions, which may be generalized functions for $b > 1$. Section 5 analyzes the interaction dynamics of the pulson solutions for any positive b and any g. Section 6 specializes the analysis of the pulson solutions to the peakons, for which $g(x) = e^{-|x|/\alpha}$ is a peaked pulse of width α, and b is taken to be arbitrary. In section 7, we add viscosity to the peakon equation and describe our numerical methods for illustrating the different types of behavior that may arise in the initial value problems for peakon solutions with $b > 0$, $b = 0$, and $b < -1$. Section 8 uses these numerical methods to determine how viscosity affects the fate of the peakons. Section 9 provides a synopsis of the figures. Section 10 summarizes the paper's main conclusions, and section 11 provides the outlook for generalizing the present work to higher dimensions.

2. History and general properties of the b-equation. Camassa and Holm [7] derived the following equation for unidirectional motion of shallow water waves in a particular Galilean frame:

$$m_t + \underbrace{u m_x}_{\text{convection}} + \underbrace{2\, u_x m}_{\text{stretching}} = \underbrace{-\, c_0 u_x - \gamma\, u_{xxx}}_{\text{dispersion}} \quad \text{with} \quad m = u - \alpha^2 u_{xx}. \tag{2.1}$$

Here $m = u - \alpha^2 u_{xx}$ is a momentum variable, partial derivatives are denoted by subscripts, the constants α^2 and γ/c_0 are squares of length scales, and $c_0 = \sqrt{g'h}$ is the linear wave speed for undisturbed water of depth h at rest under gravity g' at spatial infinity, where u and m are taken to vanish. Any constant value $u = u_0$ is also a solution of (2.1).

Equation (2.1) was derived using Hamiltonian methods in [7] and was shown in [17] also to appear as a water wave equation at quadratic order in the asymptotic expansion for unidirectional shallow water waves in terms of their two small parameters (aspect ratio and wave height). The famous Korteweg–de Vries equation (KdV) appears at linear order in this asymptotic expansion and is recovered from (2.1) when $\alpha^2 \to 0$. Both KdV at linear order and its nonlocal nonlinear generalization in (2.1) at quadratic order in this expansion have the remarkable property of being completely integrable by the isospectral transform (IST) method. The IST properties of KdV solitons are well known. The IST properties of (2.1)

were discovered in [7] and were analyzed completely for the case that the initial distribution of momentum $m(x,0)$ does not change sign in [6]. See also [10] for additional analysis of the scattering problem for the Camassa–Holm (CH) equation for this class of initial conditions. For an initial distribution of momentum that changes sign, the IST properties of the CH equation (2.1) remains an open problem.

3. Deriving b-equation for shallow water.

3.1. Linear and nonlinear balances in shallow water waves. The primary physical mechanism for the unidirectional propagation of solitary shallow water waves is the balance between nonlinear steepening and linear dispersion. This balance is *nearly* unique at linear order in an asymptotic expansion in the KdV

$$\underbrace{u_t + c_0 u_x}_{\text{Propagation}} + \underbrace{\frac{3\epsilon_1}{2} u u_x + \frac{3\epsilon_2}{20} u_{xxx}}_{\text{Balance}} = 0. \tag{3.1}$$

Here the expansion parameters satisfy $\epsilon_1 \geq \epsilon_2 > \epsilon_1^2$ and are defined by $\epsilon_1 = a/h$ and $\epsilon_2 = h^2/l^2$ in terms of wave amplitude a, mean water depth h, and typical horizontal length scale l (e.g., a wavelength). KdV possesses the famous sech^2 solitary wave solution $u(x,t) = u_0 \,\text{sech}^2((x-ct)\sqrt{u_0/\gamma}/2\,)$ for $u_0 = 2(c-c_0)/\epsilon_1$ and $\gamma = 3\epsilon_2/(5\epsilon_1)$; see [1].

The Benjamin–Bona–Mahoney equation (BBM)

$$\underbrace{u_t + c_0 u_x}_{\text{Propagation}} + \underbrace{\frac{3\epsilon_1}{2} u u_x - \frac{3\epsilon_2}{20} c_0^{-1} u_{xxt}}_{\text{Balance}} = 0 \tag{3.2}$$

has a solitary wave with the same sech^2 shape, but with γ replaced by $\gamma' = c\gamma/c_0$; see [1]. BBM is asymptotically equivalent to KdV at order $O(\epsilon_1, \epsilon_2)$. In comparison, the linear dispersion relation for BBM matches the exact relation for shallow water waves better than KdV. However, KdV is Galilean invariant, while BBM is not. And, of course, KdV is a completely integrable soliton equation, while BBM is not.

Beyond KdV and BBM at linear order, the asymptotic expansion at quadratic order in the small parameters ϵ_1 and ϵ_2 produces an infinite *family* of shallow water wave equations that are asymptotically equivalent to each other at quadratic order in the shallow water expansion parameters [18]. The equations in this family are related to each other by a continuous three-parameter group of nonlinear nonlocal transformations of variables introduced in [33, 34, 35],

$$u = v + \epsilon_1(a_1 v^2 + a_2 v_x \partial_x^{-1} v) + \epsilon_2 a_3 v_{xx}, \tag{3.3}$$

in which (a_1, a_2, a_3) are the three real parameters of the group. This transformation group was first introduced for determining normal forms of asymptotic shallow water wave equations by Kodama in [33, 34, 35].

Among the family of asymptotically equivalent shallow water wave equations at quadratic order accuracy in the small parameters $\epsilon_1 = a/h$ and $\epsilon_2 = h^2/l^2$ are several equations that are completely integrable. As for KdV at linear order, these integrable shallow water equations at quadratic order possess soliton solutions that interact via elastic collisions. In particular,

the equation in the KdV hierarchy with fifth-order derivatives (KdV5) appears amongst these integrable equations, as shown in [37].

The family of asymptotically equivalent shallow water equations that emerges at quadratic order accuracy also contains the following subfamily derived in [18] in which the constant parameter b depends on the group parameters (a_1, a_2, a_3) appearing in the Kodama transformation (3.3),

$$m_t + c_0 u_x + \epsilon_1 (u m_x + b\, m u_x) + \frac{3\epsilon_2}{20} u_{xxx} = 0, \tag{3.4}$$

where $m = u - (19\epsilon_2/60)u_{xx}$. For any $b \neq -1$, an asymptotically equivalent shallow water equation may be achieved by a Kodama transformation. However, the case $b = -1$ violates the asymptotic ordering, and the corresponding Kodama transformation is singular for $b = -1$ [18]. Notice that (3.4) is *not* Galilean invariant.

The cases $b = 2$ and $b = 3$ are special values for the b-equation (3.4). The case $b = 2$ restricts (3.4) to the integrable CH [7]. The case $b = 3$ in (3.4) recovers the Degasperis–Procesi equation (DP) [16], which was shown to be integrable in [15]. These two cases exhaust the integrable candidates for (3.4), as was shown using Painlevé analysis in [15]. The b-family of equations (3.4) was also shown in [38] to admit the symmetry conditions necessary for integrability only in the cases $b = 2$ for CH and $b = 3$ for DP.

3.2. Other work related to (3.4) with $b = 2$. Equation (3.4) with $b = 2$ was first derived by using asymptotic expansions directly in the Hamiltonian for Euler's equations in the shallow water regime and was thereby shown to be bi-Hamiltonian and IST-integrable in [7]. Additional details of its derivation and the analysis of its peakon solutions, as well as a numerical solution of the initial value problem for (3.4) with $b = 2$, were provided in [8]. Its periodic solutions were treated in [2, 3, 4, 5] and references therein.

Equation (3.4) with $b = 2$ was recently rederived as a shallow water equation using asymptotic methods via three different approaches in [21, 17, 32]. These three derivations used different variants of the method of asymptotic expansions for shallow water waves. A recent paper [14] also obtains (3.4) with $b = 2$ as a model for waves in hyperelastic rods.

Hereditary symmetries. The paper [19] rederives (3.4) with $b = 2$ from general asymptotic considerations and claims that the equation was already obtained in [24, 23] by the theory of hereditary symmetries. Indeed, before [7], families of integrable equations similar to (3.4) with $b = 2$ *could have been obtained* amidst a comprehensive list of other integrable equations provided by the theory of hereditary symmetries [20]. However, the integrable equation (3.4) with $b = 2$ was not written explicitly, nor was it derived physically as a water wave equation, and its solution properties were not studied before [7]. See [25] for an insightful discussion of how the integrable equation (3.4) with $b = 2$ relates to the theory of hereditary symmetries.

3.3. Higher-order nonlinear nonlocal integrable balance. KdV in (3.1) and the cases $b = 2$ (CH) and $b = 3$ (DP) of (3.4) are three completely integrable Hamiltonian equations that possess solitons as traveling waves. In all of these equations, the leading-order balance that confines the traveling wave soliton occurs between nonlinear steepening and linear dispersion. Physically, this is also the leading-order asymptotic balance for shallow water waves. However, the parameter b in (3.4) introduces additional possibilities for higher-order balances, including

the nonlinear nonlocal balance occurring in the following dispersionless case of CH that was studied previously for $b = 2$ in [7]:

$$m_t + \underbrace{um_x + b\,mu_x}_{\text{Nonlinear balance}} = 0, \text{ with } m = u - \alpha^2 u_{xx} \text{ and } \lim_{|x|\to\infty} u = 0. \tag{3.5}$$

To obtain (3.5) from (3.4), one absorbs the linear dispersion terms in (3.4) by a Galilean transformation and a velocity shift followed by a rescaling. Recall that (3.4) is not Galilean invariant; so Galilean transformations do not preserve the functional form of its solutions. Even in the absence of linear dispersion, the nonlinear and nonlocal terms that remain in (3.5) can still balance to produce a confined solitary traveling wave pulse $u(x,t) = ce^{-|x-ct|/\alpha}$, called the *peakon* [7].

The peakon solutions and other properties of dispersionless DP for the case $b = 3$ in (3.5) were studied in [15]. The properties of the class of dispersionless equations consisting of (3.5) for $b = 2$ with the more general defining relation $u = g * m$ in (1.2) for any even kernel $g(x) = g(-x)$ were studied in [22]. The CH peakon case is recovered for $g(x) = e^{-|x|/\alpha}$.

The peakon solution of (3.5) moves at a speed equal to its amplitude and has a jump in derivative at its peak. Peakons for either $b = 2$ or $b = 3$ are true solitons that interact via elastic collisions under CH dynamics or DP dynamics, respectively [7], [15]. In addition, the CH and DP initial value problems are both completely integrable as Hamiltonian systems using the IST method for an isospectral linear eigenvalue problem whose purely discrete spectrum gives the asymptotic speeds of the peakons [7], [15]. Figure 2 shows the evolution under dispersionless CH for the case $b = 2$ in (3.5) of a Gaussian initial velocity distribution of unit area and width 5α. In fact, peakon solutions exist for (3.5) with any value of b. However, we shall find numerically that the stability of these peakon solutions requires $b > 1$. For the case $b = 2$, stability of the single peakon was proved in [13].

Burgers equation. The dispersionless limit of KdV in (3.1) upon rescaling velocity u is the Burgers equation

$$u_t + uu_x - \nu\, u_{xx} = 0, \tag{3.6}$$

in which we have added constant viscosity ν. The Burgers solution is the classic ramp and cliff shown arising from a Gaussian initial condition in Figure 1. In the ramp/cliff solution, nonlinear steepening is balanced by linear viscosity to produce the "cliff" whose width is controlled by the magnitude of ν. The "ramp" is the self-similar $u \approx x/t$ part of the solution for which the viscous term vanishes.

Total momentum. The equations KdV, BBM, CH, and DP and the other b-equations (3.5) with $u = g * m$ and Burgers all preserve the area $M = \int_{-\infty}^{\infty} u\,dx$ (total momentum) for a solution u that vanishes at spatial infinity.

3.3.1. CH peakon solutions. Linear dispersion was absorbed in (3.4) by a Galilean transformation and a velocity shift. For $b = 2$, this procedure reduces the CH (2.1) with linear dispersion to the following purely nonlinear evolution equation that contains competing quadratically nonlinear terms representing convection and stretching:

$$m_t + \underbrace{um_x}_{\text{convection}} + \underbrace{2\,u_x m}_{\text{stretching}} = 0, \quad \text{with} \quad m = u - \alpha^2 u_{xx}. \tag{3.7}$$

This is a special case of (1.1) or (3.5), for which $b = 2$ and $g(x) = e^{-|x|/\alpha}$ in the defining relation (1.2). The traveling wave solution of (3.7) is the "peakon," $u(x,t) = ce^{-|x-ct|/\alpha}$ found in [7], where $e^{-|x|/\alpha}$ is the Green function for the Helmholtz operator that relates m and u. The interactions among N peakons are governed by the $2N$-dimensional dynamical system for the speeds $p_i(t)$ and positions $q_i(t)$, $i = 1, \ldots, N$, appearing in the *superposed* solution

$$u(x,t) = \sum_{i=1}^{N} p_i(t)\, e^{-|x-q_i(t)|}. \tag{3.8}$$

As shown in Camassa and Holm [7], a closed integrable Hamiltonian system of ordinary differential equations for the speeds $p_i(t)$ and positions $q_i(t)$ results upon substituting the superposition of peakons (3.8) into (3.7). This integrable system governs the dynamics of the peakon interactions.

A variant of (3.7) with coefficient $b = 2 \to b = 3$,

$$m_t + \underbrace{um_x}_{\text{convection}} + \underbrace{3\,u_x m}_{\text{stretching}} = 0, \quad \text{with} \quad m = u - \alpha^2 u_{xx}, \tag{3.9}$$

was first singled out for further analysis by Degasperis and Procesi [16]. Degasperis, Holm, and Hone [15] discovered that this $b = 2 \to b = 3$ variant of (3.7) also possesses superposed peakon solutions (3.8) and is completely integrable by the isospectral transform method. Thus, the N-peakon solution (3.8) is a completely integrable dynamical system under the evolution of either (3.7) or (3.9), but these two integrable cases have *different* dynamics for the speeds $p_i(t)$ and positions $q_i(t)$ of the peakons. The proof that N-peakon solutions (3.8) exist for any b in (3.5) follows by direct substitution of the solution ansatz into the equation.

3.3.2. FH pulson solutions. Fringer and Holm [22] extended the zero-dispersion shallow water equation (3.7) for the peakons to the "pulson" equation, which is contained in the family of equations (1.1) for $b = 2$,

$$m_t + \underbrace{um_x}_{\text{convection}} + \underbrace{2\,u_x m}_{\text{stretching}} = 0, \quad \text{with} \quad u = g * m. \tag{3.10}$$

Here $u = g * m$ denotes the convolution (or filtering)

$$u(x) = \int_{-\infty}^{\infty} g(x-y)m(y)\,dy \tag{3.11}$$

that relates velocity u to momentum density m by integration against the kernel $g(x)$. Fringer and Holm [22] chose $g(x)$ to be an even function so that u and m have the same parity. They studied the effects of the shape of the traveling wave $u(x,t) = cg(x-ct)$ on its interactions with other traveling waves in the superposed solution,

$$u(x,t) = \sum_{i=1}^{N} p_i(t)\, g(x - q_i(t)). \tag{3.12}$$

This superposed solution of traveling wave forms with time-dependent speeds $p_i(t)$ and positions $q_i(t)$, $i = 1, \ldots, N$, revealed that the nonlinear interactions among these pulsons occur by elastic two-pulson scattering even though the Fringer–Holm pulson equation (3.10) is not integrable for an arbitrary choice of the kernel g. When $g(x) = e^{-|x|/\alpha}$ is assumed, the pulson equation for $b = 2$ in (3.10) specializes to the peakon equation for $b = 2$ in (3.7). The proof that N-pulson solutions (3.12) exist for any b in (1.1) or (3.10) follows by direct substitution of the solution ansatz into the equation. The resulting reduced, or collective, dynamics for the speeds $p_i(t)$ and positions $q_i(t)$ of the pulsons depends upon the value of b in (3.10) and the choice of the function g in the defining relation (3.11).

3.4. Discrete symmetries: Reversibility, parity, and signature. Equation (1.1) for m is reversible, or invariant under $t \to -t$ and $u \to -u$. The latter implies $m \to -m$. Hence the transformation $u(x,t) \to -u(x,-t)$ takes solutions into solutions, and, in particular, it reverses the direction and amplitude of the traveling wave $u(x,t) = cg(x-ct)$.

We chose $g(x)$ to be an even function so that m and $u = g * m$ would both have odd parity under mirror reflections. Hence (1.1) is invariant under the parity reflections $u(x,t) \to -u(-x,t)$, and the solutions of even and odd parity form invariant subspaces.

Equation (1.1) implies a similar reversible parity invariant equation for the *absolute value* $|m|$:

$$\partial_t |m| + u|m|_x + b\, u_x |m| = 0, \quad \text{with} \quad u = g * m. \tag{3.13}$$

So the positive and negative components $m_\pm = \frac{1}{2}(m \pm |m|)$ satisfy (1.1) separately. Also, if m is initially zero, it remains so. This is conservation of the signature of m.

3.5. Lagrangian representation. If $m^{1/b}$ is well defined, (1.1) may be written as the conservation law

$$\partial_t\, m^{1/b} + \partial_x (m^{1/b} u) = 0, \tag{3.14}$$

and (3.13) for the absolute value implies

$$\partial_t\, |m|^{1/b} + \partial_x (|m|^{1/b} u) = 0. \tag{3.15}$$

Adding and subtracting (3.14) and (3.15) implies

$$\partial_t\, (m^{1/b})_\pm + \partial_x \big((m^{1/b})_\pm u \big) = 0 \quad \text{with} \quad (m^{1/b})_\pm = \frac{1}{2}(m^{1/b} \pm |m|^{1/b}). \tag{3.16}$$

Consequently, regions of positive and negative m are both transported by the velocity $u = g * m$, and their boundaries propagate so as to separately preserve the two integrals

$$\int_{-\infty}^{\infty} (m^{1/b})_\pm \, dx. \tag{3.17}$$

The shared velocity relation $u = g * m$ allows a transformation to Lagrangian coordinates $X_\pm$ defined by

$$dX_\pm = (m^{1/b})_\pm (dx - u dt) \quad \text{so that} \quad \partial_t X_\pm + u \partial_x X_\pm = 0. \tag{3.18}$$

This formal transformation is not strictly defined where $(m^{1/b})_\pm$ vanishes. However, by (3.16), regions where $(m^{1/b})_\pm$ vanishes do not propagate and do not contribute to the integrated value of $X_\pm = \int_{-\infty}^{x} (m^{1/b})_\pm(y,0)\,dy$. Hence these regions may be identified and excluded initially. The formal inverse relation holding in the remaining regions,

$$dx = (m^{1/b})_\pm^{-1}\, dX_\pm + u dt, \tag{3.19}$$

implies that

$$\left.\frac{dx}{dt}\right|_{X_\pm} = u(x,t), \tag{3.20}$$

so the Lagrangian trajectories $x = x(X_\pm, t)$ of positive and negative integrated initial values of $X_\pm = \int_{-\infty}^{x} (m^{1/b})_\pm(y,0)\,dy$ are transported by the same velocity $u = g * m$.

3.6. Preservation of the norm $\|m\|_{L^{1/b}}$ for $0 \le b \le 1$. If $|m|^{1/b}$ is well defined, the continuity equation form (3.15) of (1.1) implies conservation of

$$\int_{-\infty}^{\infty} |m|^{1/b}\,dx = \int_{-\infty}^{\infty} |m_0|^{1/b}\,dx, \quad \text{where} \quad m_0(x) = m(x,0). \tag{3.21}$$

This integral is conserved for all b but defines only a norm (the $L^{1/b}$ norm $\|m\|_{L^{1/b}}$) in the closed interval $0 \le b \le 1$. In the limit $b \to 0$ this becomes the L_∞ norm, $|m|_{max}$. Hence, when $b = 0$, (1.1) has both a maximum principle and a minimum principle for m. Such a principle is meaningful only if $m^{1/b}$ is an ordinary function, e.g., if m is not a generalized function, such as the delta functions that occur for the peakons we shall discuss below.

Thus the $L^{1/b}$ norm $\|m\|_{L^{1/b}}$ is conserved by (1.1), provided $|m|^{1/b}$ is well defined for the closed interval $0 \le b \le 1$. One may also define the corresponding conserved norm for $1/m$ in the closed interval $-1 \le b \le 0$, provided $|1/m|^{1/b}$ is well defined on this interval.

3.7. Lagrangian representation for integer b. Fluid convection means transport of a quantity by the fluid motion. Examples of transported fluid quantities are circulation (a one-form) in Kelvin's theorem for the Euler equations and its exterior derivative the vorticity (a two-form, by Stokes theorem) in the Helmholtz equation. For a Lagrangian trajectory $x(X,t)$ satisfying $x(X,0) = X$ and

$$dx = (m^{1/b})^{-1} dX + u dt, \tag{3.22}$$

we have seen that the conservation law (3.14) implies

$$m^{1/b}(x,t)dx = m^{1/b}(X,0)dX, \tag{3.23}$$

provided that $m^{1/b}$ is a well-defined function. The last issue may be avoided when b is an integer, as follows. In one dimension, higher-order differential forms may be created by using the direct, or tensor, product, e.g., $dx \otimes dx = dx^{\otimes 2}$. Consequently, the tensor product of each side of (3.23) b times gives[1]

$$m(x,t)dx^{\otimes b} = m(X,0)dX^{\otimes b}. \tag{3.24}$$

[1]Cases with positive integer values of b will allow m to be a generalized function. Cases with noninteger values of b will revert to (3.23), for which m is required to be a classical function.

Taking the partial time derivative of this equation at constant Lagrangian coordinate X and using $dx/dt|_X = u$ yield (1.1) in the form

$$\frac{d}{dt}\Big|_X \Big(m(x,t)dx^{\otimes b}\Big) = (m_t + um_x + b\,u_x m)dx^{\otimes b} = 0. \tag{3.25}$$

Thus, when the parameter b in (1.1) is an integer, it may be regarded geometrically as the number of dimensions that are brought into play by coordinate transformations of the quantity $m\,dx^{\otimes b}$ associated with m. Cases of (1.1) with negative integer $b < 0$ may be interpreted as

$$\frac{d}{dt}\Big|_X \Big(m\big(\partial_x\big)^{\otimes(-b)}\Big) = 0. \tag{3.26}$$

For example, the case $b = -1$ may be written as[2]

$$\frac{d}{dt}\Big|_X (m\partial_x) = (m_t + um_x - u_x m)\partial_x = 0, \tag{3.27}$$

in which the difference of terms $(um_x - u_x m)\partial_x$ is the commutator of the vector fields $u\partial_x$ and $m\partial_x$ on the real line. The rest of the paper will remain in the Eulerian (spatial) representation.

3.8. Reversibility and Galilean covariance. Equation (1.1) is reversible; i.e., it is invariant under the discrete transformation $u(x,t) \to -u(x,-t)$. Equation (1.1) is also Galilean covariant for all b. In fact, (1.1) keeps its form under transformations to an *arbitrarily* moving reference frame for all b. This includes covariance under transforming to a *uniformly* moving Galilean frame. However, only in the case $b = 0$ is (1.1) Galilean *invariant*, assuming that m Galileo-transforms in the same way as u. If so, then (1.1) transforms under

$$t \to t + t_0, \quad x \to x + x_0 + ct, \quad u \to u + c + u_0, \quad m \to m + c + u_0 \tag{3.28}$$

to the form

$$m_t + \, um_x + bu_x\, m + u_0 m_x + bu_x(c + u_0) = 0, \quad \text{with} \quad u = g * m. \tag{3.29}$$

Thus (1.1) is invariant under space and time translations (constants x_0 and t_0) and covariant under Galilean transforms (constant c) and acquires linear dispersion terms under velocity shifts (constant u_0). Equation (1.1) regains Galilean invariance if m is Galilean invariant. However, the dispersive term $u_0 m_x$ introduced by the constant velocity shift $u_0 \neq 0$ breaks the reversibility of (1.1) even if m is invariant under this shift.

3.9. Integral momentum conservation. Equation (3.14) implies that $M = \int_{-\infty}^{\infty} m\,dx$ is conserved for any g when $b = 1$. However, when $g(x)$ is even, the family of equations (1.1) also conserves the total momentum integral M for any b. This is shown by directly calculating from (1.1) that

$$\frac{d}{dt}\int_{-\infty}^{\infty} m(x)\,dx = (1-b)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m(x)g'(x-y)m(y)\,dx\,dy = 0, \tag{3.30}$$

[2]Remarkably, this nonlinear equation for $b = -1$ has stationary plane wave solutions, $u(x) = \cos(kx)$, and exponential solutions, $u(x) = e^{\pm\kappa x}$.

in which the double integral vanishes as the product of an even function and an odd function under interchange of x and y, when $g'(-x) = -g'(x)$. Hence, for even $g(x)$, $M = \int_{-\infty}^{\infty} m\,dx$ is conserved for either periodic or vanishing boundary conditions and for any b. We shall assume henceforth that $g(x)$ is even and, moreover, that the integral $\int m\,g * m\,dx$ is sign-definite so that it defines a norm (the kinetic energy)

$$\|m\|_g^2 = \int_{-\infty}^{\infty} m\,g * m\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} m(x)\,g(x-y)\,m(y)\,dx\,dy \geq 0. \tag{3.31}$$

This kinetic energy norm is conserved by (1.1), provided $b = 2$.

4. Traveling waves and generalized functions. Its invariance under space and time translations ensures that (1.1) admits traveling wave solutions for any b. Let us write the traveling wave solutions as

$$u = u(z) \quad \text{and} \quad m = m(z), \quad \text{where} \quad z = x - ct, \tag{4.1}$$

and let prime $'$ denote d/dz.

4.1. Case $b = 0$.

4.1.1. Pulsons for $b = 0$. For $b = 0$, (1.1) is Galilean invariant and its traveling wave solutions satisfy

$$(u(z) - c)m'(z) = 0, \quad z = x - ct, \tag{4.2}$$

where prime $'$ denotes d/dz. Equation (4.2) admits generalized functions $m'(z) \simeq \delta(z)$ matched by $u - c = 0$ at $z = 0$. The velocity u is given by the integral of the Green function that relates m and $u = g * m$,

$$u - c \simeq c\left[\int g(y)\,dy\right]_0^z. \tag{4.3}$$

4.1.2. Peakons, ramps, and cliffs for $b = 0$. When $g(x) = e^{-|x|/\alpha}$ (the Green function for the one-dimensional Helmholtz operator), we have $m = u - \alpha^2 u_{xx}$. Consequently, the equation $m' = u' - \alpha^2 u''' = \pm 2\delta(z)$ with $u - c = 0$ at $z = 0$ is satisfied by

$$u - c = \pm c\left[\int e^{-|y|/\alpha}\,dy\right]_0^z = \pm c\,\mathrm{sgn}(z)\big(1 - e^{-|z|/\alpha}\big). \tag{4.4}$$

This represents a rightward moving traveling wave that connects the left states $u - c = \pm c$ to the same two right states.

Definition 4.1 (peakons). *The symmetric solutions $u = \pm c e^{-|z|/\alpha}$, with a jump in derivative at $z = 0$, are the peakons, for which $m = u - \alpha^2 u_{xx}$ and $g(x) = e^{-|x|/\alpha}$.*

Definition 4.2 (cliffs). *The antisymmetric solutions $u = \pm c\,\mathrm{sgn}(z)(1 - e^{-|z|/\alpha})$ (with $u - c = \pm c$ connecting to $u - c = \mp c$), with no jump in derivative at $z = 0$, are the regularized shocks (cliffs). These propagate rightward but may face either leftward or rightward, because (1.1) in the absence of viscosity has no entropy condition that would distinguish between leftward and*

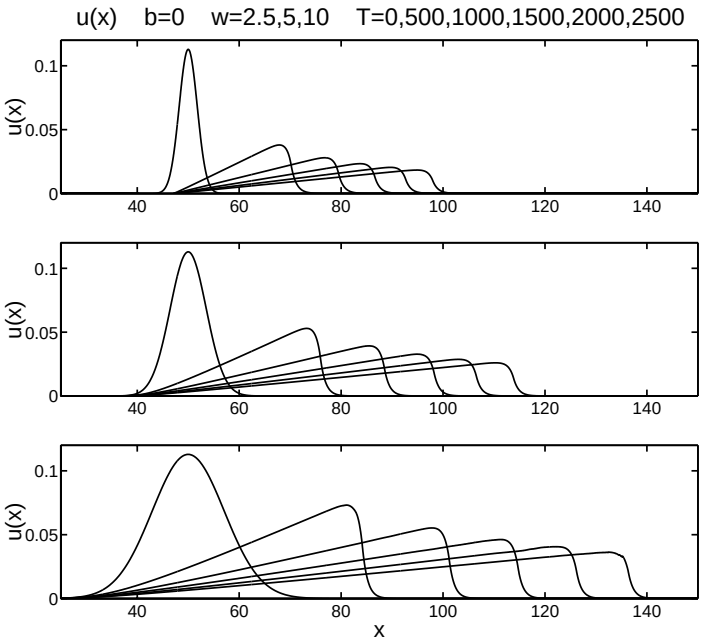


Figure 1. *Ramps and cliffs for $b = 0$. Inviscid b-family, $b = 0$, $\alpha = 1$, initial width $w = 2.5, 5, 10$.*

rightward facing solutions. For more discussion of entropy conditions for the cliff solutions, see [28].

Definition 4.3 (ramps). *Equation* (1.1) *also has ramp-like similarity solutions $u \simeq x/t$ when $g(x) = e^{-|x|/\alpha}$ for any b. These may emerge in the initial value problem for the peakon case of* (1.1) *and interact with the peakons and cliffs.*

Remark 4.1 (first integral for $b = 0$ traveling waves). For $b = 0$, the traveling wave equation (4.2) apparently has only the first integral for $m = u - \alpha^2 u_{xx}$,

$$(u - c)(u - \alpha^2 u'') - \frac{u^2}{2} + \frac{\alpha^2}{2} u'^2 = K. \tag{4.5}$$

Thus, perhaps surprisingly, we have been unable to find a second integral for the traveling wave equation for peakons when $b = 0$.

Remark 4.2 (reversibility). Reversibility means that (1.1) is invariant under the transformation $u(x, t) \to -u(x, -t)$. Consequently, the rightward traveling waves have leftward moving counterparts under the symmetry $c \to -c$. The case of constant velocity $u = \pm c$ is also a solution.

Figure 1 shows that the ramp and cliff pattern develops in the velocity profile under the peakon equation (1.1) with $g(x) = e^{-|x|/\alpha}$ for a set of Gaussian initial conditions $(5\sqrt{\pi})^{-1} \exp(-(x - 50)^2/w)$ of increasing widths $w = 2.5, 5, 10$ for $\alpha = 1$ and $b = 0$. Apparently, the ramp solution is numerically stable, but the coexisting peakon solution is not stable in this case. A complete stability analysis of these various solutions is outside the scope of the present paper. Instead we shall investigate the solutions of (1.1) by numerically integrating selected examples.

4.2. Case $b \neq 0$. For $b \neq 0$, the conservation law (3.14) for traveling waves becomes

$$\big((u-c)m^{1/b}\big)' = 0\,, \tag{4.6}$$

which yields after one integration

$$(u-c)^b m = K\,, \tag{4.7}$$

where K is the first integral. For $g(x) = e^{-|x|/\alpha}$, so that $m = u - \alpha^2 u_{xx}$, this becomes

$$(u-c)^b(u-\alpha^2 u'') = K\,. \tag{4.8}$$

For $u - c \neq 0$, we rewrite this as

$$\alpha^2 u'' = u - K(u-c)^{-b} \tag{4.9}$$

and integrate again to give the second integral in two separate cases,

$$\alpha^2 u'^{\,2} = \begin{cases} u^2 - \frac{2K}{1-b}(u-c)^{1-b} + 2H & \text{for} \quad b \neq 1\,, \\ u^2 - 2K\log(u-c) + 2H & \text{for} \quad b = 1\,. \end{cases}$$

We shall rearrange this into quadratures:

$$\pm\frac{dz}{\alpha} = \frac{du}{\Big[u^2 - \frac{2K}{1-b}(u-c)^{1-b} + 2H\Big]^{1/2}} \quad \text{for} \quad b \neq 1 \tag{4.10}$$

and

$$\pm\frac{dz}{\alpha} = \frac{du}{\Big[u^2 - 2K\log(u-c) + 2H\Big]^{1/2}} \quad \text{for} \quad b = 1\,. \tag{4.11}$$

For $b = 1$ and $K \neq 0$, the integral in (4.11) is transcendental.

4.2.1. Special cases of traveling waves for $b \neq 0$.

- For $K = 0$ the two quadratures (4.10) and (4.11) are equal, independent of b, and elementary, thereby yielding the traveling wave solutions

$$e^{-|z|/\alpha} = \frac{u + \sqrt{u^2+2H}}{c + \sqrt{c^2+2H}}\,, \tag{4.12}$$

  with $u - c = 0$ at $z = 0$.
- For $H = 0$ (4.12) recovers the peakon traveling wave.
- For $H > 0$ (4.12) gives a rightward moving traveling wave that is a continuous deformation of the peakon.
- For $H > 0$ and $c = 0$ (4.12) gives stationary solutions of the form

$$u + \sqrt{u^2+2H} \simeq e^{-|z|/\alpha}\,. \tag{4.13}$$

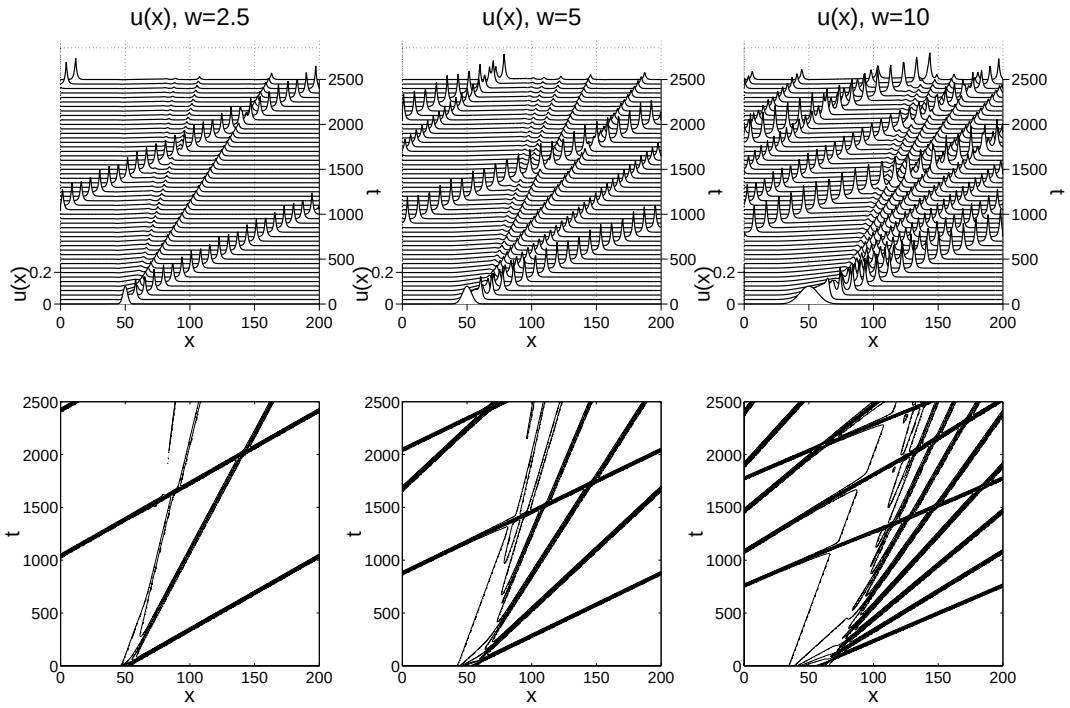


Figure 2. *Peakons for $b = 2$. Inviscid b-family, $b = 2$, $\alpha = 1$, initial width $w = 2.5, 5, 10$.*

4.3. Case $b > 0$.

4.3.1. Pulsons for $b > 0$. Equation (1.1) for $b > 0$ has nontrivial solutions vanishing as $|z| \to \infty$ that allow $K = 0$ in (4.7) so that

$$(u - c)^b m = 0\,. \tag{4.14}$$

This admits the generalized function solutions

$$m = c\delta(z) \quad \text{and} \quad u = g * m = cg(z)\,, \tag{4.15}$$

matched by $u - c = 0$ at $z = 0$. This is the *pulson traveling wave*, whose shape in u is given by the kernel g. The constant velocity case $u = c$ is a trivial traveling wave.

Remark 4.3 (pulson and peakon traveling waves). The pulson solution (4.15) requires $g(0) = 1$ and $g'(0) = 0$. We shall assume for definiteness that the even function $g(z)$ achieves its maximum at $g(0) = 1$ so that the symmetric pulson traveling wave $u(x,t) = cg(x - ct)$ moves at the speed of its maximum, which occurs at its center of symmetry. For example, the peakon $u(x,t) = ce^{-|x-ct|}$ moves at the speed of its peak.

4.3.2. Peakons for $b > 1$. Equation (4.8), for which $g(x) = e^{-|x|/\alpha}$, yields the peakon traveling wave

$$u(z) = ce^{-|z|/\alpha} \quad \text{and} \quad m(z) = u - \alpha^2 u'' = 2c\delta(z/\alpha) \tag{4.16}$$

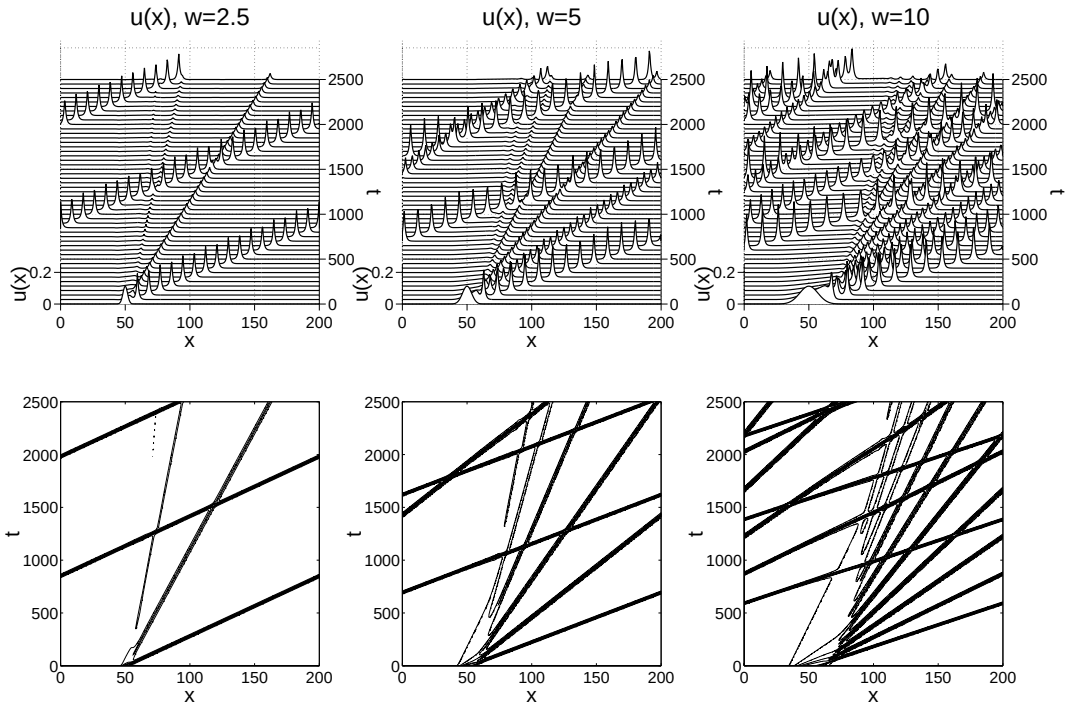


Figure 3. *Peakons for $b = 3$. Inviscid b-family, $b = 3$, $\alpha = 1$, initial width $w = 2.5, 5, 10$.*

when $K = 0$.

Figures 2 and 3 show the effects of varying the width w of a Gaussian initial condition $(5\sqrt{\pi})^{-1}\exp(-(x-50)^2/w)$ for the peakon equation in a periodic domain when $\alpha = 1$ and $b = 2, 3$. As the width of the initial Gaussian increases, the figures show that more peakons of width $\alpha = 1$ are emitted. (This is consistent with conservation of momentum.) The peakons are observed to be stable for $b > 1$, they propagate as solitary traveling waves, and they interact elastically. We shall discuss the peakon interactions in more detail in sections 6 through 9.

4.4. Case $b < 0$. We shall examine the cases $b = -0.5, -1, -2, -3, -4$. Numerical results for $b = -2$ and $b = -3$ are described in section 4.4.6. For other values of $b < 0$, the analysis is similar, but it involves less elementary considerations such as transcendental or hyperelliptic functions. The numerics shown later will demonstrate that the elementary solutions discussed here, many of them stationary, do tend to emerge in numerical integrations of the initial value problem for (1.1) with $b \le -1$.

4.4.1. Case $b = -1/2$. Figure 4 shows that a ramp and cliff pattern develops in the velocity profile under the peakon equation (1.1) with $g(x) = e^{-|x|/\alpha}$ for a set of Gaussian initial conditions $(5\sqrt{\pi})^{-1}\exp(-(x-100)^2/w)$ of increasing width $w = 10, 15, 20$ for $\alpha = 1$ and $b = -1/2$. Apparently, the ramp solution is numerically stable for $b = -1/2$.

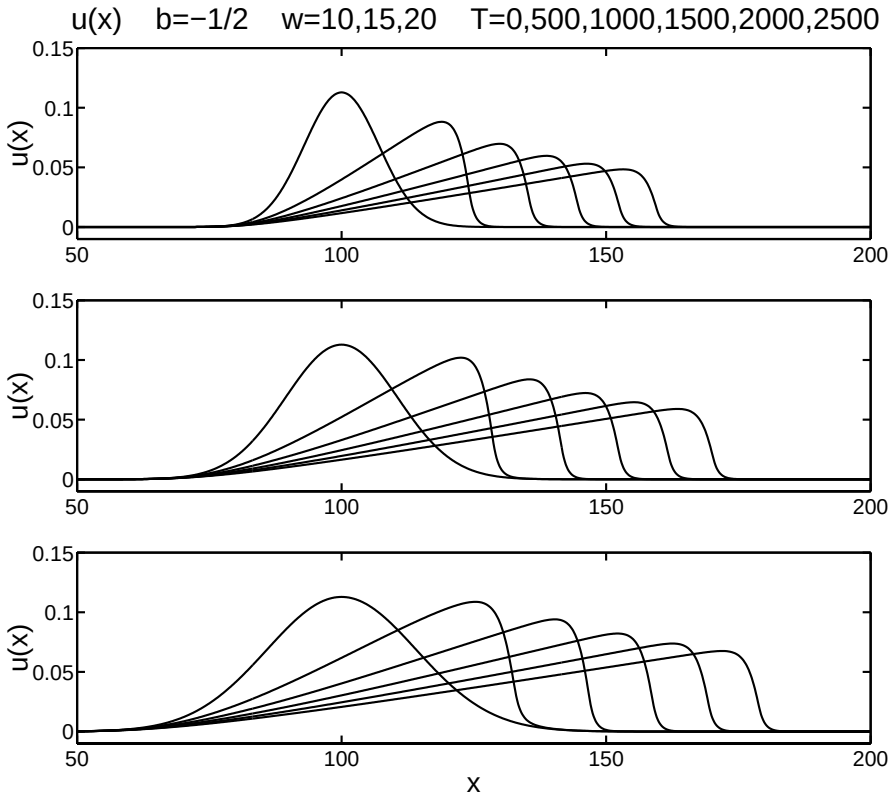


Figure 4. *Ramps and cliffs for $b = -1/2$. Inviscid b-family, $b = -1/2$, $\alpha = 1$, initial width $w = 10, 15, 20$.*

4.4.2. Case $b = -1$.

For $b = -1$, (4.10) becomes

$$\pm \frac{dz}{\alpha} = \frac{du}{\left[u^2 - K(u-c)^2 + 2H\right]^{1/2}}\,, \tag{4.17}$$

which integrates to

$$e^{-|z|/\alpha} = \frac{u + \sqrt{u^2 - K(u-c)^2 + 2H} + Kc}{c + \sqrt{c^2 + 2H} + Kc}\,, \tag{4.18}$$

with $u - c = 0$ at $z = 0$. ($K = 0$ and $H = 0$ recover the peakon traveling wave.)

Remark 4.4 (stationary plane wave solutions for $b = -1$). Equation (1.1) for $b = -1$ is satisfied for any wavenumber k by

$$m = \cos(k(x - ct) + \phi_0) \quad \text{and} \quad u = \hat{g}(k)\cos(k(x - ct) + \phi_0)\,, \tag{4.19}$$

where $\hat{g}(k)$ is the Fourier transform of the kernel $g(x)$ and ϕ_0 is a constant phase shift. In the absence of linear dispersion, these solutions are stationary, $c = 0$. When linear dispersion is added to (1.1), these solutions are the one-dimensional analogues of Rossby waves in the two-dimensional quasi-geostrophic equations.

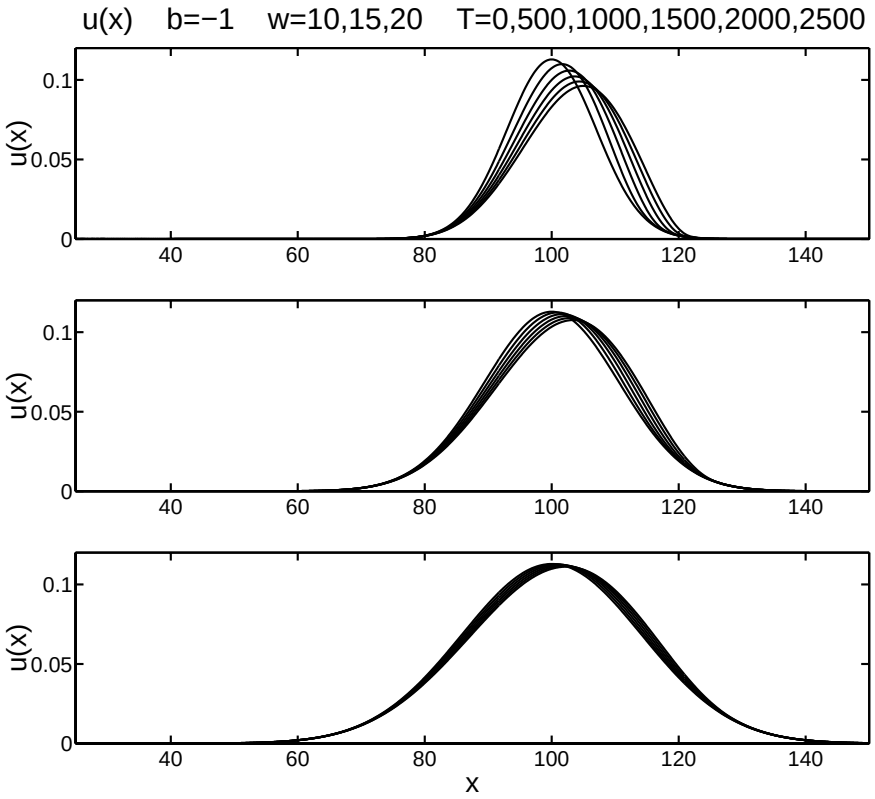


Figure 5. *Nearly stationary solutions for $b = -1$. Inviscid b-family, $b = -1$, $\alpha = 1$, initial width $w = 10, 15, 20$.*

Figure 5 shows the velocity profiles under evolution by the peakon equation, (1.1) with $g(x) = e^{-|x|/\alpha}$, for a set of Gaussian initial conditions of increasing width $w = 10, 15, 20$ for $\alpha = 1$ and $b = -1$. Evidently, the coexisting peakon solution for $b = -1$ does not emerge because $K \neq 0$ and $H \neq 0$ for this initial condition. Instead, the stable solution is essentially stationary with a slight rightward drift and leaning slightly to the right. The reason for this lethargic propagation becomes clear upon writing the b-equation solely in terms of the velocity $u(x,t)$ as

$$
\begin{aligned}
u_t + (b+1)uu_x &= \alpha^2(u_{xxt} + uu_{xxx} + bu_xu_{xx}) \\
&= \alpha^2\partial_x\left(u_{xt} + uu_{xx} + \frac{b-1}{2}u_x^2\right) \\
&= \alpha^2\partial_x^2\left(u_t + uu_x + \frac{b-3}{2}u_x^2\right).
\end{aligned}
\tag{4.20}
$$

Remark 4.5 ($b = -1$ is a turning point). When $b = -1$ the nonlinear steepening term $(b+1)uu_x$ vanishes in (4.20) and the residual propagation is due to its nonlinear "curvature terms" with higher-order derivatives. In the parameter regime $b > -1$ (resp., $b < -1$) the solutions of (1.1) or (4.20) move rightward (resp., leftward), provided the curvature terms on the right-

hand side of (4.20) are either negative or sufficiently small.

Remark 4.6 (short wave limit equation). The high wavenumber, or short wave, limit of (4.20) is

$$\partial_x^2\left(u_t + uu_x + \frac{b-3}{2}u_x^2\right) = 0\,. \tag{4.21}$$

For $b = 2$ and $\lim_{x\to-\infty} u_x = 0$, this integrates to become the Hunter–Saxton equation [31]. For $b = 3$, it is the second derivative of the Burgers equation.

4.4.3. Case $b = -2$ stationary solutions. For $b = -2$, the traveling wave quadrature (4.10) becomes an elliptic integral,

$$\pm\frac{dz}{\alpha} = \frac{d(u-c)}{\left[u^2 - \frac{2K}{3}(u-c)^3 + 2H\right]^{1/2}}\,. \tag{4.22}$$

The hyperbolic limit of this equation for $H = 0$ vanishes at infinity for the stationary solution ($c = 0$) to give

$$u(z) = \frac{3}{2K}\mathrm{sech}^2\frac{z}{2\alpha}\,. \tag{4.23}$$

4.4.4. Case $b = -3$ stationary solutions. For $b = -3$, the hyperbolic limit $H = 0$ of (4.10) is

$$\pm\frac{dz}{\alpha} = \frac{du}{\left[u^2 - \frac{K}{2}(u-c)^4\right]^{1/2}}\,, \tag{4.24}$$

which for $c = 0$ is

$$\pm\frac{dz}{\alpha} = \frac{du}{u\left[1 - \frac{K}{2}u^2\right]^{1/2}} \tag{4.25}$$

and may be integrated in closed form to obtain a continuous deformation of the peakon,

$$\frac{e^{-|z|/\alpha}}{1+\sqrt{1-K/2}} = \frac{u}{1+\sqrt{1-Ku^2/2}} \text{ for } b=-3\,,\ c=0\,,\ \text{and } H=0. \tag{4.26}$$

Rearranging (4.26) and scaling u by u_0 give

$$u(z) = \frac{u_0}{\frac{A}{2}e^{|z|/\alpha} + (1-\frac{A}{2})e^{-|z|/\alpha}}\,, \tag{4.27}$$

with $A = 1+\sqrt{1-K/2}$ so that $A \in [1,2]$ for $K \in [0,2]$. For $A = 1$, we have $u(z) = u_0\,\mathrm{sech}\,(z/\alpha)$, and for $A = 2$, we recover the stationary peakon, $u(z) = u_0\,e^{-|z|/\alpha}$.

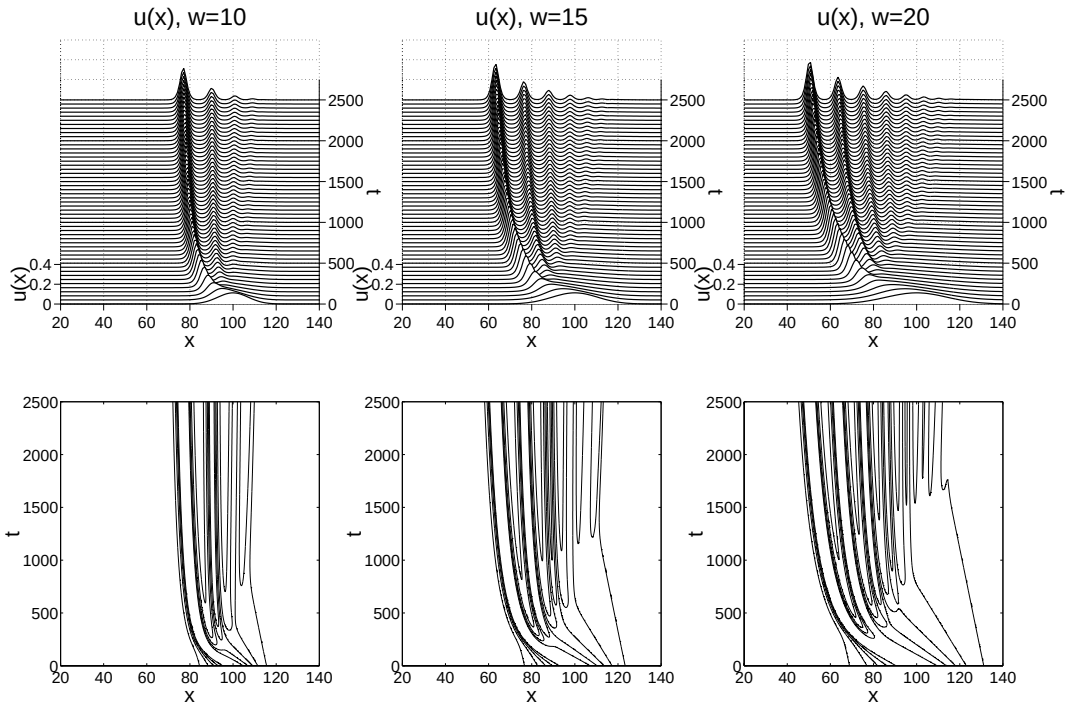


Figure 6. *Stationary solutions for $b = -2$. Inviscid b-family, $b = -2$, $\alpha = 1$, initial width $w = 10, 15, 20$.*

4.4.5. Case $b = -4$ stationary solutions. For $b \leq -4$, the analytical expressions for the cnoidal waves become less elementary, because the integral in (4.10) is then hyperelliptic. However, our numerics show that the dynamical behavior for $b = -4$ is similar to that of the cases $b = -2$ and $b = -3$ shown in Figures 6–7. Namely, a series of transient leftward propagating pulses, or *leftons*, of width alpha emerge and tend to a nearly steady state. Consistent with momentum (area) conservation and the tendency toward pulses of width alpha, the number of emerging leftons increases with the width of the initial Gaussian. At a longer time scale, this train of pulses appears to tend toward stationary ($c \to 0$).

4.4.6. Numerical results for $b = -2$ and $b = -3$. Figures 6 and 7 show that a series of leftons in the velocity profile emerges under the peakon equation for a set of Gaussian initial conditions of increasing width $w = 10, 15, 20$ for $\alpha = 1$ and $b = -2, -3$. Apparently these are not peakons, because the velocity at which they move is not equal to their height. The leftons emerge from the initial Gaussian in order of height and then tend toward a nearly stationary state. The number of emerging pulses increases with the width of the initial Gaussian, as expected from momentum (area) conservation and the tendency toward pulses of width alpha, and the leftward speed of the emerging pulses increases with the magnitude of b. The latter is consistent with the coefficient $(b+1)$ of the nonlinearity in (4.20) as b becomes more negative.

Figure 8 shows the leftons at time $T = 2500$, versus $u(x) \simeq \operatorname{sech}^2(x/(2\alpha))$ for $b = -2$, and versus $u(x) \simeq \operatorname{sech}(x/\alpha)$ for $b = -3$. By this time, the leftons have become stationary solutions with $c = 0$ for both $b = -2$ and $b = -3$.

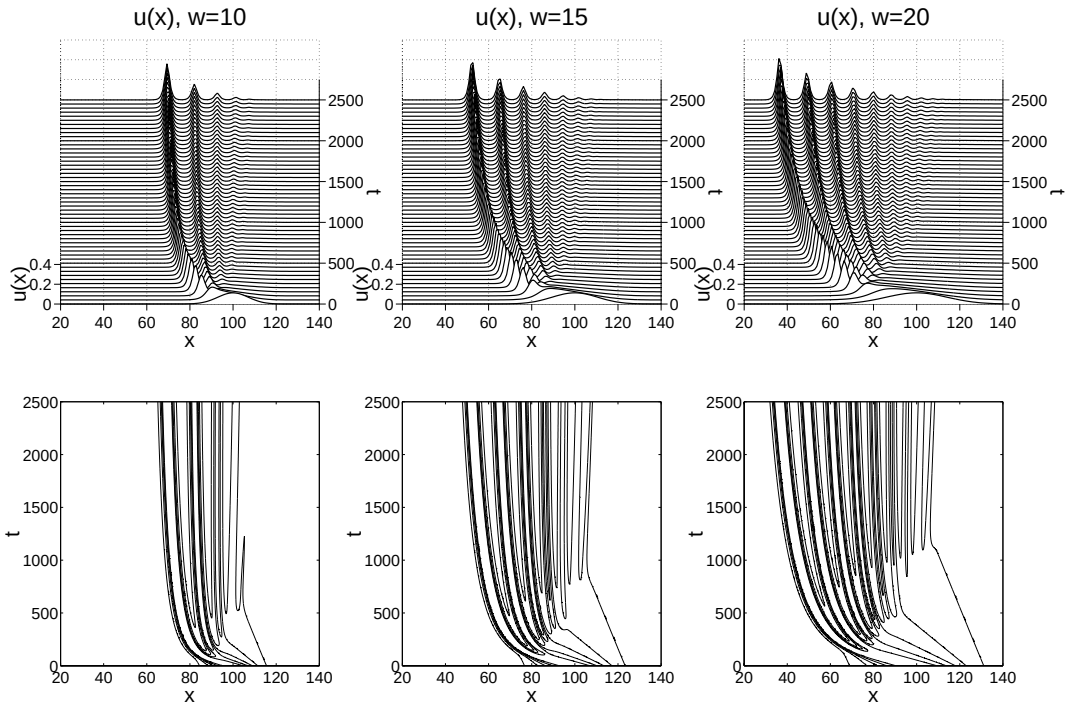


Figure 7. *Stationary solutions for $b = -3$. Inviscid b-family, $b = -3$, $\alpha = 1$, initial width $w = 10, 15, 20$.*

5. Pulson interactions for $b > 0$. As we have seen in section 4.3.1, the b-family of equations (1.1) admits the pulson traveling wave solution (4.15) for $b > 0$. The interaction dynamics among N of these pulsons is obtained by superposing the traveling wave solutions $u(x,t) = cg(x - ct)$ as

$$u(x,t) = \sum_{i=1}^{N} p_i(t) g(x - q_i(t)) \quad \text{and} \quad m(x,t) = \sum_{i=1}^{N} p_i(t) \delta(x - q_i(t)) \tag{5.1}$$

for any $b > 0$ and $u = g * m$, where the function g is even so that $g'(0) = 0$ and is bounded and we may set $g(0) = 1$. For these superpositions of pulsons to be exact solutions, the time-dependent parameters $p_i(t)$ and $q_i(t)$ must satisfy the following N-dimensional particle dynamics equations obtained by substituting (5.1) into (1.1):

$$\dot{p}_i = (1-b)\, p_i \sum_{j=1}^{N} p_j\, g'(q_i - q_j) = (1-b)\, \frac{\partial G_N}{\partial q_i}\,, \tag{5.2}$$

$$\dot{q}_i = \sum_{j=1}^{N} p_j\, g(q_i - q_j) = \frac{\partial G_N}{\partial p_i}\,. \tag{5.3}$$

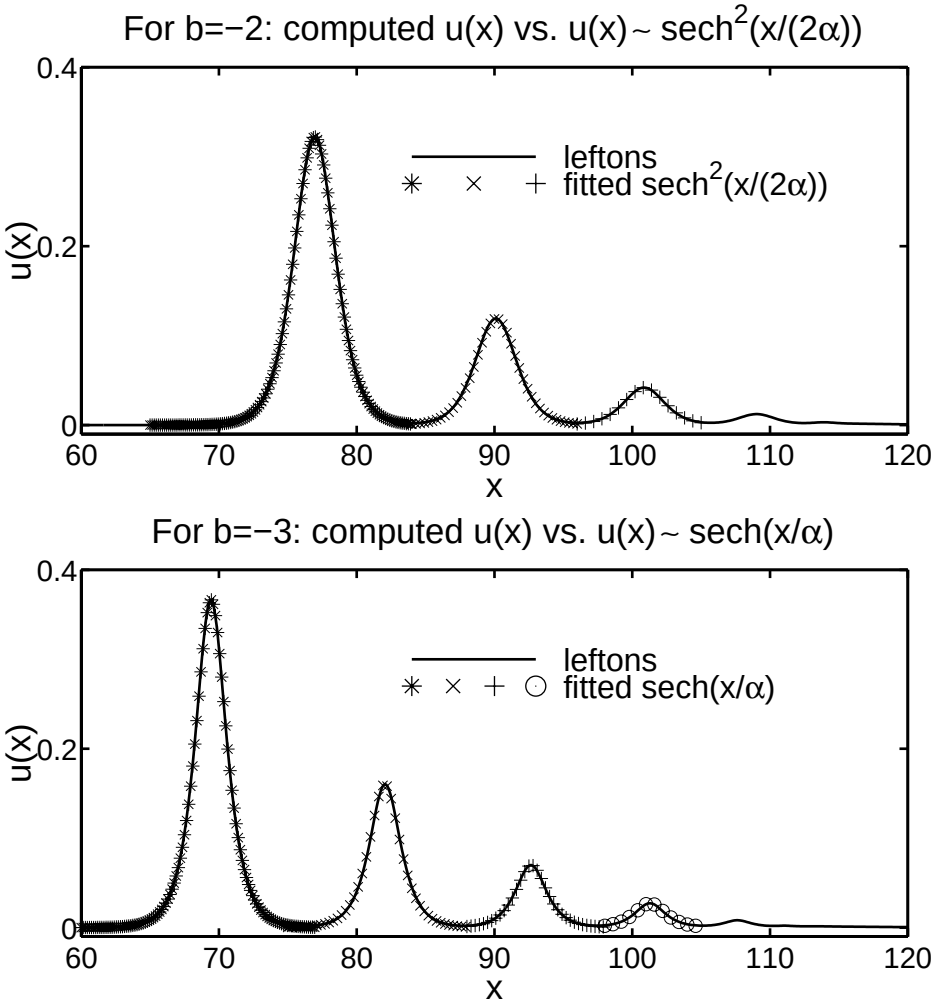


Figure 8. *Stationary solutions for* $b = -2$ *versus* $u(x) \simeq \mathrm{sech}^2(x/(2\alpha))$, *and for* $b = -3$ *versus* $u(x) \simeq \mathrm{sech}(x/\alpha)$. *Inviscid b-family,* $b = -2, -3$, $\alpha = 1$, *initial width* $w = 10$.

Here the generating function G_N is obtained by restricting the norm $\|m\|_g^2$ in (3.31) to the class of superposed traveling wave solutions (5.1), as

$$G_N = \frac{1}{2} \sum_{i,j=1}^{N} p_i p_j \, g(q_i - q_j) \, . \tag{5.4}$$

Thus the symmetric kernel $g(x)$ determines the shape of the traveling wave solutions (5.1), and these traveling waves interact nonlinearly via the pulson dynamics of $p_i(t)$ and $q_i(t)$ with $i = 1, \ldots, N$ in (5.2) and (5.3) for $b > 0$. We shall see that the character of these interactions depends vitally on the value of b.

5.1. Pulson interactions for $b = 2$. When $b = 2$, (5.2) and (5.3) describe the canonical dynamics of a Hamiltonian system with N degrees of freedom. These are the geodesic pulson

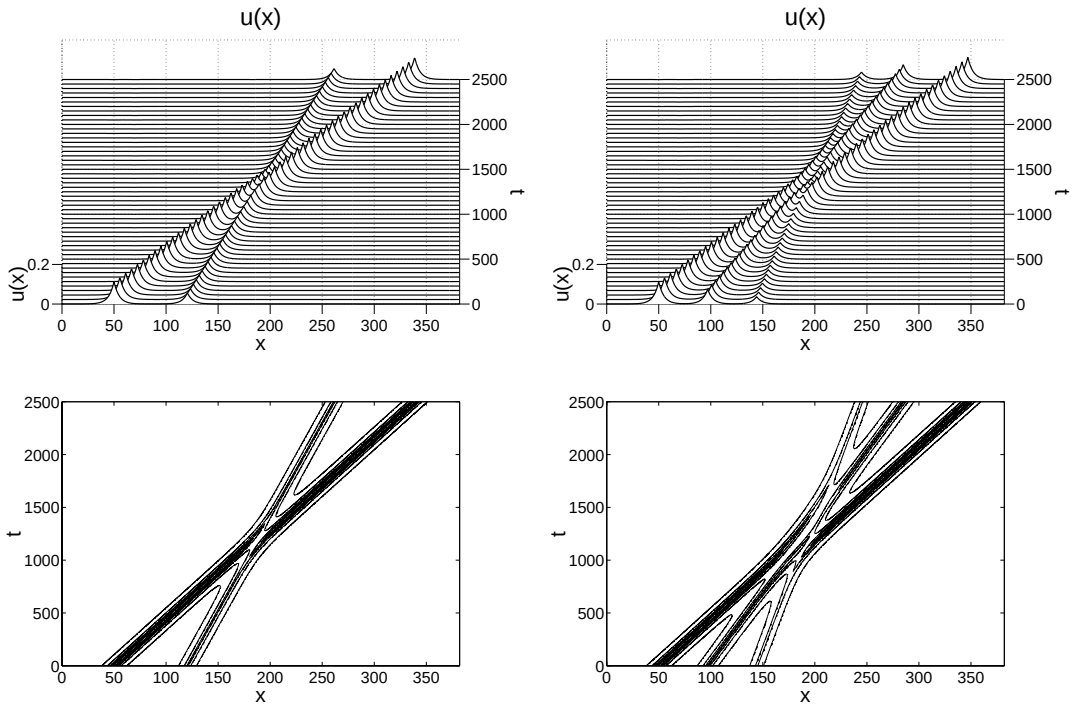


Figure 9. *Peakons of width α for $b = 2$: collisions. Inviscid b-family, $b = 2$, $\alpha = 5$, initial width $w = 5$.*

equations studied in Fringer and Holm [22], in which the following results are obtained:

- Equation (1.1) conserves the kinetic energy $\frac{1}{2}\|m\|_g^2 = \frac{1}{2}\int_{-\infty}^{\infty} m\, g * m\, dx$.
- Equations (5.2) and (5.3) describe canonical geodesic motion in an N-dimensional configuration space whose cometric is $g^{ij}(q) = g(q_i - q_j)$.
- The generating function G_N is the kinetic energy Hamiltonian for the canonical geodesic motion.
- The solutions in (5.1) behave as particle-like pulses whose pairwise interactions as determined by (5.2) and (5.3) comprise nonlinear, but elastic, scattering events.
- The pairwise interactions for the pulsons can be solved analytically for any symmetric function $g(x)$.

Remark 5.1. As we shall show, the last two statements also hold for any $b > 1$.

5.2. Peakon interactions for $b = 2$ and $b = 3$: Numerical results.

- Figure 9 shows the evolution of the velocity profiles in the 2-peakon and 3-peakon interactions for $b = 2$, with $g(x) = e^{-|x|/\alpha}$ and a periodic domain. The 3-peakon interaction decomposes into a series of 2-peakon interactions. These simulations verify the analytical results for the 2-peakon interaction to three significant figures over propagation distances of about sixty peakon widths.
- Figure 10 shows the evolution of the velocity profiles in the 2-peakon and 3-peakon interactions for $b = 3$, with $g(x) = e^{-|x|/\alpha}$ and a periodic domain.
- Figure 11 shows that peakons of initial width greater than α break up into peakons of width α under the evolution of the peakon equation in a periodic domain at fixed

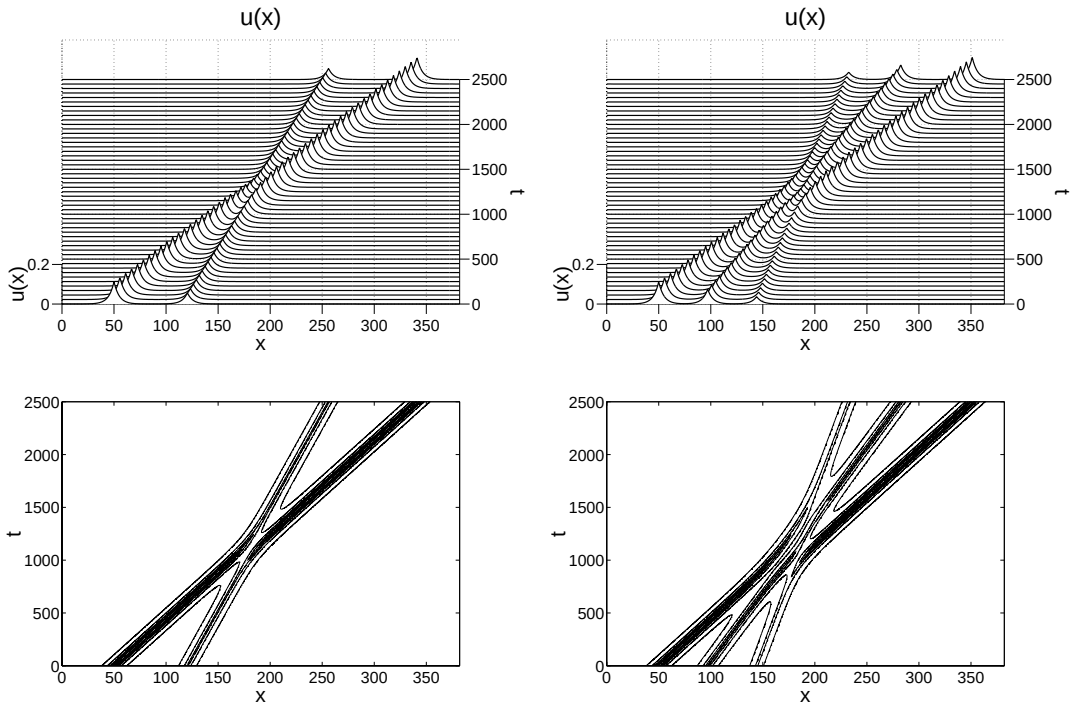


Figure 10. *Peakons of width α for $b = 3$: collisions. Inviscid b-family, $b = 3$, $\alpha = 5$, initial width $w = 5$.*

values of $\alpha = 1$ and $b = 2, 3$. The emitted peakons are stable, propagate as solitary traveling waves, and interact elastically. Conversely, a peakon or other initial condition that is narrower than α will decompose into two oppositely moving trains of peakons and antipeakons, each of width α.

5.3. Pulson-pulson interactions for $b > 0$ and symmetric g. For $N = 2$, the pulson dynamics in (5.2) and (5.3) for $b > 0$ reduces to

$$\frac{dp_1}{dt} = (1-b)\frac{\partial G}{\partial q_1}, \quad \frac{dp_2}{dt} = (1-b)\frac{\partial G}{\partial q_2}, \tag{5.5}$$

$$\frac{dq_1}{dt} = \frac{\partial G}{\partial p_1}, \quad \frac{dq_2}{dt} = \frac{\partial G}{\partial p_2}, \tag{5.6}$$

and the generating function from (5.4) is given by

$$G = \frac{1}{2}(p_1^2 + p_2^2) + p_1 p_2\, g(q_1 - q_2)\,. \tag{5.7}$$

The equations are canonically Hamiltonian only for $b = 2$, which includes the CH case for which $g(q_1 - q_2) = e^{-|q_1 - q_2|}$ gives the peakon solutions.

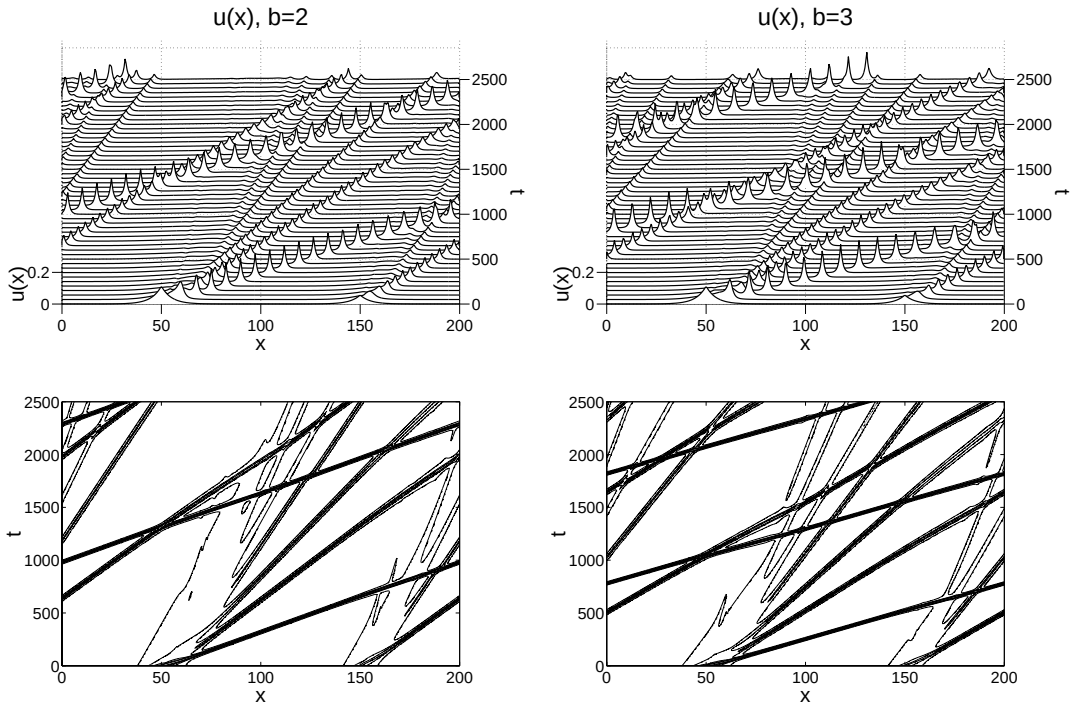


Figure 11. *Peakons of width $> \alpha$ for $b = 2, 3$: emergence of width-α peakons. Inviscid b-family, $b = 2, 3$, $\alpha = 1$, initial width $w = 5$.*

Conservation laws and reduction to quadrature. Besides the total momentum

$$P = p_1 + p_2 \,, \tag{5.8}$$

the 2-pulson system for $b > 0$ and symmetric g also conserves a second quantity that is quadratic in p_1 and p_2, namely,

$$H = p_1 p_2 \big(1 - g(q_1 - q_2)\big)^{b-1} \,. \tag{5.9}$$

For a Hamiltonian system with two degrees of freedom, this second conservation law would be enough to ensure integrability by Liouville's theorem. Even in the present case of $b > 0$ without a Hamiltonian structure, this will be sufficient to reduce the 2-pulson system to quadratures.[3]

Following the analysis for the case $b = 2$ and arbitrary g in Fringer and Holm [22], we introduce sum and difference variables as

$$P = p_1 + p_2 \,, \quad Q = q_1 + q_2 \,, \quad p = p_1 - p_2 \,, \quad q = q_1 - q_2 \,. \tag{5.10}$$

In these variables, the generating function (5.7) becomes

$$G = \frac{1}{4} P^2 (1 + g(q)) + \frac{1}{4} p^2 (1 - g(q)) \,, \tag{5.11}$$

[3]When $b = 1$, the momenta p_1 and p_2 are separately conserved, and the problem immediately reduces to quadratures in $q = q_1 - q_2$ and $Q = q_1 + q_2$.

and the second constant of motion (5.9) becomes

$$H = \frac{1}{4}(P^2 - p^2)\big(1 - g(q)\big)^{b-1}. \tag{5.12}$$

Likewise, the 2-pulson equations of motion transform to sum and difference variables as

$$\begin{aligned}
\frac{dP}{dt} &= 2(1-b)\frac{\partial G}{\partial Q} = 0\,,\\
\frac{dQ}{dt} &= 2\frac{\partial G}{\partial P} = P(1+g(q))\,,\\
\frac{dp}{dt} &= 2(1-b)\frac{\partial G}{\partial q} = \frac{1}{2}(1-b)(P^2-p^2)g'(q)\,,\\
\frac{dq}{dt} &= 2\frac{\partial G}{\partial p} = p(1-g(q)).
\end{aligned}$$

Eliminating p^2 between the formula for H and the equation of motion for q yields

$$\left(\frac{dq}{dt}\right)^2 = P^2\big(1-g(q)\big)^2 - 4H\big(1-g(q)\big)^{3-b}. \tag{5.13}$$

We rearrange this into the following quadrature:

$$dt = \frac{dg(q)}{g'(q)\sqrt{Z}}, \qquad Z = P^2\big(1-g(q)\big)^2 - 4H\big(1-g(q)\big)^{3-b}. \tag{5.14}$$

This simplifies to the quadratic $Z = P^2(1-g(q))^2 - 4H$ when $b = 3$. For the peakon case, we have $g(q) = e^q$ so that $g'(q) = g(q)$ and the quadrature (5.14) simplifies to an elementary integral for $b = -1, 0, 1, 2, 3$. Having obtained $q(t)$ from the quadrature, the momentum difference $p(t)$ is found from (5.12) via the algebraic expression

$$p^2 = P^2 - \frac{4H}{\big(1-g(q)\big)^{b-1}}\,, \tag{5.15}$$

in terms of q and the constants of motion P and H. Finally, the sum $Q(t)$ is found by a further quadrature. The remainder of the solution for arbitrary b and g closely follows Fringer and Holm [22] for the case $b = 2$.

Upon writing the quantities H, P, and G as

$$H = c_1c_2, \quad P = c_1 + c_2, \quad G = \frac{1}{2}c_1^2 + \frac{1}{2}c_2^2 = \frac{1}{2}P^2 - H, \tag{5.16}$$

in terms of the asymptotic speeds of the pulsons, c_1 and c_2, we find the relative momentum relation

$$p^2 = (c_1+c_2)^2 - \frac{4c_1c_2}{\big(1-g(q)\big)^{b-1}}\,. \tag{5.17}$$

This equation has several implications for the qualitative properties of the 2-pulson collisions.

Definition 5.1. *Overtaking, or rear-end, pulson collisions satisfy* $c_1c_2 > 0$*, while head-on pulson collisions satisfy* $c_1c_2 < 0$.

The pulson order $q_1 < q_2$ is preserved in an overtaking, or rear-end, collision when $b > 1$. This follows, as stated in Proposition 5.2.

Proposition 5.2 (preservation of pulson order). *For overtaking, or rear-end, collisions when* $b > 1$*, the 2-pulson dynamics preserves the sign condition* $q = q_1 - q_2 < 0$.

Proof. Suppose the peaks were to overlap in a collision for $b > 1$, thereby producing $q = 0$ during a collision. The condition $g(0) = 1$ implies that the second term in (5.17) diverges for $b > 1$ when the overlap occurs. However, this divergence would contradict $p^2 \geq 0$. ■

Consequently, seen as a collision between two initially well-separated "particles" with initial speeds c_1 and c_2, the separation $q(t)$ reaches a nonzero distance of closest approach q_{min} in an overtaking, or rear-end, collision that may be expressed in terms of the pulse shape, as in Corollary 5.3.

Corollary 5.3 (minimum separation distance). *The minimum separation distance reachable in 2-pulson collisions with* $c_1c_2 > 0$ *is given by*

$$\big(1 - g(q_{min})\big)^{b-1} = \frac{4c_1c_2}{(c_1 + c_2)^2}\,. \tag{5.18}$$

Proof. Set $p^2 = 0$ in (5.17). ■

Remark 5.2. We shall use result (5.18) later in checking the accuracy of our numerical simulations of these 2-pulson interactions.

Proposition 5.4 (head-on collisions admit $q \to 0$). *The 2-pulson dynamics allows the overlap* $q \to 0$ *when* $b > 1$ *in head-on collisions.*

Proof. Because $p^2 \geq 0$, the overlap $q \to 0$ implying $g \to 1$ is only possible in (5.17) with $b > 1$ for $c_1c_2 < 0$, that is, for the case of head-on collisions. ■

Remarks about head-on collisions. For $b > 1$, (5.17) implies that $p^2 \to \infty$ diverges when $q \to 0$ in head-on collisions. The case $b = 1$ is regular, and (5.17) reduces to the constant relation $p^2 = (c_1 - c_2)^2$. For $0 < b < 1$, the quantity p^2 no longer diverges when $q \to 0$, and the solution for the relative momentum in head-on collisions is again regular.

5.4. Pulson-antipulson interactions for $b > 1$ and symmetric g.

Head-on pulson-antipulson collision. We consider the special case of *completely antisymmetric* pulson-antipulson collisions, for which $p_1 = -p_2 = p/2$ and $q_1 = -q_2 = q/2$ (so that $P = 0$ and $Q = 0$). In this case, the quadrature formula (5.14) reduces to[4]

$$\pm(t - t_0) = \frac{1}{\sqrt{-4H}} \int_{q(t_0)}^{q(t)} \frac{dq'}{\big(1 - g(q')\big)^{(3-b)/2}}\,, \tag{5.19}$$

and the second constant of motion in (5.12) satisfies

$$-4H = p^2\big(1 - g(q)\big)^{b-1}\,. \tag{5.20}$$

[4]For $b = 3$, the quadrature formula (5.19) for the separation distance in the pulson-antipulson collision reduces to straight-line motion, $q(t) - q(t_0) = \pm 2c(t - t_0)$.

After the collision, the pulson and antipulson separate and travel oppositely apart so that asymptotically in time $g(q) \to 0$, $p \to 2c$, and $H \to -c^2$, where c (or $-c$) is the asymptotic speed (and amplitude) of the pulson (or antipulson). Setting $H = -c^2$ in (5.20) gives a relation for the pulson-antipulson (p, q) phase trajectories for any kernel,

$$p = \pm \frac{2c}{\big(1 - g(q)\big)^{(b-1)/2}} \,. \tag{5.21}$$

Notice that p diverges for $b > 1$ (and switches branches of the square root) when $q \to 0^+$ because $g(0) = 1$. In contrast, p remains constant for $b = 1$ and vanishes for $b < 1$ (and again switches branches of the square root) when $q \to 0^+$. Note that our convention for switching branches of the square root allows us to keep $q > 0$ throughout, so the particles retain their order.

Remark about preservation of particle identity in collisions. The relative separation distance $q(t)$ in pulson-antipulson collisions is determined by following a phase point along a level surface of the second constant of motion H in the phase space with coordinates (q, p). Because H is quadratic, the relative momentum p has two branches on such a level surface, as indicated by the $\pm$ sign in (5.21). At the pulson-antipulson collision point, both $q \to 0^+$ and either $1/p \to 0^+$ or $p \to 0^+$, so following a phase point through a collision requires that one must choose a convention for which a branch of the level surface is taken after the collision. Taking the convention that p changes sign (corresponding to a "bounce") but q does not change sign (so the "particles" keep their identity) is convenient, because it allows the phase points to be followed more easily through multiple collisions. This choice is also consistent with the pulson-pulson and antipulson-antipulson collisions. In these other "rear end" collisions, as implied by (5.17), the separation distance always remains positive, and again the particles retain their identity.

Theorem 5.5 (pulson-antipulson exact solution). *The exact analytical solution for the pulson-antipulson collision for any b and any symmetric g may be written as a function of position x and the separation between the pulses q for any pulse shape or kernel $g(x)$ as*

$$u(x, q) = \frac{c}{\big(1 - g(q)\big)^{(b-1)/2}} \big[g(x + q/2) - g(x - q/2)\big], \tag{5.22}$$

where c is the pulson speed at sufficiently large separation and the dynamics of the separation $q(t)$ is given by the quadrature (5.19) with $\sqrt{-4H} = 2c$.

Proof. The solution (5.1) for the velocity $u(x, t)$ in the head-on pulson-antipulson collision may be expressed in this notation as

$$u(x, t) = \frac{p}{2} g(x + q/2) - \frac{p}{2} g(x - q/2) \,. \tag{5.23}$$

In using (5.21) to eliminate p, this solution becomes (5.22). ■

Figure 12 shows the exact solutions for the peakon-antipeakon collision in the cases $b = 1$, $b = 2$, and $b = 3$. The positive and negative peaks approach each other until the solution develops a negative vertical slope in finite time. As the separation $q \to 0$, the positive and negative peaks "bounce," thereby reversing polarity, after which they separate in opposite directions.

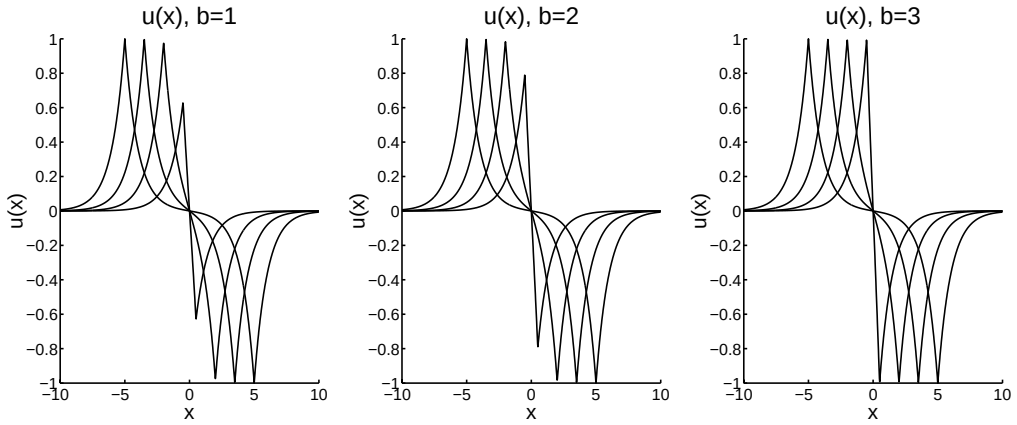


Figure 12. *Peakon-antipeakon collisions for $b = 1, 2, 3$. Inviscid b-family, $b = 1, 2, 3$, $\alpha = 1$, initial width $w = 1$.*

5.5. Specializing pulsons to peakons for $b = 2$ and $b = 3$. We now restrict to $g(x) = e^{-|x|}$, the Green function for the one-dimensional Helmholtz operator satisfying

$$\left(1 - \frac{d^2}{dx^2}\right) e^{-|x|} = 2\delta(x)\,. \tag{5.24}$$

In this case, $m = u - \alpha^2 u_{xx}$, and the pulson traveling wave solution is given by $u(x,t) = cg(x - ct) = ce^{-|x-ct|}$, has a discontinuity in derivative at its peak, and is called the peakon. For $b = 2$ and $b = 3$ in the peakon case, the main results are the following:

- For $b = 2$ and $g(x) = e^{-|x|}$, (1.1) becomes the zero-dispersion limit of the integrable CH for shallow water waves discovered in Camassa and Holm [7]. Upon restoring its linear dispersion, this equation was recently proved to be a higher-order accurate asymptotic description of shallow water waves in Dullin, Gottwald, and Holm [17].
- For $b = 3$ and $g(x) = e^{-|x|}$, (1.1) becomes the integrable PDE studied in Degasperis, Holm, and Hone [15].
- When $g(x) = e^{-|x|}$, the N-peakon dynamics for both $b = 2$ and $b = 3$ turns out to be integrable—as is the dynamics of the original PDE (1.1). The solutions of the initial value problem for (1.1) for both $b = 2$ and $b = 3$ may be found analytically by using the IST method.
- The two cases $b = 2$ and $b = 3$ have quite different isospectral eigenvalue problems. These are discussed in Camassa and Holm [7] and in Dullin, Gottwald, and Holm [17] for the case $b = 2$, and in Degasperis, Holm, and Hone [15] for the case $b = 3$. See also Beals, Sattinger, and Szmigielski [6] for a discussion of solving the inverse isospectral problem using classical methods for the case $b = 2$ for initial momentum distributions $m(x, 0)$ of a single sign.

6. Peakons of width α for arbitrary b. When $g = e^{-|x|/\alpha}$, we may invert the velocity-momentum relation $u = g*m$ by using the Green function expression (5.24) with the Helmholtz operator to find $m = u - \alpha^2 u_{xx}$. Hence (1.1) may be rearranged into the local *momentum*

conservation law,

$$m_t = -\frac{\partial}{\partial x}\left(mu + \frac{b-1}{2}u^2 - \frac{b-1}{2}\alpha^2 u_x^2\right). \tag{6.1}$$

This conservation law for peakons may also be rewritten in *convection form*:

$$u_t + uu_x = -\tau_x \quad \text{with} \quad (1-\alpha^2\partial_x^2)\tau = \frac{b}{2}u^2 + \frac{3-b}{2}\alpha^2 u_x^2\,. \tag{6.2}$$

The two forms (6.1) and (6.2) of the b-family of equations (1.1) suggest that values $b = 0, 1, 3$ are special. These values of b are natural candidates for boundaries, or bifurcation points for changes in solution behavior.

Equation (6.2) describes peakons of shape $g(x) = e^{-|x|/\alpha}$. This peakon equation will form the basis of the rest of our study.

6.1. Slope dynamics for peakons: Inflection points and the steepening lemma when $1 < b \leq 3$. We shall consider solution dynamics of (6.2) in the peakon case satisfying (4.16) or, equivalently, (1.1) with $g(x) = e^{-|x|/\alpha}$, which satisfies

$$(1-\alpha^2\partial_x^2)e^{-|x|/\alpha} = 2\alpha\delta(x)\,. \tag{6.3}$$

For this case, and with vanishing boundary conditions at spatial infinity, (6.2) and (6.3) imply the peakon equation on the real line,

$$u_t + uu_x = -\frac{1}{2\alpha}\int_{-\infty}^{\infty} e^{-|x-y|/\alpha}\big(buu_y + (3-b)\alpha^2 u_y u_{yy}\big)\,dy\,. \tag{6.4}$$

Taking the x-derivative gives the equation for the slope $u_x(x,t)$

$$\begin{aligned} u_{xt} + uu_{xx} + u_x^2 &= -\frac{1}{2\alpha}\frac{\partial}{\partial x}\int_{-\infty}^{\infty} e^{-|x-y|/\alpha}\big(buu_y + (3-b)\alpha^2 u_y u_{yy}\big)\,dy \\ &= \frac{1}{2\alpha^2}\int_{-\infty}^{\infty} e^{-|x-y|/\alpha}\,\mathrm{sgn}\,(x-y)\big(buu_y + (3-b)\alpha^2 u_y u_{yy}\big)\,dy\,. \end{aligned} \tag{6.5}$$

We shall use these expressions to prove the following.

Proposition 6.1 (peakon steepening lemma). *For b in the range $1 < b \leq 3$, a sufficiently negative slope at an inflection point of u will become vertical in finite time under the dynamics of the peakon equation* (6.4).

Proof. Following Camassa and Holm [7], we shall consider the evolution of the slope u_x at an inflection point $x = \bar{x}(t)$. Define the slope at the inflection point as $s(t) = u_x(\bar{x}(t), t)$ and note that $u_{xx}(\bar{x}(t), t) = 0$. Then (6.5) yields the following evolution equation for $s(t)$:

$$\frac{ds}{dt} + s^2 = \frac{1}{2\alpha^2}\int_{-\infty}^{\infty} e^{-|\,\bar{x}(t)-y\,|/\alpha}\,\mathrm{sgn}\,(\bar{x}(t)-y)\frac{1}{2}\frac{\partial}{\partial y}\big(bu^2 + (3-b)\alpha^2 u_y^2\big)\,dy. \tag{6.6}$$

Integrating by parts using the definition $|y| = y\,\mathrm{sgn}\,(y)$ so that $d|y|/dy = \mathrm{sgn}\,(y) + 2y\delta(y)$ and recalling that $y\delta(y) = 0$ give

$$\frac{ds}{dt} = -\left(\frac{b-1}{2}\right)s^2 + \frac{b}{2\alpha^2}\,u^2 - \frac{1}{2\alpha^2}\int_{-\infty}^{\infty} e^{-|\,\bar{x}(t)-y\,|/\alpha}\frac{1}{2\alpha}\big(bu^2 + (3-b)\alpha^2 u_y^2\big)\,dy.$$

Hence, in the range $0 \leq b \leq 3$, the last term is negative and we have the slope inequality

$$\frac{ds}{dt} \leq -\left(\frac{b-1}{2}\right)s^2 + \frac{b}{2\alpha^2}\,u^2 \quad \text{for} \quad 0 \leq b \leq 3\,. \tag{6.7}$$

We suppose the solution satisfies $(bu^2/\alpha^2) < M$ for some constant M.[5] Then

$$\frac{ds}{dt} \leq -\left(\frac{b-1}{2}\right)s^2 + \frac{M}{2} \quad \text{for} \quad 0 \leq b \leq 3\,. \tag{6.8}$$

Consequently, if $b > 1$,

$$\frac{dX}{1-X^2} = d\coth^{-1}(X) \leq \sqrt{M} \quad \text{for} \quad X = \sqrt{\frac{b-1}{M}}\,s\,. \tag{6.9}$$

This implies, for $s \leq -\sqrt{M}$ initially negative, that

$$s \leq \sqrt{M}\coth\left(\sigma + \sqrt{\frac{b-1}{M}}\frac{M}{2}\,t\right) \quad \text{for} \quad 1 < b \leq 3\,, \tag{6.10}$$

where the dimensionless integration constant $\sigma < 0$ determines the initial slope, which is negative. Under these circumstances, the slope at the inflection point must become vertical by time $t = -2\sigma/\sqrt{M(b-1)}$. ■

Remarks for $1 < b \leq 3$.

- If the initial condition is antisymmetric for $1 < b \leq 3$, then the inflection point at $u = 0$ is fixed and $d\bar{x}/dt = 0$ due to the mirror reflection symmetry $(u, x) \to (-u, -x)$ admitted by (6.4). In this case, $M = 0$, and (6.8) implies

$$\frac{ds}{dt} \leq -\left(\frac{b-1}{2}\right)s^2 \quad \Rightarrow \quad s(t) \leq \frac{-2}{b-1}\left(\frac{1}{t_0 - t}\right). \tag{6.11}$$

  Hence verticality $s = -\infty$ will develop in finite time, regardless of how small the initial slope $|s(0)|$, provided it is negative, $s(0) < 0$, as in Figure 12. If the initial slope is positive, then under this evolution it will relax to zero from above.
- Consequently, traveling wave solutions of (6.4) cannot have the usual sech-like shape for solitons because inflection points with sufficiently negative slope can produce unsteady changes in the shape of the solution profile.
- In this context, for $b = 2$, a result in [11] shows that the slope of the solution $u(x,t)$ is always bounded if the initial distribution of momentum $m(x,0)$ does not change sign.

Caveat. Strictly speaking, the formal proof of the peakon steepening lemma in Proposition 6.1 requires additional regularity assumptions on the initial data, since the slope function $s(t)$ is generally not differentiable. However, provided that the initial data is regular enough, e.g., initial velocity in the Sobolev space $u(x,0) \in H^2(R)$, this formal argument can be made rigorous. See Theorem 2.1 in [12] for an example. Under this additional regularity assumption, one finds that (6.7) and (6.8) in the proof hold almost everywhere in time, but the result of the Proposition (6.10) holds everywhere.

[5] If this inequality is violated, we have another type of singularity. However, for $b = 2$, the constant M can be estimated using a Sobolev inequality. In fact, $M = 4H_1(u)/\alpha^2$ because for this case we have

$$\max_{x\in\mathbf{R}}[u^2(x,t)] \leq \frac{1}{\alpha}\int_{-\infty}^{\infty}(u^2 + \alpha^2 u_x^2)\,dx = 2H_1 = \text{const} \quad \text{for} \quad b = 2\,.$$

6.2. Cases $0 \le b \le 1$. In the range $0 \le b \le 1$, we have from (3.21) that

$$\int_{-\infty}^{\infty} |m|^{1/b}\, dx = \int_{-\infty}^{\infty} |m_0|^{1/b}\, dx\,, \quad \text{where} \quad m_0(x) = m(x,0)\,. \tag{6.12}$$

This conservation law implies an elliptic regularity estimate showing that the slope $s = u_x$ is always bounded under the dynamics of the peakon equation (6.4). See [30] for a proof of this result and more discussion of its implications.

7. Adding viscosity to peakon dynamics. In the remainder of this paper, we shall restrict our one-dimensional considerations to the peakon case $g(x) = e^{-|x|/\alpha}$ with length scale α and investigate the fate of the peakon solutions when viscosity is introduced for given values of b and α. For purposes of comparison with previous results in the literature, we shall also extend (1.1) to a new family of equations that includes the Burgers equation by introducing two additional real parameters. These are the viscosity ν and a multiplier β for the stress, or pressure gradient.

First, we shall introduce constant viscosity $\nu > 0$ into (1.1) to form the viscous b-family of equations for the peakon case $g(x) = e^{-|x|/\alpha}$ as follows:

$$m_t + \underbrace{u m_x}_{\text{convection}} + \underbrace{b\, u_x m}_{\text{stretching}} = \underbrace{\nu\, m_{xx}}_{\text{viscosity}}, \quad \text{with} \quad m = u - \alpha^2 u_{xx}\,. \tag{7.1}$$

As in (4.20), this equation with viscosity may be expressed solely in terms of the velocity $u(x,t)$ as

$$\begin{aligned} u_t + (b+1)uu_x - \nu u_{xx} &= \alpha^2(u_{xxt} + uu_{xxx} + bu_x u_{xx} - \nu u_{xxxx}) \\ &= \alpha^2 \partial_x \left(u_{xt} + uu_{xx} - \nu u_{xxx} + \frac{b-1}{2} u_x^2 \right) \\ &= \alpha^2 \partial_x^2 \left(u_t + uu_x - \nu u_{xx} + \frac{b-3}{2} u_x^2 \right). \end{aligned} \tag{7.2}$$

Thus the nonlinear steepening term increases with b as $(b+1)uu_x$. When $\alpha \to 0$, the previous equation reduces to

$$u_t + (b+1)uu_x - \nu u_{xx} = 0, \tag{7.3}$$

and one then recovers the usual *Burgers equation* either by rescaling dimensions or by setting $b = 0$. For $b = 2$, (7.1) is the one-dimensional version of the three-dimensional Navier–Stokes-alpha model for turbulence [9]. For $b = -1$, the evolution in (7.2) occurs only by higher-order terms.

The viscous b-family of peakon equations (7.1) may be rearranged into two other equivalent forms that are convenient for introducing a stress multiplier. These are either its equivalent *conservative form*,

$$m_t = -\frac{\partial}{\partial x} \left(mu + \frac{b-1}{2} u^2 - \frac{b-1}{2} \alpha^2 u_x^2 \right) + \nu\, m_{xx}\,, \tag{7.4}$$

or its equivalent *convective form*,

$$(1 - \alpha^2 \partial_x^2)(u_t + uu_x - \nu u_{xx}) = -\partial_x \left(\frac{b}{2} u^2 + \frac{3-b}{2} \alpha^2 u_x^2 \right). \tag{7.5}$$

Stress multiplier β. Next, we shall introduce a stress multiplier β as a second parameter that for $\beta \neq 1$ deforms the convective form of the viscous b-family of equations (7.5) into the following family of Burgers-like equations with four parameters b, α, ν, and β:

$$u_t + uu_x - \nu u_{xx} = -\beta\,\tau_x \quad \text{with} \quad (1-\alpha^2\partial_x^2)\tau = \frac{b}{2}u^2 + \frac{3-b}{2}\alpha^2 u_x^2\,. \tag{7.6}$$

When $\beta = 0$, the Burgers-$\alpha\beta$ equation (7.6) recovers the usual Burgers equation. When $\beta = 1$, (7.6) recovers the viscous b-family of peakon equations (7.1).

We shall seek solutions of the Burgers-$\alpha\beta$ equation (7.6), either on the real line and vanishing at spatial infinity or in a periodic domain, for various values of its four parameters b, α, ν, and β. Under these boundary conditions, when $\beta \to 1$, (7.6) recovers the convective form (7.5) of the viscous b-family for peakons with $g(x) = e^{-|x|/\alpha}$. Thus the viscous b-family of equations (7.1)–(7.5) deforms into the Burgers-$\alpha\beta$ equation (7.6) when $\beta \neq 1$, and the Burgers-$\alpha\beta$ equation (7.6) reduces to the usual Burgers equation when $\beta = 0$. We shall be interested in the effects of the four parameters b, α, ν, and β on the solutions of the Burgers-$\alpha\beta$ equation (7.6). We shall be interested especially in the fate of the peakon solutions upon introducing the parameters ν and β so as to retain H^1_α control of the velocity. As we shall see, such control requires a special relation between the parameters b and β, namely, $(3-b)\beta = 1$.

7.1. Burgers-$\alpha\beta$ equation: Analytical estimates.

Proposition 7.1 (H^1_α control of the velocity). *The Burgers-$\alpha\beta$ equation* (7.6) *controls the α-weighted H^1 norm of the velocity*

$$\|u\|^2_{H^1_\alpha} = \int_{-\infty}^{\infty} \left(u^2 + \alpha^2 u_x^2\right) dx$$

for $\alpha^2 \neq 0$, provided $(3-b)\beta = 1$.

Proof. The spatial derivative of the Burgers-$\alpha\beta$ equation (7.6) yields the dynamics for the slope $s = u_x$ as

$$\begin{aligned} u_t + uu_x - \nu u_{xx} &= -\beta\tau_x\,, \\ s_t + us_x + s^2 - \nu s_{xx} &= -\beta\,\tau_{xx}\,, \quad \text{with} \quad s = u_x\,, \\ -\alpha^2\tau_{xx} &= \frac{b}{2}\,u^2 + \frac{3-b}{2}\alpha^2\,u_x^2 - \tau\,. \end{aligned}$$

In turn, these slope dynamics equations imply the following evolution of the α-weighted H^1 density (cf. (6.6)):

$$\begin{aligned} \frac{\partial}{\partial t}\left(\frac{1}{2}u^2 + \frac{\alpha^2}{2}s^2\right) = \frac{\partial}{\partial x}\Bigg(\frac{1}{3}\left(1-\frac{b\beta}{2}\right)u^3 + \beta u\tau + \frac{\alpha^2}{2}us^2 + \nu uu_x + \nu\alpha^2 ss_x\Bigg) \\ - \nu u_x^2 \;-\; \nu\alpha^2 s_x^2 \;+\; \frac{\alpha^2}{2}\big((3-b)\beta - 1\big)s^3\,. \end{aligned}$$

Thus, provided

$$(3-b)\beta = 1\,,$$

the last term will vanish. Under this condition, for periodic or vanishing boundary conditions, the α-weighted H^1 norm

$$\|u\|^2_{H^1_\alpha} = \int_{-\infty}^{\infty} \left(u^2 + \alpha^2 u_x^2\right) dx$$

will decay monotonically under the Burgers-$\alpha\beta$ dynamics for $\alpha^2 \neq 0$. ■

Remarks.

- When $\nu \to 0$ in the Burgers-$\alpha\beta$ equation, the α-weighted H^1 norm is conserved for $(3-b)\beta = 1$. This relation cannot be satisfied for $b = 3$. Thus the proof of decay of the α-weighted H^1 norm under the Burgers-$\alpha\beta$ dynamics is inconclusive for $\nu \neq 0$ when $b = 3$. However, one can expect on physical grounds that this norm will also decay for $b = 3$ if ν is sufficiently large.
- We shall restrict our remaining considerations to those values of b and β for which the α-weighted H^1 norm is bounded or decays monotonically. In one dimension, this control of the α-weighted H^1 norm implies the solution for the velocity will be *continuous*.
- Namely, we shall consider the following cases with $(3-b)\beta = 1$: $(b = 0,\ \beta = 1/3)$, $(b = 1,\ \beta = 1/2)$, and $(b = 2,\ \beta = 1)$.

Proposition 7.2 (Burgers-$\alpha\beta$ steepening lemma). *For b and β in the range $(3-b)\beta \leq 2$ a sufficiently negative slope at an inflection point of velocity u will become vertical in finite time under the dynamics of the inviscid Burgers-$\alpha\beta$ equation (7.6) with $\nu = 0$.*

Proof. The proof follows that for the peakon steepening lemma in Proposition 6.1 and uses the slope equation following from the Burgers-$\alpha\beta$ equation (7.6) with $\nu = 0$ that corresponds to (6.5) for the peakons, modified to include β,

$$\begin{aligned} u_{xt} + uu_{xx} + u_x^2 &= -\frac{\beta}{2\alpha}\frac{\partial}{\partial x}\int_{-\infty}^{\infty} e^{-|x-y|/\alpha}\left(buu_y + (3-b)\alpha^2 u_y u_{yy}\right) dy \\ &= \frac{\beta}{2\alpha^2}\int_{-\infty}^{\infty} e^{-|x-y|/\alpha}\,\mathrm{sgn}\,(x-y)\left(buu_y + (3-b)\alpha^2 u_y u_{yy}\right) dy\,. \end{aligned} \tag{7.7}$$

Equation (7.7) yields the inviscid Burgers-$\alpha\beta$ evolution of the slope $s(t) = u_x(\bar{x}(t), t)$ at an inflection point $x = \bar{x}(t)$ as

$$\frac{ds}{dt} \leq -\left(2 - (3-b)\beta\right)\frac{s^2}{2} + \frac{\beta M}{2} \quad \text{for} \quad 0 \leq b \leq 3\,. \tag{7.8}$$

This holds provided we assume the solution satisfies $(bu^2/\alpha^2) < M$ for some constant M. Consequently, if $2 - (3-b)\beta > 0$, we have

$$\frac{dX}{1-X^2} = d\coth^{-1}(X) \leq \sqrt{M} \quad \text{for} \quad X = \sqrt{\frac{2-(3-b)\beta}{M\beta}}\, s\,. \tag{7.9}$$

For $s \leq -\sqrt{M}$ initially negative and $\beta > 0$, this implies

$$s \leq \sqrt{M}\coth\left(\sigma + \sqrt{\frac{2-(3-b)\beta}{M\beta}}\,\frac{M}{2}\,t\right) \quad \text{for} \quad 0 \leq 3 - \frac{2}{\beta} < b \leq 3\,, \tag{7.10}$$

where the dimensionless integration constant $\sigma < 0$ determines the initial slope, which is negative. Under these circumstances, provided the inflection point continues to exist, its negative slope must become vertical by time $t = \frac{-2\sigma}{M}\sqrt{\frac{M\beta}{2-(3-b)\beta}}$. ■

Caveat. As for the peakon steepening lemma in Proposition 6.1, this formal proof for Proposition 7.2 via the slope dynamics for $s(t)$ requires an additional assumption of regularity on the initial data to make it rigorous. Under the additional regularity assumption $u(x,0)\in H^2(R)$, one may show as in [12] that (7.8) holds almost everywhere in time, while the result (7.10) holds everywhere.

Corollary 7.3 (inviscid Burgers-$\alpha\beta$ shocks). *Solutions of the inviscid Burgers-$\alpha\beta$ equation (7.6) with $\nu = 0$ that remain continuous in velocity must develop negative vertical slope in finite time.*

Proof. According to Proposition 7.1, continuity of the velocity and hence control of the H^1 norm $\|u\|_{H^1_\alpha}$ require that $(3-b)\beta = 1$. This is in the parameter range where Proposition 7.2 applies. Consequently, verticality will form at an inflection point of negative slope under the dynamics of the inviscid Burgers-$\alpha\beta$ equation (7.6) with $\nu = 0$ for $(3-b)\beta \le 2$. ■

Remark 7.1. Hence, to remain continuous without viscosity, the solution of the inviscid Burgers-$\alpha\beta$ equation must either develop verticality at an inflection point of negative slope or evolve to eliminate such points entirely.

7.2. Burgers-$\alpha\beta$ traveling waves for $\beta(3-b) = 1$ and $\nu = 0$. For $\nu = 0$, the Burgers-$\alpha\beta$ equation (7.6) has traveling waves given by

$$(u-c)u' + \beta\tau' = 0 \quad \text{and} \quad \tau - \alpha^2\tau'' = \frac{b}{2}u^2 + \frac{3-b}{2}\alpha^2(u')^2\,, \tag{7.11}$$

which yield after one integration

$$\frac{u^2}{2} - c\,u + \beta\tau = K, \tag{7.12}$$

where K is the first integral. Consequently, we find

$$\tau - \alpha^2\tau'' = \frac{1}{\beta}\left(K + c\,u - \frac{u^2}{2} + \alpha^2\big((u-c)\,u'' + (u')^2\big)\right). \tag{7.13}$$

The second equation in (7.11) integrates for the special case of $\beta(3-b) = 1$,

$$2Ku + cu^2 - \beta u^3 + \alpha^2(u-c)(u')^2 = 2H. \tag{7.14}$$

For the special case $K = 0 = H$, this becomes

$$\alpha^2(u-c)(u')^2 = (\beta u - c)u^2 \quad \text{for} \quad \beta(3-b) = 1, \tag{7.15}$$

and we recover the peakon solution $u(z) = ce^{-|z|/\alpha}$ for $\beta = 1$. In the general case that $K \neq 0$ and $H \neq 0$, we rearrange (7.14) into the following quadrature for inviscid Burgers-$\alpha\beta$ traveling waves:

$$\pm\frac{dz}{\alpha} = \frac{(u-c)^{1/2}\,du}{\big[2H - 2Ku - cu^2 + \beta u^3\big]^{1/2}} \quad \text{for} \quad \beta(3-b) = 1 \text{ and } \nu = 0\,. \tag{7.16}$$

In what follows, we shall consider the cases ($b = 0$, $\beta = 1/3$), ($b = 1$, $\beta = 1/2$), and ($b = 2$, $\beta = 1$) when $\nu \neq 0$.

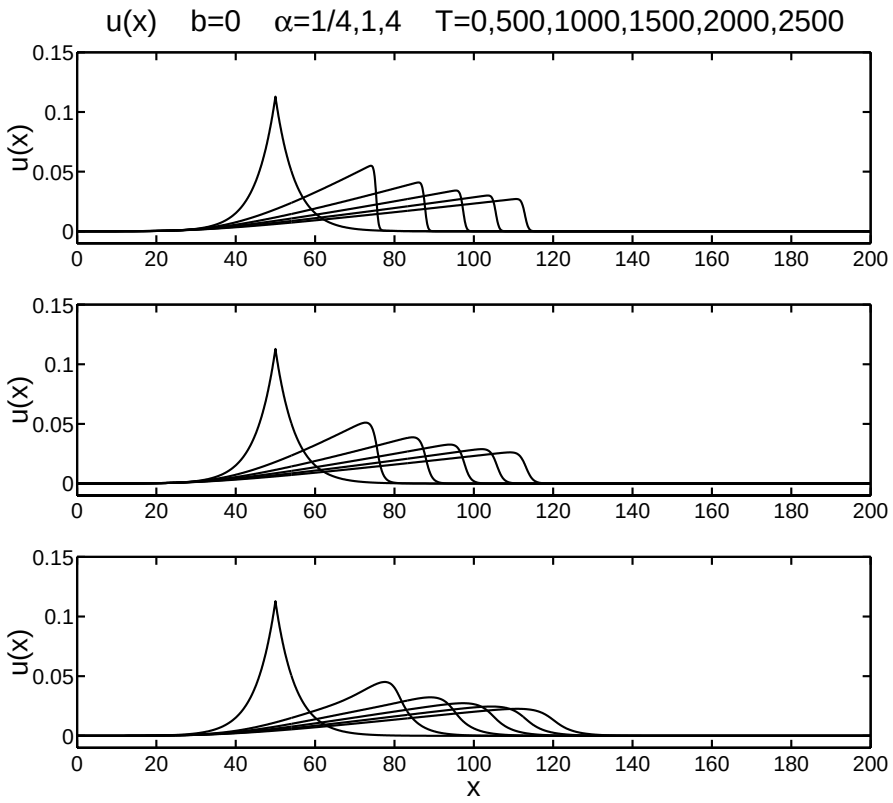


Figure 13. *Effect of increasing α for $b = 0$. Viscous b-family, $b = 0$, $\alpha = 1/4, 1, 4$, $\nu = 0.005$, initial width $w = 5$.*

8. The fate of peakons under (1) adding viscosity and (2) Burgers-$\alpha\beta$ evolution.

8.1. The fate of peakons under adding viscosity. The four figures mentioned below show the effects on the initial value problem for the viscous b-equation (7.1) of varying α and b at fixed viscosity for an initial velocity distribution given by a peakon of width $w = 5$ and initial height $U \simeq 0.1$. The parameter b takes the values $b = 0, 1, 2, 3$. In these four figures, the resolution is 2^{13} points on a domain size of 200 with viscosity $\nu = 0.005$. This corresponds to a grid-scale Reynolds number of $Re_{\Delta x} = U\Delta x/\nu = O(1)$ for velocity $U \simeq 0.1$. The pair of figures after these four then shows the effects on the same problem of increasing viscosity ν at fixed α for $b = 2$ and $b = 3$.

Figure 13 shows three plots of the evolution of the velocity profile under the viscous b-equation (7.1) of an initial peakon of width five, as a function of increasing $\alpha = 1/4, 1, 4$ at fixed viscosity $\nu = 0.005$ for $b = 0$. The peakon leans to the right and develops a Burgers-like triangular shock, or ramp and cliff, whose width increases and peak height decreases as α increases. These three plots show no discernable differences for $b = 0$ as the viscosity is decreased to $\nu = 10^{-6}$. Hence the width of the cliff in the ramp and cliff structure for $b = 0$ is set by the value of α in this range of parameters.

Figure 14 shows three plots of the same type of evolution from a peakon initial condition

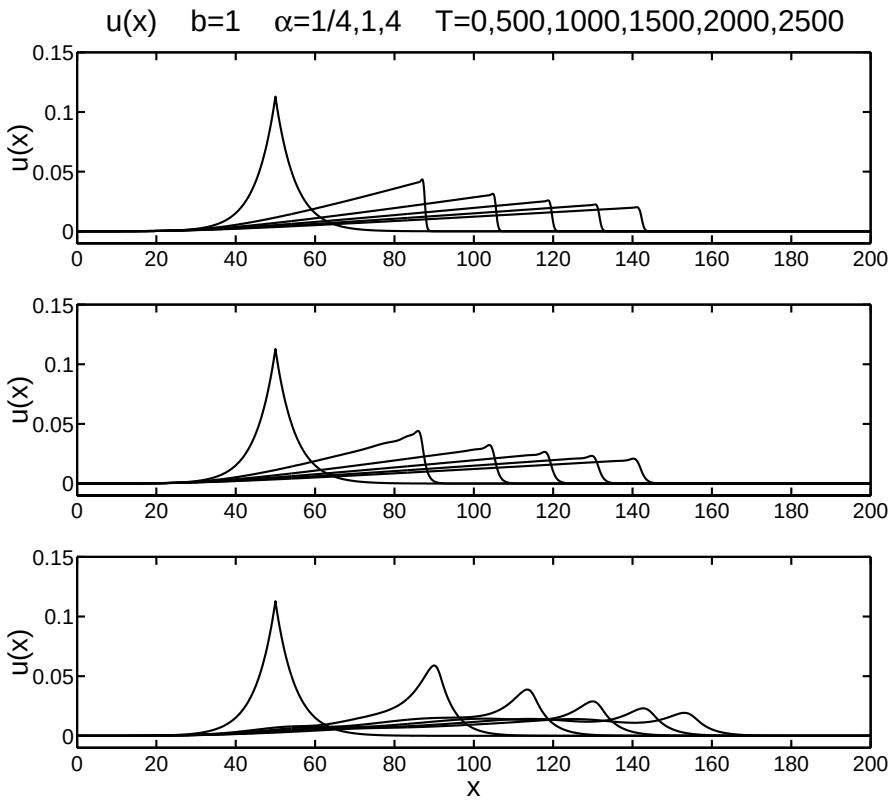


Figure 14. *Effect of increasing α for $b = 1$. Viscous b-family, $b = 1$, $\alpha = 1/4, 1, 4$, $\nu = 0.005$, initial width $w = 5$.*

of width $w = 5$, as α is varied for $b = 1$. The front of the ramp and cliff structure propagates faster and is sharper for $b = 1$ than for $b = 0$ when $\alpha = 1/4$ and $\alpha = 1$. This increase in speed appears to occur because the coefficient increases in the steepening term $(b + 1)uu_x$ in (7.2). A nascent peakon begins to form close behind the front at the top of the ramp and then eventually gets absorbed into the ramp and cliff. For $\alpha = 4$, however, this nascent peakon forms more completely and nearly escapes.

Figure 15 again shows three plots of the evolution from a peakon initial condition of width $w = 5$ as α is varied, this time for $b = 2$. The ramp and cliff structure is faster for $b = 2$ than for $b = 1$ when $\alpha = 1/4$. When $\alpha = 1$, a series of three nascent peakons forms close behind the front and then overtakes the ramp and cliff structure and slightly affects its propagation before eventually being absorbed. For $\alpha = 4$, however, the initial peakon simply propagates and decays under viscosity, although it is slightly rounded at the top.

Figure 16 also shows three plots of the evolution from a peakon initial condition of width $w = 5$ as α is varied, this time for $b = 3$. The ramp and cliff structure moves faster yet, and a single nascent peakon appears just behind the front already for $\alpha = 1/4$. When $\alpha = 1$, a series of three nascent peakons forms initially close behind the front, and they nearly escape before being slowed by viscosity. The leading peakon decays and slows due to viscosity. Then

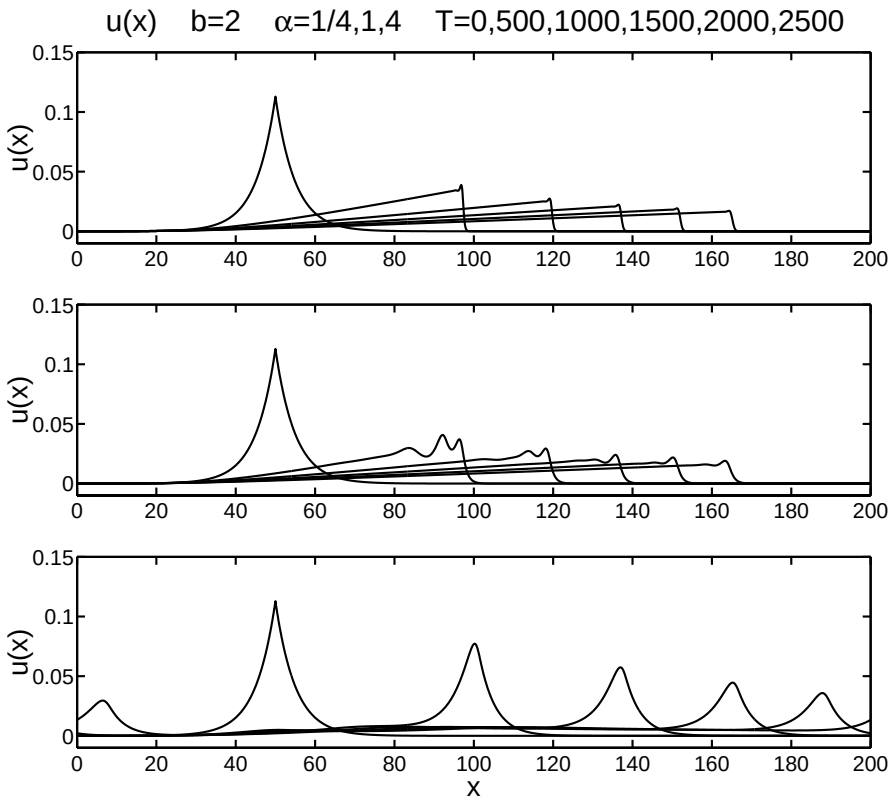


Figure 15. *Effect of increasing α for $b = 2$. Viscous b-family, $b = 2$, $\alpha = 1/4, 1, 4$, $\nu = 0.005$, initial width $w = 5$.*

the following ones overtake and collide with the ones ahead as the ramp and cliff structure forms. These collisions occur at higher relative velocity for $b = 3$ than for $b = 2$, and they significantly affect the propagation and eventual formation of the ramp and cliff. In contrast, for $\alpha = 4$, the initial peakon keeps its integrity and simply propagates rightward and decays under viscosity. The propagating peakon for $\alpha = 4$ at this viscosity decays more slowly and is sharper at the top for $b = 3$ than for $b = 2$.

Remark 8.1 (exchange of stability). To see the exchange of stability between the ramp and cliff structure and the peakon as b changes, we perform the following numerical experiment. First, we run the viscous b-equation (7.1) with $b = 0$, $\alpha = 1$, $\nu = 10^{-5}$, and an initial peakon of width $w = 5$. As we see in Figures 17 and 18, this evolves into the ramp and cliff formation even for nearly zero viscosity. Once the final ramp and cliff state is formed, we then use it as the new initial condition for (7.1) with either $b = 2$ or $b = 3$. The new evolution breaks the ramp and cliff structure into peakons, and the new final state is a rightward moving train of peakons ordered by height.

For Figures 19 and 20, we ran the same numerical experiment, this time with a value $\alpha = 5$ equal to the width of the initial peakon. The initial peakon "borrows from the negative" to form a ramp, which is not quite antisymmetric because the total area of the initial peakon

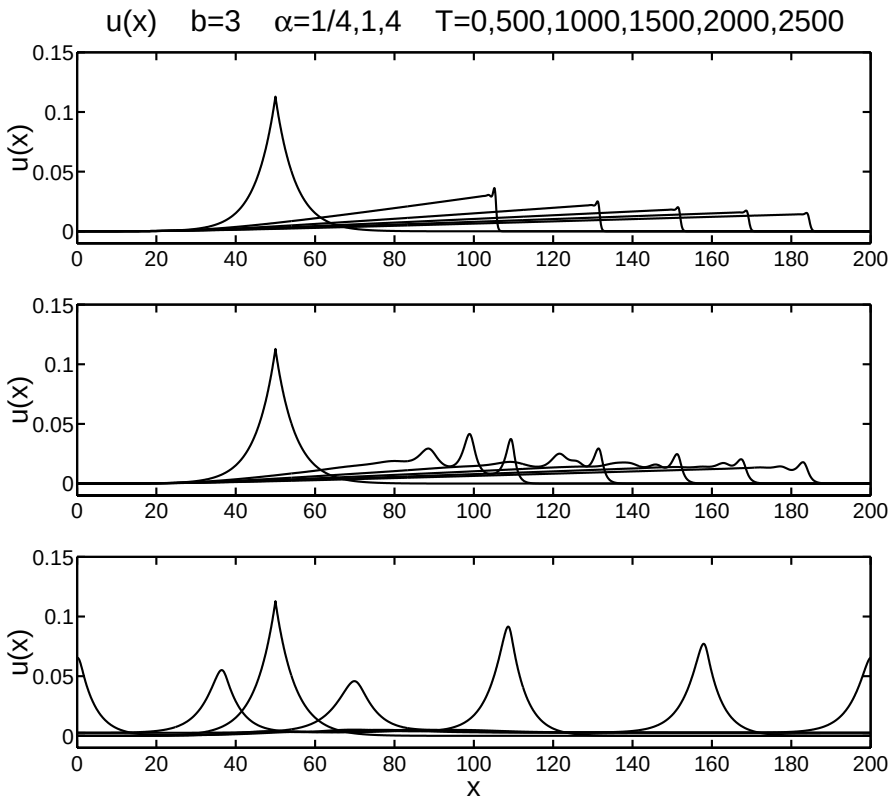


Figure 16. *Effect of increasing α for $b = 3$. Viscous b-family, $b = 3$, $\alpha = 1/4, 1, 4$, $\nu = 0.005$, initial width $w = 5$.*

must be preserved. At time $T = 150$, we switch to $b = 2$ (top plot) or $b = 3$ (bottom plot) and again observe a train of stable peakons emerging from the now unstable ramp.

Finally, for Figures 21 and 22, we again run the numerical experiment with $\alpha = 1$ and an initial peakon width $w = 5$ but this time changing to $b = -2$ or $b = -3$ after the ramp has formed. The new evolution breaks the ramp and cliff structure into leftons like those in Figures 6 and 7.

Remark 8.2 (increasing viscosity). The effect of increasing viscosity on the evolution of the peakon initial condition can be estimated from the α-scale Reynolds number defined by

$$Re_\alpha = U\alpha/\nu = (\alpha/\Delta x)Re_{\Delta x}\,.$$

For $(\alpha/\Delta x) = 40$, $U = 0.1$, and increasing viscosity ν, the Reynolds numbers $Re_{\Delta x}$ and Re_α decrease as

$$Re_{\Delta x} \simeq 2, 0.2, 0.02 \quad \text{and} \quad Re_\alpha = \simeq 80, 8, 0.8 \quad \text{for} \quad \nu = 0.01, 0.1, 1.0.$$

Perhaps not surprisingly, when $Re_\alpha = O(1)$, the viscosity will diffuse through the initial peakon before it can fully form. Figures 23 and 24 show that this effect increases as Re_α decreases.

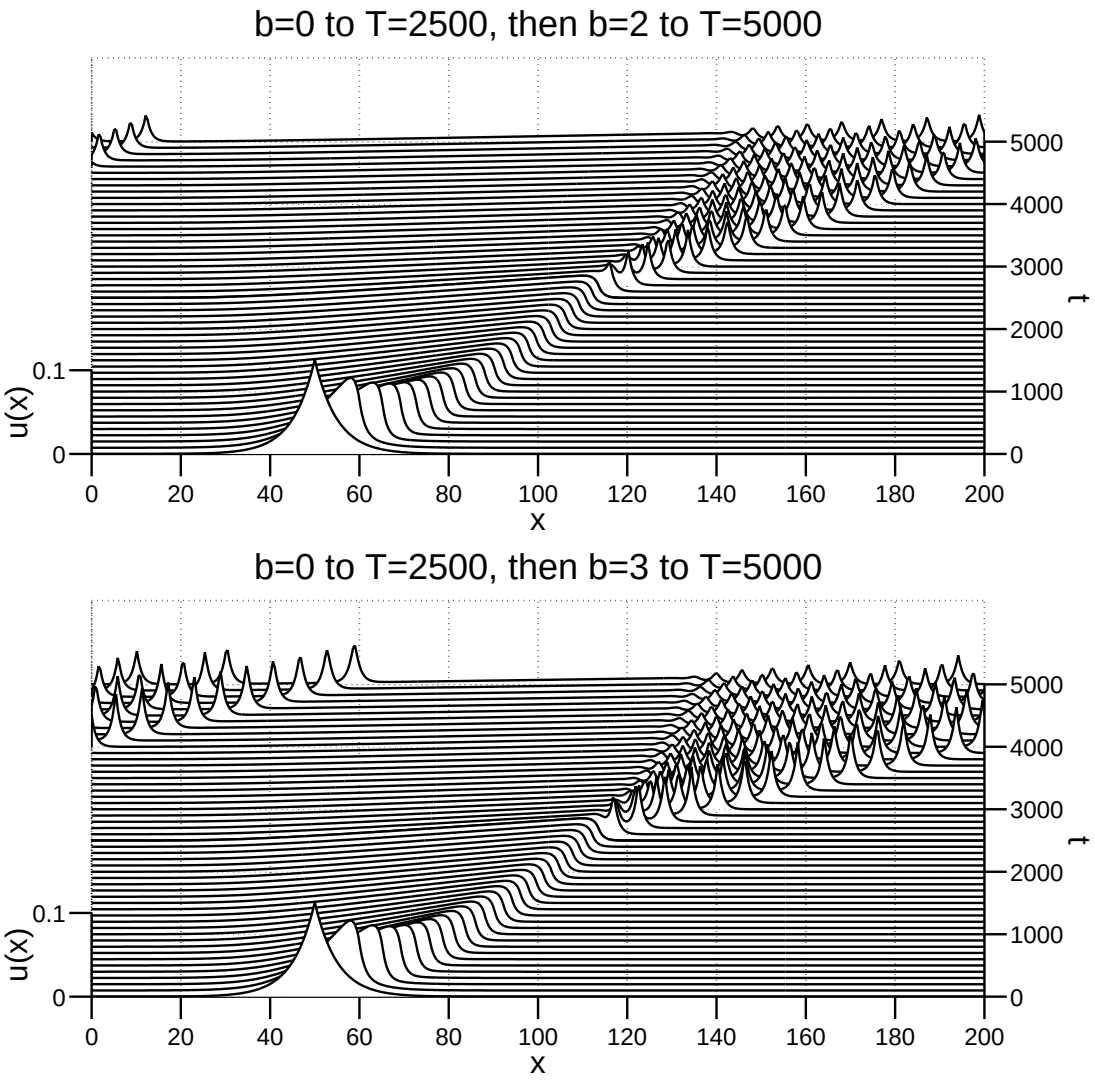


Figure 17. *Exchange of stability between ramps and peakons for $b = 0, 2, 3$, when width $> \alpha$. Viscous b-family, $b = 0 \to 2, 3$, $\alpha = 1$, $\nu = 10^{-5}$, initial width $w = 5$.*

8.2. The fate of peakons under Burgers-$\alpha\beta$ evolution. Figures 25 and 26 show the effects on the peakon initial value problem for the Burgers-$\alpha\beta$ evolution of varying α and b with $(3-b)\beta = 1$ at constant viscosity. We shall consider the following cases with $(3-b)\beta = 1$:

$$b = 0,\ \beta = 1/3,\ \nu = 0.005,\ \alpha = 1/4, 1, 4, \text{ and}$$
$$b = 1,\ \beta = 1/2,\ \nu = 0.005,\ \alpha = 1/4, 1, 4.$$

Remark 8.3 (lowering β has little effect on the ramp and cliff). Lowering β to follow $(3 - b)\beta = 1$ instead of keeping $\beta = 1$ has little effect on the development of the ramp and cliff solution for $b = 0$ and $b = 1$. Lowering β for these cases only makes the activity slightly less

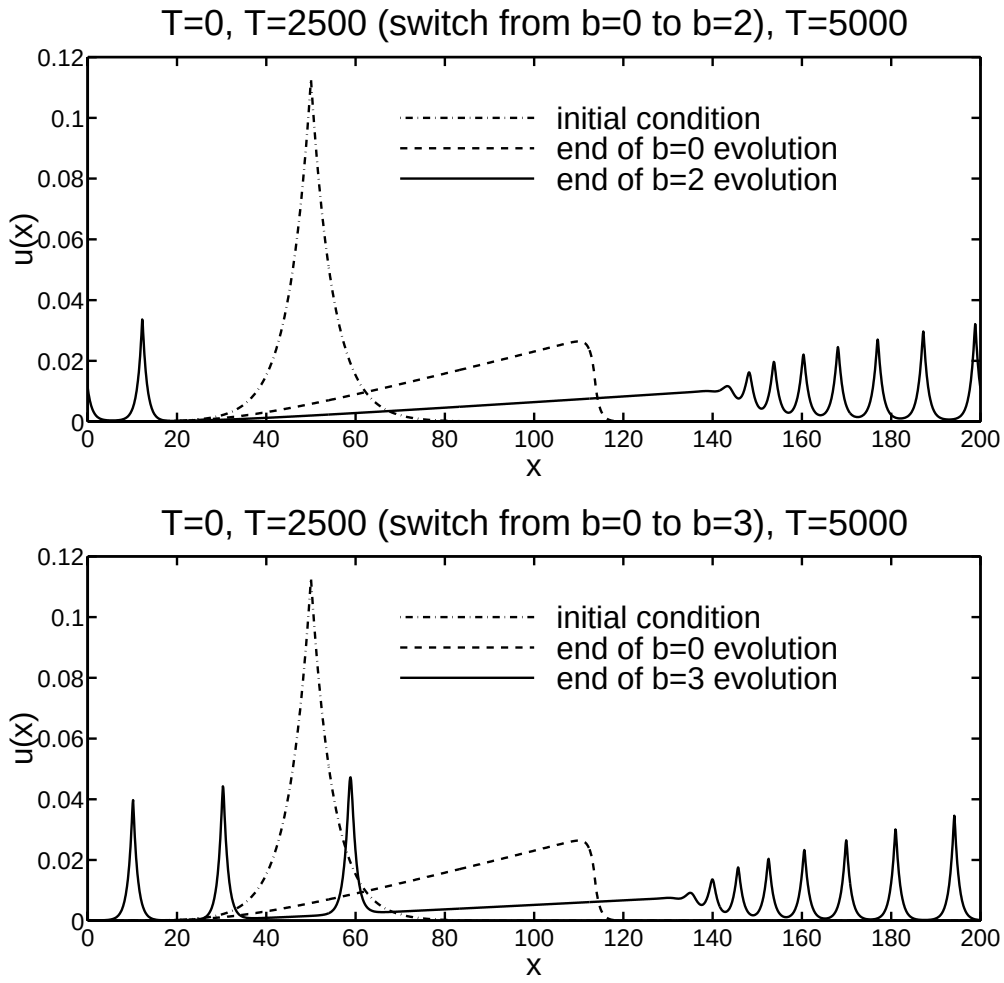


Figure 18. *Exchange of stability between ramps and peakons for $b = 0, 2, 3$, when width $> \alpha$: profiles. Viscous b-family, $b = 0 \to 2, 3$, $\alpha = 1$, $\nu = 10^{-5}$, initial width $w = 5$.*

lively at the front for ($b = 0$, $\beta = 1/3$) and ($b = 1$, $\beta = 1/2$) than for the corresponding cases of $b = 0$ and $b = 1$ with $\beta = 1$ in Figures 13 and 14. This lessened activity at lower β can only be discerned in the solution for the largest value $\alpha = 4$. The remaining case ($b = 2$, $\beta = 1$) recovers the viscous b-equation (7.1) for $b = 2$ in Figure 15, in which the larger b produces much livelier steepening and hence more activity at the front of the rightward moving pulses.

9. Numerical results for peakon scattering and initial value problems. We shall begin by summarizing the results in the figures given earlier, and then we shall describe the numerical methods used in producing them and discuss some of the ways we verified and validated the results.

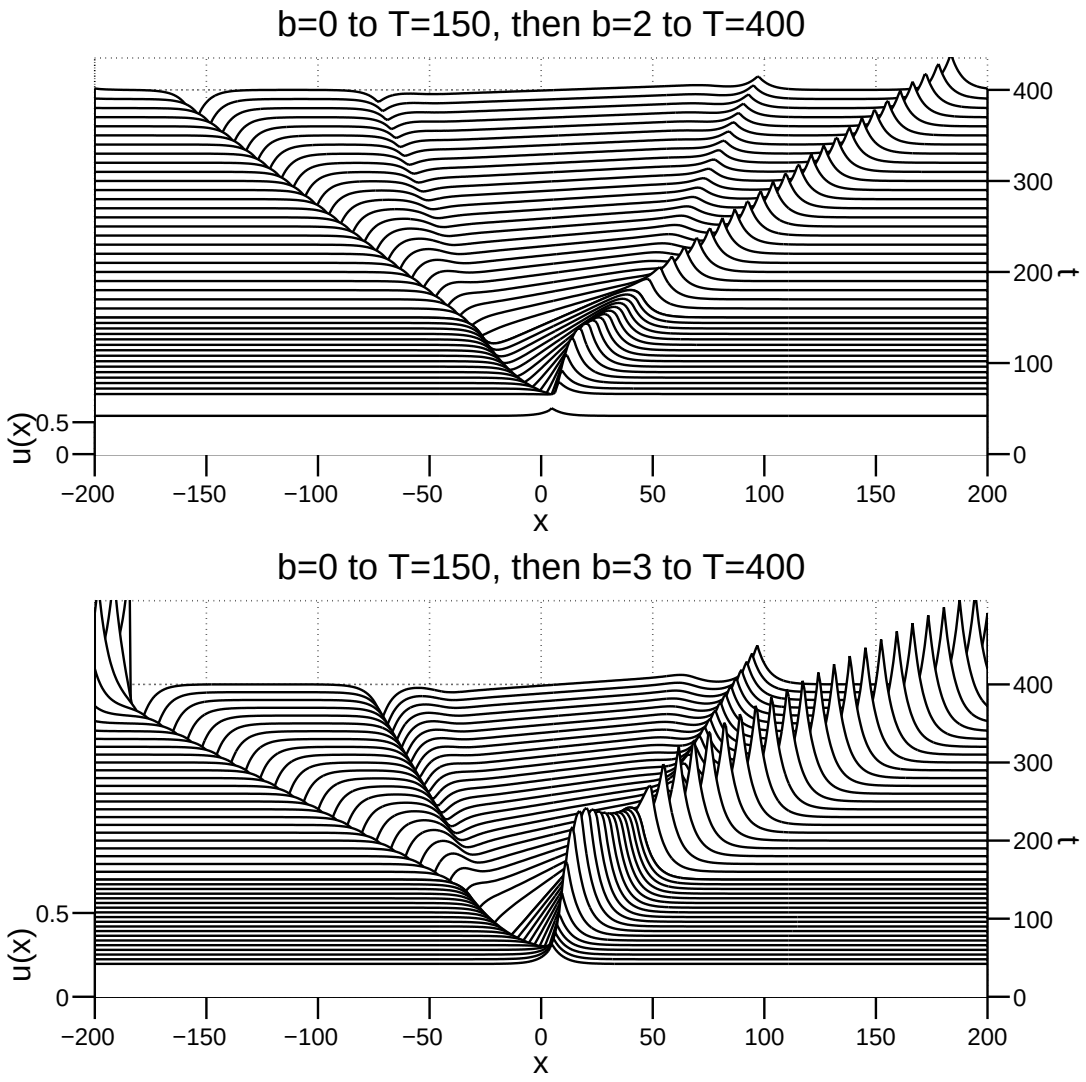


Figure 19. *Exchange of stability between ramps and peakons for $b = 0, 2, 3$, when width $= \alpha$. Viscous b-family, $b = 0 \to 2, 3$, $\alpha = 5$, $\nu = 10^{-5}$, initial width $w = 5$.*

9.1. Peakon initial value problems.

9.1.1. Inviscid b-family of equations.

Ramps and cliffs for $b = 0$. Figure 1 shows the formation of a ramp and cliff pattern for $b = 0$, $\alpha = 1$, and a set of Gaussian initial conditions of increasing width $w = 2.5, 5, 10$.

Peakons for $b = 2, 3$. Figures 2 and 3 show the formation of peakons for $b = 2$, $b = 3$, $\alpha = 1$, and a set of Gaussian initial conditions of increasing width $w = 2.5, 5, 10$.

Ramps and cliffs for $b = -1/2$. Figure 4 shows the formation of a ramp and cliff pattern for $b = -1/2$, $\alpha = 1$, and a set of Gaussian initial conditions of increasing width $w = 10, 15, 20$.

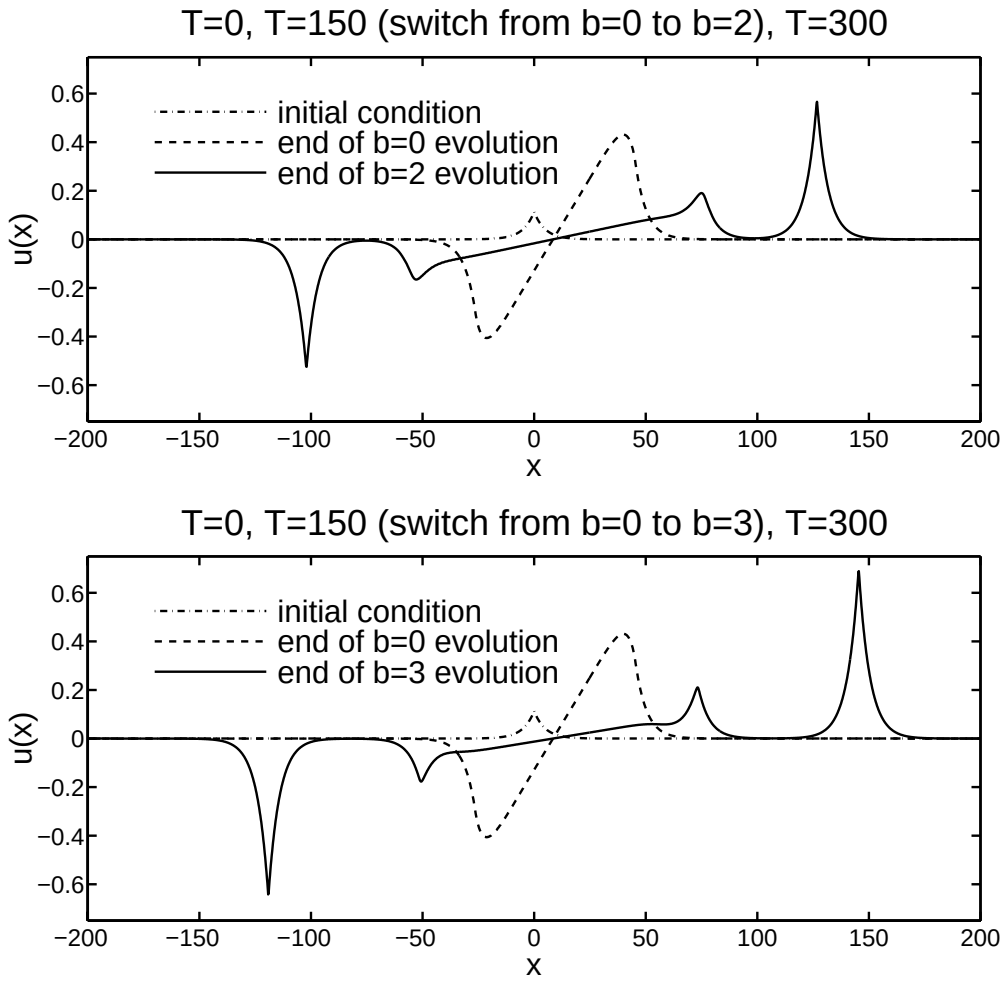


Figure 20. *Exchange of stability between ramps and peakons for $b = 0, 2, 3$, when width $= \alpha$: profiles. Viscous b-family, $b = 0 \to 2, 3$, $\alpha = 5$, $\nu = 10^{-5}$, initial width $w = 5$.*

Stationary solutions for $b \leq -1$. Figure 5 shows an essentially stationary solution with a slight rightward drift and leaning slightly to the right due to nonlinear curvature terms with higher-order derivatives in (4.20) for $\alpha = 1$ and a set of Gaussian initial conditions of increasing width $w = 10, 15, 20$. For the same α and the same set of initial conditions, Figures 6 and 7 show the emergence of leftons.

Figure 8 shows the leftons at time $T = 2500$ for the $b = -2$ case versus the analytical $u(x) \simeq \text{sech}^2(x/(2\alpha))$ from (4.23), and for the $b = -3$ case versus the analytical $u(x) \simeq \text{sech}(x/\alpha)$.

Peakons of width α for $b = 2, 3$. Figures 9 and 10 show 2-peakon and 3-peakon interactions for $b = 2$ and $b = 3$ beginning with initial peakons of width $w = \alpha = 5$.

Peakons of width $> \alpha$ for $b = 2, 3$. Figure 11 shows the emergence of peakons of width $\alpha = 1$ when we begin with peakons of width $w = 5$ greater than α for $b = 2$ and $b = 3$.

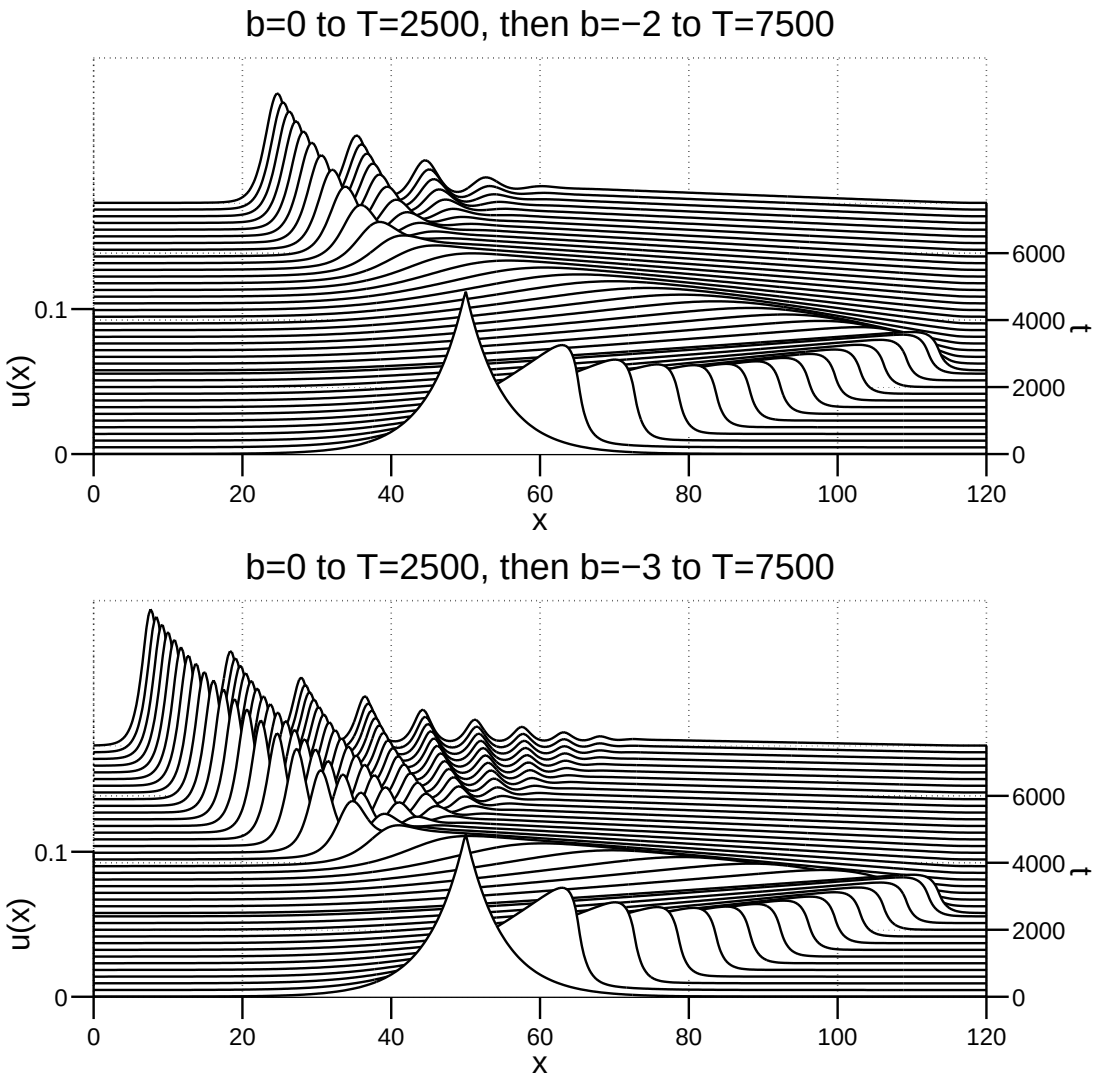


Figure 21. *Exchange of stability between ramps and leftons for $b = 0, -2, -3$, when width $> \alpha$. Viscous b-family, $b = 0 \to -2, -3$, $\alpha = 1$, $\nu = 10^{-5}$, initial width $w = 5$.*

Peakon-antipeakon collisions for $b = 1, 2, 3$. Figure 12 shows the dynamics of a peakon-antipeakon collision for $b = 1$, $b = 2$, and $b = 3$ for $\alpha = 1$ at four successive times.

9.1.2. Viscous b-family of equations.

Effect of α for $b = 0, 1, 2, 3$. Figures 13–16 show the evolution of an initial peakon of width $w = 5$ as a function of increasing $\alpha = 1/4, 1, 4$ at fixed viscosity $\nu = 0.005$ for $b = 0$, $b = 1$, $b = 2$, and $b = 3$.

Exchange of stability between ramps and peakons. Figures 17 and 18 show the exchange of stability between ramps and peakons suggested in the previous four figures, with $\alpha = 1$ and an initial peakon of width $w = 5$, but this time with a very small viscosity $\nu = 10^{-5}$ so that

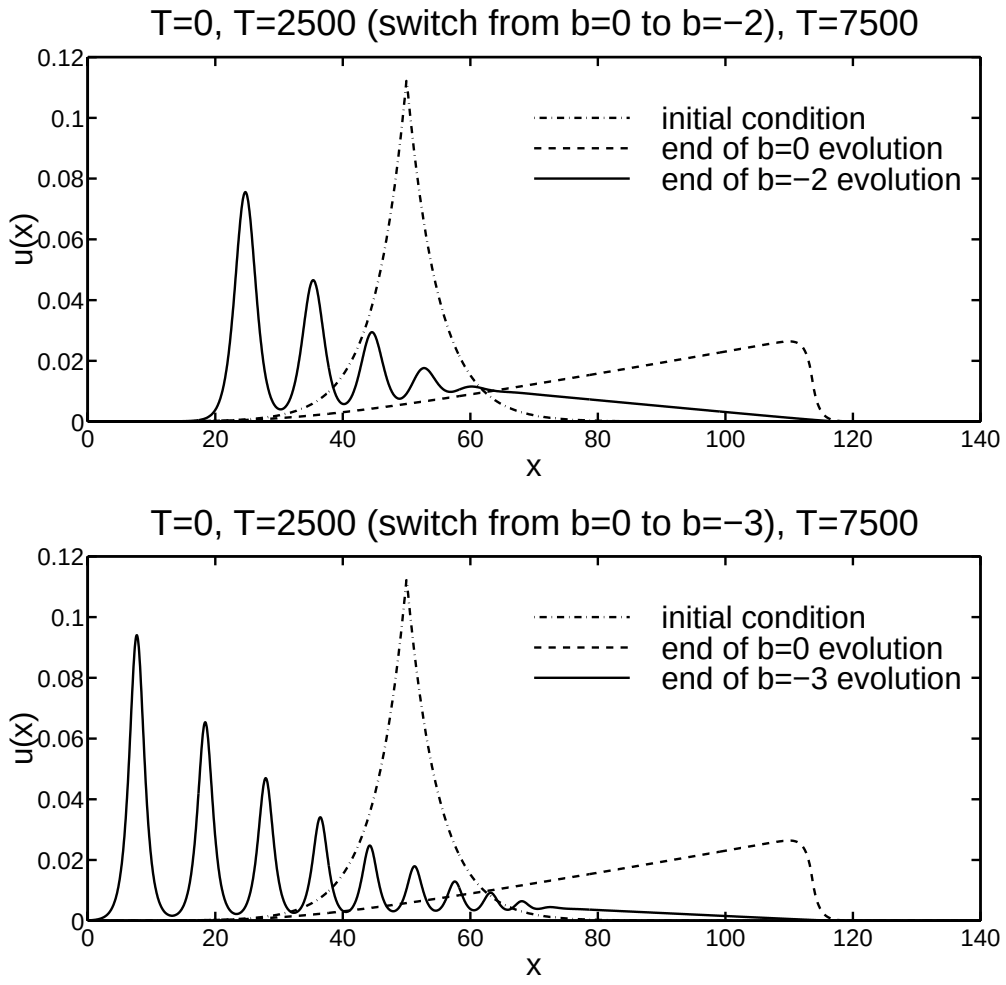


Figure 22. *Exchange of stability between ramps and leftons for $b = 0, -2, -3$, when width $> \alpha$: profiles. Viscous b-family, $b = 0 \to -2, -3$, $\alpha = 1$, $\nu = 10^{-5}$, initial width $w = 5$.*

the peakons, when stable, do not noticeably decay. The exchange of stability occurs when we switch from $b = 0$ to $b = 2$ or $b = 3$. Figures 19 and 20 again show the exchange of stability, this time using $\alpha = 5$ so that the initial peakon has width α.

Exchange of stability between ramps and leftons. Figures 21 and 22 show the exchange of stability as in Figures 17 and 18, but we switch to $b = -2$ or $b = -3$ instead and see the emergence of stable leftons from the ramp.

Effect of viscosity for $b = 2, 3$. Figures 23 and 24 show the effect of increasing viscosity $\nu = 0.01, 0.1, 1$ on the evolution of an initial peakon of width $w = 5$, with $\alpha = 1$ and $b = 2$ or $b = 3$.

9.1.3. Burgers-$\alpha\beta$ equation.

Effect of α when $(3-b)\beta = 1$. Figures 25 and 26 show the effect of increasing $\alpha = 1/4, 1, 4$ on the evolution of an initial peakon of width $w = 5$ for fixed $\nu = 0.005$ and two sets of values

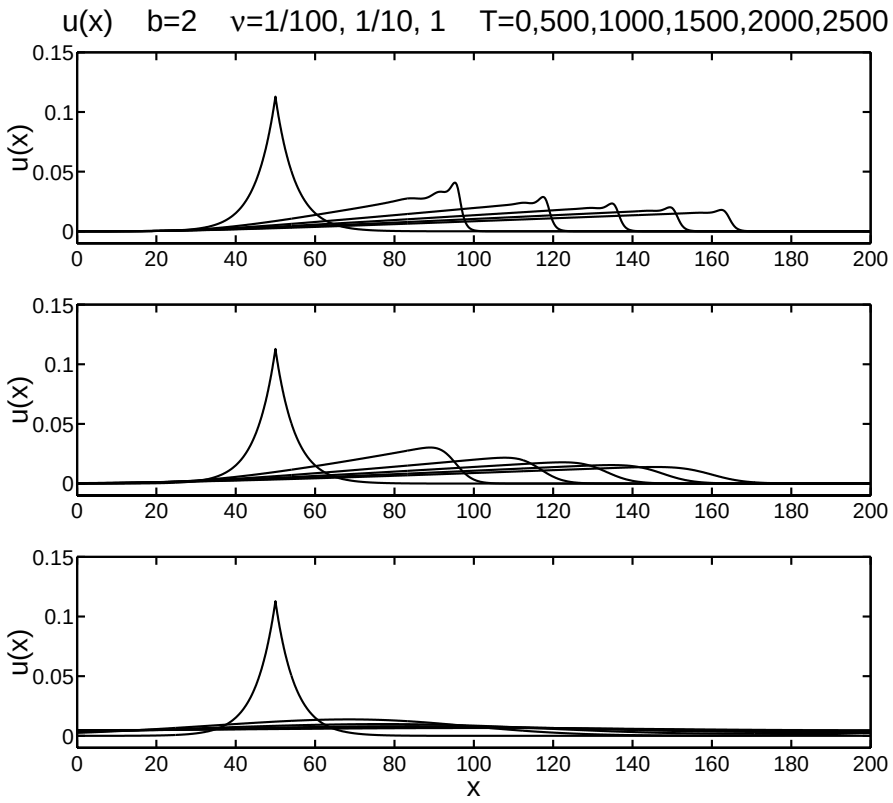


Figure 23. *Effect of increasing viscosity for $b = 2$. Viscous b-family, $b = 2$, $\alpha = 1$, $\nu = 1/100, 1/10, 1$, initial width $w = 5$.*

for b and β for which $(3 - b)\beta = 1$: $b = 0$, $\beta = 1/3$ for the first figure, and $b = 1$, $\beta = 1/2$ for the second figure.

9.2. Description of our numerical methods. For our numerical runs we advanced (6.2), (7.5), and (7.6) with an explicit, variable timestep, fourth/fifth-order Runge–Kutta–Fehlberg (RKF45) predictor/corrector. We selected the timestep for numerical stability by trial and error, while our code selected the timestep for numerical accuracy (not to exceed the timestep for numerical stability) according to the well-known formula from numerical analysis,

$$h_i = \gamma h_{i-1} \left(\frac{\epsilon |h_{i-1}|}{||\bar{u}_i - \hat{u}_i||} \right)^{1/p} . \tag{9.1}$$

This is used in the following way. At step i of the calculation, we know the predicted solution $\bar{u}_i$, the corrected solution $\hat{u}_i$, and the previous timestep h_{i-1}. The predictor's order of accuracy is $p = 4$, while the corrector's order of accuracy is $p+1$. A new timestep h_i is chosen from (9.1) based on the old timestep h_{i-1} and the norm of the difference between the current predicted and corrected solutions. We used a very strict relative error tolerance per timestep, $\epsilon = 10^{-8}$, a safety factor $\gamma = 0.9$, and an L_2 norm $|| \cdot ||_2$.

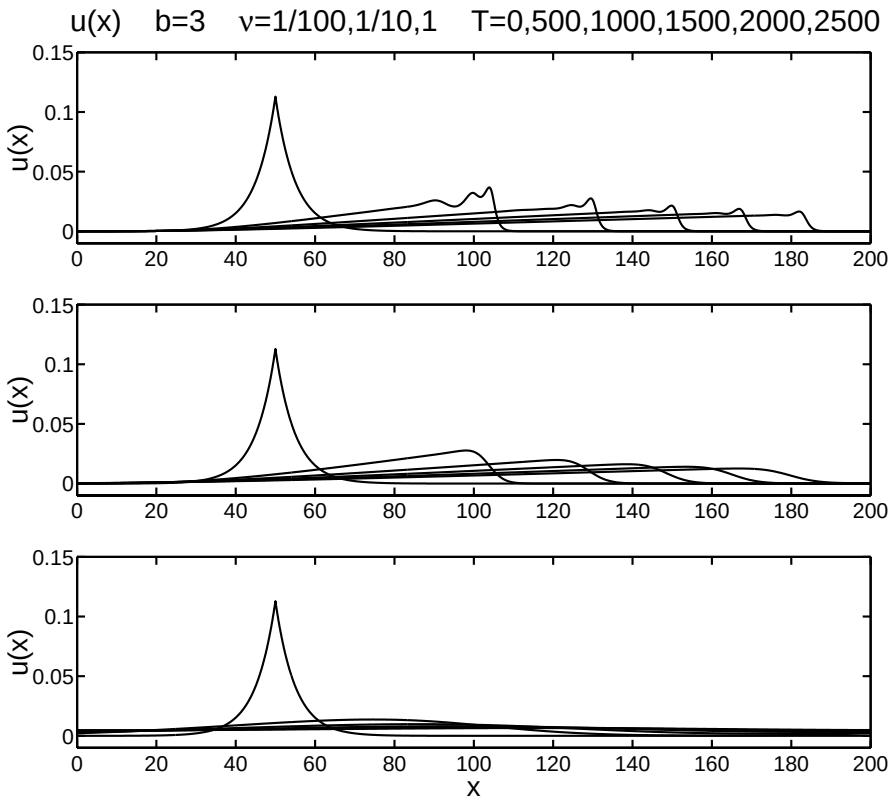


Figure 24. *Effect of increasing viscosity for $b = 3$. Viscous b-family, $b = 3$, $\alpha = 1$, $\nu = 1/100, 1/10, 1$, initial width $w = 5$.*

We computed spatial derivatives using fourth-order finite differences, generally at resolutions of $2^{13} = 8192$ or $2^{14} = 16384$ zones. To invert the Helmholtz operator in transforming between $m(x,t)$ and $u(x,t)$, we convolved $m(x,t)$ with the Green function in Fourier space. When the numerical approximation of the nonlinear terms had aliasing errors in the high wavenumbers, we applied the high pass filtered artificial viscosity

$$\nu(k) = \begin{cases} 0 & \text{if } \ 0 \le k \le \frac{N}{3}, \\ \frac{3\delta}{N}\left(k - \frac{N}{3}\right) & \text{if } \ \frac{N}{3} < k < \frac{2N}{3}, \\ \delta & \text{if } \ \frac{2N}{3} \le k \le N, \end{cases} \tag{9.2}$$

where $\delta = 0.01$ for the present simulations. N is one-half the number of zones, because for each zone we have both a Fourier sine coefficient and a Fourier cosine coefficient.

The quality of the numerical convergence may be checked analytically in the case of rear-end 2-pulson collisions, for which (5.18) in Corollary 5.3 yields

$$g(q_{min}) = g(q)\Big|_{p=0} = 1 - \left(\frac{4c_1c_2}{(c_1+c_2)^2}\right)^{1/(b-1)}. \tag{9.3}$$

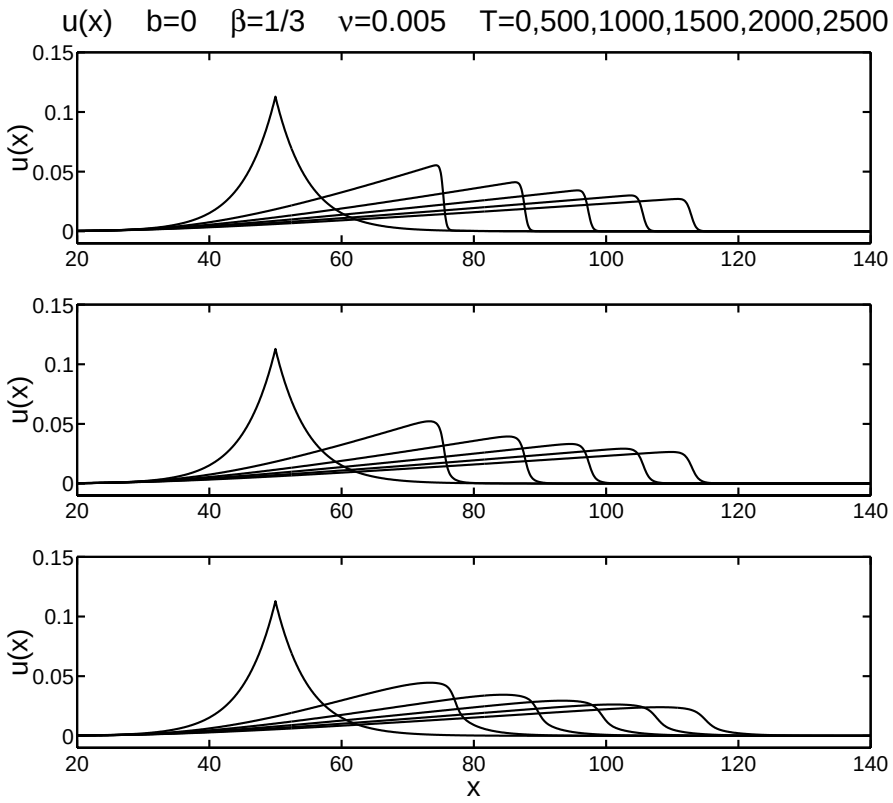


Figure 25. *Effect of increasing α when $(3-b)\beta = 1$, for $b = 0$ and $\beta = 1/3$. Burgers-$\alpha\beta$, $b = 0$, $\alpha = 1/4, 1, 4$, $\beta = 1/3$, $\nu = 0.005$, initial width $w = 5$.*

For peakons with $b = 2$ and $g(x) = e^{-|x|/\alpha}$, this formula gives the minimum separation,

$$q_{min} = -2\alpha \ln\left(\frac{c_1 - c_2}{c_1 + c_2}\right) > 0\,. \tag{9.4}$$

When $c_1 = 1$, $c_2 = 1/2$, and $\alpha = 5$, as in Figure 9, this formula implies $q_{min} = 10 \ln 3 = 10.9861$. Our numerical results with the resolution of 2^{14} zones yield $q_{min} = 11.0049$. The very small discrepancy, less than 0.2%, occurs largely because our numerical measurement of q_{min} is obtained by examining the peakon positions at each internal timestep in the code, while the code's time discretization effectively means we are unlikely to land exactly on the time at which the minimum separation occurs. The code's true accuracy is thus better than the above measure indicates, because the intermediate steps involved in advancing the solution from one discrete time to the next with an RKF45 method cancel the higher-order discretization errors.

Likewise, for peakons with $b = 3$ and $g(x) = e^{-|x|/\alpha}$, formula (9.3) gives the minimum separation,

$$q_{min} = -\alpha \ln\left(1 - \frac{\sqrt{c_1 c_2}}{(c_1 + c_2)/2}\right) > 0\,. \tag{9.5}$$

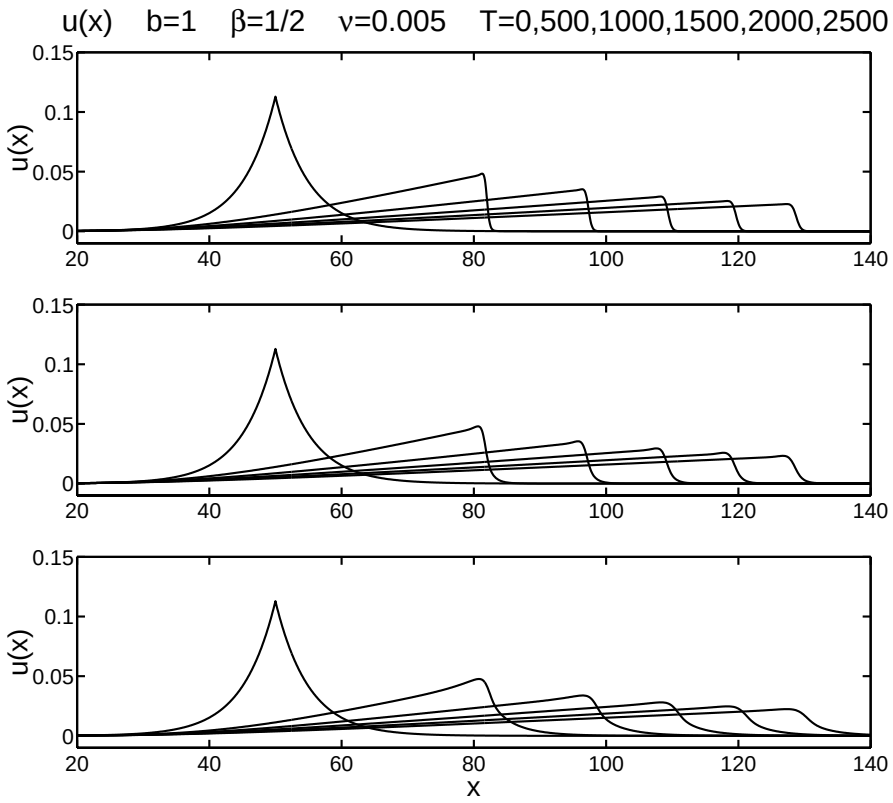


Figure 26. *Effect of increasing α when $(3-b)\beta = 1$, for $b = 1$ and $\beta = 1/2$. Burgers-$\alpha\beta$, $b = 1$, $\alpha = 1/4, 1, 4$, $\beta = 1/2$, $\nu = 0.005$, initial width $w = 5$.*

When $c_1 = 1$, $c_2 = 1/2$, and $\alpha = 5$, as in Figure 10, this formula implies $q_{min} = 5\ln(3/(3-\sqrt{8})) = 14.3068$. This time our numerical results yield $q_{min} = 14.2924$, a discrepancy of only 0.1%.

Of course, the two-body collision is rather simple compared to the plethora of other multi-wave dynamics that occurs in this problem. For this reason, we also checked the convergence of our numerical algorithms by verifying that the relative phases of the peakons in the various figures remained invariant under grid refinement. Moreover, the integrity of the waveforms in our figures attests to the convergence of the numerical algorithm—after scores of collisions, the waveforms given by the Green function for each case are still extremely well preserved. The preservation of these soliton waveforms after so many collisions would not have occurred unless the numerics had converged well.

10. Conclusions.

Equation (1.1) introduced a new family of reversible, parity invariant, evolutionary 1+1 PDEs describing motion by convection and stretching,

$$m_t + \underbrace{u m_x}_{\text{convection}} + \underbrace{b\, u_x m}_{\text{stretching}} = 0\,, \quad \text{with} \quad u = g * m\,. \tag{10.1}$$

We analyzed the transformation properties and conservation laws of this family of equations, which led us to choose g to be an even function. Then we classified its traveling waves and numerically identified the bifurcations of its traveling wave solutions as a function of the balance parameter b. For some choices of the convolution kernel $g(x)$ we studied its particle-like solutions and their interactions when $b > 1$. These were obtained by superposing N traveling wave solutions $u(x,t) = cg(x-ct)$ as

$$u(x,t) = \sum_{i=1}^{N} p_i(t) g(x - q_i(t)) \quad \text{and} \quad m(x,t) = \sum_{i=1}^{N} p_i(t)\delta(x - q_i(t)) \tag{10.2}$$

for any real constant b and $u = g * m$, in which the function g is even $g(-x) = g(x)$, so that $g'(0) = 0$, and is bounded, so we may set $g(0) = 1$.

Following [22], we call these solutions "pulsons." We have shown that for any $b > 1$, once they are initialized on their invariant manifold (which may be finite-dimensional), the pulsons undergo particle-like dynamics in terms of the moduli variables $p_i(t)$ and $q_i(t)$, with $i = 1, \ldots, N$. The pulson dynamics we studied for $b > 1$ in this framework on a finite-dimensional invariant manifold displayed all of the classical soliton interaction behavior for pulsons found in [22] for the case $b = 2$. This behavior included pairwise elastic scattering of pulsons, dominance of the initial value problem by confined pulses, and asymptotic sorting according to height—all without requiring complete integrability. Thus the "emergent pattern" for $b > 1$ in the nonlinear evolution governed by the convection equation (1.1) was the rightward moving pulson train, ordered by height. The moduli variables $p_i(t)$ and $q_i(t)$ are collective coordinates on an invariant manifold for the PDE motion governed by (1.1). Once initialized for $b > 1$, these collective degrees of freedom persist and emerge as a train of stable pulses, arranged in order of their heights, that then undergo particle-like collisions.

In contrast, the emergent pattern in the Burgers parameter region $0 \le b < 1$ is the classic ramp and cliff structure as in Figure 13. That the behavior should depend on the value of b is clear from the velocity form of (1.1) written in (7.2),

$$\begin{aligned} u_t + (b+1)uu_x - \nu u_{xx} &= \alpha^2(u_{xxt} + uu_{xxx} + bu_x u_{xx} - \nu u_{xxxx}) \\ &= \alpha^2 \partial_x \left(u_{xt} + uu_{xx} - \nu u_{xxx} + \frac{b-1}{2} u_x^2 \right) \\ &= \alpha^2 \partial_x^2 \left(u_t + uu_x - \nu u_{xx} + \frac{b-3}{2} u_x^2 \right). \end{aligned} \tag{10.3}$$

Thus nonlinear terms in this equation change sign at four integer values of the parameter b. Nonlinear α^2-terms change sign when $b = 0, 1, 3$. Also, the nonlinear steepening term increases with b as $(b+1)uu_x$. So this term changes sign when $b = -1$. In the parameter regime $b > -1$ (resp., $b < -1$), the solutions of (1.1) move rightward (resp., leftward), provided the terms on the right-hand side of (10.3) are sufficiently small.

Three regions of b. We found that the solution behavior for (1.1) changes its character near the boundaries of the following three regions in the balance parameter b.

(B1) In the stable pulson region $b > 1$, the steepening lemma for peakons proven for $1 < b \le 3$ in Proposition 6.1 allows inflection points with negative slopes to escape verticality

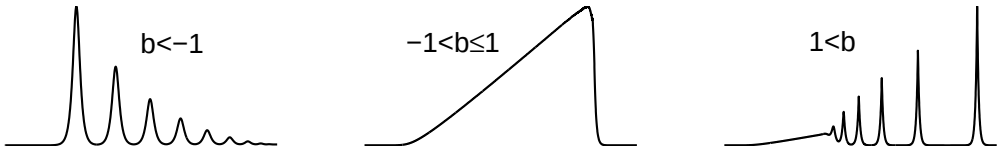


Figure 27. *Leftons* ($b < -1$)*, ramps and cliffs* ($-1 < b \leq 1$)*, and peakons* ($b > 1$) *are solutions of the b-family of equations,* $m_t + um_x + bu_x m = 0$*, with* $m = u - \alpha^2 u_{xx}$.

by producing a jump in spatial derivative at the peak of a traveling wave that eliminates the inflection points altogether. Pulson behavior dominates this region, although ramps of positive slope are also seen to coexist with the pulsons. When $b \leq 1$, we found the solution behavior of the convection equation (1.1) changed its character and excluded the pulsons entirely.

(B2) In the Burgers region $0 \leq b \leq 1$, the $L^{1/b}$ norm of the variable m is controlled[6] and the solution behavior is dominated by ramps and cliffs, as for the usual Burgers equation. Similar ramp and cliff solution properties hold for the region $-1 \leq b \leq 0$, for which the $L^{1/b}$ norm of the variable $1/|m|$ is controlled. At the boundary of the latter region, for $b = -1$, the convection equation (1.1) admits stationary plane waves as exact nonlinear solutions.

(B3) In the steady pulse region $b < -1$, pulse trains form that move leftward from a positive velocity initial condition (instead of moving rightward, as for $b > -1$). These pulse trains seem to approach a steady state.

Figure 27 illustrates this solution behavior in the region (B3) at left, (B2) in the center, and (B1) at right. For illustrative purposes, the curves are drawn at different scales; recall that the evolution of (1.1) preserves the area under the curve.

Effects of viscosity. Almost any numerical investigation will introduce some viscosity or other dissipation. Consequently, we studied the fate of the peakons when viscosity was added to the b-family in (7.1). Viscous solutions of (7.1) for the peakon case $g(x) = e^{-|x|/\alpha}$ with $\alpha = 1$ were studied in each of the three solution regions (B1)–(B3). In the Burgers region (B2) near $b = 0$, we focused on the shock-capturing properties of the solutions of (1.1), and this family of equations was extended for $\beta \neq 1$ to the Burgers-$\alpha\beta$ equation (7.6),

$$u_t + uu_x - \nu u_{xx} = -\beta\,\tau_x \quad \text{with} \quad (1 - \alpha^2\partial_x^2)\tau = \frac{b}{2}u^2 + \frac{3-b}{2}\alpha^2 u_x^2\,. \tag{10.4}$$

According to Proposition 7.1, the Burgers-$\alpha\beta$ equation (10.4) controls the α-weighted H^1 norm of the velocity for $\alpha^2 \neq 0$, provided $(3-b)\beta = 1$. This analytical property guided our study of this new equation by identifying a class of equations for which a priori estimates guarantee continuity of the solution $u(x,t)$. The shock-capturing properties of the Burgers-$\alpha\beta$ equation (10.4) and its $\alpha \to \infty$ limit will be reported in a later paper [28].

11. Outlook: The vector *b*-equation for compressible motion of momentum filaments and surfaces in *n* dimensions. We shall generalize to n dimensions the one-dimensional pulson solutions of (1.1) with defining relation (1.2) studied earlier in section 5.

[6]For $b = 0$, this is a maximum principle for $|m|$.

11.1. n-dimensional vector b-equation. In n-dimensional vector notation, the b-equation (1.1) may be written explicitly in Euclidean coordinates as a PDE for a covector function $\mathbf{m}(\mathbf{x},t):\, R^n \times R^1 \to R^n$. Namely,

$$\frac{\partial}{\partial t}\mathbf{m} + \underbrace{\mathbf{u}\cdot\nabla\mathbf{m}}_{\text{convection}} + \underbrace{\nabla\mathbf{u}^T\cdot\mathbf{m} + (b-1)\,\mathbf{m}(\operatorname{div}\mathbf{u})}_{\text{stretching}} = 0 \tag{11.1}$$

for a defining relation

$$\mathbf{u} = G * \mathbf{m} \equiv \int G(|\mathbf{x}-\mathbf{y}|)\,\mathbf{m}(\mathbf{y})\,d^n y, \tag{11.2}$$

in which we assume the function (filter) $G(|\mathbf{x}|)$ is isotropic. To interpret the stretching terms in the vector b-equation (11.1), we shall recall the interpretation of the scalar b-equation (1.1) as preservation at constant Lagrangian coordinate X of the differential form (3.24),

$$m(x,t)dx^{\otimes b} = m(X,0)dX^{\otimes b}. \tag{11.3}$$

Taking the partial time derivative of this equation at constant Lagrangian coordinate X and using $dx/dt|_X = u$ yield (1.1), as shown earlier in (3.25).

To lift this interpretation of the vector b-equation to n dimensions, we regard the momentum $\mathbf{m}(\mathbf{x},t)$ as the vector coefficient in an invariant one-form density expressed as

$$\mathbf{m}(\mathbf{x},t)\cdot d\mathbf{x}\otimes(dV)^{\otimes(b-1)} = \mathbf{m}(\mathbf{X},0)\cdot d\mathbf{X}\otimes(dV_0)^{\otimes(b-1)}, \tag{11.4}$$

where $dV = d^n x$ and $dV_0 = d^n X$ are, respectively, the Eulerian and Lagrangian volume elements. The symbol $\otimes$ denotes tensor product. Taking the partial time derivative of this equation at constant Lagrangian coordinate $\mathbf{X}$ and using $\frac{d}{dt}|_{\mathbf{X}}\mathbf{x} = \mathbf{u}$ yield

$$\begin{aligned}\frac{d}{dt}\Big|_{\mathbf{X}}\big(\mathbf{m}\cdot d\mathbf{x}\otimes(dV)^{\otimes(b-1)}\big) = {}& \frac{d\mathbf{m}}{dt}\Big|_{\mathbf{X}}\cdot d\mathbf{x}\otimes(dV)^{\otimes(b-1)} \\ & + \mathbf{m}\cdot d\mathbf{u}\otimes(dV)^{\otimes(b-1)} \\ & + (b-1)\,\mathbf{m}\cdot d\mathbf{x}\otimes(\nabla\cdot\mathbf{u})(dV)^{\otimes(b-1)} = 0.\end{aligned} \tag{11.5}$$

We have used the identities, $\frac{d}{dt}|_{\mathbf{X}}d\mathbf{x} = d\mathbf{u} = \mathbf{u}_{,j}dx^j$ and $\frac{d}{dt}|_{\mathbf{X}}dV = (\operatorname{div}\mathbf{u})dV$ for the Lagrangian time derivatives of the line element $d\mathbf{x}$ and the volume element dV. Collecting coefficients in (11.4) yields the vector b-equation (11.1) and explains the sources of its convection and stretching terms.

Equivalently, in terms of the operators div, grad and curl, in two dimensions and three dimensions, the Euclidean coordinate vector b-equation (11.1) becomes

$$\frac{\partial}{\partial t}\mathbf{m} - \mathbf{u}\times\operatorname{curl}\mathbf{m} + \nabla(\mathbf{u}\cdot\mathbf{m}) + (b-1)\,\mathbf{m}(\operatorname{div}\mathbf{u}) = 0. \tag{11.6}$$

Euler–Poincaré (EP) equation. Setting $b = 2$ in the vector b-equation in either of its equivalent forms, (11.1) or (11.6), yields the pressureless Euler–Poincaré (EP) equation, whose abstract form is [26, 27]

$$\frac{\partial}{\partial t}\mathbf{m} + \mathrm{ad}^*_{\mathbf{u}}\mathbf{m} = 0\,, \quad \mathbf{m} = \frac{\delta \ell[\mathbf{u}]}{\delta \mathbf{u}}\,, \tag{11.7}$$

and $\mathrm{ad}^*_{\mathbf{u}}\mathbf{m}$ is defined as in (11.1) for $b = 2$. Thus we have the EP equation for pressureless compressible motion in vector notation:

$$\frac{\partial}{\partial t}\mathbf{m} + \underbrace{\mathbf{u}\cdot\nabla\mathbf{m}}_{\text{convection}} + \underbrace{\nabla\mathbf{u}^T\cdot\mathbf{m} + \mathbf{m}(\operatorname{div}\mathbf{u})}_{\text{EP stretching}} = 0\,, \quad \mathbf{m} = \frac{\delta \ell[\mathbf{u}]}{\delta \mathbf{u}}\,. \tag{11.8}$$

When the Lagrangian $\ell[\mathbf{u}]$ is taken as the kinetic energy, which is also a norm, $\ell[\mathbf{u}] = \frac{1}{2}\|\mathbf{u}\|^2$, then the EP equation (11.8) describes geodesic motion on the diffeomorphism group, with respect to this kinetic energy norm.

For $b = 2$, the scalar product of the EP equation in div-grad-curl form (11.6) with the velocity $\mathbf{u}$ shows that evolution under this equation preserves the kinetic energy,

$$\int \mathbf{u}\cdot\mathbf{m}\, d^n x \equiv \langle \mathbf{u}\,,\, \mathbf{m}\rangle, \tag{11.9}$$

as a constant of the motion, only for this value of b. For the case that velocity is defined in terms of momentum as $\mathbf{u} = G * \mathbf{m}$ in (11.2), we shall assume that the convolution kernel G may be taken as the Green function for a symmetric positive definite operator Q_{op}. We may then write

$$Q_{op}\mathbf{u} = \mathbf{m}, \tag{11.10}$$

where $\mathbf{u}$ satisfies appropriate boundary conditions for the relation $\mathbf{u} = G * \mathbf{m}$ to be satisfied. When Q_{op} is assumed to be a symmetric positive definite operator, the kinetic energy in (11.9) will define a kinetic energy norm $\|\mathbf{u}\|^2$ that may be written as

$$\|\mathbf{u}\|^2 = \int \mathbf{u}\cdot\mathbf{m}\, d^n x = \int \mathbf{u}\cdot Q_{op}\mathbf{u}\, d^n x = \langle \mathbf{u}\,,\, Q_{op}\mathbf{u}\rangle \tag{11.11}$$

for appropriate boundary conditions at spatial infinity. Hence we have the following proposition.

Proposition 11.1 ($b = 2$ vector b-equation for geodesic motion). *When the defining relation* (11.2) *between momentum* $\mathbf{m}$ *and velocity* $\mathbf{u}$ *may be written equivalently in operator form as* (11.10) *with a symmetric positive definite operator* Q_{op} *for which* (11.11) *defines a norm, then for* $b = 2$ *the vector b-equation* (11.1)*, or* (11.6)*, describes geodesic motion on the diffeomorphism group, with respect to this kinetic energy norm.*

EP equation for computational anatomy. Thus the EP equation (11.8) describes geodesic motion when the Lagrangian $\ell[\mathbf{u}]$ is the kinetic energy, which is also a norm. This EP equation has the *same form* as the template matching equation [40], which is used in computational anatomy [39].

11.2. n-dimensional analogues of pulsons for the vector b-equation. The momentum for the one-dimensional pulson solution (5.1) on the real line is supported at points via the Dirac delta measures in its solution ansatz,

$$m(x,t) = \sum_{i=1}^{N} p_i(t)\,\delta\big(x - q_i(t)\big)\,, \quad m \in R^1\,. \tag{11.12}$$

For the vector b-equation (11.1), or (11.6), we shall develop n-dimensional analogues of these one-dimensional pulson solutions by generalizing the solution ansatz to allow measure-valued n-dimensional vector solutions. In these solutions, the momentum $\mathbf{m} \in R^n$ is supported on codimension-k subspaces R^{n-k} moving with the flow, where $k \in [1, n]$ is an integer. In an example in section 11.2.3, we shall consider a two-dimensional vector momentum $\mathbf{m} \in R^2$ in the plane that is supported on one-dimensional curves (momentum fronts). Likewise, in three dimensions, one could consider two-dimensional momentum surfaces (sheets) and one-dimensional momentum filaments, each moving with the local flow velocity. The corresponding vector momentum ansatz that we shall use is a vector version of the pulson solutions (11.12), namely,

$$\mathbf{m}(\mathbf{x},t) = \sum_{i=1}^{N} \int \mathbf{P}_i(s,t)\,\delta\big(\,\mathbf{x} - \mathbf{Q}_i(s,t)\,\big)ds\,, \quad \mathbf{m} \in R^n\,. \tag{11.13}$$

Here, $\mathbf{P}_i, \mathbf{Q}_i \in R^n$ for $i = 1, 2, \ldots, N$. For example, when $n - k = 1$ so that the Lagrangian coordinate $s \in R^1$ is one-dimensional, the delta function in solution (11.13) supports an evolving family of vector-valued curves, which we shall call *momentum filaments*. (For simplicity of notation, we suppress the implied subscript i in the arclength s for each $\mathbf{P}_i$ and $\mathbf{Q}_i$.) The defining relation $\mathbf{u} = G*\mathbf{m}$ in (11.2) implies that the velocity corresponding to the momentum filament ansatz (11.13) is

$$\mathbf{u}(\mathbf{x},t) = G * \mathbf{m} = \sum_{j=1}^{N} \int \mathbf{P}_j(s',t)\,G\big(\,\mathbf{x} - \mathbf{Q}_j(s',t)\,\big)ds'. \tag{11.14}$$

The function (filter) $G(|\mathbf{x}|)$ defined in (11.2) is isotropic, so $G'(0) = 0$. We shall choose $G(0) = 1$, so $\mathbf{P}_j(s',t)$ is the velocity of the jth filament at $\mathbf{x} = \mathbf{Q}_j(s',t)$, and $s' \in R^1$ is a Lagrangian coordinate along the filament. Just as for the one-dimensional case of the pulsons in (5.2) and (5.3), we shall show that substitution of the n-dimensional solution ansatz (11.13) and (11.14) into the vector b-equation (11.1) produces a closed system of equations for the n-dimensional vector parameters $\mathbf{Q}_i(s,t)$ and $\mathbf{P}_i(s,t)$, $i = 1, 2, \ldots, N$. When $b = 2$, these equations will become geodesic Hamiltonian equations for canonically conjugate variables, $\mathbf{Q}_i(s,t)$ and $\mathbf{P}_i(s,t)$.

11.2.1. Momentum filaments in R^n. For definiteness, we shall consider the example of momentum filaments, for which the parameter s is one-dimensional. Such filaments have vector-valued momenta $\mathbf{m} \in R^n$ supported on one-dimensional space curves in R^n, so $s \in R^1$ in (11.13) is the arclength parameter of one of these curves. This solution ansatz is reminiscent of the Biot–Savart law for vortex filaments, although the flow is not incompressible. The

dynamics of momentum surfaces, for $s \in R^k$ with $k < n$, follow essentially the same analysis as that given below for $k = 1$.

Substituting the momentum filament ansatz (11.13) for $s \in R^1$ and its corresponding velocity (11.14) into the vector b-equation (11.1) and then integrating against a smooth test function $\phi(\mathbf{x})$ imply the following equations, in which summation is explicit on $j \in 1, 2, \ldots, N$ and there is no sum on i:

$$
\begin{aligned}
(11.15)\quad \frac{\partial}{\partial t}\mathbf{Q}_i(s,t) &= \sum_{j=1}^{N} \int \mathbf{P}_j(s',t)\, G\big(\mathbf{Q}_i(s,t) - \mathbf{Q}_j(s',t)\big) ds', \\
\frac{\partial}{\partial t}\mathbf{P}_i(s,t) &= -\sum_{j=1}^{N} \int \big(\mathbf{P}_i(s,t)\cdot\mathbf{P}_j(s',t)\big) \frac{\partial}{\partial \mathbf{Q}_i(s,t)} G\big(\mathbf{Q}_i(s,t) - \mathbf{Q}_j(s',t)\big)\, ds' \\
&\quad - (b-2)\mathbf{P}_i(s,t) \sum_{j=1}^{N} \int \mathbf{P}_j(s',t) \cdot \frac{\partial}{\partial \mathbf{Q}_i(s,t)} G\big(\mathbf{Q}_i(s,t) - \mathbf{Q}_j(s',t)\big)\, ds'.
\end{aligned}
$$

The dot product $\mathbf{P}_i \cdot \mathbf{P}_j$ denotes the inner, or scalar, product of the two vectors $\mathbf{P}_i$ and $\mathbf{P}_j$ in R^n. Thus, the solution ansatz (11.13) yields a closed set of integro-partial-differential equations (IPDEs) given by (11.15) for the vector parameters $\mathbf{Q}_i(s,t)$ and $\mathbf{P}_i(s,t)$ with $i = 1, 2, \ldots, N$. Equations (11.15) for the n-dimensional dynamics of the momentum filaments should be compared with the one-dimensional pulson equations (5.2) and (5.3). Note that $b = 2$ is a simplifying special case.

11.2.2. Canonical Hamiltonian dynamics of momentum filaments in R^n for $b = 2$. For $b = 2$, the momentum filament equations (11.15) simplify to canonical Hamiltonian equations,

$$
(11.16)\qquad \frac{\partial}{\partial t}\mathbf{Q}_i(s,t) = \frac{\delta H_N}{\delta \mathbf{P}_i}\,, \qquad \frac{\partial}{\partial t}\mathbf{P}_i(s,t) = -\frac{\delta H_N}{\delta \mathbf{Q}_i}\,.
$$

The corresponding Hamiltonian function $H_N : (R^n \times R^n)^{\otimes N} \to R$ is

$$
(11.17)\qquad H_N = \frac{1}{2} \iint \sum_{i\,,\,j=1}^{N} \big(\mathbf{P}_i(s,t) \cdot \mathbf{P}_j(s',t)\big)\, G\big(\mathbf{Q}_i(s,t) - \mathbf{Q}_j(s',t)\big)\, ds\, ds'.
$$

This Hamiltonian arises by substituting the momentum ansatz (11.13) into the kinetic energy norm (11.11). Thus, for $b = 2$, the evolutionary IPDE system (11.15) represents canonically Hamiltonian motion on the space of curves in R^n. Moreover, this Hamiltonian motion for $b = 2$ is geodesic with respect to the cometric given on these curves in (11.17) by the Green function G. The Hamiltonian $H_N = \frac{1}{2}\|\mathbf{P}\|^2$ in (11.17) for this motion defines the norm $\|\mathbf{P}\|$ in terms of this cometric.

Summary. The momentum filament ansatz (11.14) reduces or *collects* the solution of the evolutionary vector b-equation (11.1) for $b = 2$ in n spatial dimensions into the system (11.15) of $2N$ evolutionary IPDEs in one spatial dimension (arclength along each filament). For $b = 2$, the collective equations describe geodesic flow on the space of vector-valued curves in R^n with respect to the cometric given on these curves in (11.17).

The momentum filament equations (11.15) are not typical. As far as we know, the IPDEs for momentum filaments in (11.15) have never been considered before in the literature, even for the Hamiltonian case $b = 2$. Even the Hamiltonian evolution of a single momentum filament interacting with itself has not appeared in the literature to our knowledge. There is a faint similarity of this system to vortex dynamics for the incompressible Euler equations. However, there are also fundamental differences. The main difference from the Hamiltonian motion of vortex filaments is that the momentum filaments possess *inertia*, while vortex filaments do not. Thus N vortex filaments in an incompressible flow are described by N first-order equations, while N momentum filaments in a pressureless compressible flow are described by the $2N$ first-order equations in (11.15). The main question one would like to answer is, "When do the momentum filament solutions represent the *dominant* emergent pattern in the initial value problem for the vector b-equation (11.1)?" The next subsection reports an example in which this occurs, for momentum filaments in the plane, in a certain approximation for two-dimensional shallow water waves.

11.2.3. Zero-dispersion shallow water waves in two dimensions: Two interesting choices for the operator Q_{op} when $b = 2$. The operator Q_{op} in the momentum relation $\mathbf{m} = Q_{op}\mathbf{u}$ in (11.10) corresponding to $m = u - \alpha^2 u_{xx}$ in the one-dimensional CH (3.7) for zero-dispersion shallow water waves may be defined in two dimensions as *either* of two natural choices,

$$\mathbf{m} = \mathbf{u} - \alpha^2\Delta\mathbf{u} \quad \text{or} \quad \mathbf{m} = \mathbf{u} - \alpha^2\nabla \operatorname{div}\mathbf{u}. \tag{11.18}$$

For the first choice of momentum definition in (11.18), the vector EP equation (11.8) corresponds to the (pressureless) Euler-alpha model, whose Lagrangian $\ell[\mathbf{u}] = \frac{1}{2}\|\mathbf{u}\|^2$ is the conserved H^1 norm[7]

$$\|\mathbf{u}\|^2_{H^1} = \int \mathbf{u}\cdot(1-\alpha^2\Delta)\mathbf{u}\,dx\,dy = \int |\mathbf{u}|^2 + \alpha^2(\operatorname{div}\mathbf{u})^2 + \alpha^2|\operatorname{curl}\mathbf{u}|^2\,dx\,dy\,.$$

The last equality assumes either homogeneous or periodic boundary conditions so that boundary terms may be neglected upon integrating by parts.

For the second natural choice of momentum in (11.18), the conserved kinetic energy norm becomes instead

$$\|\mathbf{u}\|^2_{KS} = \int \mathbf{u}\cdot(1-\alpha^2\nabla \operatorname{div})\mathbf{u}\,dx\,dy = \int |\mathbf{u}|^2 + \alpha^2(\operatorname{div}\mathbf{u})^2\,dx\,dy,$$

and kinetic energy conservation no longer controls $\operatorname{curl}\mathbf{u}$. This is the norm associated with vertically averaged kinetic energy that arises when one approximates the Green–Naghdi equations for shallow water motion by neglecting variations in surface elevation in the potential

[7]When incompressibility ($\operatorname{div}\mathbf{u} = 0$) is imposed as an additional constraint in this Lagrangian via a Lagrange multiplier (the pressure), then the corresponding vector EP equation (11.8) becomes the two-dimensional Lagrangian averaged Euler equation derived in [26] as a generalization of the one-dimensional CH for incompressible flow. The Hamilton principles defined by the kinetic energy norms $\|\mathbf{u}\|^2_{H^1}$ and $\|\mathbf{u}\|^2_{KS}$ have no pressure constraint, so their corresponding EP equations (11.8) allow compressible motion.

energy and in the Lagrange-to-Euler Jacobian.[8] The second term proportional to α^2 approximates (twice) the vertically averaged kinetic energy due to vertical motion. For more details of the latter approximation for the two-dimensional CH shallow water equation, see Kruse, Schreule, and Du [36].

The EP equation (11.8) was numerically integrated in [29] in the form

$$\frac{\partial}{\partial t}\mathbf{m} - \mathbf{u} \times \operatorname{curl} \mathbf{m} + \nabla(\mathbf{u} \cdot \mathbf{m}) + \mathbf{m}(\operatorname{div} \mathbf{u}) = 0 \tag{11.19}$$

for both choices of the momentum-velocity relation in (11.18). This numerical integration was performed by using a difference scheme that preserved the properties of the operators div, grad, and curl (divcurl = 0 and curlgrad = 0). The main discovery in the numerical results of [29] was that the evolution of the geodesic PDE (11.19) was found to be dominated by the emergent dynamics of momentum filaments, arising from confined initial conditions for *either* choice of momentum-velocity relation in (11.18). Thus the momentum filament solutions in both of these cases were stable, and *no other types of solution* were observed in the numerical evolution of (11.19) in the periodic plane. The dynamics of the momentum filaments that emerged was quasi-one-dimensional, with greater variation of the solution in the direction transverse to the filaments than along the tangential direction. Thus the interaction dynamics for the momentum filaments was found to be dominantly in the direction transverse to the filaments. This meant the filament interaction was governed primarily by elastic-scattering dynamics reminiscent of the one-dimensional solutions, as seen in soliton dynamics. In fact, the one-dimensional soliton collision rules were found to provide a good interpretation of the interactions among the momentum filaments. These interactions were found to allow reconnection of the quasi-one-dimensional momentum filaments. For more information and discussion of numerical results, see [29].

Acknowledgments. We are grateful to A. Degasperis, A. N. W. Hone, J. M. Hyman, S. Kurien, C. D. Levermore, R. Lowrie, and E. S. Titi for their thoughtful remarks, careful reading, and attentive discussions that provided enormous help and encouragement during the course of writing this paper. We are also grateful to the referees for their thoughtful comments. One referee pointed out a technical caveat that appeared in the original proofs of both the peakon steepening lemmas in Propositions 6.1 and 7.2 and showed how to overcome the caveat without changing the validity of the results. The caveat and its solution are stated after the formal proofs of these propositions.

[8] In this approximation for two-dimensional shallow water waves, $\operatorname{curl}\mathbf{m} = \operatorname{curl}\mathbf{u}$ and $\operatorname{div}\mathbf{m} = (1-\alpha^2\Delta)\operatorname{div}\mathbf{u}$. Thus setting $\mathbf{u} = \hat{\mathbf{z}} \times \nabla\psi - \nabla\phi$ allows one to solve for the stream function ψ and velocity potential ϕ from the momentum $\mathbf{m}$ via,

$$\hat{\mathbf{z}} \cdot \operatorname{curl}\mathbf{m} = -\Delta\psi \quad \text{and} \quad \operatorname{div}\mathbf{m} = -\Delta(1-\alpha^2\Delta)\phi\,.$$

These two relations allow one to update the potentials ψ and ϕ for the velocity $\mathbf{u}$, given the momentum $\mathbf{m}$ at each time step, provided these potentials satisfy boundary conditions that allow inversion of the Laplacian operator for ψ and the Helmholtz–Laplace operator for ϕ. Whether these boundary conditions are consistent with the diffeomorphism group is an open question.

REFERENCES

[1] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, Cambridge University Press, Cambridge, UK, 1991.

[2] M. Alber, R. Camassa, D. D. Holm, and J. E. Marsden, *The geometry of peaked solitons and billiard solutions of a class of integrable PDEs*, Lett. Math. Phys., 32 (1994), pp. 137–151.

[3] M. Alber, R. Camassa, D. D. Holm, and J. E. Marsden, *On the link between umbilic geodesics and soliton solutions of nonlinear PDE's*, Proc. Roy. Soc. London, 450 (1995), pp. 677–692.

[4] M. Alber, R. Camassa, V. N. Fedorov, D. D. Holm, and J. E. Marsden, *On billiard solutions of nonlinear PDE's*, Phys. Lett. A, 264 (1999), pp. 171–178.

[5] M. Alber, R. Camassa, V. N. Fedorov, D. D. Holm, and J. E. Marsden, *The complex geometry of piecewise solutions of integrable nonlinear PDE's of shallow water and Dym type*, Comm. Math. Phys., 221 (2001), pp. 197–227.

[6] R. Beals, D. H. Sattinger, and J. Szmigielski, *Multipeakons and the classical moment problem*, Adv. Math., 154 (2000), pp. 229–257.

[7] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664; also available online from http://xxx.lanl.gov/abs/patt-sol/9305002.

[8] R. Camassa, D. D. Holm, and J. M. Hyman, *A new integrable shallow water equation*, Adv. Appl. Mech., 31 (1994), pp. 1–33.

[9] S. Chen, C. Foias, D. D. Holm, E. J. Olson, E. S. Titi, and S. Wynne, *The Camassa-Holm equations as a closure model for turbulent channel and pipe flows*, Phys. Rev. Lett., 81 (1998), pp. 5338–5341; also available online from http://xxx.lanl.gov/abs/chao-dyn/9804026.

[10] A. Constantin, *On the scattering problem for the Camassa-Holm equation*, Proc. Roy. Soc. London, 457 (2001), pp. 953–970.

[11] A. Constantin and J. Escher, *Global existence and blow-up for a shallow water equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 303–328.

[12] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.

[13] A. Constantin and W. A. Strauss, *Stability of peakons*, Comm. Pure Appl. Math., 53 (2000), pp. 603–610.

[14] H.-H. Dai and Y. Huo, *Solitary shock waves and other travelling waves in general compressible hyperelastic rods*, Proc. Roy. Soc. London, 456 (2000), pp. 331–363.

[15] A. Degasperis, D. D. Holm, and A. N. W. Hone, *A new integrable equation with peakon solutions*, Theoret. Math. Phys., 133 (2002), pp. 1463–1474; also available online from http://xxx.lanl.gov/abs/nlin.SI/0205023.

[16] A. Degasperis and M. Procesi, *Asymptotic integrability*, in Symmetry and Perturbation Theory, A. Degasperis and G. Gaeta, eds., World Scientific, River Edge, NJ, 1999, pp. 23–37.

[17] H. Dullin, G. Gottwald, and D. D. Holm, *An integrable shallow water equation with linear and nonlinear dispersion*, Phys. Rev. Lett., 87 (2001), pp. 194501–194504.

[18] H. R. Dullin, D. D. Holm, and G. Gottwald, *Camassa-Holm, Korteweg-de Vries-5 and other asymptotically equivalent equations for shallow water waves*, Fluid Dynam. Res., 33 (2003), pp. 73–95.

[19] A. S. Fokas, *On a class of physically important integrable equations*, Phys. D, 87 (1995), pp. 145–150.

[20] A. S. Fokas and B. Fuchssteiner, *Backlund-transformations for hereditary symmetries*, Nonlinear Anal., 5 (1981), pp. 423–432.

[21] A. S. Fokas and Q. M. Liu, *Asymptotic integrability of water waves*, Phys. Rev. Lett., 77 (1996), pp. 2347–2351.

[22] O. Fringer and D. D. Holm, *Integrable vs. nonintegrable geodesic soliton behavior*, Phys. D, 150 (2001), pp. 237–263; also available online from http://xxx.lanl.gov/abs/solv-int/9903007.

[23] B. Fuchssteiner and A. S. Fokas, *Symplectic structures, their Bäcklund transformations and hereditary symmetries*, Phys. D, 4 (1981/82), pp. 47–66.

[24] B. Fuchssteiner, *The Lie algebra structure of nonlinear evolution equations admitting infinite-dimensional abelian symmetry groups*, Progr. Theoret. Phys., 65 (1981), pp. 861–876.

[25] B. Fuchssteiner, *Some tricks from the symmetry-toolbox for nonlinear equations: Generalizations of the Camassa-Holm equation*, Phys. D, 95 (1996), pp. 229–243.

[26] D. D. Holm, J. E. Marsden, and T. S. Ratiu, *The Euler–Poincaré equations and semidirect products with applications to continuum theories*, Adv. Math., 137 (1998), pp. 1–81.
[27] D. D. Holm, J. E. Marsden, and T. S. Ratiu, *Euler–Poincaré models of ideal fluids with nonlinear dispersion*, Phys. Rev. Lett., 80 (1998), pp. 4173–4177.
[28] D. D. Holm, R. B. Lowrie, and M. F. Staley, *Shock-Capturing Properties of the Burgers–$\alpha\beta$ Equation*, manuscript.
[29] D. D. Holm and M. F. Staley, *Momentum Filament Dynamics in a Family of $2+1$ Evolutionary PDEs*, manuscript.
[30] D. D. Holm and E. S. Titi, *PDE Results for Peakon Dynamics*, manuscript.
[31] J. K. Hunter and R. H. Saxton, *Dynamics of director fields*, SIAM J. Appl. Math., 51 (1991), pp. 1498–1521.
[32] R. S. Johnson, *Camassa-Holm, Korteweg-de Vries and related models for water waves*, J. Fluid. Mech., 455 (2002), pp. 63–82.
[33] Y. Kodama, *On integrable systems with higher order corrections*, Phys. Lett. A, 107 (1985), pp. 245–249.
[34] Y. Kodama, *Normal forms for weakly dispersive wave equations*, Phys. Lett. A, 112 (1985), pp. 193–196.
[35] Y. Kodama, *On solitary wave interaction*, Phys. Lett. A, 123 (1987), pp. 276–282.
[36] H. P. Kruse, J. Schreule, and W. Du, *A two-dimensional version of the CH equation*, in Symmetry and Perturbation Theory: SPT 2001, D. Bambusi, G. Gaeta, and M. Cadoni, eds., World Scientific, New York, 2001, pp. 120–127.
[37] Z. Li and N. R. Sibgatullin, *An improved theory of long waves on the water surface*, J. Appl. Math. Mech., 61 (1997), pp. 177–182.
[38] A. V. Mikhailov and V. S. Novikov, *Perturbative symmetry approach*, J. Phys. A, 35 (2002), pp. 4775–4790.
[39] M. I. Miller, A. Trouvé, and L. Younes, *On the metrics and Euler-Lagrange equations of computational anatomy*, Ann. Rev. Biomed. Engrg., 4 (2002), pp. 375–405.
[40] D. Mumford, *Pattern theory and vision*, in Questions Mathématiques En Traitement Du Signal et de L'Image, Institut Henri Poincaré, Paris, 1998, pp. 7–13.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 3, pp. 381–416

Nonsmooth Lagrangian Mechanics and Variational Collision Integrators*

R. C. Fetecau†, J. E. Marsden‡, M. Ortiz§, and M. West¶

Abstract. Variational techniques are used to analyze the problem of rigid-body dynamics with impacts. The theory of smooth Lagrangian mechanics is extended to a nonsmooth context appropriate for collisions, and it is shown in what sense the system is symplectic and satisfies a Noether-style momentum conservation theorem.

Discretizations of this nonsmooth mechanics are developed by using the methodology of variational discrete mechanics. This leads to variational integrators which are symplectic-momentum preserving and are consistent with the jump conditions given in the continuous theory. Specific examples of these methods are tested numerically, and the long-time stable energy behavior typical of variational methods is demonstrated.

Key words. discrete mechanics, variational integrators, collisions

AMS subject classifications. 37M15, 70F35, 58E30

DOI. 10.1137/S1111111102406038

1. Introduction.

In this paper, we investigate nonsmooth Lagrangian mechanics and its discretization by means of variational, numerical, and geometric methods. In particular, we are interested in the problem of rigid-body collisions, for which the velocity, acceleration, and forces are all nonsmooth or even discontinuous.

We shall begin with a survey of some history and literature to put our own work into context. The literature and history is of course quite complex with many points of view, so we focus on selected highlights only.

History and literature: Theory. The problem of collisions has been extensively treated in the literature since the early days of mechanics. More recently, much work has been done on the rigorous mathematical foundation of impact problems, in particular, by generalizing Newton's law to include forces which are measure-valued and hence can include impulses at the point of impact. The contact dynamics is thus governed by a *measure differential inclusion*, a general

*Received by the editors April 23, 2002; accepted for publication (in revised form) by M. Golubitsky May 1, 2003; published electronically August 23, 2003.

http://www.siam.org/journals/siads/2-3/40603.html

†Applied Mathematics 217-50, California Institute of Technology, Pasadena, CA 91125 (van@acm.caltech.edu). The research of this author was partially supported by the California Institute of Technology and the National Science Foundation.

‡Control and Dynamical Systems 107-81, California Institute of Technology, Pasadena, CA 91125 (marsden@cds.caltech.edu). The research of this author was partially supported by the California Institute of Technology and the National Science Foundation.

§Graduate Aeronautical Laboratories 105-50, California Institute of Technology, Pasadena, CA 91125 (ortiz@aero.caltech.edu).

¶Department of Mathematics, University of California Davis, One Shields Avenue, Davis, CA 95616 (mwest@math.ucdavis.edu).

formulation that can directly incorporate impulsive forces and nonsmooth solutions. In this context, a measure differential inclusion has the form

$$\frac{dv}{dt} \in F(t,x), \ \frac{dx}{dt} = g(t,x,v),$$

where $v(t)$ and $x(t)$ denote the velocity and the position, F is a set-valued function, and $v(\cdot)$ is required only to have bounded variation.

The extension of the concept of a differential equation to that of a *differential inclusion* was first considered in [12, 13, 14]. These works provide a deep study of ODEs with a discontinuous right-hand side, but the fact that solutions are required to be continuous in the phase space makes the theory inapplicable to collisions. Measure differential inclusions can be found in different contexts in [51, 52], and the use of this concept in rigid-body dynamics was further developed in [38, 39], where the (unilateral) contact between rigid bodies received a formulation (called by the author a *sweeping process*) that combines differential inclusions with convex analysis. Since then, an extensive literature has been devoted to the theoretical and numerical study of nonsmooth dynamics within the mathematical framework of measure differential inclusions.

Substantial progress has been made in the last two decades on the existence and uniqueness theory for the generalized solutions of rigid-body dynamics. The first rigorous results in this area were produced in [33] for the case of an inelastic collision with a single convex constraint. Further results generalized the existence theory to more general contacts in [42, 43], to more general (nonconvex, but of class C^1) constraints in [34], or even to a less regular constraint for an arbitrary frictionless impact in [32]. The recent works [55, 56] consider the impact dynamics with friction and give a rigorous mathematical solution to the famous problem of Painlevé.

In the same elegant framework of differential inclusions, but oriented toward the control and stability of nonsmooth dynamical systems, we mention the works of Brogliato [6, 7] and Brogliato, Niculescu, and Orhant [5].

History and Literature: Computations. The measure differential inclusion has also been proved to be an excellent mathematical foundation for the study of numerical methods for discontinuous ODEs. It is not within our scope to give a complete account of these methods, but we refer the reader to the excellent overviews of numerical methods for differential inclusions in [11] and [30]. In particular, such numerical approaches have been pursued to develop efficient numerical methods for rigid-body dynamics in the sweeping process formalism in [38, 40, 44, 56].

Various other numerical methods for rigid-body systems have been studied extensively in the engineering and mathematics literature. We refer to the excellent book [47] for a comprehensive account of some of these methods. We particularly note the approach that reduces the contact to a *complementarity problem*, a concept frequently used in constrained optimization, to decide at each step which constraints are active.

However, most existing practical codes are based on smoothing techniques, a class of methods which use a penalty formulation to regularize the problem. This approach relies on the definition of a proper gap function as a means to detect and penalize the interpenetration; see, for example, [54, 8, 60, 57]. An obvious weakness of the penalty methods is that they

cannot handle collisions of irregularly shaped bodies (bodies with corners), where neither normals nor gap functions can be defined. An elegant solution to this problem is offered by the *nonsmooth analysis* approach from [25], where new robust contact algorithms are derived using the powerful tools of nonsmooth calculus (see [10]).

An important issue in contact dynamics is how to formulate physically correct friction models, and an extensive body of literature has addressed this problem. Frictional effects are generally accounted for by introducing a friction law (Coulomb's law is an example) which relates the sliding velocity to the contact forces. An alternative approach uses the *maximum dissipation principle*, where the friction force c_f is required to maximize the rate of energy dissipation $-c_f^T v_{rel}$, where v_{rel} is the relative velocity at the contact, out of all possible friction forces allowed by a given contact force c_n. However, the correct modeling of friction still has many open questions which generate controversy in various engineering and mathematical communities. All the various numerical methods for contact that we mentioned above have introduced friction in the dynamics, and we refer to [2] and [39] for measure differential inclusion methods, [23, 28, 45, 48, 58, 59, 46, 3] for the complementarity and gap function formulation, and [41] for the nonsmooth analysis approach.

Variational methodology. Our approach, in contrast, is based on a variational methodology that goes back to [61] which allows the direct handling of the nonsmooth nature of contact problems. We also use a variational approach to develop numerical integrators for nonsmooth rigid-body dynamics. The procedure is based on a discrete Lagrangian principle and automatically generates a symplectic-momentum preserving integrator. Near impact, we introduce a collision point and a collision time and solve for them using a variational method.

Variational integrators are known to have remarkable near-energy preserving properties, and we will recover this excellent energy behavior even in the nonsmooth case. We want to emphasize that the variational point of view is not confined to conservative systems but also applies to forced and dissipative systems, as demonstrated in [26]. In future works, we will investigate how forces and friction can be added to our collision algorithm and also how to incorporate other dissipative effects (inelastic collisions).

Issues addressed in this paper. We first show that, by introducing a space of configuration trajectories extended by introducing curve parameterizations as variables, the traditional approach to the calculus of variations can be applied. Moreover, the formulation in the extended setting enables us to address and give a rigorous interpretation of the sense in which the flow map of a mechanical system subjected to dissipationless impact dynamics is symplectic. The nonautonomous variational approach also leads to Weierstrass–Erdmann-type conditions for impact, in terms of energy and momentum conservation at the contact point (see [19] and [61]).

On the discrete side, the variational formalism leads to symplectic-momentum preserving integrators that are consistent with the jump conditions and the continuous theory.

The theory of geometric integration (see, for example, [50] and [18]) is typically concerned with smooth Hamiltonian or Lagrangian systems posed on smooth spaces. These techniques do not immediately apply to nonsmooth settings, and naive applications can result in extremely bad behavior, as demonstrated in [56].

Our methods answer an important question posed by [56]: *How can geometric integrators be formulated and implemented for collision problems?* In fact, the algorithms developed in

the present paper show how a symplectic method can be constructed for nonsmooth systems so that it retains the good behavior normally associated with symplectic methods.

Some work on extensions of geometric integration to collision problems exists. In particular, [4] has constructed time-symmetric methods for contact, and [21] (see also [20]) has developed methods for impacts of hard spheres. To date there have been no symplectic methods for collisions presented, in part due to difficulties with understanding symplecticity in a nonsmooth setting. However, the variational formulation of continuous time nonsmooth systems that we develop here is a key which allows us to understand the geometric structure of the problem, both before and after discretization. Our methods can be considered extensions of the large body of work on geometric integration of ODEs (see, for example, [16, 17, 18, 22, 29, 37]).

We caution that the algorithm presented in this paper is implicit and very expensive and thus may not be appropriate for use with large collision systems. Nonetheless, it is the first geometric integrator for collision problems and thus serves as a basis for the construction of more efficient methods in the future. In fact, the methods of this paper have already led to the development of more computationally feasible collision integrators (see [9]).

We also discuss how nonsmooth analysis techniques [25] can be incorporated into the variational procedure such that the integrator can cope with nonsmooth contact geometries (such as corner-to-corner collisions). As we mentioned before, this is the case which most existing algorithms cannot handle (the standard penalty methods simply fail since no proper gap function can be defined for such geometries).

Organization of the paper. In section 2, we first consider the time-continuous situation and extend the conventional setting of geometric Lagrangian mechanics (see, for example, [36]) to include nonsmooth but still continuous trajectories. This allows us to recover the standard jump conditions at impact and to prove that the flow map of the system is symplectic in the extended sense.

To apply the standard geometric mechanical tools in nonsmooth situations, it is necessary to formulate the problem so that the space of admissible trajectories of the system has a smooth manifold structure. To do this, we work in the extended framework, where both configuration variables and time are considered as functions of a fixed parameter space. This is the same approach as that used in multisymplectic mechanics (see [15, 35]), where it was introduced to allow the consideration of right, or horizontal, transformations of the system.

Next, in section 3, we discretize the variational structure of the system, based on the concept of discrete mechanics (see [37] for an overview and history), to obtain variational integrators for collision problems. By discretizing the variational structure, rather than some generalized equations of motion, we are able to show that our methods have various geometric properties, including the preservation of momentum maps and symplectic structures.

Finally, in section 4, we consider particular examples of our variational integrators for collision problems and investigate their behavior on a number of sample problems of rigid-body collisions. In the appendix, we lay the foundations of some future work by briefly discussing possible uses of the nonsmooth calculus approach (see [25]) in the context of variational collision integrators.

1.1. Overview of the continuous model. Before we begin with the body of the paper, we will first give a brief overview of the main ideas and techniques used. This is not a rigorous treatment, but everything stated here will be precisely defined and proven later.

Consider the system defined by the Lagrangian $L(q,\dot{q}) = \frac{1}{2}\dot{q}^T M\dot{q} - V(q)$, where M is a mass matrix and V is a potential function. Here $q = (q^1,\dots,q^n)$ is a vector of configuration variables which lives in the configuration manifold $q \in Q$. We now consider a subset $C \subset Q$, which we call the admissible set and which represents those configurations for which no contact is occurring. The boundary ∂C of the admissible set is all of those points at which contact has just occurred but for which no interpenetration is taking place.

We now consider a trajectory $q(t)$ which maps $q : [0,T] \to Q$ such that $q(t) \in C$, except at a particular time t_i for which $q(t_i) \in \partial C$. The time t_i is thus the time at which contact occurs, and we allow the trajectory $q(t)$ to be nonsmooth but still continuous at this time.

Proceeding in the standard way for Lagrangian mechanics, we construct the action of the trajectory by integrating the Lagrangian along $q(t)$. We then compute variations of the action with respect to variations in both the curve $q(t)$ and the impact time t_i, holding the endpoints of the curve fixed, to give

$$\begin{aligned}
&\delta \int_0^T L(q(t),\dot{q}(t))\,dt \\
&\quad = \int_0^{t_i} \left[\frac{\partial L}{\partial q}\cdot\delta q + \frac{\partial L}{\partial \dot{q}}\cdot\delta\dot{q}\right] dt + \int_{t_i}^{T} \left[\frac{\partial L}{\partial q}\cdot\delta q + \frac{\partial L}{\partial \dot{q}}\cdot\delta\dot{q}\right] dt \\
&\qquad - \Big[L(q,\dot{q})\cdot\delta t_i\Big]_{t_i^-}^{t_i^+} \\
&\quad = \int_0^{t_i} \left[\frac{\partial L}{\partial q} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)\right]\cdot\delta q\;dt + \int_{t_i}^{T} \left[\frac{\partial L}{\partial q} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)\right]\cdot\delta q\;dt \\
&\qquad - \left[\frac{\partial L}{\partial \dot{q}}\cdot\delta q + L\right]_{t_i^-}^{t_i^+},
\end{aligned}$$

where we have used integration by parts and the condition $\delta q(T) = \delta q(0) = 0$. Requiring that the variations of the action be zero for all δq implies that on the intervals away from t_i the integrand must be zero, giving the well-known Euler–Lagrange equations

$$\frac{\partial L}{\partial q}(q,\dot{q}) - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}(q,\dot{q})\right) = 0.$$

For the particular form of the Lagrangian chosen above, this is simply

$$M\ddot{q} = -\nabla V(q),$$

which is Newton's equation of mass times acceleration equals force, and this equation describes the motion of the system away from impact.

Not only must the two integrals in the variation equation be zero, but the jump term at t_i must also be zero. Here it is necessary to recall that the curve at time t_i must lie in the boundary ∂C of the admissible set, and differentiating this relationship $q(t_i) \in \partial C$ gives the

condition

$$\delta q(t_i) + \dot{q}(t_i) \cdot \delta t_i \in T\partial C,$$

which states that the combined variation on the left-hand side must be in the tangent plane to ∂C at the impact point. The space of allowable $\delta q(t_i)$ and δt_i is spanned by the set of $\delta q(t_i) \in T\partial C$ with $\delta t_i = 0$, together with the additional variation $\delta q(t_i) = -\dot{q}(t_i)$ with $\delta t_i = 1$. Substituting each of these into the jump term in the variation equation gives the two relations

$$\left[\frac{\partial L}{\partial \dot{q}}\bigg|_{t_i^+} - \frac{\partial L}{\partial \dot{q}}\bigg|_{t_i^-} \right] \cdot \delta q(t_i) = 0 \text{ for all } \delta q(t_i) \in T\partial C,$$
$$\left[\frac{\partial L}{\partial \dot{q}} \cdot \dot{q} - L \right]_{t_i^+} - \left[\frac{\partial L}{\partial \dot{q}} \cdot \dot{q} - L \right]_{t_i^-} = 0.$$

When the Lagrangian is of the form kinetic minus potential, as above, these can be written as

$$\dot{q}(t_i^+) - \dot{q}(t_i^-) \in N_C(q_i(t)), \tag{1a}$$
$$E_L(t_i^+) - E_L(t_i^-) = 0, \tag{1b}$$

where the energy is $E_L(q, \dot{q}) = \frac{1}{2}\dot{q}^T M \dot{q} + V(q)$ and $N_C(q)$ is the normal cone to ∂C at q. The first of these two equations states that the jump in velocity at the impact point must be orthogonal to the boundary ∂C, while the second equation states that energy must be conserved during the impact. Together these two equations constitute a system of n equations which describe the evolution of the system during the collision.

It is well known that the system described by the Euler–Lagrange equations has many special properties. In particular, the flow on state space is symplectic, meaning that it conserves a particular two-form, and if there are symmetry actions on phase space, then there are corresponding conserved quantities of the flow, known as momentum maps. All of these geometric properties can be proven directly from the variational principle used above, and so they also hold for nonsmooth systems. Later we will see how this can be precisely formulated.

1.2. Overview of the discrete model. Discrete variational mechanics is based on replacing the position q and velocity $\dot{q}$ with two nearby positions q_0 and q_1 and a timestep h. These positions should be thought of as being two points on a curve at time h apart so that $q_0 \approx q(0)$ and $q_1 \approx q(h)$ for some short curve segment $q(t)$.

We next consider a discrete Lagrangian $L_d(q_0, q_1, h)$, which we think of as approximating the action integral along the curve segment between q_0 and q_1. For concreteness, consider the very simple approximation given by

$$L_d(q_0, q_1, h) = h \left[\left(\frac{q_1 - q_0}{h} \right)^T M \left(\frac{q_1 - q_0}{h} \right) - V(q_0) \right].$$

This is simply the rectangle rule applied to approximate the action integral, with the velocity being approximated by the difference operator.

Now consider a discrete curve of points $\{q_k\}_{k=0}^N$ in C and corresponding times $t_k = kh$, together with a special impact point $\tilde{q} \in \partial C$ and an impact time $\tilde{t} = \alpha t_{i-1} + (1-\alpha)t_i$. Here $\alpha \in [0,1]$ is a parameter which interpolates $\tilde{t}$ with the interval $[t_{i-1}, t_i]$. Given such a discrete trajectory

$$(q_0, t_0), \dots, (q_{i-1}, t_{i-1}), (\tilde{q}, \tilde{t}), (q_i, t_i), \dots, (q_N, t_N),$$

we calculate the discrete action along this sequence by summing the discrete Lagrangian on each adjacent pair, with the timestep being the difference between the pair of times. Following the continuous derivation above, we compute variations of this action sum with respect to variations in the q_k as well as $\tilde{q}$ and α (and hence $\tilde{t}$), with the boundary points q_0 and q_N held fixed. This gives

$$\begin{aligned}
\delta \Bigg[\sum_{k=0}^{i-2} & L_d(q_k, q_{k+1}, h) + L_d(q_{i-1}, \tilde{q}, \alpha h) \\
& + L_d(\tilde{q}, q_i, (1-\alpha)h) + \sum_{k=i}^{N-1} L_d(q_k, q_{k+1}, h) \Bigg] \\
&= \sum_{k=0}^{N-1} \Big[D_1 L_d(q_k, q_{k+1}, h) \cdot \delta q_k + D_2 L_d(q_k, q_{k+1}, h) \cdot \delta q_{k+1} \Big] \\
&= \sum_{k=1}^{i-2} \Big[D_2 L_d(q_{k-1}, q_k, h) + D_1 L_d(q_k, q_{k+1}, h) \Big] \cdot \delta q_k \\
&\qquad + \sum_{k=i+1}^{N-1} \Big[D_2 L_d(q_{k-1}, q_k, h) + D_1 L_d(q_k, q_{k+1}, h) \Big] \cdot \delta q_k \\
&\qquad + \Big[D_2 L_d(q_{i-2}, q_{i-1}, h) + D_1 L_d(q_{i-1}, \tilde{q}, \alpha h) \Big] \cdot \delta q_{i-1} \\
&\qquad + \Big[D_2 L_d(q_{i-1}, \tilde{q}, \alpha h) + D_1 L_d(\tilde{q}, q_i, (1-\alpha)h) \Big] \cdot \delta \tilde{q} \\
&\qquad + \Big[D_2 L_d(\tilde{q}, q_i, (1-\alpha)h) + D_1 L_d(q_i, q_{i+1}, h) \Big] \cdot \delta q_i \\
&\qquad + \Big[D_3 L_d(q_{i-1}, \tilde{q}, \alpha h) - D_3 L_d(\tilde{q}, q_i, (1-\alpha)h) \Big] \cdot h \delta \alpha,
\end{aligned}$$

where we have rearranged the summation and we have used the fact that $\delta q_0 = \delta q_N = 0$. This calculation is illustrated graphically in Figure 1.

If we now require that the variations of the action be zero for any choice of δq_k, then we obtain the discrete Euler–Lagrange equations

$$D_2 L_d(q_{k-1}, q_k, h) + D_1 L_d(q_k, q_{k+1}, h) = 0,$$

which must hold for each k away from the impact time. For the particular L_d chosen above, we compute

$$\begin{aligned}
D_2 L_d(q_{k-1}, q_k, h) &= M\left(\frac{q_k - q_{k-1}}{h}\right), \\
D_1 L_d(q_k, q_{k+1}, h) &= -\left[M\left(\frac{q_{k+1} - q_k}{h}\right) + h \nabla V(q_k) \right],
\end{aligned}$$

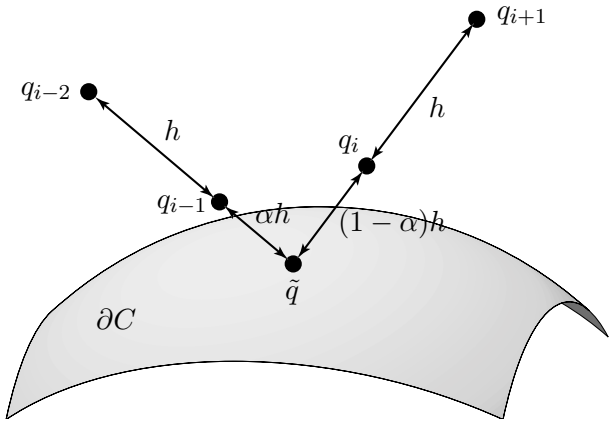


Figure 1. *The discrete variational principle for collisions.*

and so the discrete Euler–Lagrange equations are

$$M\left(\frac{q_{k+1}-2q_k+q_{k-1}}{h^2}\right) = -\nabla V(q_k).$$

This is clearly a discretization of Newton's equations, using a simple finite difference rule for the derivative.

If we take initial conditions (q_0, q_1), then the discrete Euler–Lagrange equations define a recursive rule for calculating the sequence $\{q_k\}_{k=0}^N$. Regarded in this way, they define a map $F_{L_d} : (q_k, q_{k+1}) \mapsto (q_{k+1}, q_{k+2})$, which we can think of as a one-step integrator for the system defined by the continuous Euler–Lagrange equations, away from impact.

Near impact, we must consider the other equations which are implied by the discrete variation equation being zero. Assume that we have used the discrete Euler–Lagrange equations to compute the trajectory up until the pair (q_{i-2}, q_{i-1}), just before impact. Now we have the equation

$$D_2L_d(q_{i-2}, q_{i-1}, h) + D_1L_d(q_{i-1}, \tilde{q}, \alpha h) = 0,$$

which becomes

$$M\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right) - M\left(\frac{q_{i-1}-q_{i-2}}{h}\right) = -\alpha h \nabla V(q_{i-1}).$$

Combining this with the condition that $\tilde{q} \in \partial C$ we obtain $n+1$ equations to be solved for the $n+1$ unknowns $\tilde{q}$ and α. We thus now know the point and time of contact. Next, we recall that $\tilde{q} \in \partial C$, and so its variations must lie in the tangent space. This means that we have the pair of equations

$$\begin{aligned}\Big[D_2L_d(q_{i-1}, \tilde{q}, \alpha h) + D_1L_d(\tilde{q}, q_i, (1-\alpha)h)\Big] \cdot \delta\tilde{q} = 0 \text{ for all } \delta\tilde{q} \in T\partial C,\\ D_3L_d(q_{i-1}, \tilde{q}, \alpha h) - D_3L_d(\tilde{q}, q_i, (1-\alpha)h) = 0,\end{aligned}$$

which become

$$M\left(\frac{q_i-\tilde{q}}{(1-\alpha)h}\right)-M\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right)+(1-\alpha)h\nabla V(\tilde{q})\in N_C(\tilde{q}),$$

$$\left[\frac{1}{2}\left(\frac{q_i-\tilde{q}}{(1-\alpha)h}\right)^T M\left(\frac{q_i-\tilde{q}}{(1-\alpha)h}\right)+V(\tilde{q})\right]$$

$$-\left[\frac{1}{2}\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right)^T M\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right)+V(q_{i-1})\right]=0.$$

These are discrete versions of (1), and they give n equations to be solved for q_i. Finally, we use the equation

$$D_2L_d(\tilde{q},q_i,(1-\alpha)h)+D_1L_d(q_i,q_{i+1},h)=0,$$

which is

$$M\left(\frac{q_{i+1}-q_i}{h}\right)-M\left(\frac{q_i-\tilde{q}}{(1-\alpha)h}\right)=-\nabla V(q_i),$$

to solve for q_{i+1}, and then we can revert to using the standard discrete Euler–Lagrange equations to continue away from the impact.

The power of the variational approach becomes apparent when we consider the geometric properties of the discrete system. Just as in the continuous case, we can derive conservation laws of the discrete system directly from the variational principle. In particular, we will see that there is a conserved discrete symplectic form, and conserved discrete momentum maps arise from symmetries. In addition, in section 4, we will investigate the numerical behavior of the discrete system, regarded as an integrator for the continuous problem, and we will see that it also has excellent long-time stable energy behavior.

To understand the geometry and properties of both the continuous and discrete nonsmooth mechanics, however, we now need to return to the beginning and develop a more rigorous treatment of the variational procedure.

2. Continuous model. As noted in the introduction, the basic methodology used here is that of variational mechanics and variational discretizations. Clearly, a generalization to the nonsmooth setting of the autonomous, smooth variational mechanics cannot be done in a straightforward way. One of the major obstacles is that the lack of smoothness for the mappings prevents us from using the differential calculus on the manifold of mappings, as one essentially does in the smooth case (see [36]).

The main issue addressed in this section is how to overcome this difficulty and how to derive the conservation of quantities such as energy, momentum maps, and the symplectic form using a variational approach. The approach we use is to extend the problem to the nonautonomous case so that both configuration variables and time are functions of a separate parameter τ. This allows the impact to be fixed in τ space while remaining variable in both configuration and time spaces, and it means that the relevant space of configurations will indeed be a smooth manifold, as we shall prove.

To make our variational procedure clear, we initially consider only the frictionless, purely elastic impact problem. In the last subsection, however, we show how the results can be extended to deal with friction and nonelastic impacts.

2.1. Lagrangian mechanics in a nonsmooth setting. Consider a configuration manifold Q and a submanifold with boundary $C \subset Q$ which represent the subset of admissible configurations. Let ∂C be called the contact set, and let $L : TQ \to \mathbb{R}$ be a regular Lagrangian.

Remark. Similar results are obtained if we considered the configuration Q a manifold with boundary and the contact set to be ∂Q.

Let us now consider the *path space* defined by

$$\mathcal{M} = \mathcal{T} \times \mathcal{Q}([0,1], \tau_i, \partial C, Q),$$

where

$$\begin{aligned} \mathcal{T} &= \{c_t \in C^\infty([0,1], \mathbb{R}) \mid c_t' > 0 \text{ in } [0,1]\}, \\ \mathcal{Q}([0,1], \tau_i, \partial C, Q) &= \{c_q : [0,1] \to Q \mid c_q \text{ is a } C^0, \text{ piecewise } C^2 \text{ curve}, \\ &\qquad c_q(\tau) \text{ has only one singularity at } \tau_i, c_q(\tau_i) \in \partial C\}. \end{aligned}$$

A path $c \in \mathcal{M}$ is thus a pair $c = (c_t, c_q)$. Given a path, we can form the *associated curve* $q : [c_t(0), c_t(1)] \to Q$ by

$$q(t) = c_q(c_t^{-1}(t)),$$

and we denote by $\mathcal{C}$ the space of all these paths $q(t) \in Q$.

The theory we will develop applies to rigid-body impact problems, such as a particle bouncing on a rigid wall or two rigid bodies colliding, where the submanifold ∂C is obtained from the condition that interpenetration of matter cannot occur. The moment of impact τ_i is fixed in the τ space but is allowed to vary in the t space according to $t_i = c_t(\tau_i)$; thus the setting we suggest is not restrictive in this sense.

We use a nonautonomous formulation of an autonomous mechanical system in order to achieve smoothness of the manifold of mappings, as one can see from the following lemmas.

Lemma 2.1. *$\mathcal{T}$ is a smooth manifold.*

Proof. $\mathcal{T}$ is an open set in $C^\infty([0,1], \mathbb{R})$, which is a smooth manifold (see [36]). Then $\mathcal{T}$ is a submanifold of $C^\infty([0,1], \mathbb{R})$ and thus a manifold. ∎

Lemma 2.2. *$\mathcal{Q}([0,1], \tau_i, \partial C, Q)$ is a smooth manifold.*

Proof. Fix a chart U in Q such that $U \cap \partial C \neq \emptyset$ and $U \cap \partial C$ is a chart in ∂C. Consider the set

$$\mathcal{Q}_U = \mathcal{Q}([0, \tau_i], U) \times \mathcal{Q}([\tau_i, 1], U) \times (U \cap \partial C),$$

where

$$\begin{aligned} \mathcal{Q}([0, \tau_i], U) &= \{q : [0, \tau_i] \to Q \mid q \text{ is a } C^\infty \text{ curve}, q(\tau_i) \in U\}, \\ \mathcal{Q}([\tau_i, 1], U) &= \{q : [\tau_i, 1] \to Q \mid q \text{ is a } C^\infty \text{ curve}, q(\tau_i) \in U\}. \end{aligned}$$

An element $c \in \mathcal{Q}([0,1], \tau_i, \partial C, Q)$ is the inverse image of the origin for some map $g_U : \mathcal{Q}_U \to \mathbb{R}^{2n}$ given by

$$g_U(q_1(\tau), q_2(\tau), q_i) = \begin{pmatrix} q_1(\tau_i) - q_i \\ q_2(\tau_i) - q_i \end{pmatrix},$$

where we denote the dimension of Q by n. One can prove that 0 is a regular value of g_U, and then the set $g_U^{-1}(0)$ is a submanifold of $\mathcal{Q}_U$, and thus it has a manifold structure.

Now $\{g_U^{-1}(0)\}_U$ represents a covering of $\mathcal{Q}([0,1],\tau_i,\partial C,Q)$, where each element of the covering is a manifold. The elements of the covering satisfy the compatibility conditions necessary to ensure that $\mathcal{Q}([0,1],\tau_i,\partial C,Q)$ itself is a manifold (see [1]). ■

Corollary 2.3. *$\mathcal{M}$ is a smooth manifold.*

Remark. The theory can be easily extended to a problem involving more than one impact by simply taking multiple points τ_i at which the trajectory is nonsmooth.

Note that the tangent space at $q \in \mathcal{Q}$ can be written as

$$T_q\mathcal{Q} = \{v : [0,1] \to TQ \mid v \text{ is a } C^0 \text{ piecewise } C^2 \text{ map }, v(\tau_i) \in T_{q(\tau_i)}\partial C\},$$

which will be a convenient form below when we consider variations of trajectories. The tangent space to the path space $\mathcal{M}$ is then given by $T\mathcal{M} = T\mathcal{T} \times T\mathcal{Q}$.

Remark. As we have noted above, fixing the impact point τ_i in τ space allows us to rigorously define what we mean by a variation of the impact point in t space. This is similar to the introduction of a parameterized spacetime in [35] and [31].

The *action map* $\mathfrak{G} : \mathcal{M} \to \mathbb{R}$ is given by

$$\mathfrak{G}(c_t, c_q) = \int_0^1 L\left(c_q(\tau), \frac{c_q'(\tau)}{c_t'(\tau)}\right) c_t'(\tau)\, d\tau, \tag{2}$$

where c' denotes the derivative with respect to τ.

Remark. $c_q'(\tau)$ does not exist at τ_i, but the definition makes sense nonetheless.

If q is the associated curve for $c \in \mathcal{M}$, by the change of coordinates $s = c_t(\tau)$ we can also write $\mathfrak{G}$ as

$$\mathfrak{G}(q) = \int_{c_t(0)}^{c_t(1)} L(q(s), \dot{q}(s))ds, \tag{3}$$

where $\dot{q}$ denotes the derivative with respect to t.

Define the *extended configuration manifold* to be $Q_e = \mathbb{R} \times Q$ and the *second order submanifold* of $T(TQ_e)$ to be

$$\ddot{Q}_e = \left\{ \frac{d^2c}{d\tau^2}(0) \in T(TQ_e) \mid c : [0,1] \to Q_e \text{ is a } C^2 \text{ curve} \right\}. \tag{4}$$

Now we can derive the equations of motion and the jump conditions in a purely variational way, by taking variations of the actions with respect to the path. This leads to the following fundamental theorem.

Theorem 2.4. *Given a C^k Lagrangian L, $k \geq 2$, there exist a unique C^{k-2} mapping $EL : \ddot{Q} \to T^*Q_e$ and a unique C^{k-1} one-form Θ_L on TQ_e such that for all variations $\delta c \in T_c\mathcal{M}$ of c we have*

$$d\mathfrak{G}(c) \cdot \delta c = \int_0^{\tau_i} EL(c'') \cdot \delta c\, d\tau + \int_{\tau_i}^1 EL(c'') \cdot \delta c\, d\tau + \Theta_L(c') \cdot \hat{\delta} c|_0^{\tau_i^-} + \Theta_L(c') \cdot \hat{\delta} c|_{\tau_i^+}^1, \tag{5}$$

where

$$\hat{\delta}c(\tau) = \left(\left(c(\tau), \frac{\partial c}{\partial \tau}(\tau) \right), \left(\delta c(\tau), \frac{\partial \delta c}{\partial \tau}(\tau) \right) \right).$$

The mapping EL is called the Euler–Lagrange derivative *and the one-form* Θ_L *is called the* Lagrangian one-form. *In coordinates they have the expression*

$$EL(c'') = \left[\frac{\partial L}{\partial q} c_t' - \frac{d}{d\tau} \left(\frac{\partial L}{\partial \dot{q}} \right) \right] dc_q + \left[\frac{d}{d\tau} \left(\frac{\partial L}{\partial \dot{q}} \frac{c_q'}{c_t'} - L \right) \right] dc_t, \tag{6}$$

$$\Theta_L(c') = \left[\frac{\partial L}{\partial \dot{q}} \right] dc_q - \left[\frac{\partial L}{\partial \dot{q}} \frac{c_q'}{c_t'} - L \right] dc_t. \tag{7}$$

Proof. Consider $\delta c \in T_c\mathcal{M}$. We calculate $d\mathfrak{G}(c) \cdot \delta c$ using the definition (see [36])

$$d\mathfrak{G}(c) \cdot \delta c = \frac{d}{d\lambda} \mathfrak{G}(c^\lambda) \Big|_{\lambda=0}, \tag{8}$$

where c^λ is a curve in $\mathcal{M}$ with $c^0 = c$ and $\frac{dc^\lambda}{d\lambda}|_{\lambda=0} = \delta c$. Splitting c^λ into components $c^\lambda = (c_t^\lambda, c_q^\lambda)$, we then have $(\frac{d}{d\lambda} c_t^\lambda|_{\lambda=0}, \frac{d}{d\lambda} c_q^\lambda|_{\lambda=0}) = (\delta c_t, \delta c_q)$, and we can calculate

$$d\mathfrak{G} \cdot \delta c = \int_0^1 \left[\frac{\partial L}{\partial q} \delta c_q + \frac{\partial L}{\partial \dot{q}} \left(\frac{\delta c_q'}{c_t'} - \frac{c_q' \delta c_t'}{(c_t')^2} \right) \right] c_t' \, d\tau + \int_0^1 L \delta c_t' \, d\tau.$$

Now we split the integral $\int_0^1$ into $\int_0^{\tau_i} + \int_{\tau_i}^1$ in order to integrate the $\delta c_q'$ and $\delta c_t'$ terms by parts. Some straightforward algebra then leads to (5). ∎

2.2. Hamilton's principle of critical action. Hamilton's principle of critical action tells us that we should consider critical points of the action function. Therefore, let us define the *space of solutions* $\mathcal{M}_L \subset \mathcal{M}$ to be the set of all paths $c \in \mathcal{M}$ which satisfy $d\mathfrak{G}(c) \cdot \delta c = 0$ for all variations $\delta c \in T_c\mathcal{M}$ which are zero at the boundary points 0 and 1.

Using (5), we can see that c is a solution if it satisfies

$$\int_0^{\tau_i} EL(c'') \cdot \delta c \, d\tau + \int_{\tau_i}^1 EL(c'') \cdot \delta c \, d\tau + \Theta_L(c')|_{\tau_i^-}^{\tau_i^+} \cdot \hat{\delta} c(\tau_i) = 0 \tag{9}$$

for all variations $\delta c \in T_c\mathcal{M}$.

From (9) it is clear that c is a solution iff the Euler–Lagrange derivative is zero on smooth portions and the Lagrangian one-form has a zero jump at τ_i. Splitting $EL(c'')$ into the two components, we obtain

$$\frac{\partial L}{\partial q} - \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right) = 0 \quad \text{in} \quad [t_0, t_i) \cup (t_i, t_1], \tag{10}$$

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \dot{q} - L \right) = 0 \quad \text{in} \quad [t_0, t_i) \cup (t_i, t_1], \tag{11}$$

where $t_0 = c_t(0)$, $t_1 = c_t(1)$, and $t_i = c_t(\tau_i)$.

In fact, (11) is redundant, as it is a consequence of (10). Indeed, if c is a path satisfying (10) for all $t \in (t_0, t_i) \cup (t_i, t_1)$, then the second component (11) of the Euler–Lagrange equations is identically satisfied. To see this, we may calculate

$$\begin{aligned}\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\dot{q} - L\right) &= \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)\dot{q} + \frac{\partial L}{\partial \dot{q}}\ddot{q} - \frac{dL}{dt} \\ &= \left[\frac{\partial L}{\partial q}\dot{q} + \frac{\partial L}{\partial \dot{q}}\ddot{q}\right] - \frac{dL}{dt} \\ &= 0,\end{aligned}$$

where we used (10) to pass from the first to the second line.

The second part (11) of the Euler–Lagrange equations represents the conservation of energy for an autonomous system, provided the motion is smooth. The *energy* $E : TQ \to \mathbb{R}$ is defined to be

$$E(q, \dot{q}) = \frac{\partial L}{\partial \dot{q}}(q, \dot{q}) \cdot \dot{q} - L(q, \dot{q}).$$

It is not surprising that the second part of the Euler–Lagrange equations (11) is redundant, since the first part (10) already has the energy evolution built into it.

The previous definition of the energy function allows us to write the Lagrangian one-form in the compact notation

$$\Theta_L = \frac{\partial L}{\partial \dot{q}} dq - E dt, \tag{12}$$

where we use (q, t) to refer to the two components of c. The conservation of the Lagrangian one-form at the impact time reads

$$\Theta_L|_{\tau_i^-} = \Theta_L|_{\tau_i^+} \quad \text{on} \quad TQ_e|(\mathbb{R} \times \partial C). \tag{13}$$

Splitting this into the two components gives

$$\left.\frac{\partial L}{\partial \dot{q}}\right|_{t=t_i^-} \cdot \delta q = \left.\frac{\partial L}{\partial \dot{q}}\right|_{t=t_i^+} \cdot \delta q \tag{14}$$

for any $\delta q \in T_{q(t_i)}\partial C$ and

$$E(q(t_i^-), \dot{q}(t_i^-)) = E(q(t_i^+), \dot{q}(t_i^+)). \tag{15}$$

These equations are the Weierstrass–Erdmann-type conditions for impact. That is, (14) states that the linear momentum must be conserved in the tangent direction to ∂C, while (15) states that the energy must be conserved during an elastic impact.

The system of (14) and (15) must be solved for $\dot{q}(t_i^+)$. An obvious solution is $\dot{q}(t_i^+) = \dot{q}(t_i^-)$, but this is ruled out since the resulting trajectory would no longer lie in the admissible set. That is, it would violate the physical noninterpenetration condition.

Remark. Of course, existence and uniqueness for nonsmooth systems are very deep questions. Here, we will simply remark that for a codimension-one smooth boundary ∂C and quadratic kinetic energy, solutions to the system (14), (15) exist and are unique locally. The questions of global existence and uniqueness of solutions for more general Lagrangians is left for future works.

2.3. Lagrangian flows and conservation of the symplectic form. As we have already seen, a path $c \in \mathcal{M}$ is a solution of the variational principle if its associated curve $q(t)$ satisfies the Euler–Lagrange equations (10) and the jump conditions (14) and (15). It is a well-known fact that, in the smooth case, such a trajectory is uniquely determined by an initial condition in TQ. Since we work in a nonsmooth context, we must assume uniqueness of the physical trajectory at the impact point; we have already discussed in the previous subsection some conditions under which this actually occurs.

Under this hypothesis, the space $\mathcal{C}_L$, defined to be the space of curves $q(t)$ that satisfy (10), (14), and (15), may be identified with the space of initial conditions $(t_0, q_0, \dot{q}_0)$ on $\mathbb{R} \times TQ$.

Based on these remarks, we can define a flow $F_t : \mathbb{R} \times TQ \to \mathbb{R} \times TQ$ as

$$F_t(t_0, q_0, \dot{q}_0) = (t_0 + t, q(t_0 + t), \dot{q}(t_0 + t)), \tag{16}$$

where $q(t)$ is the unique trajectory in $\mathcal{C}_L$ corresponding to $(t_0, q_0, \dot{q}_0) \in \mathbb{R} \times TQ$. The mapping F_t is called the *Lagrangian flow*. In the nonsmooth setting, F_t will not necessarily be a smooth map on the whole of its domain. Later, we will restrict our attention to the parts of the domain on which F_t is smooth in order to use the derivatives of F_t with respect to the initial conditions and to time.

Remark. Even though we have worked within an extended configuration manifold formulation up until this point, here we have defined a flow on TQ rather than taking a flow on TQ_e with initial conditions in TQ_e. The reason for doing this is that the derivative t_0' has no physical meaning, and no mechanical problem has the derivative of time with respect to some parameter as an initial condition.

Next, we will show in which sense the Lagrangian flow F_t is symplectic. We begin by relating the previous approach to the one used in the rest of the section.

As we noted above, to any initial condition $(t_0, q_0, \dot{q}_0)$ in $\mathbb{R} \times TQ$ there corresponds a unique trajectory $q(t) \in \mathcal{C}_L$ such that $(q(t_0), \dot{q}(t_0)) = (q_0, \dot{q}_0)$. Trajectories in $\mathcal{C}_L$ are unique up to reparameterization in τ. Accordingly, we can define an equivalence relation in $\mathcal{M}_L$ by

$$c^0 \sim c^1 \text{ iff } c_q^0 \circ (c_t^0)^{-1} = c_q^1 \circ (c_t^1)^{-1}, \tag{17}$$

where $c^0, c^1 \in \mathcal{M}_L, c^0 = (c_q^0, c_t^0), c^1 = (c_q^1, c_t^1)$. That is, two paths are equivalent if they have the same associated curve, and so to a given trajectory $q(t)$ in $\mathcal{C}_L$ there corresponds an equivalence class $\hat{c}$ of curves in the extended space.

In a similar manner, we can define an equivalence relation on TQ_e by

$$(t_0, q_0, t_0', q_0') \sim (t_1, q_1, t_1', q_1') \text{ iff } t_0 = t_1,\ q_0 = q_1, \text{ and } \frac{q_0'}{t_0'} = \frac{q_1'}{t_1'}, \tag{18}$$

which is a pointwise version of the previous equivalence relation (17).

Now, the quotient space $TQ_e/\sim$ may be identified with the product $\mathbb{R}\times TQ$, and the flow F_t may be regarded not as a flow on TQ_e (which would not be desirable, as explained in the previous remark) but as a flow on the equivalence classes of TQ_e.

To prove symplecticity for the flow F_t in a precise sense, we must reinterpret Theorem 2.4 by slightly modifying the definition of the Lagrangian one-form Θ_L.

That is, Theorem 2.4 stands with the same statement and fundamental relation (5) if we replace Θ_L with the one-form $\bar\Theta_L$ on $\mathbb{R}\times TQ\cong TQ_e/\sim$, where $\bar\Theta_L$ is given by the same coordinate expression as Θ_L, i.e., relation (7). More precisely, (5) becomes

$$d\mathfrak{G}(c)\cdot\delta c=\int_0^{\tau_i} EL(c'')\cdot\delta c\,d\tau+\int_{\tau_i}^1 EL(c'')\cdot\delta c\,d\tau+\bar\Theta_L(\tilde c)\cdot\delta\tilde c\big|_0^{\tau_i^-}+\bar\Theta_L(\tilde c)\cdot\delta\tilde c\big|_{\tau_i^+}^1, \tag{19}$$

where

$$\begin{aligned}\tilde c(\tau)&=\left(c_t(\tau),c_q(\tau),\frac{c_q'(\tau)}{c_t'(\tau)}\right),\\ \delta\tilde c(\tau)&=\left(\left(c_t(\tau),c_q(\tau),\frac{c_q'(\tau)}{c_t'(\tau)}\right),\left(\delta c_t(\tau),\delta c_q(\tau),\left(\frac{\delta c_q'}{c_t'}-\frac{c_q'\delta c_t'}{(c_t')^2}\right)(\tau)\right)\right).\end{aligned}$$

It is exactly this one-form $\bar\Theta_L$ on $\mathbb{R}\times TQ$ which is preserved by the flow F_t, as we will now show.

To any fixed $(t_0,q_0,\dot q_0)\in\mathbb{R}\times TQ$ we associate the integral curve $s\mapsto F_s(t_0,q_0,\dot q_0)$ for $s\in[0,t]$; the value of $\mathfrak{G}$ on that curve is denoted by $\mathfrak{G}_t$ and again called the *action*. Thus we define the map $\mathfrak{G}_t:\mathbb{R}\times TQ\to\mathbb{R}$ by

$$\mathfrak{G}_t(t_0,q_0,\dot q_0)=\int_{t_0}^{t_0+t} L\left(q(s),\dot q(s)\right)ds, \tag{20}$$

where $q(t)\in\mathcal{C}_L$ is the solution corresponding to $(t_0,q_0,\dot q_0)$.

If $c=(c_t,c_q)$ is any representative in the equivalence class $\hat c$ corresponding to q, we can write

$$F_t(t_0,q_0,\dot q_0)=\left(c_t(\tau),c_q(\tau),\frac{c_q'(\tau)}{c_t'(\tau)}\right), \tag{21}$$

where $\tau=c_t^{-1}(t_0+t)$.

Consider now an arbitrary curve $\lambda\mapsto(t_0^\lambda,q_0^\lambda,\dot q_0^\lambda)$ in $\mathbb{R}\times TQ$ which passes through $(t_0,q_0,\dot q_0)$ at $\lambda=0$. Denote by $q^\lambda(t)$ the unique trajectories in $\mathcal{C}_L$ corresponding to $(t_0^\lambda,q_0^\lambda,\dot q_0^\lambda)$ and by $\hat c^\lambda$ their equivalence classes in $\mathcal{M}_L$. (At $\lambda=0$ they reduce to $q(t)$ and $\hat c$, respectively.) We pick representatives $(c_t^\lambda,c_q^\lambda)$ in $\hat c^\lambda$ such that for any $\lambda\ge0$ we have

$$(c_t^\lambda)^{-1}(t_0^\lambda+t)=\text{const} \tag{22}$$

for some $t>0$; we denote this common value by τ.

Then, using (21) and (22), the fundamental (19) becomes

$$d\mathfrak{G}_t((t_0,q_0,\dot{q}_0))\cdot(\delta t_0,\delta q_0,\delta\dot{q}_0)=\bar{\Theta}_L(F_t(t_0,q_0,\dot{q}_0))\cdot\frac{d}{d\lambda}F_t(t_0^\lambda,q_0^\lambda,\dot{q}_0^\lambda)\Big|_{\lambda=0}$$
$$-\bar{\Theta}_L(t_0,q_0,\dot{q}_0)\cdot\frac{d}{d\lambda}(t_0^\lambda,q_0^\lambda,\dot{q}_0^\lambda)\Big|_{\lambda=0}, \tag{23}$$

where $(\delta t_0,\delta q_0,\delta\dot{q}_0)=\frac{d}{d\lambda}|_{\lambda=0}(t_0^\lambda,q_0^\lambda,\dot{q}_0^\lambda)$.

Taking the exterior derivative of (23), we derive

$$0=dd\mathfrak{G}_t=F_t^*(d\bar{\Theta}_L)-d\bar{\Theta}_L. \tag{24}$$

Defining the *Lagrangian symplectic form* by $\Omega_L=-d\bar{\Theta}_L$, we now see that relation (24) gives the symplecticity of the flow in the extended sense

$$F_t^*\Omega_L=\Omega_L. \tag{25}$$

Thus we derived conservation of the canonical symplectic structure in the extended sense (see [24]), namely,

$$\Omega_L=\omega_L+dE\wedge dt, \tag{26}$$

where $\omega_L=-d\theta_L$ is the canonical symplectic form. Here, θ_L represents the component of the Lagrangian one-form given by (12)

$$\theta_L=\frac{\partial L}{\partial\dot{q}}dq. \tag{27}$$

It is the term $dE\wedge dt$ that distinguishes the nonautonomous structure used here from the autonomous approach, for which the symplectic structure is given only by the canonical symplectic form ω_L.

2.4. Noether's theorem. Suppose that a Lie group G, with Lie algebra $\mathfrak{g}$, acts on Q by the (left or right) action $\Phi: G\times Q\to Q$. Consider the tangent lift of this action to $T\Phi: G\times TQ\to TQ$ given by $(T\Phi)_g(v_q)=T(\Phi_g)\cdot v_q$, and for $\xi\in\mathfrak{g}$ define the *infinitesimal generators* $\xi_Q: Q\to TQ$ and $\xi_{TQ}: TQ\to T(TQ)$ by

$$\xi_Q(q)=\frac{d}{dt}_{|t=0}\exp(t\xi)\cdot q,$$
$$\xi_{TQ}(v_q)=\frac{d}{dt}_{|t=0}T_q\phi_t(v_q),$$

where ϕ_t is the flow of the vector field ξ_Q.

In this subsection, we will not use the extended configuration manifold setting with variations in both time and configuration variables, as is done in the rest of the paper. This means that we are restricted to symmetries of the configuration variables, which do not involve altering the time variable. This allows us to deal with most of the interesting physical problems,

while still keeping the theory relatively simple. For a full account of the conservation of momentum maps in the extended setting, see [37].

For a fixed initial time $t_0 \in \mathbb{R}$, define the flow map $\tilde{F}_t : TQ \to TQ$ by

$$\tilde{F}_t(q_0, \dot{q}_0) = (q(t_0 + t), \dot{q}(t_0 + t)), \tag{28}$$

where $q(t)$ is the unique trajectory in $\mathcal{C}_L$ corresponding to $(q_0, \dot{q}_0) \in TQ$, as initial condition at t_0.

In the autonomous setting, the Lagrangian one-form $\bar{\Theta}_L$ reduces to the configuration component θ_L given by (27), and the action $\mathfrak{G}_t$ from (20) becomes the map $\bar{\mathfrak{G}}_t : TQ \to \mathbb{R}$ defined by

$$\bar{\mathfrak{G}}_t(q_0, \dot{q}_0) = \int_{t_0}^{t_0+t} L\left(q(s), \dot{q}(s)\right) ds. \tag{29}$$

Define the *Lagrangian momentum map* $J_L : TQ \to \mathfrak{g}^*$ to be

$$J_L(v_q) \cdot \xi = \theta_L \cdot \xi_{TQ}(v_q).$$

We will now show that when the group action is a symmetry of both the Lagrangian and the submanifold ∂C, then the momentum maps are conserved quantities of the flow.

A Lagrangian $L : TQ \to \mathbb{R}$ is said to be *infinitesimally invariant* under the lift of the group action $\Phi : G \times Q \to Q$ if $dL \cdot \xi_{TQ} = 0$ for all $\xi \in \mathfrak{g}$, and in this case the group action is said to be a *symmetry* of the Lagrangian.

In proving the following theorem, we will essentially use the assumption that the group action Φ leaves the boundary ∂C of the collision set invariant (locally). An example where this assumption is valid is the case of two or more irregular bodies (for example, binary astroids) moving in space under gravitational forces. In this case, the collision set is invariant to translations and rotations ($G = SE(3)$).

Theorem 2.5 (Noether's theorem). *Consider a Lagrangian system $L : TQ \to \mathbb{R}$ which is infinitesimally invariant under the lift of the (left or right) group action $\Phi : G \times Q \to Q$. Under the assumption that the group action leaves ∂C invariant (locally), the corresponding Lagrangian momentum map $J_L : TQ \to \mathfrak{g}^*$ is a conserved quantity of the flow so that $J_L \circ \tilde{F}_t = J_L$ for all times t.*

Proof. The group action of G on Q induces a group action of G on the space $\mathcal{C}$ of paths $q(t)$ in Q by pointwise action so that $\Phi_g(q)(t) = \Phi_g(q(t))$. The tangent lift of Φ acting on $\mathcal{C}$ will thus be the pointwise group action of the tangent lift of Φ group action on Q. From this we derive

$$d\mathfrak{G}(q) \cdot \xi_{\mathcal{C}}(q) = \int_{t_0}^{t_1} dL \cdot \xi_{TQ}\, dt,$$

and so, symmetries of the Lagrangian induce symmetries of the action. This implies that Φ_g leaves the space of solutions $\mathcal{C}_L$ of the Euler–Lagrange equations invariant, and so we may restrict Φ_g to $\mathcal{C}_L$.

Furthermore, the flow map $\tilde{F}_t : TQ \to TQ$ commutes with the tangent lift of Φ on $\mathcal{C}$: $\tilde{F}_t \circ T\Phi_g = T\Phi_g \circ \tilde{F}_t$. Differentiating this with respect to g in the direction ξ gives

$$T(\tilde{F}_t) \cdot \xi_{TQ} = \xi_{TQ} \circ \tilde{F}_t.$$

We now follow the same idea used to prove symplecticity of the flow map F_t and identify the space of solutions $\mathcal{C}_L$ with the space of initial conditions TQ. For an initial condition $v_q \in TQ$ and corresponding solution curve $q \in \mathcal{C}_L$, we thus have

$$\begin{aligned} d\mathfrak{G}(q) \cdot \xi_{\mathcal{C}}(q) &= d\bar{\mathfrak{G}}_t(v_q) \cdot \xi_{TQ}(v_q) \\ &= ((\tilde{F}_t)^*(\theta_L) - \theta_L)(v_q) \cdot \xi_{TQ}(v_q) \end{aligned} \tag{30}$$

from (23).

To derive (30), one uses the assumption that the group action Φ leaves ∂C invariant (locally). More precisely, it is essential that the path curves $q^\lambda \in \mathcal{C}$ corresponding to $v_q^\lambda = \eta_\lambda(v_q)$ (by η_λ we denote the flow of ξ_{TQ} on TQ) have exactly the same impact time t_i as the curve q. We conclude this from the relation

$$q^\lambda(t_i) = \Phi_{\exp(\lambda\xi)}(q)(t_i) = \exp(\lambda\xi) \cdot q(t_i)$$

as well as the assumption on the group action and the condition that $q(t_i) \in \partial C$.

As the left-hand side of (30) is always zero, the previous identity gives

$$(\theta_L \cdot \xi_{TQ}) \circ \tilde{F}_t = \theta_L \cdot \xi_{TQ},$$

which is the definition of conservation of the momentum map. We complete the proof by noting that the argument above is valid for all times $t \in \mathbb{R}$. ∎

2.5. Forcing and friction. In this subsection, we extend the theory developed so far to include forcing and friction. To do this in the variational framework, we turn from using Hamilton's principle to the Lagrange–d'Alembert extension of it.

The usual force field description of impact dynamics contains a given external force, a normal contact force field over the area in contact, and a friction force field required to be self-equilibrated and tangential to the surfaces in contact.

Following [36], we define the *exterior force field* as a fiber-preserving map $F : TQ_e \to T^*Q_e$ over the identity, which we write in coordinates as

$$F : (c, c') \mapsto (c, F(c, c')). \tag{31}$$

We use a unified treatment of contact forces (the normal and the frictional forces) by defining the *contact force field* to be a map $f^{\text{con}} : TQ_e|(\partial C \times \mathbb{R}) \to T^*(\partial C \times \mathbb{R})$.

Given a Lagrangian L and the exterior and contact force fields defined as above, the *integral Lagrange–d'Alembert principle* for a curve $c \in \mathcal{M}$ states that

$$\begin{aligned} \delta \int_0^1 L\left(c_q(\tau), \frac{c_q'(\tau)}{c_t'(\tau)}\right) c_t'(\tau)\, d\tau + \int_0^1 F(c(\tau), c'(\tau)) \cdot \delta c(\tau)\, d\tau \\ + f^{\text{con}}(c(\tau_i), c'(\tau_i)) \cdot \delta c(\tau_i) = 0 \end{aligned} \tag{32}$$

for all admissible variations δc vanishing at the endpoints.

Using integration by parts and notation from section 2.1, one can show that (32) is equivalent to

$$\int_0^{\tau_i} \left[EL(c'') + F(c')\right] \cdot \delta c \, d\tau + \int_{\tau_i}^1 (EL(c'') + F(c')) \cdot \delta c \, d\tau + \Theta_L(c')\big|_{\tau_i^-}^{\tau_i^+} \cdot \hat{\delta} c(\tau_i) + f^{\mathrm{con}}(c(\tau_i), c'(\tau_i)) \cdot \delta c(\tau_i) = 0. \tag{33}$$

From (33) we obtain the *extended forced Euler–Lagrange equations*, which have coordinate expressions

$$\frac{d}{d\tau}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} c_t' = F_q \text{ in } [0, \tau_i) \cup (\tau_i, 1], \tag{34}$$

$$-\frac{d}{d\tau}\left(\frac{\partial L}{\partial \dot{q}}\dot{q} - L\right) = F_t \text{ in } [0, \tau_i) \cup (\tau_i, 1], \tag{35}$$

where (F_t, F_q) denote the corresponding components of F.

However, the first part (34) of the extended forced Euler–Lagrange equations has the energy evolution built into it, as can be seen from

$$\begin{aligned} \frac{dE}{dt} &= \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\dot{q} - L\right) \\ &= \left(\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q}\right)\dot{q} \\ &= \frac{F_q}{c_t'}\dot{q}, \end{aligned} \tag{36}$$

where we used (34) to pass from the first to the second line.

Therefore, from (35), the time component F_t of the exterior force field must necessarily be of the form

$$F_t = -F_q \cdot \dot{q}. \tag{37}$$

This compatibility condition is a consequence of the fact that the mechanical system is autonomous and the equations must depend only on the associated curve $q(t)$. The nonautonomous approach is relevant only in the context of nonsmooth mechanics, and it is not surprising that there is no particular gain from this approach wherever the motion is smooth.

Now we turn to (33) and write the remaining terms on the left-hand side in components to obtain

$$\frac{\partial L}{\partial \dot{q}}\bigg|_{t_i^-}^{t_i^+} \cdot \delta q + f_q^{\mathrm{con}} \cdot \delta q = 0 \tag{38}$$

for any $\delta q \in T_{q(t_i)}\partial C$, and

$$E(q(t_i^+), \dot{q}(t_i^+)) - E(q(t_i^-), \dot{q}(t_i^-)) - f_t^{\mathrm{con}} = 0. \tag{39}$$

Equations (38) and (39) represent the standard jump conditions for an inelastic impact with friction. Equation (38) gives the jump in the tangential component of the linear momentum due to the frictional forces acting on the tangent plane of the contact submanifold ∂C. The energy dissipation, given by (39), is due to the tangential frictional forces as well as to the normal reaction force exerted by the constraint. For frictionless collisions, f_t^{con} plays the same role as the coefficient of restitution from the measure differential inclusion formulation of contact dynamics [27, 56].

3. Discrete model. We now turn to considering discrete models of contact problems, in which the continuous time variable is replaced with a discrete time index. The equations of motion are thus algebraic rather than differential equations, and they can be regarded as an integrator for the continuous system.

The approach we use is based on discrete variational mechanics (see [37]), in which the variational principle is discretized and the discrete equations and their conservation properties are derived as in the continuous case. This has the advantage of automatically capturing much of the geometric structure of the true problem even in the approximate discrete setting.

3.1. Discrete configurations and equations of motion. Disregard for the moment the continuous formulation of the previous section, and introduce a fixed *timestep* $h \in \mathbb{R}$. Consider a *discrete Lagrangian* $L_d : Q \times Q \to \mathbb{R}$, which is a function of two configuration points and the timestep, so that $L_d = L_d(q_0, q_1, h)$. The discrete Lagrangian will be chosen to approximate the continuous action integral over an interval of length h so that

$$L_d(q_0, q_1, h) \approx \int_0^h L(q, \dot{q}) dt,$$

where $q : [0, h] \to \mathbb{R}$ is an exact solution of the Euler–Lagrange equations, for L satisfies the boundary conditions $q(0) = q_0$ and $q(h) = q_1$.

We now consider an increasing sequence of times

$$t_k = kh \text{ for } k = 0, \dots, N,$$

we also fix $\tilde{\alpha} \in [0, 1]$, and we let $\tilde{\tau} = t_{i-1} + \tilde{\alpha} h$ denote the fixed impact time (corresponding to τ_i from the continuous model) and $\tilde{t} = t_{i-1} + \alpha h$ denote the actual impact time (corresponding to t_i). We take $\alpha = t_d(\tilde{\alpha})$, where t_d is some strictly increasing function which maps $[0, 1]$ onto $[0, 1]$. Thus we assumed only that the step at which the impact occurs is known and not the impact time $\tilde{t}$, which is allowed to vary according to variations in α.

The *discrete path space* is defined by

$$\mathcal{M}_d = \mathcal{T}_d \times \mathcal{Q}_d(\tilde{\alpha}, \partial C, Q), \tag{40}$$

where

$$\mathcal{T}_d = \{t_d(\tilde{\alpha}) \mid t_d \in C^\infty([0, 1], [0, 1]), t_d \text{ onto}, t_d' > 0 \text{ in } [0, 1]\}, \tag{41}$$

$$\mathcal{Q}_d(\tilde{\alpha}, \partial C, Q) = \{q_d : \{t_0, \dots, t_{i-1}, \tilde{\tau}, t_i, \dots, t_N\} \to Q, q_d(\tilde{\tau}) \in \partial C\}. \tag{42}$$

Remark. The set $\mathcal{T}_d$ is actually the real interval $[0,1]$, but we used (41) to define it in order to emphasize the analogy with the continuous case.

We identify the discrete trajectory with its image

$$(\alpha, q_d) = (\alpha, \{q_0, \ldots, q_{i-1}, \tilde{q}, q_i, \ldots, q_N\}),$$

where $q_k = q_d(t_k)$ for $k \in \{0, \ldots, N\}$, $\tilde{q} = q_d(\tilde{\tau})$, and $\alpha = t_d(\tilde{\alpha})$. Thus a discrete trajectory can be regarded as a sequence of points in Q, one of which must be in ∂C, and a single real number $\alpha \in [0,1]$.

The *discrete action map* $\mathfrak{G}_d : \mathcal{M}_d \to \mathbb{R}$ is defined by

$$\mathfrak{G}_d(\alpha, q_d) = \sum_{k=0}^{i-2} L_d(q_k, q_{k+1}, h) + \sum_{k=i}^{N-1} L_d(q_k, q_{k+1}, h) + L_d(q_{i-1}, \tilde{q}, \alpha h) + L_d(\tilde{q}, q_i, (1-\alpha)h). \tag{43}$$

As the discrete path space $\mathcal{M}_d$ is isomorphic to $[0,1] \times Q \times \cdots \times \partial C \times \cdots \times Q$ (N copies of Q), it can be given a smooth manifold structure.

For $q_d \in \mathcal{Q}_d(\tilde{\alpha}, \partial C, Q)$, the tangent space $T_{q_d}\mathcal{Q}_d(\tilde{\alpha}, \partial C, Q)$ is the set of all maps $v_{q_d} : \{t_0, \ldots, t_{i-1}, \tilde{\tau}, t_i, \ldots, t_N\} \to TQ$ such that $\pi_Q \circ v_{q_d} = q_d$ and $v_{q_d}(\tilde{\tau}) \in T_{\tilde{q}}\partial C$. For simplicity we will identify v_{q_d} with its image in TQ.

The tangent space to the full discrete path space is now $T\mathcal{M}_d = T\mathcal{T}_d \times T\mathcal{Q}_d$. At a given point $(\alpha, q_d) \in \mathcal{M}_d$ we will write a tangent vector in $T_{(\alpha,q_d)}\mathcal{M}_d$ as

$$(\delta\alpha, \delta q_d) = (\delta\alpha, \{\delta q_0, \ldots, \delta q_{i-1}, \delta\tilde{q}, \delta q_i, \ldots, \delta q_N\}).$$

Define the *discrete second order manifold* to be

$$\ddot{Q}_d = Q \times Q \times Q,$$

which has the same information content as the continuous second order manifold $\ddot{Q}$.

We now proceed, as in the continuous case, to derive the discrete equations of motion and the conservation laws from Hamilton's principle of critical action. We take variations of the discrete action sum with respect to the discrete path and to the parameter α, as stated in the following theorem.

Theorem 3.1. *Given a C^k discrete Lagrangian $L_d : Q \times Q \times \mathbb{R} \to \mathbb{R}$, $k \geq 1$, there exist a unique C^{k-1} mapping $EL_d : \ddot{Q}_d \to T^*Q$ and unique C^{k-1} one-forms $\Theta^-_{L_d}$ and $\Theta^+_{L_d}$ on the discrete Lagrangian phase space $Q \times Q$ such that, for all variations $(\delta\alpha, \delta q_d) \in T_{(\alpha,q_d)}\mathcal{M}_d$ of (α, q_d), we have*

$$\begin{aligned}
&d\mathfrak{G}_d(\alpha, q_d) \cdot (\delta\alpha, \delta q_d) \\
&= \sum_{k=1}^{i-2} EL_d(q_{k-1}, q_k, q_{k+1}) \cdot \delta q_k + \sum_{k=i+1}^{N-1} EL_d(q_{k-1}, q_k, q_{k+1}) \cdot \delta q_k \\
&\quad + \Theta^+_{L_d}(q_{N-1}, q_N) \cdot (\delta q_{N-1}, \delta q_N) - \Theta^-_{L_d}(q_0, q_1) \cdot (\delta q_0, \delta q_1) \\
&\quad + [D_2 L_d(q_{i-2}, q_{i-1}, h) + D_1 L_d(q_{i-1}, \tilde{q}, \alpha h)] \cdot \delta q_{i-1} \\
&\quad + h\,[D_3 L_d(q_{i-1}, \tilde{q}, \alpha h) - D_3 L_d(\tilde{q}, q_i, (1-\alpha)h)] \cdot \delta\alpha \\
&\quad + i^*(D_2 L_d(q_{i-1}, \tilde{q}, \alpha h) + D_1 L_d(\tilde{q}, q_i, (1-\alpha)h)) \cdot \delta\tilde{q} \\
&\quad + [D_2 L_d(\tilde{q}, q_i, (1-\alpha)h) + D_1 L_d(q_i, q_{i+1}, h)] \cdot \delta q_i,
\end{aligned} \tag{44}$$

where $i^* : T^*Q \to T^*\partial C$ is the cotangent lift of the embedding $i : \partial C \to Q$.

The map EL_d is called the discrete Euler–Lagrange derivative *and the one-forms $\Theta^+_{L_d}$ and $\Theta^-_{L_d}$ are the* discrete Lagrangian one-forms. *In coordinates these have the expressions*

$$EL_d(q_{k-1}, q_k, q_{k+1}) = [D_2L_d(q_{k-1}, q_k, h) + D_1L_d(q_k, q_{k+1}, h)]\, dq_k \tag{45}$$

for $k \in \{1, \dots, i-2, i, \dots, N-1\}$ and

$$\Theta^+_{L_d}(q_k, q_{k+1}) = D_2L_d(q_k, q_{k+1}, h)\, dq_{k+1},$$
$$\Theta^-_{L_d}(q_k, q_{k+1}) = -D_1L_d(q_k, q_{k+1}, h)\, dq_k.$$

Proof. The formula is derived by straightforward algebra, by computing the derivative of the discrete action map, and by some rearrangement of the summation. This rearrangement corresponds to a discrete version of integration by parts, resulting in two boundary terms which are interpreted as the discrete Lagrangian one-forms. ■

By using the discrete version of Hamilton's principle, we consider the paths (α, q_d), which are critical points of the discrete action. Therefore, we define the *discrete space of solutions* to be the set of all paths which satisfy $d\mathfrak{G}_d(\alpha, q) \cdot (\delta\alpha, \delta q) = 0$ for all variations $(\delta\alpha, \delta q_d) \in T_{(\alpha,q_d)}\mathcal{M}_d$ which are zero at the boundary points 0 and N.

From (44) we conclude that (α, q_d) is a solution iff the discrete Euler–Lagrange derivative is zero at all k other than $\{0, i-1, i, N\}$. This statement at an arbitrary k reads

$$D_2L_d(q_{k-1}, q_k, h) + D_1L_d(q_k, q_{k+1}, h) = 0 \tag{46}$$

and is known as *discrete Euler–Lagrange equations*. These describe the motion of the system away from the impact point by implicitly defining a map $(q_{k-1}, q_k) \mapsto (q_k, q_{k+1})$.

Near the point of impact, the discrete Hamilton's principle gives three additional sets of equations, namely,

$$D_2L_d(q_{i-2}, q_{i-1}, h) + D_1L_d(q_{i-1}, \tilde{q}, \alpha h) = 0, \tag{47a}$$
$$\tilde{q} \in \partial C, \tag{47b}$$

which is a system of $n+1$ equations to be solved for $\tilde{q}$ and α, and

$$D_3L_d(q_{i-1}, \tilde{q}, \alpha h) - D_3L_d(\tilde{q}, q_i, (1-\alpha)h) = 0, \tag{48a}$$
$$i^*(D_2L_d(q_{i-1}, \tilde{q}, \alpha h) + D_1L_d(\tilde{q}, q_i, (1-\alpha)h)) = 0, \tag{48b}$$

which is a system of n equations for the unknown q_i. Finally, we also have

$$D_2L_d(\tilde{q}, q_i, (1-\alpha)h) + D_1L_d(q_i, q_{i+1}, h) = 0, \tag{49}$$

which gives n equations to be solved for q_{i+1}.

A discrete trajectory can thus be formed by starting from an initial condition (q_0, q_1), using (46) to solve successively for the q_k until the impact time is reached, and then solving the systems (47), (48), and (49) in turn to obtain $\tilde{q}$, α and then q_i and q_{i+1}, before once again continuing with (46) to complete the trajectory.

Remark. The discrete energy conservation through the collision depends critically on exactly resolving the collision time with the parameter α. This is also the key feature of the recent improvements of [41] to the nonsmooth collision methods developed by [25].

3.2. Relationship between discrete and continuous models. Having established the basic discrete variational mechanics, we now consider how the discrete model can be regarded as an approximation to the continuous model.

At first glance it appears that the discrete Euler–Lagrange equations are defined only in terms of pairs of configuration positions. We will now see, however, that they can also be interpreted as defining a mapping on the cotangent bundle T^*Q. Define the *discrete Legendre transforms* or *discrete fiber derivatives* $\mathbb{F}^+L_d, \mathbb{F}^-L_d : Q \times Q \to T^*Q$ as given by

$$\begin{aligned}\mathbb{F}^+L_d(q_0, q_1) \cdot \delta q_1 &= D_2L_d(q_0, q_1, h) \cdot \delta q_1,\\ \mathbb{F}^-L_d(q_0, q_1) \cdot \delta q_0 &= -D_1L_d(q_0, q_1, h) \cdot \delta q_0,\end{aligned}$$

where h is the timestep in between q_0 and q_1. We note the implicit dependence on the timestep of the definition above. This dependence is completely neglected in the constant timestep discrete variational mechanics or rigorously treated in the nonautonomous setting (using adaptive timesteps), and we refer to [37] for a complete account of these ideas.

These also can be written

$$\mathbb{F}^+L_d : (q_0, q_1) \mapsto (q_1, p_1) = (q_1, D_2L_d(q_0, q_1, h)), \tag{50a}$$

$$\mathbb{F}^-L_d : (q_0, q_1) \mapsto (q_0, p_0) = (q_0, -D_1L_d(q_0, q_1, h)). \tag{50b}$$

If both discrete fiber derivatives are locally isomorphisms, then we say that L_d is *regular*. We will generally assume that we are working with regular discrete Lagrangians.

We introduce the notation

$$p^+_{k,k+1} = p^+(q_k, q_{k+1}, h) = \mathbb{F}^+L_d(q_k, q_{k+1}), \tag{51a}$$

$$p^-_{k,k+1} = p^-(q_k, q_{k+1}, h) = \mathbb{F}^-L_d(q_k, q_{k+1}) \tag{51b}$$

for the momentum at the two endpoints of each interval $[k, k + 1]$.

We can now use definitions (50a) and (51) of the discrete fiber derivatives and of the discrete momenta to see that the discrete Euler–Lagrange equations (46) can be written as

$$\mathbb{F}^+L_d(q_{k-1}, q_k) = \mathbb{F}^-L_d(q_k, q_{k+1}) \tag{52}$$

or simply

$$p^+_{k-1,k} = p^-_{k,k+1}. \tag{53}$$

That is, the discrete Euler–Lagrange equations enforce the condition that the momentum at time k should be the same when evaluated from the lower interval $[k - 1, k]$ or the upper interval $[k, k + 1]$.

In this interpretation, (48b) represents conservation of the projection of momentum (by i^*, on $T^*\partial C$) at the moment of impact

$$i^*p^+(q_{i-1}, \tilde{q}, \alpha h) = i^*p^-(\tilde{q}, q_i, (1 - \alpha)h), \tag{54}$$

which is a discrete version of the jump condition (14) from the continuous case.

To give an interpretation of the discrete equations around the impact time, we define the *discrete energy* to be

$$E_d(q_k, q_{k+1}, h) = -D_3 L_d(q_k, q_{k+1}, h). \tag{55}$$

Using this, we can write (48a) as

$$E_d(q_{i-1}, \tilde{q}, \alpha h) = E_d(\tilde{q}, q_i, (1-\alpha)h), \tag{56}$$

so this equation simply represents conservation of discrete energy at the impact time, a discrete analogue of (15).

Remark. The discrete energy defined in this way is used in [24] and can be motivated in several ways; first, for Lagrangians of the form of kinetic minus potential energy, and with the choice of discrete Lagrangians given by

$$L_d(q_0, q_1, h) = L\left(\gamma q_0 + (1-\gamma)q_1, \frac{q_1 - q_0}{h}\right), \tag{57}$$

where $\gamma \in [0, 1]$ is an interpolation parameter, the discrete energy gets the usual expression

$$E_d(q_0, q_1, h) = \frac{1}{2}\left(\frac{q_1 - q_0}{h}\right)^T M \left(\frac{q_1 - q_0}{h}\right) + V(\gamma q_0 + (1-\gamma)q_1). \tag{58}$$

A second motivation is the fact that the discrete energy becomes exactly the Hamiltonian when one uses the exact discrete Lagrangian L_d^E—that is, the discrete Lagrangian is equal to the action integral taken along exact solutions of the Euler–Lagrange equations.

3.3. Symplecticity of the flow. Define the *discrete Lagrangian map* $F_{L_d} : Q \times Q \to Q \times Q$ by

$$(q_0, q_1) \mapsto (q_1, q_2), \tag{59}$$

where q_2 is obtained by using the algorithm from section 3.1. A solution $(\alpha, q_d) \in \mathcal{M}_d$ is formed by iteration of the map F_{L_d}, and it is uniquely determined by the initial condition $(q_0, q_1) \in Q \times Q$ and the choice of timestep h. Hence we parameterize the discrete solutions of the variational principle by the initial conditions (q_0, q_1), and we consider the restriction of $\mathfrak{G}_d$ to that solution space.

The discrete fiber derivatives enable us to push the discrete Lagrangian map $F_{L_d} : Q \times Q \to Q \times Q$ forward to T^*Q. We define the *discrete Hamiltonian map* $\tilde{F}_{L_d} : T^*Q \to T^*Q$ by

$$\tilde{F}_{L_d} = \mathbb{F}^+ L_d \circ F_{L_d} \circ (\mathbb{F}^+ L_d)^{-1}, \tag{60}$$

with the coordinate expression

$$\tilde{F}_{L_d} : (q_0, p_0) \mapsto (q_1, p_1). \tag{61}$$

We note that the discrete Hamiltonian map can be equivalently defined using the other discrete Legendre transform

$$\tilde{F}_{L_d} = \mathbb{F}^- L_d \circ F_{L_d} \circ (\mathbb{F}^- L_d)^{-1}. \tag{62}$$

Define the *restricted discrete action map* $\hat{\mathfrak{G}}_d : Q \times Q \to \mathbb{R}$ to be $\hat{\mathfrak{G}}_d(q_0, q_1) = \mathfrak{G}_d(\alpha, q_d)$, where (α, q_d) is the corresponding solution in $\mathcal{M}_d$ such that $(q_d(t_0), q_d(t_1)) = (q_0, q_1)$. Then (44) becomes

$$d\hat{\mathfrak{G}}_d = (F^N_{L_d})^* \Theta^+_{L_d} - \Theta^-_{L_d}. \tag{63}$$

Taking a further derivative of this expression and using the fact that $d^2\hat{\mathfrak{G}}_d = 0$, we obtain

$$(F^N_{L_d})^*(\Omega_{L_d}) = \Omega_{L_d}, \tag{64}$$

where $\Omega_{L_d} = d\Theta^+_{L_d} = d\Theta^-_{L_d}$ is the unique *discrete Lagrangian symplectic form*, with coordinate expression

$$\Omega_{L_d}(q_0, q_1) = \frac{\partial^2 L_d}{\partial q_0^i \partial q_1^j} dq_0^i \wedge dq_1^j. \tag{65}$$

We have thus proven that the discrete evolution map exactly preserves a discrete symplectic structure, so, regarding F_d as an integrator for the continuous system, we see that it is automatically a symplectic method.

Note that the discrete Lagrangian symplectic form is the pullback under either discrete Legendre transform of the canonical symplectic form on T^*Q. The discrete Hamiltonian map $\tilde{F}_{L_d} : T^*Q \to T^*Q$ thus preserves the canonical symplectic form and the canonical momentum maps on T^*Q.

3.4. Discrete Noether theorem. Consider the (left or right) group action $\Phi : G \times Q \to Q$ of a Lie group G on Q, with infinitesimal generator as defined in section 2.4. This action can be lifted to $Q \times Q$ by the product $\Phi_g(q_0, q_1) = (\Phi_g(q_0), \Phi_g(q_1))$, which has the *infinitesimal generator* $\xi_{Q\times Q} : Q \times Q \to T(Q \times Q)$ given by

$$\xi_{Q\times Q}(q_0, q_1) = (\xi_Q(q_0), \xi_Q(q_1)). \tag{66}$$

The two *discrete Lagrangian momentum maps* $J^+_{L_d}, J^-_{L_d} : Q \times Q \to g^*$ are

$$J^+_{L_d}(q_0, q_1) \cdot \xi = \Theta^+_{L_d} \cdot \xi_{Q\times Q}(q_0, q_1), \tag{67a}$$

$$J^-_{L_d}(q_0, q_1) \cdot \xi = \Theta^-_{L_d} \cdot \xi_{Q\times Q}(q_0, q_1). \tag{67b}$$

As in the continuous approach to Noether's theorem from section 2.4, we are restricted to symmetries of the configuration variables only. We consider symmetries which do not involve altering the time variable and thus consider the timestep h to be a fixed constant.

If a discrete Lagrangian $L_d : Q \times Q \to \mathbb{R}$ is such that $dL_d \cdot \xi = 0$, then L_d is said to be *infinitesimally invariant* under the group action, and Φ is said to be a *symmetry* of the discrete Lagrangian. Note that

$$dL_d \cdot \xi = (\Theta^+_{L_d} - \Theta^-_{L_d}) \cdot \xi_{Q\times Q},$$

and so, when L_d is infinitesimally invariant under the group action Φ, the two discrete momentum maps are equal. In such cases, we will use the notation $J_{L_d} : Q \times Q \to \mathfrak{g}^*$ for the unique single *discrete Lagrangian momentum map*.

Theorem 3.2 (discrete Noether's theorem). *Consider a discrete Lagrangian system $L_d : Q \times Q \times \mathbb{R} \to \mathbb{R}$ which is infinitesimally invariant under the lift of the (left or right) action $\Phi : G \times Q \to Q$. If we assume that the action leaves ∂C invariant (locally), then the corresponding discrete Lagrangian momentum map $J_{L_d} : Q \times Q \to \mathfrak{g}^*$ is a conserved quantity of the discrete Lagrangian map $F_{L_d} : Q \times Q \to Q \times Q$ so that $J_{L_d} \circ F_{L_d} = J_{L_d}$.*

Proof. We introduce an action of G on the discrete path space $\mathcal{M}_d$ by pointwise action on the configuration components so that $\Phi_g : \mathcal{M}_d \to \mathcal{M}_d$ is given by $\Phi_g(\alpha, q_d) = (\alpha, \Phi_g(q_d))$. Then the infinitesimal generator $\xi_{\mathcal{M}_d} : \mathcal{M}_d \to T\mathcal{M}_d$ is given by

$$\xi_{\mathcal{M}_d}(\alpha, q_d) = (0, \xi_Q(q_0), \dots, \xi_Q(q_{i-1}), \xi_Q(\tilde{q}), \xi_Q(q_i), \dots, \xi_Q(q_N)).$$

From (43) we derive

$$d\mathfrak{G}_d(\alpha, q_d) \cdot \xi_{\mathcal{M}_d}(\alpha, q_d) = \sum_{k=0}^{N-1} dL_d \cdot \xi, \tag{68}$$

and so the space of solutions of the discrete Euler–Lagrange equations is invariant under the action of G, and the Lagrangian map $F_{L_d} : Q \times Q \to Q \times Q$ commutes with the lifted action $\Phi_g : Q \times Q \to Q \times Q$.

Identifying the space of solutions with the space of initial conditions $Q \times Q$ and using (63), we obtain

$$\begin{aligned} d\mathfrak{G}_d(\alpha, q_d) \cdot \xi_{\mathcal{M}_d}(\alpha, q_d) &= d\hat{\mathfrak{G}}_d(q_0, q_1) \cdot \xi_{Q\times Q}(q_0, q_1) \\ &= ((F_{L_d}^N)^*(\Theta_{L_d}^+) - \Theta_{L_d}^-)(q_0, q_1) \cdot \xi_{Q\times Q}(q_0, q_1). \end{aligned}$$

From (68) and the invariance of the discrete Lagrangian, the left-hand side of the previous equation is zero, and so we have

$$(\Theta_{L_d}^+ \cdot \xi_{Q\times Q}) \circ F_{L_d}^N = \Theta_{L_d}^- \cdot \xi_{Q\times Q}. \tag{69}$$

The last relation is simply the statement of preservation of the discrete momentum map, given that for symmetry actions there is only a single unique discrete momentum map and that the above argument holds for all subintervals, including a single timestep. ∎

Observe that J_{L_d} is the pullback under $\mathbb{F}^{\pm}L_d$ of the canonical momentum map J_H on T^*Q and that J_H is thus preserved by $\tilde{F}_{L_d}$.

4. Numerical examples.

In this section, we will choose a particular discrete Lagrangian and illustrate the performance of the algorithm from the previous section on two simple conservative systems. Here we are particularly interested in the extent to which the variational integrator preserves the energy for very long time simulations.

The examples that we present very much simplify the issues regarding grazing impacts and multiple nearby solutions, such as one would encounter in complex collisions (simulation studies of powder flows, for example). Our algorithm, as presented in this paper, is limited to

relatively simple situations, when one can readily identify and resolve the impacts. However, considerable progress has already been made in extending these methods to more practical schemes which are demonstrated in examples involving very complicated collision sequences (see [9]).

4.1. The discrete algorithm. For systems of the form

$$L(q,\dot{q}) = \frac{1}{2}\dot{q}^T M \dot{q} - V(q), \tag{70}$$

where M is a mass matrix and V is a potential function, the Euler–Lagrange equations are given by

$$M\ddot{q} = -\nabla V(q),$$

which is simply Newton's equation of mass times acceleration equals force. We consider the second order discrete Lagrangian

$$L_d(q_0,q_1,h) = \frac{h}{2}\left(\frac{q_1-q_0}{h}\right)^T M\left(\frac{q_1-q_0}{h}\right) - h\left(\frac{V(q_0)+V(q_1)}{2}\right), \tag{71}$$

which is clearly an approximation to the action integral over an interval of length h. The discrete energy function for this choice of discrete Lagrangian is

$$E_d(q_0,q_1,h) = \frac{1}{2}\left(\frac{q_1-q_0}{h}\right)^T M\left(\frac{q_1-q_0}{h}\right) + \left(\frac{V(q_0)+V(q_1)}{2}\right), \tag{72}$$

and the discrete Euler–Lagrange equations are

$$M\frac{q_{k+2}-2q_{k+1}+q_k}{h^2} = -\nabla V(q_{k+1}). \tag{73}$$

Using the discrete Legendre transform (50a), we can push this algorithm forward on T^*Q and obtain a map

$$(q_k,p_k) \mapsto (q_{k+1},p_{k+1})$$

given by

$$\begin{aligned} q_{k+1} &= q_k + hM^{-1}p_k - \frac{h^2}{2}M^{-1}\nabla V(q_k), \\ p_{k+1} &= p_k - \frac{h}{2}\left(\nabla V(q_k) + \nabla V(q_{k+1})\right). \end{aligned}$$

The integrator defined by the previous set of equations is called the leap-frog/Verlet integrator and is one of the most popular integration schemes in molecular dynamics. It is a second order accurate integrator, as one can also infer from the fact that the discrete Lagrangian is second order (see [37] for details about this theory).

This equation describes the motion of the discrete system away from the point of impact. Given a point (q_{i-1}, p_{i-1}) just before impact, we must then solve (47) for $\tilde{q}$ and α, which are

$$M\frac{\tilde{q}-q_{i-1}}{\alpha h} - M\frac{q_{i-1}-q_{i-2}}{h} + (1+\alpha)\frac{h}{2}\nabla V(q_{i-1}) = 0, \tag{74a}$$

$$\tilde{q} \in \partial C. \tag{74b}$$

Next we solve (48) for q_i, which reads

$$\frac{1}{2}\left(\frac{q_i-\tilde{q}}{(1-\alpha)h}\right)^T M\left(\frac{q_i-\tilde{q}}{(1-\alpha)h}\right) - \frac{1}{2}\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right)^T M\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right)$$

$$+\frac{1}{2}(V(q_i)-V(q_{i-1})) = 0, \tag{75a}$$

$$i^*\left(M\frac{q_i-\tilde{q}}{(1-\alpha)h} - M\frac{\tilde{q}-q_{i-1}}{\alpha h} + \frac{h}{2}\nabla V(\tilde{q})\right) = 0. \tag{75b}$$

To implement the system (75), we write (75b) in a form using Lagrange multipliers. More precisely, we consider ∂C to have a local representation $\partial C = \phi^{-1}(0) \subset Q$, where 0 is a regular point of the constraint function $\phi : Q \to \mathbb{R}$. Then we solve (75a) together with the system

$$M\frac{q_i-\tilde{q}}{(1-\alpha)h} - M\frac{\tilde{q}-q_{i-1}}{\alpha h} + \frac{h}{2}\nabla V(\tilde{q}) + \lambda\nabla\phi(\tilde{q}) = 0 \tag{76}$$

for the unknowns $q_i \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.

Finally, we solve for q_{i+1} by (49), which is

$$M\frac{q_{i+1}-q_i}{h} - M\frac{q_i-\tilde{q}}{(1-\alpha)h} + (2-\alpha)\frac{h}{2}\nabla V(q_i) = 0, \tag{77}$$

and we then continue integrating with (73) above.

We can also handle multiple impacts within a single timestep by dividing the impact step into as many substeps as we need and solving (74) and (75) sequentially for any constraint involved in the impact. We will explicitly derive the equations for the case of two impacts solved within the timestep (t_{i-1}, t_i); generalization to an arbitrary number would be immediate.

If multiple impacts are realized in the timestep (t_{i-1}, t_i), then the system (75) will return a solution q_i which is not admissible ($q_i \notin C$). Let us assume that there is only one additional impact in the subinterval $(\tilde{t}, t_i)$ which occurs at the contact point $\tilde{q}' \in \partial C$ and time $\tilde{t}' = \tilde{t} + \beta h$, with $0 < \beta \leq 1-\alpha$. Then the conservation of the discrete energy and the momentum at the impact point $\tilde{q}$, in addition to the condition that $\tilde{q}'$ must lie on ∂C, give the system of $n+1$ equations

$$\frac{1}{2}\left(\frac{\tilde{q}'-\tilde{q}}{\beta h}\right)^T M\left(\frac{\tilde{q}'-\tilde{q}}{\beta h}\right) - \frac{1}{2}\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right)^T M\left(\frac{\tilde{q}-q_{i-1}}{\alpha h}\right)$$

$$+\frac{1}{2}(V(\tilde{q}')-V(q_{i-1})) = 0, \tag{78a}$$

$$i^*\left(M\frac{\tilde{q}'-\tilde{q}}{\beta h} - M\frac{\tilde{q}-q_{i-1}}{\alpha h} + (\alpha+\beta)\frac{h}{2}\nabla V(\tilde{q})\right) = 0, \tag{78b}$$

$$\tilde{q}' \in \partial C, \tag{78c}$$

to be solved for $\tilde{q}'$ and β.

Next, the analogous versions of (78a) and (78b) for the second impact point $\tilde{q}'$ give n equations for q_i:

$$\frac{1}{2}\left(\frac{q_i-\tilde{q}'}{(1-\alpha-\beta)h}\right)^T M\left(\frac{q_i-\tilde{q}'}{(1-\alpha-\beta)h}\right)-\frac{1}{2}\left(\frac{\tilde{q}'-\tilde{q}}{\beta h}\right)^T M\left(\frac{\tilde{q}'-\tilde{q}}{\beta h}\right)$$

$$+\frac{1}{2}(V(q_i)-V(\tilde{q}))=0, \tag{79a}$$

$$i^*\left(M\frac{q_i-\tilde{q}'}{(1-\alpha-\beta)h}-M\frac{\tilde{q}'-\tilde{q}}{\beta h}+(1-\alpha)\frac{h}{2}\nabla V(\tilde{q}')\right)=0. \tag{79b}$$

The case of an arbitrary number of impacts is treated in a similar manner, by dividing the timestep (t_{i-1}, t_i) into as many substeps as needed and sequentially solving systems of type (78) to find all the contact points and times. Finally, we solve a system of type (79) for q_i, and then we revert to the standard discrete Euler–Lagrange equations to continue away from the impact.

In the numerical examples, we solve the implicit sets of (74) and (75) with nested Newton loops.

4.2. Particle colliding with a rigid surface. The first example we consider consists of a particle with unit mass moving under gravity in the (x, y)-plane and successively colliding and bouncing on a horizontal rigid floor located at $y = 0$. This simple system has two degrees of freedom (the coordinates of the particle) $q = (x, y)$, the configuration manifold is $Q = \mathbb{R}^2$, and the contact submanifold ∂C is the line $y = 0$. The particle moves with trajectory $q(t) \in \mathbb{R}^2$ in the admissible set $y \geq 0$.

The Lagrangian describing this problem is in the form (70), where M is the diagonal 2×2 mass matrix with diagonal elements (m, m) (m denotes the mass of the particle) and V is the gravitational potential given by

$$V(q) = mgy. \tag{80}$$

Here, g denotes the gravitational acceleration.

The discretization we use is (71), the one for which the variational collision integrator was explicitly derived in the last subsection. The integrator is run with a step size of $h = 0.01$; the initial conditions we used in the simulation are $q_0 = (0, 1)$ and $\dot{q}_0 = (-2, 0)$. We considered a unitary mass particle ($m = 1$).

The energy behavior in this case is shown in Figure 2 for a relatively large number of impacts (1000 impacts). The same pattern is observed if the simulation is carried out for essentially arbitrarily long times. This fluctuating energy behavior is typical of symplectic methods. A detailed account on how the variational symplectic methods perform on smooth conservative systems can be found in [26].

4.3. Rotating nonconvex rigid body colliding with a rigid surface. Now consider a sequence of collisions and bounces on a horizontal rigid floor for a three-degree-of-freedom system, namely, a rotating four-point star-shaped rigid body (see Figure 3) moving in a plane. The convex hull of the star-shaped body is a square with sides of length L. The rigid

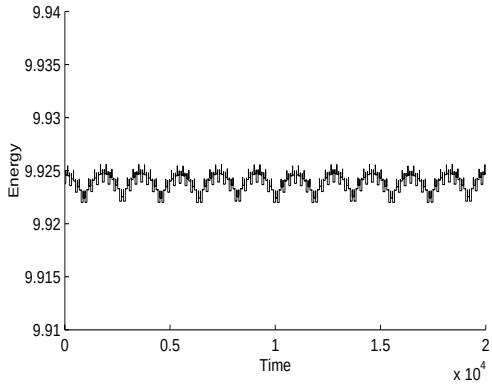


Figure 2. *The long-time energy behavior for a particle bouncing on a rigid floor.*

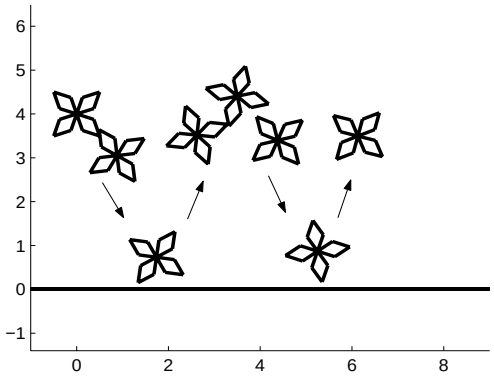


Figure 3. *A rotating four point, star shaped rigid body colliding and bouncing on a horizontal rigid floor.*

body moves under the gravitational force field in the vertical (x, y) plane. The configuration manifold Q is $SE(2)$ with local coordinates $q = (x, y, \theta)$, where $(x, y) \in \mathbb{R}^2$ stand for the coordinates of the center of mass and $\theta \in [0, 2\pi]$ stands for the oriented angle that a line moving rigidly with the body makes with the horizontal axis. The contact set ∂C given by the nonpenetration condition is given explicitly by

$$y = \frac{L}{2}\left(|\sin\theta| + |\cos\theta|\right). \tag{81}$$

The subset of points where $y \geq \frac{L}{2}\left(|\sin\theta| + |\cos\theta|\right)$ represents the admissible set $C \subset Q$, and contact occurs whenever the relation becomes an equality.

The Lagrangian describing this problem has the expression (70), where V is the gravitational potential (80) and M is the diagonal 3×3 mass matrix with diagonal elements (m, m, I), where m is the mass of the body and I is the moment of inertia of the star-shaped body with respect to the z-axis through its center of symmetry. In terms of m and L, I is given by $I = \frac{29}{192}mL^2$.

We use again the discretization given by (71) and run the variational collision integrator from section 4.1 with a timestep $h = 0.005$ and initial conditions $q_0 = (0, 3.5, 0)$ and $\dot{q}_0 =$

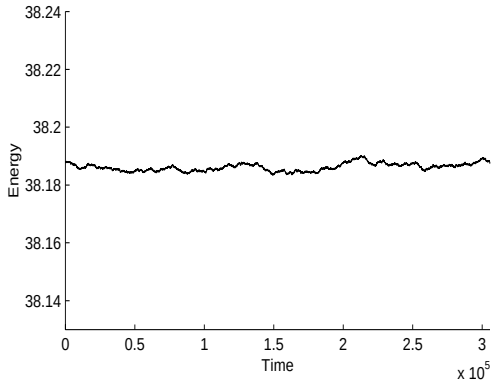


Figure 4. *The long-time energy behavior for a star-shaped rigid body bouncing on a rigid floor. Note the fluctuating energy behavior typical of symplectic methods.*

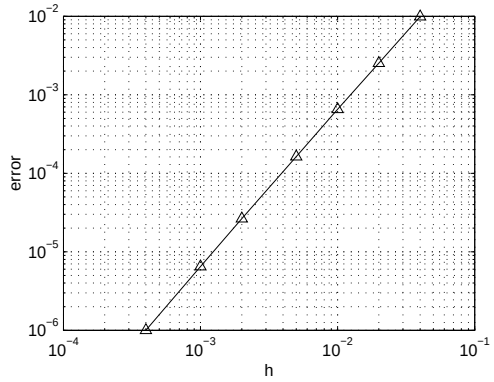


Figure 5. *Log-log error diagram for the method in the star bounce example, after one collision. The second order accuracy of the integrator from the smooth setting is preserved through collision. In this case a nonlinear gravity was used to avoid the degeneracy of the exactly integrable linear gravitational potential.*

$(-2, 0, 5)$. We considered the body to have unitary mass $m = 1$ and a square convex hull of size $L = 1$. A long-time (1,500 impacts) energy plot is shown in Figure 4. The long-time energy behavior appears to be reasonably stable. It is not clear from these numerical experiments whether this is an indication of a nearby conserved energy, as exists for variational integrators applied to smooth systems, or simply a fairly stable random walk. More numerical investigations and analytical work are needed to resolve this question.

We numerically checked the order of accuracy of the algorithm for the star bounce example. A log-log error diagram after one collision is presented in Figure 5. The numerical results show that the integrator is second order accurate; i.e., the order of the method is the same as the order of the discrete Lagrangian L_d. This is in fact a fundamental property of the variational integrators developed in smooth settings (see [37]), and we believe that it extends to the nonsmooth setting as well. In our future work on the subject, we intend to formulate and prove such results for the variational collisional algorithms presented in this paper.

5. Appendix: Nonsmooth analysis approach. The purpose of this appendix is to discuss collisions with multibody nonsmooth contact geometries when the contact set ∂C has a large number of singularities. For example, in granular flows or fragmentation of brittle solids, there are a large number of fragments undergoing complex collision sequences. For these collisions, situations like corner-to-corner contact are very likely to occur, and the variational algorithm from section 3.1 cannot cope with contact in singular points of the contact set ∂C.

However, the nonsmooth analysis (see [10]) provides an efficient analytical tool to formulate and treat algorithmically complex contact situations, as shown in [25]. The goal of this appendix is to combine discrete Lagrangian mechanics with nonsmooth calculus to derive a variational formulation of the nonsmooth contact (in the sense of nonsmooth admissible configuration sets). The symplectic nature of such an algorithm is poorly understood, but one can conjecture that future theory on that would depend on approaches like this one.

If C is the admissible set (possibly nonsmooth and nonconvex) of the system, we must have $q(t) \in C$ for all times or, in the discrete case, $q_k \in C$ for all k. These constraints may be enforced by adding to the Lagrangian the indicator function I_C of C defined by

$$I_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{otherwise.} \end{cases} \tag{82}$$

In the discrete context, this translates into defining a constrained discrete Lagrangian $\tilde{L}_d$ by adding contributions from the indicator function. One particular way to do this is

$$\tilde{L}_d(q_k, q_{k+1}, h) = L_d(q_k, q_{k+1}, h) - \frac{1}{2}\left[I_C(q_k) + I_C(q_{k+1})\right]. \tag{83}$$

We use, as before, the variational principle of Hamilton to derive the discrete equations of motions. Thus the discrete Euler–Lagrange equations become

$$D_2L_d(q_{k-1}, q_k, h) + D_1L_d(q_k, q_{k+1}, h) - \partial I_C(q_k) \ni 0, \tag{84}$$

where ∂I_C denotes the generalized gradient of the indicator function.

For points q in the interior of C, $\partial I_C(q) = \{0\}$, while for points on the boundary of C, $\partial I_C(q) = N_C(q)$, where $N_C(q)$ represents the normal cone to C at q defined in the nonsmooth analysis framework (see [10] for a complete account of the nonsmooth calculus used here). However, if q is a convex point, $N_C(q)$ reduces to the normal cone in the usual convex analysis sense.

The constrained discrete equations (84) are thus the usual discrete Euler–Lagrange equations (46) away from the impact. The generalized gradient ∂I_C is not trivial only for $\tilde{q} \in \partial C$. If we specialize (84) for points $q_{i-1}, \tilde{q}$, and q_{i+1}, then we obtain

$$D_2L_d(q_{i-1}, \tilde{q}, \alpha h) + D_1L_d(\tilde{q}, q_{i+1}, (1-\alpha)h) - \partial I_C(\tilde{q}) \ni 0, \tag{85}$$

which is a natural generalization of (48b) in the case when $\tilde{q}$ is a singular point of ∂C.

Alternatively, using the previous notation for discrete momenta (51), (85) can be written as

$$p^+(q_{i-1}, \tilde{q}, \alpha h) - p^-(\tilde{q}, q_{i+1}, (1-\alpha)h) \in N_C(\tilde{q}), \tag{86}$$

where we used $\partial I_C(\tilde{q}) = N_C(\tilde{q})$ for $\tilde{q} \in \partial C$. Therefore, (86) generalizes (54) in the case when $\tilde{q}$ is a corner of ∂C and we cannot define a tangent plane at that point.

For Lagrangians consisting of only kinetic energy and for the particular discretization (57), the momentum conservation (85) leads to a very interesting geometrical interpretation. Indeed, for a unitary mass matrix, (85) becomes

$$\frac{\tilde{q} - q_{i-1}}{\alpha h} - \frac{q_i - \tilde{q}}{(1-\alpha)h} - \partial I_C(\tilde{q}) \ni 0. \tag{87}$$

The inclusion (87) can be rewritten in the form

$$(1-\alpha)q_{i-1} + \alpha q_i \in (I - \partial I_C)(\tilde{q}). \tag{88}$$

Now we will connect (88) with the concept of the *closest point projection*. First recall the definition of the *resolvent* of the set-valued operator ∂I_C as

$$R = (I + \partial I_C)^{-1}. \tag{89}$$

It is a well-known fact that the resolvent of the subgradient of the indicator function of a convex set is the closest-point projection onto that set (see [49]). Under the assumption that $\bar{C}$, the complement of C, is a convex set (see Figure 6), the inclusion (88) can be written as

$$\tilde{q} = P_{\bar{C}}(q_{i-\alpha}), \tag{90}$$

where $P_{\bar{C}}$ represents the closest-point projection operator onto $\bar{C}$ and $q_{i-\alpha}$ is the convex combination of the points q_{i-1} and q_i

$$q_{i-\alpha} = (1-\alpha)q_{i-1} + \alpha q_i.$$

We will conclude this appendix by the following two remarks which re-emphasize the particular benefit of the nonsmooth calculus approach.

Remark. Besides its theoretical attractiveness, the nonsmooth analysis approach has a great advantage over the standard penalty formulation methods in dealing with complex nonsmooth contact geometries (see [25]) where neither normals nor gap functions may be defined. Indeed, for such problems penalty methods simply fail.

Remark. The nonsmooth approach also gives the natural framework for constructing time-adaptive variational integrators for collisions (see [24] and [37]), but we will leave the development and illustration of such contact algorithms for future work.

6. Future directions.

Order of accuracy. For systems without collisions, the order of accuracy of the discrete Lagrangian L_d and the discrete Hamiltonian map $\tilde{F}_d$ are the same. In principle, this will also be true for contact algorithms as developed in this paper (see the numerical results presented in Figure 5), but precise proofs remain to be formulated.

Elastic bodies. Although the numerical simulations presented in section 4 were all for rigid-body collisions, the discrete variational formalism applies for arbitrary potential energies, such as those for hyperelastic materials. We have not yet tested these methods for such systems, however. In this context, it will also be very interesting to use the techniques of section 2.5 to include external forces and dissipative effects.

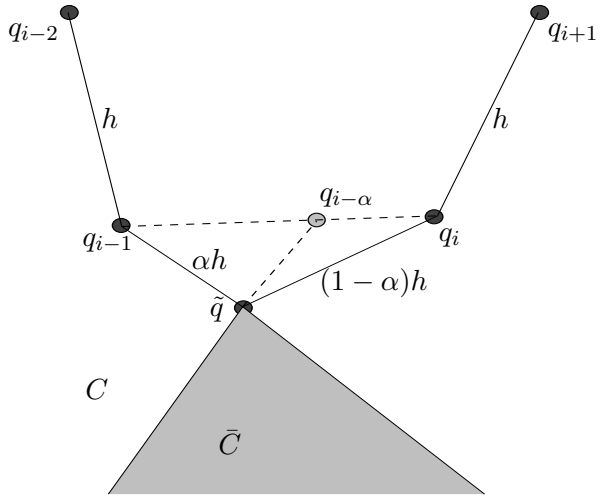


Figure 6. *Collision at a singular point of the contact set ∂C.*

Multisymplectic extensions to PDEs . While PDE contact systems can be first discretized in space and then treated as a system of contact ODEs in time with the algorithms developed in this paper, much greater understanding can be gained by a fully space-time variational formulation of both the continuous and discrete problems. The framework of multisymplectic mechanics [15] and multisymplectic discretizations [35] is particularly appropriate for this, and we will treat this subject in a forthcoming paper.

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis and Applications*, 2nd ed., Appl. Math. Sci. 75, Springer-Verlag, New York, 1988.
[2] M. Anitescu, F. A. Potra, and D. E. Stewart, *Time-stepping for three-dimensional rigid body dynamics*, Comput. Methods Appl. Mech. Engrg., 177 (1999), pp. 183–197.
[3] F. Armero and E. Petöcz, *A new dissipative time-stepping algorithm for frictional contact problems: Formulation and analysis*, Comput. Methods Appl. Mech. Engrg., 179 (1999), pp. 151–178.
[4] E. Barth, B. Leimkuhler, and S. Reich, *A time-reversible variable-stepsize integrator for constrained dynamics*, SIAM J. Sci. Comput., 21 (1999), pp. 1027–1044.
[5] B. Brogliato, S. Niculescu, and P. Orhant, *On the control of finite dimensional mechanical systems with unilateral constraints*, IEEE Trans. Automat. Control, 42 (1997), pp. 200–215.
[6] B. Brogliato, *Nonsmooth Impact Mechanics: Models, Dynamics and Control*, Lecture Notes in Control and Inform. Sci. 220, Springer-Verlag, New York, 1996.
[7] B. Brogliato, *On the control of non-smooth complementarity dynamical systems*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 359 (2001), pp. 2369–2383.
[8] N. J. Carpenter, R. L. Taylor, and M. G. Katona, *Lagrange constraints for transient finite-element surface-contact*, Internat. J. Numer. Methods Engrg., 32 (1991), pp. 103–128.
[9] F. Cirak and M. West, *Scalable Explicit Collision Response Using Momentum Decomposition*, manuscript.
[10] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.
[11] A. Dontchev and F. Lempio, *Difference methods for differential inclusions: A survey*, SIAM Rev., 34 (1992), pp. 263–294.

[12] A. F. Filippov, *Differential equations with discontinuous right-hand side*, American Mathematical Society Translations, 42 (1964), pp. 199–231.
[13] A. F. Filippov, *Classical solutions of differential equations with multivalued right-hand side*, SIAM J. Control, 5 (1967), pp. 609–621.
[14] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Side*, Kluwer Academic, Norwell, MA, 1988.
[15] M. Gotay, J. Isenberg, and J. E. Marsden, *Momentum Maps and the Hamiltonian Structure of Classical Relativistic Field Theories* I, http://www.cds.caltech.edu/~marsden/ (1997).
[16] E. Hairer and C. Lubich, *Long-time energy conservation of numerical methods for oscillatory differential equations*, SIAM J. Numer. Anal., 38 (2000), pp. 414–441.
[17] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I: *Nonstiff problems*, 2nd ed., Springer Ser. Comput. Math. 8, Springer-Verlag, New York, 1993.
[18] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II: *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer Ser. Comput. Math. 14, Springer-Verlag, New York, 1996.
[19] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley & Sons, New York, 1966.
[20] Y. A. Houndonougbo and B. B. Laird, *Constant-temperature molecular dynamics algorithms for mixed hard-core/continuous potentials*, J. Chem. Phys., 117 (2002), pp. 1001–1009.
[21] Y. A. Houndonougbo, B. B. Laird, and B. J. Leimkuhler, *A molecular dynamics algorithm for mixed hard-core/continuous potentials*, Mol. Phys., 98 (2000), pp. 309–316.
[22] A. Iserles, H. Munthe-Kaas, and A. Zanna, *Lie group methods*, in Acta Numerica, Acta Numer. 9, Cambridge University Press, Cambridge, UK, 2000, pp. 215–365.
[23] M. Jean, *Unilateral contact and dry friction: Time and space discrete variables formulation*, Arch. Mech. (Arch. Mech. Stos.), 40 (1988), pp. 677–691.
[24] C. Kane, J. E. Marsden, and M. Ortiz, *Symplectic-energy-momentum preserving variational integrators*, J. Math. Phys., 40 (1999), pp. 3353–3371.
[25] C. Kane, E. A. Repetto, M. Ortiz, and J. E. Marsden, *Finite element analysis of nonsmooth contact*, Comput. Methods Appl. Mech. Engrg., 180 (1999), pp. 1–26.
[26] C. Kane, J. E. Marsden, M. Ortiz, and M. West, *Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems*, Internat. J. Numer. Methods Engrg., 49 (2000), pp. 1295–1325.
[27] M. Kunze and D. P. Monteiro Marques, *An introduction to Moreau's sweeping process*, in Impacts in Mechanical Systems, B. Brogliato, ed., Lecture Notes in Phys. 551, Springer-Verlag, Berlin, 2000, pp. 1–61.
[28] C. Y. Lee and J. T. Oden, *Theory and approximation of quasi-static frictional contact problems*, Comput. Methods Appl. Mech. Engrg., 106 (1993), pp. 407–429.
[29] B. Leimkuhler and S. Reich, *A reversible averaging integrator for multiple time-scale dynamics*, J. Comput. Phys., 171 (2001), pp. 95–114.
[30] F. Lempio and V. Veliov, *Discrete approximations of differential inclusions*, Bayreuth. Math. Schr., 54 (1998), pp. 149–232.
[31] A. Lew, J. E. Marsden, M. Ortiz, and M. West, *Asynchronous variational integrators*, Arch. Ration. Mech. Anal., 167 (2003), pp. 85–146.
[32] M. Mabrouk, *A unified variational model for the dynamics of perfect unilateral constraints*, Eur. J. Mech. A Solids, 17 (1998), pp. 819–842.
[33] M. D. P. Monteiro Marques, *Chocs inélastiques standards: Un rèsultat d'existence*, Sèm. Anal. Convexe, 15 (1985), pp. 1–32.
[34] D. P. Monteiro Marques, *Differential Inclusions in Nonsmooth Mechanical Problems: Shocks and Dry Friction*, Progress. in Nonlinear Differential Equations and Their Applications, 9, Birkhäuser Verlag, Basel, 1993.
[35] J. E. Marsden, G. W. Patrick, and S. Shkoller, *Multisymplectic geometry, variational integrators and nonlinear PDEs*, Comm. Math. Phys., 199 (1998), pp. 351–395.
[36] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Texts Appl. Math. 17, Springer-Verlag, New York, 1994.

[37] J. E. Marsden and M. West, *Discrete mechanics and variational integrators*, in Acta Numerica, Acta Numer. 10, Cambridge University Press, Cambridge, UK, 2001, pp. 357–514.
[38] J.-J. Moreau, *Une formulation du contact à frottement sec; application au calcul numérique*, C. R. Acad. Sci. Paris Sér. II, 302 (1986), pp. 799–801.
[39] J.-J. Moreau, *Unilateral contact and dry friction in finite freedom dynamics*, in Nonsmooth Mechanics and Applications, J.-J. Moreau and P. D. Panagiotopoulos, eds., CISM Courses and Lectures 302, Springer-Verlag, Vienna, 1988, pp. 1–82.
[40] J.-J. Moreau, *Numerical aspects of the sweeping process*, Comput. Methods Appl. Mech. Engrg., 177 (1999), pp. 329–349.
[41] A. Pandolfi, C. Kane, J. E. Marsden, and M. Ortiz, *Time-discretized variational formulation of nonsmooth frictional contact*, Internat. J. Numer. Methods Engrg., 53 (2002), pp. 1801–1829.
[42] L. Paoli and M. Schatzman, *Schéma numérique pour un modèle de vibrations avec constraints unilatérale et perte d'énergie aux impacts, en dimension finie*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 211–215.
[43] L. Paoli and M. Schatzman, *Mouvement à un nombre fini de degrés de liberté avec contraintes unilatérales: Cas avec perte d'énergie*, RAIRO Modél. Math. Anal. Numér., 27 (1993), pp. 673–717.
[44] L. Paoli and M. Schatzman, *Approximation et existence en vibro-impact*, C. R. Acad. Sci. Paris Sér. I Math., 329 (1999), pp. 1103–1107.
[45] D. Peric and D. R. J. Owen, *Computational model for* 3*-d contact problems with friction based on the penalty method*, Internat. J. Numer. Methods Engrg., 35 (1992), pp. 1289–1309.
[46] F. Pfeiffer, *Unilateral problems of dynamics*, Archive of Applied Mechanics, 69 (1999), pp. 503–527.
[47] F. Pfeiffer and C. Glocker, *Dynamics of Rigid Body Systems with Unilateral Constraints*, Wiley, New York, 1996.
[48] E. B. Pires and J. T. Oden, *Analysis of contact problems with friction under oscillating loads*, Comput. Methods Appl. Mech. Engrg., 39 (1983), pp. 337–362.
[49] R. T. Rockafellar, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.
[50] J. M. Sanz-Serna and M. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.
[51] M. Schatzman, *Sur une classe de problèmes hyperboliques non linéaires*, C. R. Acad. Sci. Paris Sér. A-B, 277 (1973), pp. A671–A674.
[52] M. Schatzman, *A class of nonlinear differential equations of second order in time*, Nonlinear Anal., 2 (1978), pp. 355–373.
[53] M. Schatzman, C.-H. Lamarque, and J. Bastien, *An ill-posed mechanical problem with friction*, Eur. J. Mech. A Solids, 18 (1999), pp. 415–420.
[54] J. C. Simo, P. Wriggers, and R. L. Taylor, *A perturbed Lagrangian formulation for the finite-element solution of contact problems*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 163–180.
[55] D. E. Stewart, *Convergence of a time-stepping scheme for rigid-body dynamics and resolution of Painleve's problem*, Arch. Ration. Mech. Anal., 145 (1998), pp. 215–260.
[56] D. E. Stewart, *Rigid-body dynamics with friction and impact*, SIAM Rev., 42 (2000), pp. 3–39.
[57] R. L. Taylor and P. Papadopoulos, *On a finite-element method for dynamic contact impact problems*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 2123–2140.
[58] L. White and J. T. Oden, *Dynamics and control of viscoelastic solids with contact and friction effects*, Nonlinear Anal., 13 (1989), pp. 459–474.
[59] P. Wriggers, T. V. Van, and E. Stein, *Finite-element formulation of large deformation impact-contact problems with friction*, Comput. & Structures, 37 (1990), pp. 319–331.
[60] P. Wriggers and G. Zavarise, *Application of augmented Lagrangian techniques for nonlinear constitutive laws in contact interfaces*, Comm. Numer. Methods Engrg., 9 (1993), pp. 815–824.
[61] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders Company, Philadelphia, 1969, corrected printing, Chelsea, UK, 1980.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 3, pp. 417–430

Vortex Dynamics on a Cylinder*

James Montaldi†, Anik Soulière‡, and Tadashi Tokieda‡

Abstract. Point vortices on a cylinder (periodic strip) are studied geometrically. The Hamiltonian formalism is developed, a nonexistence theorem for relative equilibria is proved, equilibria are classified when all vorticities have the same sign, and several results on relative periodic orbits are established, including as corollaries classical results on vortex streets and leapfrogging.

Key words. point vortices, Hamiltonian systems with symmetry, geometric mechanics, conservation laws

AMS subject classifications. 37J15, 37J45, 70H33, 76B47

DOI. 10.1137/S1111111102415569

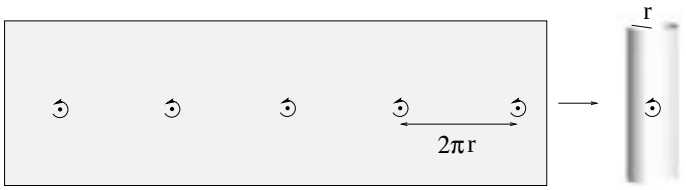


Figure 1.

1. Introduction.

Spatially periodic rows of point vortices in a two-dimensional ideal fluid have long attracted the attention of fluid dynamicists, one of the earliest and the most popular instances being Kármán's vortex street [6], [16, photos 94–98]. The general problem is as follows: analyze the motion of an infinite configuration consisting of vortices $z_1, \dots, z_N \in \mathbb{C}$ with vorticities $\Gamma_1, \dots, \Gamma_N \in \mathbb{R}$ together with their translates $\{z_k + 2\pi r m \mid k = 1, \dots, N, m \in \mathbb{Z}\}$, where $2\pi r > 0$ is the spatial period of translation. Traditionally the problem is analyzed on the plane $\mathbb{C}$, but in this paper we place the vortices on a cylinder $\mathbb{C}/2\pi r\mathbb{Z}$ (Figure 1). Though the two pictures—periodic planar and cylindrical—are for most purposes equivalent, as we shall see there are advantages, both conceptual and computational, to working on a cylinder rather than on the plane. The proviso "for most purposes" is necessary because the cylindrical picture posits that everything in the dynamics be $2\pi r$-periodic, whereas in the planar picture one could allow, for example, nonperiodic perturbations to the periodic row. Physically, however, perturbations are usually due to some small change in the mechanism generating the vortex row, and the simplest type of change generates spatially periodic

*Received by the editors October 2, 2002; accepted for publication (in revised form) by J. Marsden June 5, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/siads/2-3/41556.html
†Department of Mathematics, UMIST, PO Box 88, Manchester M60 1QD, UK (j.montaldi@umist.ac.uk). The work of this author was partially supported by the European Union through the Research Training Network MASIE.
‡Département de Mathématiques, Université de Montréal, C.P. 6128, succ. Centre-Ville, Montréal H3C 3J7, Canada (souliere@dms.umontreal.ca, tokieda@dms.umontreal.ca).

perturbations. Symmetry-breaking perturbations, which do occur in real fluids and are very interesting, arise at the next level of complexity. So it is natural to look at the cylindrical picture first.

We shall be interested in how vortices move relative to one another, more precisely in their dynamics modulo the translational action of the symmetry group $\mathbb{C}/2\pi\mathbb{Z}$. The basic objects of interest are relative equilibria and relative periodic orbits. A *relative equilibrium* is a motion of vortices that lies entirely in a group orbit (i.e., it looks stationary up to translation), and a *relative periodic orbit* is a motion that revisits the same group orbit after some time (i.e., it looks periodic in time up to translation). Equilibria and periodic orbits in the ordinary sense are special examples of relative equilibria and relative periodic orbits. When we wish to exclude ordinary equilibria or periodic orbits, we speak of relative equilibria or relative periodic orbits *with nonzero drift*.

As on the plane, the dynamics of point vortices on a cylinder lends itself to a Hamiltonian formalism. The model presented here is then a finite-dimensional Hamiltonian approximation to the vortex dynamics of the Euler equation. This approximation is mathematically very rich and in the context of the plane can claim a pedigreed history [7, chap. VII], [17]. Conversely, the motion of point vortices is amenable to desingularization to a solution of the Euler equation.

For vortices on the plane or on a sphere, an extensive theory of relative equilibria is available (especially when the vorticities are identical or opposite) [1, 9, 8]. In contrast, apart from a study on three vortices on the periodic strip [2] and a study of rings of point vortices on surfaces of revolution [4], no literature seems to exist on relative equilibria and relative periodic orbits of N vortices on a cylinder. In this paper we develop the Hamiltonian formalism for vortex dynamics on a cylinder (section 2), prove that if the vorticities do not sum to zero a cylinder supports no relative equilibrium with nonzero drift (section 3), classify equilibria when all vorticities have the same sign (section 3), show that three vortices form a relative periodic orbit for "small" initial conditions or for vorticities dependent over $\mathbb{Q}$ with zero sum, and establish several results on a class of relative periodic orbits called leapfrogging [16, photo 79] (section 4), which may be regarded as a splitting of Kármán's vortex street.

Although Noether's theorem tells us that associated to any one-parameter group of symmetries there is a corresponding first integral, there is a topological hypothesis (that certain closed 1-forms are exact) which is not fulfilled by the cylinder, and while the subgroup of horizontal translations $\mathbb{R}/2\pi\mathbb{Z} \subset \mathbb{C}/2\pi\mathbb{Z}$ does have a conserved quantity associated to it, the subgroup $i\,\mathbb{R}$ of vertical translations does not. However, since any closed 1-form is *locally* exact, this subgroup does have *locally* well-defined first integrals, and one of the novelties of the present work is to exploit these local first integrals (Theorems 2, 3, and 4).

Many of the results have analogues in the theory of vortices on a torus, i.e., for spatially biperiodic arrays of vortices.

2. Hamiltonian formalism of vortices on a cylinder. Throughout the paper, *cylinder* means the surface $\mathbb{C}/2\pi r\mathbb{Z} \simeq (\mathbb{R}/2\pi r\mathbb{Z}) \times \mathbb{R}$, where $r > 0$ is some fixed constant, the *radius* of the cylinder. The coordinate $z = x+iy$ on $\mathbb{C}/2\pi r\mathbb{Z}$ is to be read modulo $2\pi r$, i.e., $x \equiv x+2\pi rn$ for all $n \in \mathbb{Z}$; the x-axis (which is a circle) is *horizontal*, and the y-axis is *vertical*. The phase space for the motion of vortices $z_1, \dots, z_N$ with vorticities $\Gamma_1, \dots, \Gamma_N$ is the product of N copies

of the cylinder with diagonals removed (to exclude collisions). The Hamiltonian is a weighted combination $H(z_1,\dots,z_N)=\sum_{k<l}\Gamma_k\Gamma_l\psi(z_k,z_l)$ of Green's functions ψ for the Laplacian on the cylinder: $\nabla^2\psi(z,z_0)=-\delta_{z_0}(z)$ (see, e.g., [15, section 2]). Hamilton's equations are

$$\frac{dz_k}{dt}=\frac{2}{i}\frac{\partial H}{\partial(\Gamma_k\overline{z}_k)},\qquad k=1,\dots,N.$$

The quickest way to derive the Hamiltonian on a cylinder is to periodize Green's function on the plane $\psi(z_k,z_l)=-\frac{1}{2\pi}\log|z_k-z_l|$ by taking into account contributions from $2\pi\mathbb{Z}$-translates. Formally the periodized Hamiltonian becomes

$$-\frac{1}{2\pi}\sum_{n\in\mathbb{Z}}\sum_{k<l}\Gamma_k\Gamma_l\log|z_k-z_l-2\pi rn|,$$

which, as it stands, diverges. But, since additive constants in H do not affect the dynamics, we can subtract a constant divergent series to force the remaining functional part to converge. Jettisoning $-\frac{1}{2\pi}\sum_n\sum_{k<l}\Gamma_k\Gamma_l\log|2\pi rn|$ and pairing terms in n and $-n$,

(2.1)
$$H=-\frac{1}{2\pi}\sum_{k<l}\Gamma_k\Gamma_l\log\left|(z_k-z_l)\prod_{n\geqslant 1}\left(1-\left(\frac{z_k-z_l}{2\pi rn}\right)^2\right)\right|=-\frac{1}{2\pi}\sum_{k<l}\Gamma_k\Gamma_l\log\left|\sin\frac{z_k-z_l}{2r}\right|.$$

The equations of motion on a cylinder are therefore

$$\frac{dz_k}{dt}=\frac{i}{4\pi r}\sum_{l,l\neq k}\Gamma_l\operatorname{cotan}\frac{\overline{z_k}-\overline{z_l}}{2r},\qquad k=1,\dots,N. \tag{2.2}$$

For reference, we list expressions in real coordinates:

$$H=-\frac{1}{4\pi}\sum_{k<l}\Gamma_k\Gamma_l\log\left\{\sin^2\left(\frac{x_k-x_l}{2r}\right)+\sinh^2\left(\frac{y_k-y_l}{2r}\right)\right\}, \tag{2.3}$$

$$\begin{cases}\dfrac{dx_k}{dt}=-\dfrac{1}{8\pi r}\displaystyle\sum_{l,l\neq k}\Gamma_l\frac{\sinh\dfrac{y_k-y_l}{r}}{\sin^2\left(\dfrac{x_k-x_l}{2r}\right)+\sinh^2\left(\dfrac{y_k-y_l}{2r}\right)},\\[2ex] \dfrac{dy_k}{dt}=\dfrac{1}{8\pi r}\displaystyle\sum_{l,l\neq k}\Gamma_l\frac{\sin\dfrac{x_k-x_l}{r}}{\sin^2\left(\dfrac{x_k-x_l}{2r}\right)+\sinh^2\left(\dfrac{y_k-y_l}{2r}\right)},\end{cases}\qquad k=1,\dots,N. \tag{2.4}$$

One noteworthy feature of (2.4) is that as $y_k-y_l\to\infty$ (infinite vertical separation), the velocity induced by z_l on the vortex z_k does not decay to 0 but tends to $\Gamma_l/4\pi r$, as is obvious upon calculating in the planar theory the circulation around a tall window of width

$2\pi r$ enclosing z_l. Another way to interpret the feature is to note that in the planar theory, up to rescaling, stretching vertical separation amounts to narrowing the spatial period $2\pi r \to 0$; the latter limit produces a *vortex sheet* (or, more aptly, *vortex line* in this two-dimensional theory), which induces a velocity field constant above (and the opposite constant below) the sheet independently of the distance to the sheet. This is exactly as in two-dimensional electromagnetism or gravity, where the force induced by a homogeneous charge or mass distribution along an infinite line is independent of the distance to the line.

Physically, periodizing the plane with period $2\pi r$ and considering N vortices on the resulting cylinder are the same as periodizing with period $2\pi rn$ and considering nN vortices on the resulting wider cylinder. The equivalence between these periodizations is trivial yet sometimes useful.

Proposition. *Let $z_1, \dots, z_N$ be vortices with vorticities $\Gamma_1, \dots, \Gamma_N$ on a cylinder of radius r. Next let $z_1, \dots, z_N, z_1 + 2\pi r, \dots, z_N + 2\pi r, \dots, z_1 + 2\pi rn, \dots, z_N + 2\pi rn$ be their "n-fold copies" with corresponding vorticities on a cylinder of radius rn, where n is any strictly positive integer. Then the dynamics on the cylinder of radius rn covers the dynamics on the cylinder of radius r.*

In particular, given a relative equilibrium or a relative periodic orbit, we can reel off infinite families of relative equilibria or relative periodic orbits at no extra cost by replicating the configuration sideways on a wider cylinder.

Remark 1. A torus has the form $\mathbb{C}/(\pi\mathbb{Z} + \tau\pi\mathbb{Z})$, where the parameter $\tau \in \mathbb{C}$, $\mathrm{Im}\tau > 0$, controls the conformal class. The Hamiltonian is

$$H = -\frac{1}{2\pi} \sum_{k<l} \Gamma_k \Gamma_l \left\{ \log |\vartheta_1(z_k - z_l|\tau)| - \frac{(\mathrm{Im}(z_k - z_l))^2}{\pi \mathrm{Im}\tau} \right\},$$

where ϑ_1 is the first Jacobian theta function [12, 14, 15].

A cylinder has a translational symmetry of $\mathbb{C}/2\pi r\mathbb{Z}$ acting on itself and hence acting diagonally on the phase space. The plane has a supplementary rotational symmetry $z \mapsto e^{i\theta} z$, $\theta \in \mathbb{R}$; this is lost on the cylinder. Via Noether's theorem the translational symmetry of $\mathbb{C}/2\pi r\mathbb{Z}$ should give rise to a first integral, a *momentum map* $(z_1, \dots, z_N) \mapsto \sum_k \Gamma_k z_k$, but there is a rub: because z's are defined only modulo $2\pi r$ this "momentum map" is not well-defined as a map to the dual of the Lie algebra of the symmetry group $\mathbb{C}/2\pi\mathbb{Z}$. Nor is it advisable to treat this "momentum map" as a multivalued function, for generically $\Gamma_1, \dots, \Gamma_N$ are independent over $\mathbb{Q}$, and so the ambiguity $\{2\pi r \sum_k \Gamma_k n_k \mid n_1, \dots, n_N \in \mathbb{Z}\}$ in the value of the "map" is dense in $\mathbb{R}$. Nevertheless, the momentum map is *locally* (i.e., on each chart) well-defined. From now on, whenever we write $\sum_k \Gamma_k z_k$, some suitable ad hoc chart will be understood.

When $\sum_k \Gamma_k \neq 0$, the *center of vorticity* $\sum_k \Gamma_k z_k / \sum_k \Gamma_k$ is a more intuitive first integral [7, art. 154]. The next result provides a substitute for center of vorticity when $\sum_k \Gamma_k = 0$.

Theorem 1. *Let $\{z\}$ be vortices on the plane or on a cylinder whose vorticities sum to zero: $\sum \Gamma = 0$. Suppose the vortices are partitioned into two groups $\{z'\}$, $\{z''\}$ and within each group $\sum \Gamma' \neq 0$, $\sum \Gamma'' \neq 0$ so that the center of vorticity for each group is well-defined. Then the vector connecting the two centers of vorticity is a local first integral (Figure 2).*

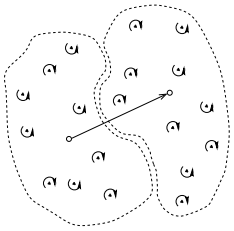


Figure 2.

Proof. Since $\sum \Gamma' + \sum \Gamma'' = 0$, the vector in question is

$$\frac{\sum \Gamma' z'}{\sum \Gamma'} - \frac{\sum \Gamma'' z''}{\sum \Gamma''} = \frac{\sum \Gamma' z'}{\sum \Gamma'} + \frac{\sum \Gamma'' z''}{\sum \Gamma'} = \frac{\sum \Gamma z}{\sum \Gamma'},$$

and $\sum \Gamma z$ is a local first integral. ∎

Theorem 1 is serviceable in many problems. The simplest illustration is the motion of a *vortex pair* z_1, z_2 with vorticities $\Gamma, -\Gamma$ [16, photos 77, 78]. Treating z_1 as one group and z_2 as the other group, we check against Theorem 1 that $z_2 - z_1$ is constant during the motion. In fact, according to (2.4), the vortex pair on a cylinder forms a relative equilibrium moving with slope

$$-\sin\frac{x_2 - x_1}{r} \Big/ \sinh\frac{y_2 - y_1}{r}.$$

When $x_2 - x_1 = 0$ or πr, the pair moves horizontally. The corresponding configurations on the plane are the unstaggered or fully staggered cases of Kármán's vortex street; see also [4]. When z_1, z_2 are in general position, the corresponding vortex street on the plane translates at an angle to the horizontal, a case studied in [11]. The "plane limit" $r \to \infty$ yields the angle of progression of a vortex pair on the plane $-(x_2 - x_1)/(y_2 - y_1)$. For a beautiful study of the stability of variants of vortex streets, see [5].

3. Relative equilibria. The first fact about relative equilibria of vortices on a cylinder is that there are not many of them.

Theorem 2. *Let $z_1, \ldots, z_N$ be vortices with vorticities $\Gamma_1, \ldots, \Gamma_N$ on a cylinder $\mathbb{C}/2\pi r\mathbb{Z}$. Suppose $\sum_k \Gamma_k \neq 0$. Then all relative equilibria are in fact equilibria. Moreover, if all Γ's have the same sign, then for each cyclic ordering there exists a unique (up to translation by $\mathbb{C}/2\pi r\mathbb{Z}$) equilibrium, and all the vortices are aligned on a single horizontal circle.*

Proof. If $z_1, \ldots, z_N$ form a relative equilibrium, then all z's move with some common drift velocity v. The local first integral should not vary:

$$0 = \frac{d}{dt} \sum_k \Gamma_k z_k = v \sum_k \Gamma_k,$$

so $\sum_k \Gamma_k = 0$ or else $v = 0$.

If the vortices are not aligned on a single horizontal circle, pick a "top vortex" (one with maximal y-coordinate) and a "bottom vortex" (one with minimal y-coordinate). If all Γ's have the same sign, then by (2.4) the velocities of the top and bottom vortices must have x-components with opposite signs, so this position cannot constitute an equilibrium.

Now suppose all the vorticities are of the same sign. Fix a cyclic ordering of the vortices, and place the vortices in order on a single horizontal circle. The Hamiltonian is given by

$$H = -\frac{1}{4\pi}\sum_{k<l}\Gamma_k\Gamma_l \log\sin^2\left(\frac{x_k - x_l}{2r}\right).$$

One readily checks that this is a convex function of $x_1, \dots, x_N$: one first checks that wherever they are defined the second derivatives satisfy $\partial^2 H/\partial x_k \partial x_l < 0$ for $k \neq l$ and $\partial^2 H/\partial x_k^2 > 0$ and $\sum_l \partial^2 H/\partial x_k \partial x_l = 0$ for each k; it then follows from a variant of Gershgorin's theorem (Lemma 1 below) that 0 is a simple eigenvalue of the Hessian of H and all other eigenvalues are strictly positive. Consequently, on each connected component of the domain of definition, there is a unique minimum and no other critical point, and different connected components correspond to different cyclic orderings. This is the same argument as for [9, Theorem 4.8]. ■

Lemma 1. *Let $A = (a_{kl})$ be a symmetric $N \times N$ matrix satisfying $a_{kl} < 0$ for $k \neq l$, and $a_{kk} > 0$, $\sum_{l=1}^N a_{kl} = 0$ for each k. Then 0 is a simple eigenvalue of A, and all other eigenvalues are strictly positive.*

Proof. Let $u = (u_1, \dots, u_n)^T$ be an eigenvector of A with eigenvalue λ, normalized so that there is an index k for which $u_k = 1$ and $|u_l| \leqslant 1$ for all l. The kth row of the equation $Au = \lambda u$ is $a_{kk} + \sum_{l,l\neq k} a_{kl}u_l = \lambda$, which in view of the hypotheses on a_{kl} may be written $\sum_l |a_{kl}|(1 - u_l) = \lambda$. But $1 - u_l \geqslant 0$ and $|a_{kl}| > 0$ for each l; it follows that $\lambda \geqslant 0$ and $\lambda = 0$ if and only if all $u_l = 1$. On the other hand, $(1, \dots, 1)^T$ is obviously an eigenvector with eigenvalue 0. ■

If the vortices are placed on a single horizontal circle so that successive vorticities have alternating signs, then we also get the existence of an equilibrium, though the uniqueness problem is open as the function is no longer convex. In full generality, if the signs are neither the same nor alternating, the argument for existence fails as $H \to +\infty$ for some collisions and $\to -\infty$ for others.

Remark 2. For $N = 2$, if $\Gamma_1 + \Gamma_2 \neq 0$, we have generically a periodic orbit and exceptionally an equilibrium of antipodal vortices $z, z + \pi r$ or a separatrix connecting equilibria. For $N > 2$, if $\sum_k \Gamma_k \neq 0$ but Γ's have mixed signs, the equilibria are less severely constrained. For example, for $N = 3$, let z_1, z_2 be vortices with vorticities $\Gamma_1, \Gamma_2 > 0$. To secure an equilibrium, the third vortex z_3 with vorticity $\Gamma_3 < 0$ must be placed at one of the two stagnation points of the velocity field induced by z_1, z_2, given in view of (2.2) as roots of

$$\Gamma_1 \operatorname{cotan}\frac{z - z_1}{2} + \Gamma_2 \operatorname{cotan}\frac{z - z_2}{2} = 0.$$

Having chosen z_3 as one of the roots and thereby immobilized z_3, adjust Γ_3 so as to immobilize z_1:

$$\Gamma_2 \operatorname{cotan}\frac{z_1 - z_2}{2} + \Gamma_3 \operatorname{cotan}\frac{z_1 - z_3}{2} = 0.$$

Then z_2 too is automatically immobilized:

$$\Gamma_3 \operatorname{cotan}\frac{z_2 - z_3}{2} + \Gamma_1 \operatorname{cotan}\frac{z_2 - z_1}{2} = 0.$$

The upshot is that given any z_1, z_2 with vorticities of the same sign, we have two positions to place z_3 with the right vorticity of the opposite sign to secure an equilibrium. For example, vortices z_1, z_2 both of vorticity Γ such that $z_2 - z_1 = 2ib$ are immobilized by the adjunction of a vortex $(z_1 + z_2)/2$ of vorticity

$$\Gamma\left(\frac{1}{2}\operatorname{sech}^2\frac{b}{2r} - 1\right).$$

This is always less than $-\Gamma/2$, and in the plane limit $r \to \infty$ tends to the corresponding value in the planar theory $-\Gamma/2$. On the other hand, in the "vortex sheet limit" $b \to \infty$, this tends to $-\Gamma$, also as it should. Similarly, vortices z_1, z_2 of vorticity Γ such that $z_2 - z_1 = 2a$ are immobilized by the adjunction of a vortex $(z_1 + z_2)/2$ of vorticity

$$\Gamma\left(\frac{1}{2}\sec^2\frac{a}{2r} - 1\right).$$

In the planar limit, this tends again to $-\Gamma/2$. On the other hand, it is 0 when $a = \pi r/2$; z_1, z_2 are antipodal on the cylinder and are stationary already by themselves. When $a \to \pi r$, z_1, z_2 nearly meet at the back, and a stronger and stronger vortex is required at the front to prevent them from moving.

Remark 3. Now suppose $\sum_k \Gamma_k = 0$. It was pointed out at the end of section 2 that a vortex pair $N = 2$ is always a relative equilibrium. For $N = 3$, Aref and Stremler [2] made a detailed study of relative equilibria; the patterns of some trajectories are surprisingly complicated. For $N > 3$ and N even, we have for any $a, b > 0$ a family of relative equilibria consisting of $n = N/2$ vortices with vorticity Γ at

$$ib, ib + \frac{2\pi r}{n}, \ldots, ib + (n-1)\frac{2\pi r}{n} \tag{3.1}$$

and n vortices with vorticity $-\Gamma$ at

$$a - ib, a - ib + \frac{2\pi r}{n}, \ldots, a - ib + (n-1)\frac{2\pi r}{n}. \tag{3.2}$$

This is merely a crowded vortex street with spatial period $2\pi r/n$ or, equivalently, a single vortex pair on a thinner cylinder of radius r/n (see stability calculations in [3]). No essentially different family of relative equilibria seems to be known for $N > 3$.

Incidentally, even the trivial equivalence between one vortex on a cylinder of radius r and n horizontally equidistributed vortices on a cylinder of radius nr leads to amusing identities [1]. For example, equating the induced velocity fields and rescaling the variables in (2.2),

$$\frac{1}{n}\sum_{l=1}^{n} \operatorname{cotan}\frac{z + \pi l}{n} = \operatorname{cotan} z \quad \forall z \in \mathbb{C}.$$

Remark 4. On the plane, equilibria do not exist when all Γ's are of the same sign (even the possibility of a horizontal circle is lost), and the nonexistence of translational relative equilibria with nonzero drift when $\sum_k \Gamma_k \neq 0$ holds also on the plane and on a torus. The proof carries over verbatim from the cylindrical theorem. A torus, however, accommodates more varied families of equilibria; for example, $n_1 n_2$ vortices with identical vorticity Γ placed on a sublattice $(\pi/n_1)\mathbb{Z} + (\tau\pi/n_2)\mathbb{Z}$ form an equilibrium [15]. Many further patterns of equilibria may be designed on a torus with identical or alternating vortices.

4. Relative periodic orbits. Once a relative equilibrium of vortices is known, a frequently successful recipe for creating relative periodic orbits consists in *splitting* the vortices. Assume the vortices $z_1, \dots, z_N$ with vorticities $\Gamma_1, \dots, \Gamma_N$ form a relative equilibrium. Let us split each z_k into a cluster, near the original position of z_k, of n_k vortices $z_{k,1}, \dots, z_{k,n_k}$ whose vorticities are of the same sign and sum to Γ_k. We expect the child vortices $z_{k,1}, \dots, z_{k,n_k}$ to orbit around one another and remain a cluster, while seen from far away they still look like the original parent vortex z_k with vorticity Γ_k. It is reasonable to conjecture that for suitable initial configurations the child vortices form a relative periodic orbit, and for perhaps generic splittings they form a relative *quasi-periodic* orbit.

A vortex pair on a cylinder, which corresponds in the planar picture to Kármán's vortex street, is a relative equilibrium. In this section, we shall create various relative periodic orbits by splitting a vortex pair; as a special case, we recover the phenomenon classically known in the planar picture as leapfrogging. In Theorem 3, we split one of the vortices, while in Theorem 4 we split both. The split is measured by a complex variable $\zeta = \xi + i\eta$ (or rather by 2ζ), and we are principally interested in small values of $|\zeta|$. In all the formulae the radius of the cylinder is normalized to $r = 1$; denormalization is a matter of dimensional analysis. Later in the section, additional classes of relative periodic orbits are described.

Take a vortex pair at $c, -c$, where $c = a + ib \in \mathbb{C}$. We split it into three or four vortices as in Figure 3: the left diagram illustrates Theorem 3, the middle one Theorem 4, case $-b(1 + \Gamma/\Gamma')/2 < \eta < b(1 + \Gamma'/\Gamma)/2$, and the right one case $b(1 + \Gamma'/\Gamma)/2 < \eta$. Theorem 4, case $\eta < -b(1 + \Gamma/\Gamma')/2$ is like the right diagram reflected laterally with Γ, Γ' interchanged.

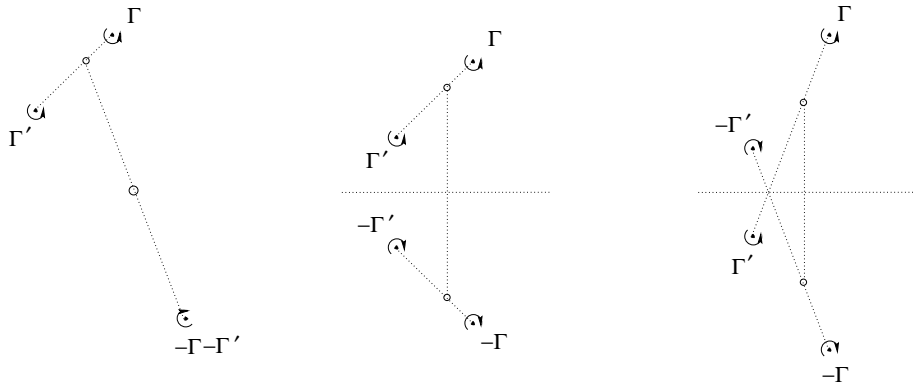


Figure 3.

Theorem 3. *Let $c \in \mathbb{C} \setminus \{0\}$. On a cylinder, consider the configuration of three vortices with vorticities $\Gamma, \Gamma', -\Gamma - \Gamma'$ (Γ and Γ' being of the same sign) at*

$$c + \frac{2\Gamma'}{\Gamma + \Gamma'}\zeta, \quad c - \frac{2\Gamma}{\Gamma + \Gamma'}\zeta, \quad -c.$$

There exists an open punctured neighborhood of $\zeta = 0$ such that for every initial condition $\zeta(0) \neq 0$ in this neighborhood, these vortices form a relative periodic orbit. If $\Gamma/\Gamma' \in \mathbb{Q}$, then for a generic choice of $\zeta(0)$ (no restriction on its size) these vortices form a relative periodic orbit, and for isolated choices of $\zeta(0)$ they form a relative equilibrium or a separatrix connecting relative equilibria.

Combined with the proposition of section 2, Theorem 3 gives relative equilibria and relative periodic orbits of $N = 3n$ vortices for all $n \geqslant 1$. The result for $N = 3$ when $\Gamma/\Gamma' \in \mathbb{Q}$ is in [2], but we give a somewhat different proof. The relative periodicity for small $\zeta(0)$ is new.

The proof invokes the following elementary lemma.

Lemma 2. *Let H be a function with only nondegenerate critical points on a compact surface with p punctures such that $|H| \to \infty$ near each puncture. Then the generic level sets of H are disjoint unions of loops. If $p > 2$, then in addition to loops there exist isolated saddles and separatrices connecting the saddles.*

Proof. By rescaling the values of H and compactifying the punctures, we reduce the problem to the situation where H is defined on a compact surface, takes values in $[-1, 1]$, and attains ± 1 at the points where the punctures used to be. The first part of the conclusion is immediate from Sard's theorem and the implicit function theorem. Moreover, from Morse theory,

$$p - \#\text{saddles} \leqslant \#\max + \#\min - \#\text{saddles} = \text{Euler characteristic } \leqslant 2,$$

whence the second part of the conclusion. ∎

The idea now for the proof of Theorem 3 is to use symmetries and Theorem 1 to rewrite the Hamiltonian as a function on a punctured two-dimensional sphere, satisfying the condition of divergence near the punctures. Applying Lemma 2 and recalling that a phase point in a Hamiltonian system moves along a level set of the Hamiltonian, we shall be home.

Proof of Theorem 3. The center of vorticity of the group Γ, Γ' is at c, and that of the singleton group $-\Gamma - \Gamma'$ is at $-c$. By Theorem 1, the vector connecting these centers is a local first integral. Hence, passing to the quotient by translations, these centers may be assumed immobile. Within the group Γ, Γ', the position of one vortex determines the position of the other. (It is at a definite ratio of distances across their center.) Hence the trajectory of the vortex with vorticity Γ determines the trajectories of all three vortices up to translation, and the Hamiltonian H may be regarded as a function of $\zeta = \xi + i\eta$ alone *as long as the trajectory of ζ lies on a single chart*. If the vortices Γ, Γ' are very close, they orbit like a binary star around their immobile center c within the chart so that sooner or later $\arg \zeta$ increases by 2π. Since $H(\zeta) \to +\infty$ as $\zeta \to 0$, for large enough $E \in \mathbb{R}$ the connected component of $\{\zeta \in \mathbb{C} \setminus 0 \mid |H(\zeta)| > E\}$ surrounding the singularity $\zeta = 0$ is topologically a punctured open disk, free of critical points of H. (The infimum of such E is the largest of the saddle values of H.) The level sets of H on this neighborhood are topologically circles, and so every ζ starting from $\zeta(0) \neq 0$ in this neighborhood returns to $\zeta(0)$, guaranteeing relative periodicity.

We must deal with the scenario where the trajectory of ζ does not lie on a single chart. Since $\Gamma/\Gamma' \in \mathbb{Q}$, the lowest common multiple L of $2, 1 + \Gamma/\Gamma', 1 + \Gamma'/\Gamma$ makes sense. To define ζ on the whole cylinder, we must swell the cylinder to $\mathbb{C}/L\pi\mathbb{Z}$. The swollen cylinder $\mathbb{C}/L\pi\mathbb{Z}$ covers the original cylinder $\mathbb{C}/2\pi\mathbb{Z}$ and H as a function of ζ lifts to a function on $\mathbb{C}/L\pi\mathbb{Z} \setminus \{\text{singularities}\}$. The singularities represent the collisions between

$$\Gamma \sim \Gamma' \text{ (front and back)}, \quad \Gamma \sim -\Gamma - \Gamma', \quad \Gamma' \sim -\Gamma - \Gamma',$$

where $|H| \to \infty$; off the singularities, by (2.1),

$$e^{2\pi H/\Gamma\Gamma'} = \frac{\left|\sin\left(c + \frac{\zeta}{1+\Gamma/\Gamma'}\right)\right|^{1+\Gamma/\Gamma'} \left|\sin\left(c - \frac{\zeta}{1+\Gamma'/\Gamma}\right)\right|^{1+\Gamma'/\Gamma}}{|\sin\zeta|}. \tag{4.1}$$

Toward the "ends" $\eta \to \pm\infty$, $|H| \to \infty$ as well. Topologically $\mathbb{C}/L\pi\mathbb{Z}\setminus\{\text{singularities}\}$ is a sphere with at least four punctures. Equation (4.1) shows that the critical points of H are all nondegenerate and $|H| \to \infty$ near each puncture. By Lemma 2, the generic level sets of H are loops, representing (putting horizontal translation back in) relative periodic orbits, and there exist values of ζ representing relative equilibria as well as separatrices (relative heteroclinic orbits) connecting relative equilibria. ■

Remark 5. In Theorem 3, relative periodicity when $\Gamma/\Gamma' \notin \mathbb{Q}$ is spoilt only for $\zeta(0)$ too large. For such $\zeta(0)$, the orbit is relative quasi-periodic. Of course, even when $\Gamma/\Gamma' \notin \mathbb{Q}$ there are questions that can be settled within a chart. Thus, for three vortices with arbitrary vorticities that sum to zero, topological reasons imply the existence of a configuration that forms a relative equilibrium.

Theorem 4. *Let $b \in \mathbb{R}\setminus\{0\}$. On a cylinder, consider the configuration of four vortices with vorticities $\Gamma, \Gamma', -\Gamma', -\Gamma$ (Γ and Γ' being of the same sign) at*

$$ib + \frac{2\Gamma'}{\Gamma+\Gamma'}\zeta, \quad ib - \frac{2\Gamma}{\Gamma+\Gamma'}\zeta, \quad -ib - \frac{2\Gamma}{\Gamma+\Gamma'}\bar{\zeta}, \quad -ib + \frac{2\Gamma'}{\Gamma+\Gamma'}\bar{\zeta}.$$

Let $\Gamma/\Gamma' \neq 1$. Then for a generic choice of the initial condition $\zeta(0)$ these vortices form a relative periodic orbit, and for isolated choices of $\zeta(0)$ they form a relative equilibrium or a separatrix connecting relative equilibria. If $\Gamma/\Gamma' = 1$, the same conclusion holds for $\zeta(0)$ such that $|\operatorname{Im}\zeta(0)| < b$ or $\pi H(\zeta(0))/\Gamma^2 < \log\sinh b$.

Combined with the proposition of section 2, Theorem 4 gives relative equilibria and relative periodic orbits of $N = 4n$ vortices for all $n \geqslant 1$.

Proof. As in the proof of Theorem 3, the positions of all four vortices are determined by those of the ones with vorticities Γ and $-\Gamma$. Thanks to a supplementary reflexive symmetry $z \mapsto \bar{z}$, the position of Γ determines that of $-\Gamma$. This time, after passing to the quotient by translations, H is a genuine function on the cylinder $\mathbb{C}/\pi\mathbb{Z}$ of $\zeta = \xi + i\eta$, $-\pi/2 < \xi \leqslant \pi/2$, with the singularities removed. Off the singularities, by (2.1),

$$e^{2\pi H/\Gamma\Gamma'} = \left|\frac{\sin\left(ib + \frac{\Gamma'\zeta + \Gamma\bar{\zeta}}{\Gamma+\Gamma'}\right)}{\sin\zeta}\right|^2 \left|\sin\left(ib + \frac{\zeta-\bar{\zeta}}{1+\Gamma/\Gamma'}\right)\right|^{\Gamma/\Gamma'} \left|\sin\left(ib - \frac{\zeta-\bar{\zeta}}{1+\Gamma'/\Gamma}\right)\right|^{\Gamma'/\Gamma}. \tag{4.2}$$

In particular, when $\Gamma/\Gamma' = 1$,

$$e^{2\pi H/\Gamma^2} = \frac{\sin^2\xi + \sinh^2 b}{\sin^2\xi + \sinh^2\eta} \left|\sinh(b+\eta)\sinh(b-\eta)\right|. \tag{4.3}$$

The isolated singularities represent simultaneous collisions between

$$\Gamma \sim \Gamma' \text{ and } -\Gamma' \sim -\Gamma,$$

where $H \to +\infty$, and, if $\Gamma/\Gamma' \neq 1$, between

$$\Gamma \sim -\Gamma' \text{ and } \Gamma' \sim -\Gamma,$$

where $H \to -\infty$. Toward the ends, $H \to +\infty$. There are also circles of singularities $\eta = -b(1+\Gamma/\Gamma')/2$, $b(1+\Gamma'/\Gamma)/2$ representing collisions between

$$\Gamma \sim -\Gamma, \quad \Gamma' \sim -\Gamma',$$

where $H \to -\infty$. Let us saw the cylinder $\mathbb{C}/\pi\mathbb{Z}$ of ζ into three trunks:

$$\begin{aligned} C_+ &= \{\zeta \mid b(1+\Gamma'/\Gamma)/2 < \eta\}, \\ C_0 &= \{\zeta \mid -b(1+\Gamma/\Gamma')/2 < \eta < b(1+\Gamma'/\Gamma)/2\}, \\ C_- &= \{\zeta \mid \eta < -b(1+\Gamma/\Gamma')/2\}. \end{aligned}$$

Topologically C_+, C_0, C_- are spheres with punctures. C_0 contains $\zeta = 0$, the simultaneous collisions between $\Gamma \sim \Gamma', -\Gamma' \sim -\Gamma$, so C_0 has at least three punctures and $|H| \to \infty$ near each puncture. Lemma 2 applies to C_0 and implies the existence of relative periodic orbits and relative equilibria.

For the moment, suppose $\Gamma/\Gamma' \neq 1$. ζ representing the simultaneous collisions between $\Gamma \sim -\Gamma'$, $\Gamma' \sim -\Gamma$ is in C_+ or C_- accordingly as $\Gamma/\Gamma' > 1$ or < 1. If $\Gamma/\Gamma' > 1$, this puts on C_+ at least three punctures near each of which $|H| \to \infty$, so Lemma 2 applies and implies the existence of relative periodic orbits and relative equilibria, whereas C_- acquires only two punctures, so we can conclude the existence of relative periodic orbits only. If $\Gamma/\Gamma' < 1$, the roles of C_+, C_- are reversed.

Note that as H is symmetric under the lateral reflection along $\xi = 0$ and along $\xi = \pi/2$, every point on either line where $\partial H/\partial\eta$ vanishes is critical. Let $\Gamma/\Gamma' > 1$ and work on C_+. The strip $0 < \xi < \pi/2$ is free of critical points, for here by (4.3) H is strictly monotone in ξ along any line $\eta =$ constant. Along $\xi = 0$, $H \to -\infty$ as $\eta \to b(1+\Gamma'/\Gamma)/2$ or $b(\Gamma+\Gamma')/(\Gamma-\Gamma')$, between which $\partial H/\partial\eta$ must vanish, signaling a saddle at, say, ζ_1. Along $\xi = \pi/2$, $H \to -\infty$ or $+\infty$ as $\eta \to b(1+\Gamma'/\Gamma)/2$ or $+\infty$. These bits of information, together with the fact that all critical points of H are nondegenerate, imply that $\partial H/\partial\eta$ vanishes twice along $\xi = \pi/2$, signaling a maximum at, say, ζ_2 and a saddle (which shall be left nameless). As a bonus we learn that two relative equilibria are represented in C_+, whereas a count of three singularities just predicts at least one relative equilibrium. The analysis works mutatis mutandis on C_- if $\Gamma/\Gamma' < 1$.

Finally, suppose $\Gamma/\Gamma' = 1$. Then the simultaneous collisions $\Gamma \sim -\Gamma'$, $\Gamma' \sim -\Gamma$ as well as ζ_1, ζ_2 escape to the ends $\eta \to \pm\infty$, and toward the ends $2\pi H/\Gamma^2$ asymptotes to $\log(\sin^2\xi + \sinh^2 b)$, which remains bounded. Hence all the critical points in C_+, C_- disappear. Relative periodic orbits are represented by compact level sets of H, i.e., those that fill the region $e^{\pi H/\Gamma^2} < \sinh b$ of C_+, C_-; there is no relative equilibrium on these trunks. ■

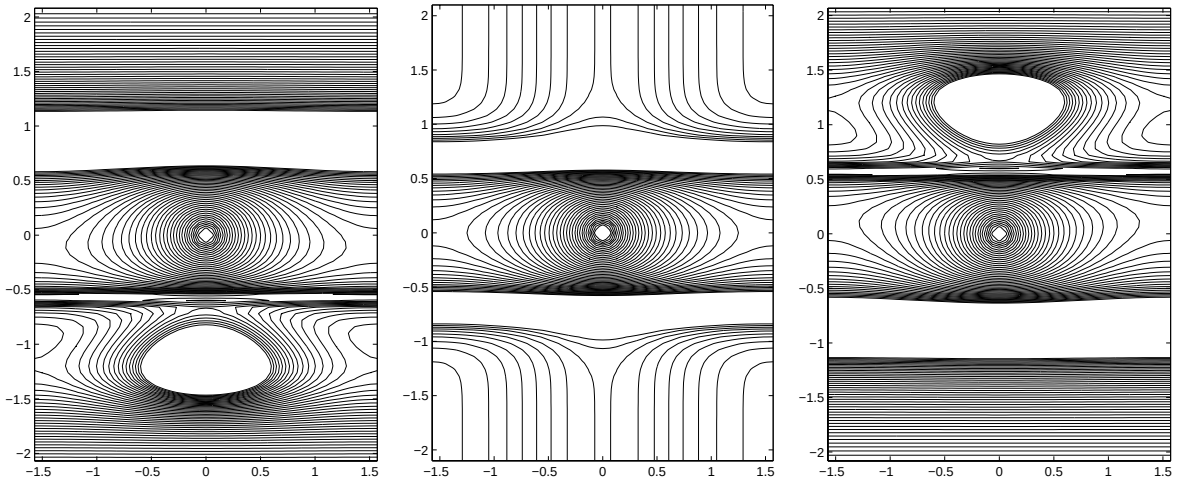


Figure 4.

The plots of Figure 4 depict the level sets of H as a function of ζ for $\Gamma/\Gamma' < 1, = 1, > 1$, respectively; they were drawn at $b = 1$. By (4.2), the levels for $\Gamma/\Gamma' < 1$ and > 1 are mirror images of each other via $\zeta \mapsto \overline{\zeta}$. The blank holes and bands indicate where H diverges to $-\infty$ too steeply, while the diamond in the middle of each plot surrounds a peak $H \to +\infty$.

Take the $N = 4$ case as in Theorem 4, and initially align the four vortices vertically: $\xi(0) = 0$. If $\eta(0)$ is sufficiently small, the vortices of the group Γ, Γ' orbit like a binary counterclockwise, and the vortices of the group $-\Gamma', -\Gamma$ orbit like a binary clockwise, while the two groups progress together like a vortex pair. The superposition produces *leapfrogging*, a relative periodic orbit whose plane limit $r \to \infty$ is observed as the motion of a cross-section of consecutive vortex rings as they overtake each other. By adjusting the parameters Γ/Γ', b, $\zeta(0)$, we can render leapfrogging on a cylinder not only relative periodic but periodic. Alternatively, if $\eta(0)$ is sufficiently close to $b(1 + \Gamma'/\Gamma)/2$ or to $-b(1 + \Gamma/\Gamma')/2$, the vortices $\Gamma', -\Gamma'$ or $\Gamma, -\Gamma$ form a pair and rush off without leapfrogging. In the planar theory, in the case $\Gamma/\Gamma' = 1$, [10] calculated the critical value of $\eta(0)$ that separates the leapfrogging and nonleapfrogging regimes. In our setup, this value may be obtained at once as follows.

In the situation of Theorem 4, denote by $\rho(b, \Gamma/\Gamma')$ the distance from the origin $\zeta = 0$ to the nearest separatrix. Then $\eta(0) = \rho(b, \Gamma/\Gamma) = \rho(b, 1)$. Denote by $\zeta_{\mathrm{re}} = \xi_{\mathrm{re}} + i\eta_{\mathrm{re}}$ a value of ζ at a saddle of $H(\zeta)$, representing a relative equilibrium. Inside the separatrices connecting the saddles, we have leapfrogging; outside, we do not. $\rho = \rho(b, 1)$ is the ordinate at which a separatrix cuts the η-axis. Since the value of H is the same along the separatrices as on the saddles, $H(0, \rho) = H(\xi_{\mathrm{re}}, \eta_{\mathrm{re}})$. It is clear that a relative equilibrium occurs when two vortex pairs are antipodal: $\xi_{\mathrm{re}} = \pm\pi/2, \eta_{\mathrm{re}} = 0$. This fixes ρ in the cylindrical theory: $\sqrt{2}\tanh\rho = \tanh b$. Restoring r and taking the plane limit $r \to \infty$, we get in the planar theory $\rho = b/\sqrt{2}$, agreeing with [10, section 3], which arrived at $(b + \rho)/(b - \rho) = 3 + 2\sqrt{2}$.

When $\Gamma/\Gamma' \neq 1$, ζ_{re} and $\rho(b, \Gamma/\Gamma')$ are difficult to pin down in closed form. At any rate

$\xi_{\mathrm{re}} = \pm\pi/2$; η_{re} is the unique root of

$$(\Gamma+\Gamma')\tanh\eta + (\Gamma-\Gamma')\tanh\left(b - \frac{\Gamma-\Gamma'}{\Gamma+\Gamma'}\eta\right) - \Gamma\coth\left(b+\frac{2\eta}{1+\Gamma/\Gamma'}\right) + \Gamma'\coth\left(b-\frac{2\eta}{1+\Gamma'/\Gamma}\right) = 0,$$

which in view of (2.4) is the condition that the vertically aligned pairs $\Gamma, -\Gamma$ and $\Gamma', -\Gamma'$, antipodal to each other, move with the same velocity. If $\Gamma/\Gamma' = 1+\epsilon$, then up to second order in ϵ,

$$\eta_{\mathrm{re}} \simeq \tanh b\,\operatorname{sech}^2 b\left(\frac{\epsilon}{2} - \left(1+\frac{\operatorname{sech}^4 b}{2}\right)\frac{\epsilon^2}{4}\right), \qquad \rho(b,1+\epsilon) = \rho(b,1) - \frac{\tanh b\,\operatorname{sech}^2 b}{1+\cosh^2 b}\,\frac{\epsilon^2}{4\sqrt{2}}.$$

Remark 6. By an argument parallel to that of Theorem 4, we see that four vortices with vorticities $\Gamma, \Gamma', -\Gamma', -\Gamma$ at

$$a + \frac{2\Gamma'}{\Gamma+\Gamma'}\zeta, \quad a - \frac{2\Gamma}{\Gamma+\Gamma'}\zeta, \quad -a + \frac{2\Gamma}{\Gamma+\Gamma'}\bar{\zeta}, \quad -a - \frac{2\Gamma'}{\Gamma+\Gamma'}\bar{\zeta}$$

leapfrog as well (Figure 5, left diagram). Unlike the $N = 4$ case of Theorem 4, however, the configuration on the right does not leapfrog.

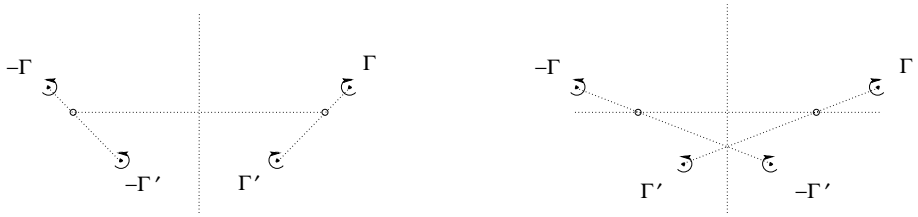


Figure 5.

Remark 7. Leapfrogging vortices and their generalizations analyzed above owe their relative periodicity to the type of symmetry compatible with the local first integral of Theorem 1. Other types of symmetry permit other types of relative periodic orbits. Thus $2n$ vortices with identical vorticity Γ at (3.1), (3.2) form a relative periodic orbit [13, section 3.2].

Remark 8. Vortex streets and leapfrogging vortices can be adapted to a torus, where they form relative periodic orbits. A torus accommodates many further types of relative periodic orbits. For example, on $\mathbb{C}/(\pi\mathbb{Z}+i\pi\mathbb{Z})$, by splitting each point of a sublattice into a rectangular quadruplet of vortices with vorticities $\Gamma, -\Gamma, \Gamma, -\Gamma$, we create a periodic orbit, the "dancing vortices" of [15].

Acknowledgments. Tadashi Tokieda thanks Yiannis Petridis and Morikazu Toda for instructive conversations and David Acheson for his gift of a copy of the paper [10]. The authors would like to thank the referees for their comments and particularly for pointing out the paper of Hally [4].

REFERENCES

[1] H. Aref, P. K. Newton, M. A. Stremler, T. Tokieda, and D. L. Vainchtein, *Vortex crystals*, in Advances in Applied Mechanics, Adv. Appl. Mech. 29, Academic Press, Boston, 2003, pp. 1–79.

[2] H. Aref and M. A. Stremler, *On the motion of three point vortices in a periodic strip*, J. Fluid Mech., 314 (1996), pp. 1–25.

[3] U. Domm, *Über die Wirbelstraßen von geringster Instabilität*, Z. Angew. Math. Mech., 36 (1956), pp. 367–371.

[4] D. Hally, *Stability of streets of vortices on surfaces of revolution with a reflection symmetry*, J. Math. Phys., 21 (1980), pp. 211–217.

[5] I. Imai, *On the stability of a double row of vortices with unequal strengths in a channel of finite breadth*, Proc. Physico-Math. Soc. Japan, 18 (1936), pp. 436–459.

[6] T. von Kármán, *Über den Mechanismus des Widerstandes, den ein bewegter Körper in einer Flüssigkeit erfährt*, Nachr. Ges. Wiss. Göttingen math.-phys. Klasse, (1911), pp. 509–517; (1912), pp. 547–556.

[7] H. Lamb, *Hydrodynamics*, 6th ed., Cambridge University Press, Cambridge, UK, 1932.

[8] F. Laurent-Polz, *Point vortices on the sphere: A case with opposite vorticities*, Nonlinearity, 15 (2002), pp. 143–171.

[9] C. Lim, J. Montaldi, and R. M. Roberts, *Relative equilibria of point vortices on the sphere*, Phys. D, 148 (2001), pp. 97–135.

[10] A. E. H. Love, *On the motion of paired vortices with a common axis*, Proc. London Math. Soc., 25 (1894), pp. 185–194.

[11] A. W. Maue, *Zur Stabilität der Kármánschen Wirbelstraße*, Z. Angew. Math. Mech., 20 (1940), pp. 129–137.

[12] K. A. O'Neil, *On the Hamiltonian dynamics of vortex lattices*, J. Math. Phys., 30 (1989), pp. 1373–1379.

[13] A. Soulière and T. Tokieda, *Periodic motions of vortices on surfaces with symmetry*, J. Fluid Mech., 460 (2002), pp. 83–92.

[14] M. A. Stremler and H. Aref, *Motion of three point vortices in a periodic parallelogram*, J. Fluid Mech., 392 (1999), pp. 101–128.

[15] T. Tokieda, *Tourbillons dansants*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 943–946.

[16] M. Van Dyke, *An Album of Fluid Motion*, Parabolic Press, Stanford, CA, 1982.

[17] H. Villat, *Leçons sur la théorie des tourbillons*, Gauthier-Villars, Paris, 1930.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 3, pp. 431–463

Effective Equations Modeling the Flow of a Viscous Incompressible Fluid through a Long Elastic Tube Arising in the Study of Blood Flow through Small Arteries*

Sunčica Čanić† and Andro Mikelić‡

Abstract. We study the flow of an incompressible viscous fluid through a long tube with compliant walls. The flow is governed by a given time-dependent pressure drop between the inlet and the outlet boundary. The pressure drop is assumed to be small, thereby introducing creeping flow in the tube. Stokes equations for incompressible viscous fluid are used to model the flow, and the equations of a curved, linearly elastic membrane are used to model the wall. Due to the creeping flow and to small displacements, the interface between the fluid and the lateral wall is linearized and supposed to be the initial configuration of the membrane. We study the dynamics of this coupled fluid-structure system in the limit when the ratio between the characteristic width and the characteristic length tends to zero. Using the asymptotic techniques typically used for the study of shells and plates, we obtain a set of Biot-type visco-elastic equations for the effective pressure and the effective displacements. The approximation is rigorously justified through a weak convergence result and through the error estimates for the solution of the effective equations modified by an outlet boundary layer.

Applications of the model problem include blood flow in small arteries. We recover the well-known law of Laplace and obtain new improved models that hold in cases when the shear modulus of the vessel wall is not negligible and the Poisson ratio is arbitrary.

Key words. fluid-structure interaction, Navier–Stokes equations, asymptotic analysis, blood flow, compliant vessels

AMS subject classifications. 76Z05, 76D07, 35Q30

DOI. 10.1137/S1111111102411286

1. Introduction.

This paper is motivated by the study of blood flow through compliant vessels. One of the open mathematical problems in this research area is rigorous derivation and error analysis of the asymptotic equations that hold in long and narrow axisymmetric vessels. In particular, it is important to analyze fluid-structure interaction resulting from the nonlinear coupling between the incompressible Navier–Stokes equations and the motion of a compliant vessel wall. Modeling of compliant vessel walls is a complex problem in its own right [11, 23]. Even in the simplified case when the anisotropic behavior of the vessel wall is ignored and angular deformations are neglected, in which case the linear Navier equations for the curved membrane can be used to model the wall [13, 16, 23], the analysis of the nonlinear coupling between the flow equations (Navier–Stokes equations) and wall behavior (Navier membrane equations) is unresolved. See [23] for the closest results. As a first step

*Received by the editors July 16, 2002; accepted for publication (in revised form) by J. Keener June 17, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/siads/2-3/41128.html

†Department of Mathematics, University of Houston, 4800 Calhoun Rd., Houston, TX 77204-3476 (canic@math.uh.edu). The research of this author was supported by the Texas Higher Education Board, ARP grant #003652-0112-2001, and by the National Science Foundation under grants DMS-9970310 and DMS-0245513.

‡UFR Mathématiques, Université Claude Bernard Lyon 1, Bât. 101, 43, bd. du 11 nov. 1918, 69622 Villeurbanne Cedex, France (amikelic@free.fr).

in analyzing this problem we focus in this paper on understanding the coupling between the Stokes equations (creeping flow) and the Navier equations for a curved elastic membrane. This is a good model for the flow of blood in small arteries. Indeed, it was noted in [20] that in small arteries, viscous effects of blood become more important than the inertia effects, and therefore Stokes equations are appropriate. We study the time-dependent flow governed by a given time-dependent pressure drop between the inlet and the outlet boundary. We refer the reader to books [5, 14, 15] for a general geometric setting of boundary-value problems in thin domains.

In this paper we derive the reduced (effective) equations that hold for this fluid-structure interaction problem when the ratio ε between the radius and the length of the vessel is small. The reduced equations are justified by showing that the solution of the original problem converges to the solution of the reduced problem as $\varepsilon \to 0$. Furthermore, we obtain the precise error estimates which show that the error in the solution of the reduced problem is of order $\mathcal{O}(\varepsilon^2)$ everywhere except at the outlet boundary, where the formation of the boundary layer deteriorates the error to $\mathcal{O}(\varepsilon^{3/2})$. Derivation of the reduced equations and the precise error estimates are the main original contribution of this work. Numerical simulations showing the dynamics of the fluid-structure interaction and the formation of the boundary layer are presented at the end of the paper.

It is interesting to point out that we recover the well-known law of Laplace [11], namely, the "independent ring model," given by (5.23) [11, 20, 4], as a special case of the parameter values, namely, in the case when the Poisson ratio of the wall structure is 0.5 and shear modulus of the vessel wall is negligible. For a general Poisson ratio, we provide a general pressure-displacement relationship. Finally, we obtain a model which holds in the situations when the shear modulus is not negligible. Such situations arise, for example, in "stented" arteries (arteries treated with prostheses to prevent occlusion or to prevent aneurysm rupture [3, 10]). Our analysis provides the reduced equations, which are a more accurate model in these scenarios. Their simple form in terms of the pressure makes the resulting equations particularly attractive for numerical simulations where they can be coupled to the two- or three-dimensional solvers in nonuniform geometries or used, for example, in the study of blood flow in small (e.g., coronary) branching arteries, with the appropriate coupling at the branching locations (see, for example, [21]).

This paper is organized as follows. We define the problem in section 2. In section 3 we obtain the energy estimates for the above-mentioned fluid-structure interaction problem and show how forcing, given in terms of the pressure drop, controls the elastic and viscous energy. From the energy equality we obtain (optimal) a priori estimates of the solution given in terms of the small parameter ε. The a priori estimates and the choice of the "correct" time-scale (to capture the oscillations induced by the outside forcing and governed by the fluid motion) provide leading order behavior of the state variables and the correct form of the asymptotic expansions. Asymptotic expansions, which are studied on a rescaled domain, are presented in section 4. By inserting the "correct" asymptotic expansions into the coupled fluid-structure system, we obtain the reduced equations in section 5. The reduced equations are a second-order approximation to the original problem in the interior of the domain. The approximation is rigorously justified by obtaining a weak convergence result and by calculating the corresponding error estimates. The weak convergence result, presented in section 6, shows

that the solution of the original problem converges weakly to the unique solution of the asymptotic equations. The error estimates, obtained in section 7, show the order of the approximation and reveal the formation of a boundary layer at the outlet boundary. See Figure 7.1. The boundary layer at the outlet boundary gives rise to major technical difficulties in the error estimates. To the authors' knowledge, there are no mathematical results on higher order approximations and error estimates for fluid-structure problems in thin domains. This is primarily due to various difficulties associated with dealing with boundary conditions. The closest results on error estimates are the ones presented in [19], where higher order approximations and error estimates for the Stokes flow through *fixed* domains with small thickness are given, with the given velocity field at the lateral boundary. This is different from our approach. In section 7 we explicitly construct the outlet boundary layer and find the error estimates for the approximate solution modified by the boundary layer solution. We show that due to the presence of the boundary layer, the error in the approximation is of order $\mathcal{O}(\varepsilon^{3/2})$. This is by $\sqrt{\varepsilon}$ order of magnitude less than the accuracy obtained in [19]. The deterioration of the order of approximation is concentrated at the outlet boundary. This can be seen in the numerical simulations, presented in Figure 7.1.

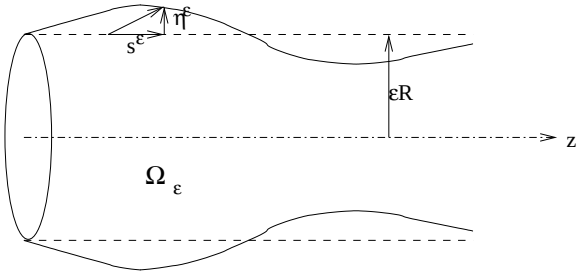


Figure 2.1. *Wall displacement.*

2. Statement of the problem. We consider the unsteady axisymmetric flow of a Newtonian incompressible fluid in a thin right cylinder whose radius is small with respect to its length. Define the ratio between the radius and the length of the cylinder to be ϵ. For each fixed $\epsilon > 0$ introduce Ω_ε to be

$$\Omega_\varepsilon = \big\{x \in \mathbb{R}^3; x = (r\cos\vartheta, r\sin\vartheta, z),\ r < \varepsilon R,\ 0 < z < L\big\}. \tag{2.1}$$

We assume that the cylinder's lateral wall $\Sigma_\varepsilon = \{r = \varepsilon R\} \times (0, L)$ is elastic and that its motion is described in Lagrangian coordinates by the Navier equations

$$F_r = -\frac{h(\varepsilon)E(\varepsilon)}{1-\sigma^2}\left(\frac{\sigma}{\varepsilon R}\frac{\partial s^\varepsilon}{\partial z} + \frac{\eta^\varepsilon}{\varepsilon^2 R^2}\right) + h(\varepsilon)G(\varepsilon)k(\varepsilon)\frac{\partial^2\eta^\varepsilon}{\partial z^2} - \rho_w h(\varepsilon)\frac{\partial^2\eta^\varepsilon}{\partial t^2}, \tag{2.2}$$

$$F_z = \frac{h(\varepsilon)E(\varepsilon)}{1-\sigma^2}\left(\frac{\partial^2 s^\varepsilon}{\partial z^2} + \frac{\sigma}{\varepsilon R}\frac{\partial\eta^\varepsilon}{\partial z}\right) - \rho_w h(\varepsilon)\frac{\partial^2 s^\varepsilon}{\partial t^2}. \tag{2.3}$$

Here η^ε is the radial and s^ε is the longitudinal displacement from the reference state (see Figure 2.1), $h = h(\varepsilon)$ is the membrane thickness, ρ_w is the wall volumetric mass, $E = E(\varepsilon)$ is the Young's modulus, $0 < \sigma < 1/2$ is the Poisson ratio, $G = G(\varepsilon)$ is the shear modulus, and

Table 2.1
Parameter values.

Parameters	Values
ε	0.04
Characteristic radius: εR	0.004 m
Dynamic viscosity: μ	$3.4 \times 10^{-3} m^2/s$
Young's modulus: E	6000 Pa, [6]
Shear modulus: G*k	5×10^5 Pa, [11]
Wall thickness: h	4×10^{-4} m
Wall density: ρ_w	$1.1 kg/m^2$, [23]
Blood density: ρ	$1050 kg/m^3$
Reference pressure: P_0	13000 Pa
Normalized pressure drop:	$\varepsilon^{1/2}$

$k = k(\varepsilon)$ is the Timoshenko shear correction factor (see [16, 23]). F_r and F_z are the radial and the longitudinal component of the external forces, coming from the stresses induced by the fluid. For the underlying blood-flow problem, the parameter values are presented in Table 2.1. Throughout the paper we will be assuming the following relationships between the parameters in the model.

Assumption 1. *The Young's modulus, the wall thickness, and the shear modulus satisfy*

$$h(\varepsilon)E(\varepsilon) > \varepsilon, \tag{2.4}$$

$$\lim_{\varepsilon \to 0} \frac{h(\varepsilon)E(\varepsilon)}{\varepsilon} = E_0 \in (0, +\infty), \tag{2.5}$$

$$\lim_{\varepsilon \to 0} \varepsilon h(\varepsilon)G(\varepsilon)k(\varepsilon) = G_0 \in [0, +\infty). \tag{2.6}$$

Initially, the cylinder is filled with fluid and the entire structure is in an equilibrium. The equilibrium state has an initial reference pressure P_0 and the initial velocity is zero. If we denote by T the (membrane) stress tensor, then in the equilibrium (unperturbed) state only the T_{zz} and $T_{\vartheta\vartheta}$ components of the stress tensor corresponding to the curved membrane Σ_ε are not zero (see [16, 23]). Their values are kG and $\varepsilon R \Delta P_0/h$, respectively, where ΔP_0 is the difference between the reference pressure in the tube and the surrounding tissue. For simplicity we assume that $\Delta P_0 = 0$; hence $T_{\vartheta\vartheta}$ is zero in the unperturbed state. A pressure difference between the inlet and the outlet boundary of Ω_ε creates a deviation from the unperturbed state. We assume that the pressure drop is small compared to the reference pressure and that the fluid acceleration is negligible compared to the effects of the fluid viscosity μ. Therefore, we can use the axially symmetric incompressible Stokes system to model fluid velocity $v^\varepsilon = (v^\varepsilon{}_r, v^\varepsilon{}_\theta, v^\varepsilon{}_z)$ and the pressure perturbation, p^ε, from the reference pressure P_0. Assuming zero angular velocity, in cylindrical coordinates the Eulerian formulation of the problem reads

$$-\mu\left(\frac{\partial^2 v_r^\varepsilon}{\partial r^2} + \frac{\partial^2 v_r^\varepsilon}{\partial z^2} + \frac{1}{r}\frac{\partial v_r^\varepsilon}{\partial r} - \frac{v_r^\varepsilon}{r^2}\right) + \frac{\partial p^\varepsilon}{\partial r} = 0 \quad \text{in} \quad \Omega_\varepsilon \times \mathbb{R}_+, \tag{2.7}$$

$$-\mu\left(\frac{\partial^2 v_z^\varepsilon}{\partial r^2} + \frac{\partial^2 v_z^\varepsilon}{\partial z^2} + \frac{1}{r}\frac{\partial v_z^\varepsilon}{\partial r}\right) + \frac{\partial p^\varepsilon}{\partial z} = 0 \quad \text{in} \quad \Omega_\varepsilon \times \mathbb{R}_+, \tag{2.8}$$

$$\frac{\partial v_r^\varepsilon}{\partial r} + \frac{\partial v_z^\varepsilon}{\partial z} + \frac{v_r^\varepsilon}{r} = 0 \quad \text{in} \quad \Omega_\varepsilon \times \mathbb{R}_+. \tag{2.9}$$

These equations are coupled with the Navier equations for the curved membrane through the lateral boundary conditions requiring continuity of velocity and continuity of forces at the wall Σ_ε. More specifically, we require

$$v_r^\varepsilon = \frac{\partial \eta^\varepsilon}{\partial t} \quad \text{on} \quad \Sigma_\varepsilon \times \mathbb{R}_+, \tag{2.10}$$

$$v_z^\varepsilon = \frac{\partial s^\varepsilon}{\partial t} \quad \text{on} \quad \Sigma_\varepsilon \times \mathbb{R}_+, \tag{2.11}$$

and we set the radial and longitudinal forces F_r and F_z in (2.2) and (2.3) equal to the radial and longitudinal component of the stress exerted by the fluid to the membrane

$$-F_r = \big(p^\varepsilon I - 2\mu D(v^\varepsilon)\big)\vec{e}_r \cdot \vec{e}_r \quad \text{on} \quad \Sigma_\varepsilon \times \mathbb{R}_+, \tag{2.12}$$

$$-F_z = \big(p^\varepsilon I - 2\mu D(v^\varepsilon)\big)\vec{e}_r \cdot \vec{e}_z \quad \text{on} \quad \Sigma_\varepsilon \times \mathbb{R}_+, \tag{2.13}$$

where $D(v^\varepsilon)$ is the rate of the strain tensor, i.e., the symmetrized gradient of the velocity

$$D(v^\varepsilon) = \frac{1}{2}(\nabla v^\varepsilon + (\nabla v^\varepsilon)^t).$$

We note that in this approximation the interface is identified with the reference elastic wall Σ_ϵ.

The initial state of the structure is unperturbed and at initial velocity zero

$$\eta^\varepsilon = s^\varepsilon = \frac{\partial \eta^\varepsilon}{\partial t} = \frac{\partial s^\varepsilon}{\partial t} = 0 \quad \text{on} \quad \Sigma_\varepsilon \times \{0\}, \tag{2.14}$$

and we consider the following boundary data, which, as we shall see in section 6.1, give rise to a well-posed initial-boundary-value problem for the limiting configuration ($\epsilon \to 0$):

$$v_r^\varepsilon = 0 \ \text{ and } \ p^\varepsilon = 0 \quad \text{on} \quad (\partial\Omega_\varepsilon \cap \{z = 0\}) \times \mathbb{R}_+, \tag{2.15}$$

$$v_r^\varepsilon = 0 \ \text{ and } \ p^\varepsilon = A(t) \quad \text{on} \quad (\partial\Omega_\varepsilon \cap \{z = L\}) \times \mathbb{R}_+, \tag{2.16}$$

$$\frac{\partial s^\varepsilon}{\partial z} = \eta^\varepsilon = 0 \ \text{ for } \ z = 0, \quad s^\varepsilon = \eta^\varepsilon = 0 \ \text{ for } \ z = L \ \text{ and } \ \forall t \in \mathbb{R}_+. \tag{2.17}$$

Notice that pressure drop $A(t)$ drives the problem. For simplicity we suppose that $A \in C_0^\infty(0, +\infty)$. Note that physically one should expect nonzero displacements at the outlet boundary. The fixed outlet boundary, required in (2.17), gives rise to the formation of a boundary layer. See sections 6.1 and 7. Periodic boundary conditions, although natural in rigid-wall geometries, do not give rise to well-posed limiting problems when compliant walls are considered.

We summarize the initial-boundary-value problem for the coupled fluid-structure interaction driven by the time-dependent pressure drop between the inlet and the outlet boundary.

Problem P^ε. *For each fixed $\varepsilon > 0$, find a solution to (2.7), (2.8), and (2.9) in domain Ω_ε defined by (2.1), with an elastic lateral boundary Σ_ε. The lateral boundary conditions are given by the continuity of the velocity (2.10) and (2.11) and by the continuity of forces (2.2) and (2.3), where the left-hand sides of (2.2) and (2.3) are substituted by (2.12) and (2.13),*

respectively. The boundary conditions at the inlet and outlet boundaries are (2.15) *and* (2.16), *and the behavior of the elastic wall there is prescribed by* (2.17). *The initial data is given by* (2.14).

In the next section we derive the weak formulation and the energy equality and obtain the a priori estimates for the solution. The existence of a unique weak solution to this problem is straightforward.

3. Weak formulation and energy estimates.

3.1. Weak formulation. We define the space of test functions V^ε and the solution space $\mathcal{V}^\varepsilon$ to be the following.

Definition 1. *The space* $V^\varepsilon \subset H^1(\Omega_\varepsilon)^3$ *consists of all the axially symmetric functions* φ *such that* $\varphi_r|_{\Sigma_\varepsilon}, \varphi_z|_{\Sigma_\varepsilon} \in H^1(0,L)$, $\varphi_r(0,r) = \varphi_z(L,\varepsilon R) = \varphi_r(L,r) = 0$ *for* $r \le \varepsilon R$, *and* $\operatorname{div}\varphi = 0$ *in* Ω_ε.

Definition 2. *The space* $\mathcal{V}^\varepsilon$ *consists of all the functions* $(w_r, w_z, d_r, d_z) \in H^1((0,T); V^\varepsilon) \times (H^1((0,L)\times(0,T))^2 \cap H^2(0,T; L^2(0,L))^2)$ *such that*

1. $\frac{\partial w_r}{\partial r} + \frac{\partial w_z}{\partial z} + \frac{w_r}{r} = 0$ *in* $\Omega_\varepsilon \times \mathbb{R}_+$,
2. $r^{-1} w_r \in L^2((0,T)\times\Omega_\varepsilon)$,
3. $d_r(t,0) = d_z(t,L) = d_r(t,L) = 0$ *on* $\mathbb{R}_+$,
4. $w_r = 0$ *on* $(\partial\Omega_\varepsilon \cap \{z=0\}) \times \mathbb{R}_+$, *and*
5. $w_r = \frac{\partial d_r}{\partial t}$, *and* $w_z = \frac{\partial d_z}{\partial t}$ *on* $\Sigma_\varepsilon \times \mathbb{R}_+$.

Recall that for an axially symmetric vector valued function $\psi = \psi_r \vec{e}_r + \psi_z \vec{e}_z$ we have

$$D(\psi) = \begin{pmatrix} \dfrac{\partial \psi_r}{\partial r} & 0 & \dfrac{1}{2}\left(\dfrac{\partial \psi_r}{\partial z} + \dfrac{\partial \psi_z}{\partial r}\right) \\ 0 & \dfrac{\psi_r}{r} & 0 \\ \dfrac{1}{2}\left(\dfrac{\partial \psi_r}{\partial z} + \dfrac{\partial \psi_z}{\partial r}\right) & 0 & \dfrac{\partial \psi_z}{\partial z} \end{pmatrix}.$$

Define the matrix norm $|\cdot|$ through the scalar product

$$\Xi : \Psi = T_r(\Xi \cdot \Psi^t),\ \Xi, \Psi \in \mathbb{R}^9.$$

Then for each fixed $\varepsilon > 0$ the variational formulation and weak solution are defined by the following.

Definition 3. *Vector function* $(v_r^\varepsilon, v_z^\varepsilon, \eta^\varepsilon, s^\varepsilon) \in \mathcal{V}^\varepsilon$ *is a weak solution of problem* P^ε *if the following variational formulation is satisfied:*

$$\begin{aligned} &2\mu \int_{\Omega_\varepsilon} D(v^\varepsilon) : D(\varphi)\ r dr dz \\ &+ \varepsilon R \int_0^L \left\{ h(\varepsilon) G(\varepsilon) k(\varepsilon) \frac{\partial \eta^\varepsilon}{\partial z} \frac{\partial \varphi_r}{\partial z} + \frac{h(\varepsilon) E(\varepsilon)}{1-\sigma^2} \left(\frac{\sigma}{\varepsilon R} \frac{\partial s^\varepsilon}{\partial z} + \frac{\eta^\varepsilon}{\varepsilon^2 R^2} \right) \varphi_r \right. \\ &\left. + \frac{h(\varepsilon) E(\varepsilon)}{1-\sigma^2} \left(\frac{\partial s^\varepsilon}{\partial z} \frac{\partial \varphi_z}{\partial z} - \frac{\sigma}{\varepsilon R} \frac{\partial \eta^\varepsilon}{\partial z} \varphi_z \right) \right\} |_{r=\varepsilon R}\ dz + \varepsilon R \rho_w h \frac{d^2}{dt^2} \int_0^L \left(\eta^\varepsilon \varphi_r + s^\varepsilon \varphi_z \right) |_{r=\varepsilon R}\ dz \\ &= -\int_0^{\varepsilon R} A(t) \varphi_z|_{z=L}\ r dr \quad in \quad \mathcal{D}'(\mathbb{R}_+) \quad \forall \varphi = \varphi_r \vec{e}_r + \varphi_z \vec{e}_z \in V^\varepsilon \end{aligned} \tag{3.1}$$

and

$$\eta^\varepsilon = s^\varepsilon = \frac{\partial \eta^\varepsilon}{\partial t} = \frac{\partial s^\varepsilon}{\partial t} = 0 \quad on \quad \Sigma_\varepsilon \times \{0\}. \tag{3.2}$$

Existence of a unique solution to problem (3.1)–(3.2) is classical. It is a consequence of the corresponding energy estimate and of Korn's inequality [9].

Our goal is to study the behavior of the solution to (3.1)–(3.2) in the limit, as $\varepsilon \to 0$. In order to do that we need an optimal energy estimate with respect to ε. Once the optimal energy estimate is obtained, a priori solution estimates will follow, and the correct asymptotic expansions will be obtained. By plugging the asymptotic expansions into the original equations, the reduced equations, second-order accurate in ε, will be derived. We will see in section 3.2 that obtaining the optimal energy estimate is nonstandard because we do not have the no-slip boundary condition at the lateral boundary and because we will have to deal with the viscous energy which is given in terms of the symmetrized gradient of the velocity $D(v^\varepsilon)$ instead of ∇v^ε.

3.2. Energy estimate. The energy of this problem, obtained by using the velocity field as a test function in (3.1), consists of the elastic energy of the membrane, the viscous energy of the fluid, and the energy due to the outside forcing. The time derivative of the elastic energy

$$\begin{aligned}\frac{d\mathcal{E}_{el}}{dt} \equiv \varepsilon R \int_0^L \Bigg\{ & h(\varepsilon)G(\varepsilon)k(\varepsilon)\frac{\partial \eta^\varepsilon}{\partial z}\frac{\partial^2 \eta^\varepsilon}{\partial z \partial t} + \frac{h(\varepsilon)E(\varepsilon)}{1-\sigma^2}\Bigg(\left(\frac{\sigma}{\varepsilon R}\frac{\partial s^\varepsilon}{\partial z} + \frac{\eta^\varepsilon}{\varepsilon^2 R^2}\right)\frac{\partial \eta^\varepsilon}{\partial t} \\ & + \left(\frac{\partial s^\varepsilon}{\partial z}\frac{\partial^2 s^\varepsilon}{\partial z \partial t} - \frac{\sigma}{\varepsilon R}\frac{\partial \eta^\varepsilon}{\partial z}\frac{\partial s^\varepsilon}{\partial t}\right)\Bigg) + \varepsilon R \rho_w h(\varepsilon)\left(\frac{\partial^2 \eta^\varepsilon}{\partial t^2}\frac{\partial \eta^\varepsilon}{\partial t} + \frac{\partial^2 s^\varepsilon}{\partial t^2}\frac{\partial s^\varepsilon}{\partial t}\right)\Bigg\} dz\end{aligned}$$

can be expressed as follows.

Lemma 3.1. *The displacements η^ε and s^ε satisfy*

$$\begin{aligned}\frac{d\mathcal{E}_{el}}{dt} = \varepsilon R \frac{d}{2dt}\Bigg\{ & \rho_w h(\varepsilon)\int_0^L \left(\left|\frac{\partial \eta^\varepsilon}{\partial t}\right|^2 + \left|\frac{\partial s^\varepsilon}{\partial t}\right|^2\right) dz + h(\varepsilon)G(\varepsilon)k(\varepsilon)\int_0^L \left|\frac{\partial \eta^\varepsilon}{\partial z}\right|^2 dz \\ & + \frac{h(\varepsilon)E(\varepsilon)}{1-\sigma^2}\left(\sigma \int_0^L \left(\frac{\eta^\varepsilon}{\varepsilon R} - \frac{\partial s^\varepsilon}{\partial z}\right)^2 + (1-\sigma)\int_0^L \left(\left|\frac{\eta^\varepsilon}{\varepsilon R}\right|^2 + \left|\frac{\partial s^\varepsilon}{\partial z}\right|^2\right)\right)\Bigg\}.\end{aligned} \tag{3.3}$$

We are interested in the oscillations of the membrane that are due to the time-dependent pressure drop $A(t)$. These occur at a different time-scale than the characteristic "physical" time. In particular, as we will see later in the text, fluid velocity is greater than the velocity of the displacement. This, in turn, gives rise to long-wavelength elastic waves. It is these waves, among other things, that we would like to keep in our asymptotic reduction. This is why we introduce a new time-scale

$$\tilde{t} = \omega^\varepsilon t, \tag{3.4}$$

where the characteristic frequency ω^ε will be specified later (see (3.11)) to include both the waves that occur at the leading order time-scale as well as the oscillations of the membrane caused by a response of the elastic material. The pressure drop is supposed to be a function of $\tilde{t}$.

From now on we use the rescaled time $\tilde{t}$ and drop the tilde. By keeping the rescaled time in mind and by using the expression for the elastic energy (3.3), we obtain the following.

Proposition 3.2. *Solution* $(v_r^\varepsilon, v_z^\varepsilon, \eta^\varepsilon, s^\varepsilon)$ *of problem* (3.1)–(3.2) *satisfies the variational equality*

(3.5)
$$\omega^\varepsilon h(\varepsilon)\frac{d}{2dt}\left\{(\omega^\varepsilon)^2\rho_w\varepsilon R\left(\left\|\frac{\partial\eta^\varepsilon(t)}{\partial t}\right\|^2_{L^2(0,L)}+\left\|\frac{\partial s^\varepsilon(t)}{\partial t}\right\|^2_{L^2(0,L)}\right)+G(\varepsilon)k(\varepsilon)\varepsilon R\left\|\frac{\partial\eta^\varepsilon(t)}{\partial z}\right\|^2_{L^2(0,L)}\right.$$
$$\left.+\frac{E(\varepsilon)\varepsilon R}{1-\sigma^2}\left(\sigma\left\|\frac{\eta^\varepsilon(t)}{\varepsilon R}-\frac{\partial s^\varepsilon(t)}{\partial z}\right\|^2_{L^2(0,L)}+(1-\sigma)\left(\left\|\frac{\eta^\varepsilon(t)}{\varepsilon R}\right\|^2_{L^2(0,L)}+\left\|\frac{\partial s^\varepsilon(t)}{\partial z}\right\|^2_{L^2(0,L)}\right)\right)\right\}$$
$$+2\mu\|D(v^\varepsilon(t))\|^2_{L^2(\Omega_\varepsilon)}=-\int_0^{\varepsilon R}A(t)v_z^\varepsilon(t,r,L)\,rdr,$$

with $v_r^\varepsilon=\omega^\varepsilon\frac{\partial\eta^\varepsilon}{\partial t}$ *and* $v_z^\varepsilon=\omega^\varepsilon\frac{\partial s^\varepsilon}{\partial t}$ *on* $\Sigma_\varepsilon\times(0,T)$.

We now investigate how the energy of the forcing term controls the elastic and the viscous energy of the coupled fluid-structure interaction. We start by transforming and estimating the right-hand side. Since we do not have the no-slip condition for the velocity at the lateral boundary, the situation is more complicated than in the derivation of Reynolds' equation. See [7] and the subsequent papers on the same subject. Furthermore, since on the left-hand side we have only the L^2-norm of $D(v^\varepsilon)$ and not the L^2-norm of ∇v^ε, a standard approach based on using the Gronwall estimate and the L^2-norm of the velocity, $\rho\int_{\Omega_\varepsilon}|v^\varepsilon(t)|^2\,rdrdz$, is insufficient to guarantee the correct order of magnitude of the velocity. To get around this difficulty we transform the right-hand side term in (3.5) to a combination of a volume term and a lateral boundary term

$$-\int_0^{\varepsilon R}A(t)v_z^\varepsilon(t,r,L)\,rdr=-\int_{\Omega_\varepsilon}\frac{A(t)}{L}v_z^\varepsilon\,rdrdz+\varepsilon R\int_0^L A(t)\frac{z}{L}v_r^\varepsilon(t,\varepsilon R,z)\,dz \tag{3.6}$$

and use the following variant of Biot law which will relate the forcing term with the volume shear stress term in the viscous energy and the elastic energy of the membrane.

Lemma 3.3. *The following estimate holds:*

$$\left|\int_{\Omega_\varepsilon}v_z^\varepsilon\,rdrdz-\omega^\varepsilon\frac{\varepsilon^2R^2}{2}\frac{\partial}{\partial t}\int_0^L s^\varepsilon(t,z)\,dz\right|\le\frac{R^2\sqrt{L}}{2}\varepsilon^2\left\|\frac{1}{2}\left(\frac{\partial v_z^\varepsilon}{\partial r}+\frac{\partial v_r^\varepsilon}{\partial z}\right)\right\|_{L^2(\Omega_\varepsilon)}. \tag{3.7}$$

Proof. The estimate follows by noticing that

$$\int_{\Omega_\varepsilon}v_z^\varepsilon r\,drdz=\int_0^L\left[\frac{r^2}{2}v_z^\varepsilon\right]_0^{\varepsilon R}dz-\int_0^L\int_0^{\varepsilon R}\frac{r^2}{2}\frac{\partial v_z^\varepsilon}{\partial r}\,drdz$$
$$=\omega^\varepsilon\frac{\partial}{\partial t}\frac{\varepsilon^2R^2}{2}\int_0^L s^\varepsilon(t,z)\,dz-\int_0^L\int_0^{\varepsilon R}\frac{r}{2}\left(\frac{\partial v_z^\varepsilon}{\partial r}+\frac{\partial v_r^\varepsilon}{\partial z}\right)r\,drdz.$$
■

Throughout the text we will be using the following notation:

$$\|A(t)\|^2_{\mathcal{H}}=\max_{0\le\tau\le t}|A(\tau)|^2+\int_0^t|\partial_\tau A(\tau)|^2\,d\tau.$$

An estimate for the axial component of the displacement in the energy equality (3.5) is given by the following lemma.

Lemma 3.4. *Axial displacement s^ε satisfies the estimate*

(3.8)
$$\begin{aligned}
&\frac{h(\varepsilon)E(\varepsilon)\varepsilon R}{2(1+\sigma)}\left\|\frac{\partial s^\varepsilon}{\partial z}(t)\right\|^2_{L^2(0,L)}+\frac{\varepsilon^2R^2}{2L}\left(A(t)\int_0^L s^\varepsilon(t,z)\,dz-\int_0^t\frac{\partial A(\tau)}{\partial\tau}\int_0^L s^\varepsilon(\tau,z)\,dzd\tau\right)\\
&\geq\frac{h(\varepsilon)E(\varepsilon)\varepsilon R}{4(1+\sigma)}\left\{\left\|\frac{\partial s^\varepsilon}{\partial z}(t)\right\|^2_{L^2(0,L)}+\left(\left\|\frac{\partial s^\varepsilon}{\partial z}(t)\right\|_{L^2(0,L)}-\frac{R(1+\sigma)\sqrt{L}}{3}\frac{A(t)\varepsilon}{h(\varepsilon)E(\varepsilon)}\right)^2\right.\\
&\left.-\int_0^t\left\|\frac{\partial s^\varepsilon}{\partial z}(\tau)\right\|^2_{L^2(0,L)}d\tau\right\}-\frac{\varepsilon^3}{h(\varepsilon)E(\varepsilon)}\frac{(1+\sigma)LR^3}{6}\|A(t)\|^2_{\mathcal{H}}.
\end{aligned}$$

Next, an estimate for the radial displacement is given by the following lemma.

Lemma 3.5. *Radial displacement η^ε satisfies the estimate*

$$\begin{aligned}
&\frac{h(\varepsilon)E(\varepsilon)}{2\varepsilon R(1+\sigma)}\|\eta^\varepsilon(t)\|^2_{L^2(0,L)}-\frac{\varepsilon R}{L}\left(A(t)\int_0^L z\eta^\varepsilon(t,z)\,dz-\int_0^t\frac{\partial A(\tau)}{\partial\tau}\int_0^L z\eta^\varepsilon(\tau,z)\,dzd\tau\right)\\
&\geq\frac{h(\varepsilon)E(\varepsilon)}{4\varepsilon R(1+\sigma)}\left\{\|\eta^\varepsilon(t)\|^2_{L^2(0,L)}-\int_0^t\|\eta^\varepsilon(\tau)\|^2_{L^2(0,L)}d\tau\right\}-\frac{\varepsilon^3}{h(\varepsilon)E(\varepsilon)}\frac{(1+\sigma)LR^3}{3}\|A(t)\|^2_{\mathcal{H}}.
\end{aligned}$$

Combining these two estimates, Lemma 3.3 and the following estimate, which relates the viscous energy with the forcing term,

$$\varepsilon^2\sqrt{\frac{R^4}{4L}}|A(t)|\left\|\frac{1}{2}\left(\frac{\partial v^\varepsilon_z}{\partial r}+\frac{\partial v^\varepsilon_r}{\partial z}\right)\right\|_{L^2(\Omega_\varepsilon)}\leq\mu\left\|\frac{1}{2}\left(\frac{\partial v^\varepsilon_z}{\partial r}+\frac{\partial v^\varepsilon_r}{\partial z}\right)\right\|^2_{L^2(\Omega_\varepsilon)}+\frac{R^4\varepsilon^4}{16\mu L}|A(t)|^2,$$

from (3.5) we obtain the following lemma.

Lemma 3.6. *The radial displacement η^ε, the axial displacement s^ε, the viscous energy $\mu\|D(v^\varepsilon)\|^2_{L^2(\Omega_\varepsilon)}$, and the energy induced by the pressure drop $A(t)$ satisfy the energy estimate*

(3.9)
$$\begin{aligned}
&\omega^\varepsilon\frac{h(\varepsilon)E(\varepsilon)}{4\varepsilon R(1+\sigma)}\left\{\|\eta^\varepsilon(t)\|^2_{L^2(0,L)}+\varepsilon^2R^2\left\|\frac{\partial s^\varepsilon}{\partial z}(t)\right\|^2_{L^2(0,L)}\right\}+\mu\int_0^t\|D(v^\varepsilon)(\tau)\|^2_{L^2(\Omega_\varepsilon)}d\tau\\
&\leq\omega^\varepsilon\frac{h(\varepsilon)E(\varepsilon)}{4\varepsilon R(1+\sigma)}\int_0^t\left\{\|\eta^\varepsilon(\tau)\|^2_{L^2(0,L)}+\varepsilon^2R^2\left\|\frac{\partial s^\varepsilon}{\partial z}(\tau)\right\|^2_{L^2(0,L)}\right\}d\tau\\
&+\frac{R^4\varepsilon^4}{16\mu L}\int_0^t|A(\tau)|^2\,d\tau+\omega^\varepsilon\frac{\varepsilon^3}{h(\varepsilon)E(\varepsilon)}\frac{(1+\sigma)LR^3}{2}\|A(t)\|^2_{\mathcal{H}}.
\end{aligned}$$

By applying the Gronwall inequality to (3.9), we get an estimate which is crucial in determining the leading order behavior in asymptotic expansions. The estimate is a basis for the a priori solution estimates in terms of the small parameter ϵ and the characteristic frequency ω^ε.

Proposition 3.7. *Solution* $(v_r^\varepsilon, v_z^\varepsilon, \eta^\varepsilon, s^\varepsilon)$ *of problem* (3.1)–(3.2) *satisfies the estimate*

$$\omega^\varepsilon \frac{h(\varepsilon)E(\varepsilon)}{4\varepsilon R(1+\sigma)}\left\{\|\eta^\varepsilon(t)\|^2_{L^2(0,L)} + \varepsilon^2 R^2 \left\|\frac{\partial s^\varepsilon}{\partial z}(t)\right\|^2_{L^2}\right\} + \mu\int_0^t \|D(v^\varepsilon)(\tau)\|^2_{L^2(\Omega_\varepsilon)}d\tau$$
$$\leq \left\{\frac{R^4\varepsilon^4}{16\mu L}\int_0^t |A(\tau)|^2\, d\tau + \omega^\varepsilon \frac{\varepsilon^3}{h(\varepsilon)E(\varepsilon)}\frac{(1+\sigma)LR^3}{2}\|A(t)\|^2_{\mathcal{H}}\right\}e^t. \tag{3.10}$$

To capture the elastic response of the membrane to the oscillations in the pressure drop between the inlet and the outlet boundary, ω^ε is chosen so that both terms on the right-hand side are of the same order in ϵ. Using assumption (2.4), we get

$$\omega^\varepsilon = \frac{\varepsilon^2}{\mu}. \tag{3.11}$$

We are now ready to obtain the a priori solution estimates in terms of ϵ. In the text that follows we denote all the constants independent of ε by C. Define

$$\|A\|^2_{\mathcal{V}} = e^T\left\{\|A\|^2_{L^\infty(0,T)} + \int_0^T \left(|\partial_\tau A(\tau)|^2 + |A(\tau)|^2\right)\, d\tau\right\}.$$

Proposition 3.8. *Solution* $(v_r^\varepsilon, v_z^\varepsilon, \eta^\varepsilon, s^\varepsilon)$ *of problem* (3.1)–(3.2) *satisfies the a priori estimates*

$$\int_0^t \left\{\left\|\frac{\partial v_r^\varepsilon}{\partial r}\right\|^2_{L^2(\Omega_\varepsilon)} + \left\|\frac{v_r^\varepsilon}{r}\right\|^2_{L^2(\Omega_\varepsilon)} + \left\|\frac{\partial v_z^\varepsilon}{\partial z}\right\|^2_{L^2(\Omega_\varepsilon)}\right\}d\tau \leq C\left(\frac{\varepsilon^2}{\mu}\right)^2\|A\|^2_{\mathcal{V}}, \tag{3.12}$$

$$\int_0^t \left\{\left\|\frac{\partial v_z^\varepsilon}{\partial r}\right\|^2_{L^2(\Omega_\varepsilon)} + \left\|\frac{\partial v_r^\varepsilon}{\partial z}\right\|^2_{L^2(\Omega_\varepsilon)}\right\}d\tau \leq C\left\{\left(\frac{\varepsilon^2}{\mu}\right)^2\|A\|^2_{\mathcal{V}} + \varepsilon^2(\omega^\varepsilon)^2\int_0^t \|\partial_\tau s^\varepsilon(\tau)\|^2_{L^2(0,L)}d\tau\right\}, \tag{3.13}$$

$$\int_0^t \|v_z^\varepsilon\|^2_{L^2(\Omega_\varepsilon)}d\tau \leq C\varepsilon^2\int_0^t \left\{(\omega^\varepsilon)^2\|\partial_\tau s^\varepsilon(\tau)\|^2_{L^2(0,L)} + \left\|\frac{\partial v_z^\varepsilon}{\partial r}\right\|^2_{L^2(\Omega_\varepsilon)}\right\}d\tau, \tag{3.14}$$

$$\frac{1}{\varepsilon^2}\|\eta^\varepsilon(t)\|^2_{L^2(0,L)} + \left\|\frac{\partial s^\varepsilon}{\partial z}(t)\right\|^2_{L^2(0,L)} \leq C\frac{\varepsilon}{h(\varepsilon)E(\varepsilon)}\|A\|^2_{\mathcal{V}}. \tag{3.15}$$

Proof. First notice that (3.12) and (3.15) are obvious consequences of Proposition 3.7. Equation (3.14) is a variant of the Poincaré inequality, and so we need only to prove (3.13).

To prove (3.13) we start from estimate (3.10) for the shear stress term in $D(v^\varepsilon)$. It reads

$$\int_0^t\int_0^L\int_0^{\varepsilon R}\left\{\left(\frac{\partial v_r^\varepsilon}{\partial z}\right)^2 + 2\frac{\partial v_r^\varepsilon}{\partial z}\frac{\partial v_z^\varepsilon}{\partial r} + \left(\frac{\partial v_z^\varepsilon}{\partial r}\right)^2\right\} r dr dz d\tau \leq C\left(\frac{\varepsilon^2}{\mu}\right)^2\|A\|^2_{\mathcal{V}}.$$

The difficulty comes from the term which is the product of two off-diagonal gradient terms $\frac{\partial v_r^\varepsilon}{\partial z}\frac{\partial v_z^\varepsilon}{\partial r}$. Estimate this term by using the boundary behavior of v^ε, $\partial_z v_z^\varepsilon = 0$ at $z = 0, L$, and

the incompressibility condition (2.9) to obtain

$$\begin{aligned}\int_{\Omega_\varepsilon} \frac{\partial v_r^\varepsilon}{\partial z}\frac{\partial v_z^\varepsilon}{\partial r}\, rdrdz &= -\int_{\Omega_\varepsilon} (v_z^\varepsilon - \omega^\varepsilon \partial_t s^\varepsilon)\frac{\partial}{\partial r}\left(r\frac{\partial v_r^\varepsilon}{\partial z}\right) drdz \\ &= \int_{\Omega_\varepsilon} (v_z^\varepsilon - \omega^\varepsilon \partial_t s^\varepsilon)\frac{\partial^2 v_z^\varepsilon}{\partial z^2}\, rdrdz \\ &= -\int_{\Omega_\varepsilon} \frac{\partial}{\partial z}(v_z^\varepsilon - \omega^\varepsilon \partial_t s^\varepsilon)\frac{\partial v_z^\varepsilon}{\partial z}\, rdrdz \\ &= -\int_{\Omega_\varepsilon} \left(\frac{\partial v_z^\varepsilon}{\partial z}\right)^2 rdrdz + \omega^\varepsilon \int_{\Omega_\varepsilon} \frac{\partial}{\partial z}(\partial_t s^\varepsilon)\frac{\partial v_z^\varepsilon}{\partial z}\, rdrdz.\end{aligned}$$

The rest of the proof is now immediate. ■

Corollary 3.9. *The time derivatives $\partial_t(v_r^\varepsilon, v_z^\varepsilon, \eta^\varepsilon, s^\varepsilon)$ satisfy all the above estimates but with A replaced by $\partial_t A$.*

The important estimates are summarized in the following theorem. Here we recall Assumption 1 and use, with a slight abuse of notation, E_0 and G_0 to denote the expressions $E(\varepsilon)h(\varepsilon)/\varepsilon$ and $G(\varepsilon)k(\varepsilon)h(\varepsilon)\varepsilon$, respectively.

Theorem 3.10. *Solution $(v_r^\varepsilon, v_z^\varepsilon, \eta^\varepsilon, s^\varepsilon)$ of problem (3.1)–(3.2) satisfies*

$$\frac{G_0}{\varepsilon^2}\left\|\frac{\partial \eta^\varepsilon}{\partial z}(t)\right\|^2_{L^2(0,L)} + \frac{1}{\varepsilon^2}\|\eta^\varepsilon(t)\|^2_{L^2(0,L)} + \left\|\frac{\partial s^\varepsilon}{\partial z}(t)\right\|^2_{L^2(0,L)} \le C\frac{1}{E_0}\|A\|^2_{\mathcal{V}}, \tag{3.16}$$

$$\left\|\frac{\partial v_z^\varepsilon}{\partial r}\right\|^2_{L^2(\Omega_\varepsilon\times(0,T))} + \left\|\frac{\partial v_r^\varepsilon}{\partial z}\right\|^2_{L^2(\Omega_\varepsilon\times(0,T))} \le C\left(\frac{\varepsilon^2}{\mu}\right)^2\|A\|^2_{\mathcal{V}}, \tag{3.17}$$

$$\|v_r^\varepsilon\|_{L^2(\Omega_\varepsilon\times(0,T))} \le C\frac{\varepsilon^3}{\mu}\|A\|_{\mathcal{V}}, \tag{3.18}$$

$$\|v_z^\varepsilon\|_{L^2(\Omega_\varepsilon\times(0,T))} \le C\frac{\varepsilon^3}{\mu}\|A\|_{\mathcal{V}}. \tag{3.19}$$

The same estimates hold for $\partial_t(v_r^\varepsilon, v_z^\varepsilon, \eta^\varepsilon, s^\varepsilon)$ but in terms of $\partial_t A$.

4. The rescaled problem and asymptotic expansions. In order to study problem P^ε in the limit as $\varepsilon \to 0$, it is convenient to use a rescaling which maps domain Ω_ε to a fixed domain $\Omega = \Omega_1$ corresponding to $\varepsilon = 1$. This, in turn, rescales the variables and their derivatives in the following manner.

Let u^ε be a sequence of axially symmetric functions defined on Ω_ε with values in $\mathbb{R}^3$, $u^\varepsilon = u_r^\varepsilon \vec{e}_r + u_z^\varepsilon \vec{e}_z$. Introduce the following scaling [14]:

$$u(\varepsilon)(r,z) = u^\varepsilon(\varepsilon r, z).$$

The sequence $\{u(\varepsilon)\}$ is now a sequence of functions defined on a fixed domain $\Omega \equiv \Omega_1$. The

rescaled differential operators D_ε and $\operatorname{div}_\varepsilon$ applied on $u(\varepsilon)$ are given by

$$D_\varepsilon(u)_{rr} = \frac{1}{\varepsilon}\frac{\partial u_r}{\partial r}, \quad D_\varepsilon(u)_{\vartheta\vartheta} = \frac{u_r}{\varepsilon r}, \quad D_\varepsilon(u)_{r\vartheta} = 0, \quad D_\varepsilon(u)_{zz} = \frac{\partial u_z}{\partial z},$$
$$D_\varepsilon(u)_{rz} = \frac{1}{2}\left(\frac{\partial u_r}{\partial z} + \frac{1}{\varepsilon}\frac{\partial u_z}{\partial r}\right), \quad D_\varepsilon(u)_{z\vartheta} = 0, \quad \text{and}$$
$$\operatorname{div}_\varepsilon u = \frac{1}{\varepsilon}\frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} + \frac{1}{\varepsilon}\frac{u_r}{r}.$$

The rescaled incompressible Stokes' equations (2.7), (2.8), and (2.9), defined on $\Omega \times \mathbb{R}_+$, read

$$-\frac{\mu}{\varepsilon^2}\left(\frac{\partial^2 v(\varepsilon)_r}{\partial r^2} + \varepsilon^2\frac{\partial^2 v(\varepsilon)_r}{\partial z^2} + \frac{1}{r}\frac{\partial v(\varepsilon)_r}{\partial r} - \frac{v(\varepsilon)_r}{r^2}\right) + \frac{1}{\varepsilon}\frac{\partial p(\varepsilon)}{\partial r} = 0, \tag{4.1}$$

$$-\frac{\mu}{\varepsilon^2}\left(\frac{\partial^2 v(\varepsilon)_z}{\partial r^2} + \varepsilon^2\frac{\partial^2 v(\varepsilon)_z}{\partial z^2} + \frac{1}{r}\frac{\partial v(\varepsilon)_z}{\partial r}\right) + \frac{\partial p(\varepsilon)}{\partial z} = 0, \tag{4.2}$$

$$\varepsilon \operatorname{div}_\varepsilon v(\varepsilon) = \frac{\partial v(\varepsilon)_r}{\partial r} + \varepsilon\frac{\partial v(\varepsilon)_z}{\partial z} + \frac{v(\varepsilon)_r}{r} = 0. \tag{4.3}$$

Since the quantities defined on the lateral boundary are invariant under this scaling, we use the same notation for the wall displacements of the rescaled problem as for the original problem, namely, η^ε and s^ε. The lateral boundary conditions then read

$$\frac{\mu}{\varepsilon^2}v(\varepsilon)_r = \frac{\partial \eta^\varepsilon}{\partial t}, \quad \frac{\mu}{\varepsilon^2}v(\varepsilon)_z = \frac{\partial s^\varepsilon}{\partial t}, \tag{4.4}$$

$$\big(p(\varepsilon)I - 2\mu D_\epsilon(v(\varepsilon))\big)\vec{e}_r \cdot \vec{e}_r = -F_r, \tag{4.5}$$

$$\big(p(\varepsilon)I - 2\mu D_\epsilon(v(\varepsilon))\big)\vec{e}_r \cdot \vec{e}_z = -F_z, \tag{4.6}$$

where F_r and F_z are given by (2.2) and (2.3), respectively. The initial conditions and the inlet and outlet boundary data are (2.14), (2.15)–(2.17) written in terms of the rescaled quantities.

To write this problem in variational form, introduce the space of all test functions $\varphi = \varphi_r\vec{e}_r + \varphi_z\vec{e}_z$ to be the space V given by Definition 1 using $\epsilon = 1$. Furthermore, let $\mathcal{V}$ be the space of rescaled functions on a fixed domain Ω defined by Definition 2 with $\epsilon = 1$, with the incompressibility condition for the velocity replaced by $\operatorname{div}_\varepsilon v(\varepsilon) = 0$ and with $\frac{\partial}{\partial t}$ replaced by $\frac{\mu}{\varepsilon^2}\frac{\partial}{\partial t}$. Let $\psi \in C_0^\infty(0,T)$ be a temporal test function. Then, after rewriting variational equality (3.1) in rescaled variables, multiplying (3.1) by $\psi(t)$, integrating with respect to time, and dividing by ε^2, we obtain the variational formulation of the rescaled problem specified in (4.7) where

$$\begin{aligned}
\mathcal{E}_\mu(v,\varphi,\psi;\varepsilon) \equiv{}& \int_0^T\int_\Omega \frac{\mu}{\varepsilon^2}\frac{\partial v(\varepsilon)_z}{\partial r}\frac{\partial \varphi_z}{\partial r}\psi(t)\, r dr dz dt \\
&+ 2\varepsilon\int_0^T\int_\Omega \frac{\mu}{\varepsilon^2}\left(\frac{1}{2}\frac{\partial v(\varepsilon)_r}{\partial z}\frac{\partial \varphi_z}{\partial r} + \frac{\partial v(\varepsilon)_r}{\partial r}\frac{\partial \varphi_r}{\partial r} + \frac{v(\varepsilon)_r\varphi_r}{r^2}\right)\psi(t)\, r dr dz dt \\
&+ 2\varepsilon^2\int_0^T\int_\Omega \frac{\mu}{\varepsilon^2}\left(\frac{1}{2}\frac{\partial v(\varepsilon)_z}{\partial r}\frac{\partial \varphi_r}{\partial z} + \frac{\partial v(\varepsilon)_z}{\partial z}\frac{\partial \varphi_z}{\partial z}\right)\psi(t)\, r dr dz dt \\
&+ \varepsilon^3\int_0^T\int_\Omega \frac{\mu}{\varepsilon^2}\frac{\partial v(\varepsilon)_r}{\partial z}\frac{\partial \varphi_r}{\partial z}\psi(t)\, r dr dz dt,
\end{aligned}$$

$$\mathcal{E}_{\rm el}(\eta^\varepsilon, s^\varepsilon, \varphi, \psi; \varepsilon) \equiv R\int_0^T \int_0^L \Big\{ h(\varepsilon)G(\varepsilon)k(\varepsilon)\frac{\partial \eta^\varepsilon}{\partial z}\frac{\partial \varphi_r}{\partial z} + \frac{h(\varepsilon)E(\varepsilon)}{1-\sigma^2}\Big(\frac{\sigma}{\varepsilon R}\frac{\partial s^\varepsilon}{\partial z} + \frac{\eta^\varepsilon}{\varepsilon^2 R^2}\Big)\varphi_r + \frac{h(\varepsilon)E(\varepsilon)}{\varepsilon(1-\sigma^2)}\Big(\frac{\partial s^\varepsilon}{\partial z}\frac{\partial \varphi_z}{\partial z} - \frac{\sigma}{\varepsilon R}\frac{\partial \eta^\varepsilon}{\partial z}\varphi_z\Big)\Big\}\psi(t)\, dzdt + \frac{R\rho_w h(\varepsilon)(\omega^\varepsilon)^2}{\varepsilon}\int_0^T \frac{d^2\psi(t)}{dt^2}\int_0^L \big(\varepsilon\eta^\varepsilon\varphi_r + s^\varepsilon\varphi_z\big)\, dzdt,$$

and

$$\mathcal{E}_{\rm src}(A, \varphi, \psi) \equiv \int_0^T \int_0^R A(t)\varphi_z\psi(t)\, rdrdt.$$

Definition 4 (weak formulation of the rescaled problem $\mathrm{P}(\varepsilon)$). *Function* $(v(\varepsilon)_r, v(\varepsilon)_z, \eta^\varepsilon, s^\varepsilon) \in \mathcal{V}$ *is a weak solution of problem* $\mathrm{P}(\varepsilon)$ *if the following variational formulation is satisfied:*

$$\mathcal{E}_\mu(v, \varphi, \psi; \varepsilon) + \mathcal{E}_{\rm el}(\eta^\varepsilon, s^\varepsilon, \varphi, \psi; \varepsilon) = -\mathcal{E}_{\rm src}(A, \varphi, \psi) \quad \forall \psi \in \mathcal{D}(\mathbb{R}_+) \text{ and } \forall \varphi \in V. \tag{4.7}$$

The initial conditions at the lateral boundary are

$$\eta^\varepsilon = s^\varepsilon = \frac{\partial \eta^\varepsilon}{\partial t} = \frac{\partial s^\varepsilon}{\partial t} = 0 \quad on \quad \Sigma \times \{0\}. \tag{4.8}$$

Later in the text we will also need the weak formulation which includes the pressure. For this purpose we consider the test functions φ which are not divergence-free. Namely, denote by

$$\begin{aligned} V_{div\neq 0} = \{\varphi \in H^1(\Omega)^3 | \varphi \text{ is axially symmetric, } \varphi_r|_\Sigma, \varphi_z|_\Sigma \in H^1(0, L), \\ \varphi_z(L, R) = \varphi_r(L, r) = \varphi_r(0, r) = 0\}. \end{aligned} \tag{4.9}$$

Then the weak formulation of the problem, cast in terms of the velocity and pressure, reads as follows.

Definition 5 (weak formulation of $\mathrm{P}(\varepsilon)$ in the pressure-velocity form). *Vector function* $(v(\varepsilon)_r, v(\varepsilon)_z, \eta^\varepsilon, s^\varepsilon) \in \mathcal{V}$ *and* $p(\varepsilon) \in L^2((0, T) \times \Omega))$ *form a weak solution of problem* $\mathrm{P}(\varepsilon)$ *if*

$$\begin{aligned} \mathcal{E}_\mu(v(\varepsilon), \varphi, \psi; \varepsilon) - \int_0^T \int_\Omega p(\varepsilon)\Big(\frac{\partial \varphi_z}{\partial z} + \frac{\partial \varphi_r}{\partial r} + \frac{\varphi_r}{r}\Big)\psi(t)\, rdrdzdt \\ + \mathcal{E}_{\rm el}(\eta^\varepsilon, s^\varepsilon, \varphi, \psi; \varepsilon) = -\mathcal{E}_{\rm src}(A, \varphi, \psi) \quad in\ \mathcal{D}'(\mathbb{R}_+)\ \forall \varphi \in V_{div\neq 0}. \end{aligned} \tag{4.10}$$

Proposition 4.1. *Variational problem* (3.1), (3.2) *is equivalent to the rescaled problem* (4.3), (4.7), *and* (4.8).

The a priori estimates for the rescaled functions follow from Proposition 3.8, Theorem 3.10, and the following result.

Lemma 4.2. *For* $v \in L^2(\Omega_\varepsilon)$ *the rescaled function* $v(\varepsilon)$ *satisfies*

$$\|v(\varepsilon)\|_{L^2(\Omega)} = \varepsilon^{-1}\|v\|_{L^2(\Omega_\varepsilon)}, \tag{4.11}$$

$$\Big\|\frac{\partial v(\varepsilon)}{\partial r}\Big\|_{L^2(\Omega)} = \Big\|\frac{\partial v}{\partial r}\Big\|_{L^2(\Omega_\varepsilon)}, \tag{4.12}$$

$$\Big\|\frac{\partial v(\varepsilon)}{\partial z}\Big\|_{L^2(\Omega)} = \varepsilon^{-1}\Big\|\frac{\partial v}{\partial z}\Big\|_{L^2(\Omega_\varepsilon)}. \tag{4.13}$$

Corollary 4.3. *Solution* $(v(\varepsilon)_r, v(\varepsilon)_z, \eta^\varepsilon, s^\varepsilon)$ *of the rescaled problem satisfies the a priori estimates*

$$\left\|\frac{v(\varepsilon)_r}{r}\right\|_{L^2} + \|v(\varepsilon)_z\|_{L^2} + \left\|\frac{\partial v(\varepsilon)_r}{\partial r}\right\|_{L^2} + \left\|\frac{\partial v(\varepsilon)_z}{\partial r}\right\|_{L^2} \leq C\frac{\varepsilon^2}{\mu}\|A\|_{\mathcal{V}}, \tag{4.14}$$

$$\left\|\frac{\partial v(\varepsilon)_r}{\partial z}\right\|_{L^2} + \left\|\frac{\partial v(\varepsilon)_z}{\partial z}\right\|_{L^2} \leq C\frac{\varepsilon}{\mu}\|A\|_{\mathcal{V}}, \tag{4.15}$$

$$\|s^\varepsilon\|_{H^1} \leq C\|A\|_{\mathcal{V}}, \tag{4.16}$$

where all the norms are defined on $\Omega \times (0,T)$.

From the rescaled momentum equation relating the gradient of the pressure with the viscous effects of the fluid, we get an a priori estimate for the pressure $p(\varepsilon)$ corresponding to $v(\varepsilon)$. More precisely, we have the following proposition.

Proposition 4.4. *The scaled pressure* $p(\varepsilon)$ *satisfies the estimates*

$$\|p(\varepsilon)\|_{L^2(\Omega\times(0,T))} \leq C\|A\|_{\mathcal{V}}, \tag{4.17}$$

$$\int_0^T \left\|\frac{\partial}{\partial z}p(\varepsilon)\right\|^2_{H^{-1}(\Omega)} dt + \frac{1}{\varepsilon^2}\int_0^T \left\|\frac{\partial}{\partial r}p(\varepsilon)\right\|^2_{H^{-1}(\Omega)} dt \leq C\|A\|^2_{\mathcal{V}}. \tag{4.18}$$

Proof. Let φ be an axially symmetric function in $H^1(\Omega)^3$ such that $\varphi = 0$ on $\partial\Omega \setminus \{z = 0\}$. Then, for all such φ the scaled momentum equation gives

$$\begin{aligned}\langle \nabla_\epsilon p(\varepsilon), \varphi\rangle_\Omega &= \left\langle \frac{1}{\varepsilon}\frac{\partial}{\partial r}p(\varepsilon), \varphi_r\right\rangle_\Omega + \left\langle \frac{\partial}{\partial z}p(\varepsilon), \varphi_z\right\rangle_\Omega \\ &= -\mu\int_\Omega D_\varepsilon(v(\varepsilon)) : D_\varepsilon(\varphi).\end{aligned}$$

Now let $g \in L^2(\Omega)$. Due to the surjectivity of the divergence operator between any subspace of $H^1(\Omega)^3$ bigger than $H^1_0(\Omega)^3$ and $L^2(\Omega)$, there exists a $\tilde{\varphi} \in H^1(\Omega)^3$, axially symmetric with $\tilde{\varphi} = 0$ on $\partial\Omega \setminus \{z = 0\}$, such that $\operatorname{div}\tilde{\varphi} = g$. Let $\varphi_r = \varepsilon\tilde{\varphi}_r$ and $\varphi_z = \tilde{\varphi}_z$. Then $g = \operatorname{div}_\varepsilon \varphi$ and $|\int_\Omega p(\varepsilon)g| = |\int_\Omega p(\varepsilon)\operatorname{div}_\varepsilon \phi| = |\langle \nabla_\epsilon p(\varepsilon), \varphi\rangle_\Omega| \leq C\|g\|_{L^2(\Omega)}\ \forall g \in L^2(\Omega)$, implying (4.18). ∎

After obtaining the uniform estimates for $\{v(\varepsilon), p(\varepsilon)\}$, which are valid for their time derivatives as well, we are in a situation where we can define the "correct" asymptotic expansions for $v(\varepsilon)$, $p(\varepsilon)$, η^ε, and s^ε. The usual difficulty with asymptotic expansions is to determine the "optimal" choice of the leading order powers of ε. In general, they follow from the a priori estimates. More precisely, for problem (3.1)–(3.2), results from Corollary 4.3 and Proposition 4.4 imply that the L^2-norms of $\frac{\mu}{\varepsilon^2}v(\varepsilon)$ and $p(\varepsilon)$ are bounded independently of ε and μ. Thus it is natural to use the following asymptotic expansions:

$$v(\varepsilon)(z,r,t) = \frac{\varepsilon^2}{\mu}\sum_{i\geq 0}\varepsilon^i v^i(z,r,t), \tag{4.19}$$

$$p(\varepsilon)(z,r,t) = \sum_{i\geq 0}\varepsilon^i p^i(z,r,t), \tag{4.20}$$

$$\eta^{\varepsilon}(z,t) = \varepsilon \sum_{i\geq 0} \varepsilon^i \eta^i(z,t), \tag{4.21}$$

$$s^{\varepsilon}(z,t) = \sum_{i\geq 0} \varepsilon^i s^i(z,t). \tag{4.22}$$

In the next section we will use these expansions to derive the reduced problem, which, as we shall see, will be second-order accurate in ϵ. In section 6.1 we prove the existence of a unique solution of the limiting problem, as $\epsilon \to 0$. This shows, in particular, that the a priori estimates and the asymptotic expansions given above are optimal.

5. The reduced problem. We derive the effective (reduced) equations, second-order accurate in ε, that hold for small ε. The reduced equations are obtained by inserting the asymptotic expansions above into (4.1), (4.2), and (4.3) and by collecting the powers of ε. The incompressibility condition (4.3) implies

$$\varepsilon^{-1}\frac{\partial}{r\partial r}(rv_r^0) + \frac{\partial v_z^0}{\partial z} + \frac{\partial}{r\partial r}(rv_r^1) + \varepsilon \sum_{i\geq 0} \varepsilon^i \left\{ \frac{\partial v_z^{i+1}}{\partial z} + \frac{\partial}{r\partial r}(rv_r^{i+2}) \right\} = 0. \tag{5.1}$$

Relation (5.1) gives

$$v_r^0 = 0 \quad \text{and} \tag{5.2}$$

$$\frac{\partial (v_z^0 + \varepsilon v_z^1)}{\partial z} + \frac{\partial}{r\partial r}(r(v_r^1 + \varepsilon v_r^2)) = 0 \quad \text{in } \Omega \times (0,T). \tag{5.3}$$

Notice that (5.2) indicates that in this coupled fluid-structure problem for creeping flow, the radial component of the velocity is by one order of magnitude smaller than the axial component.

Insert expansions (4.19)–(4.22) into the momentum equations (4.1) and (4.2), and collect the powers of ε to obtain

$$-\frac{1}{\varepsilon}\frac{\partial (p^0 + \varepsilon p^1)}{\partial r} + \varepsilon \left\{ \frac{\partial}{r\partial r}\left(r\frac{\partial v_r^1}{\partial r} \right) - \frac{v_r^1}{r^2} - \frac{\partial p^2}{\partial r} \right\} = \mathcal{O}(\varepsilon^2), \tag{5.4}$$

$$\frac{\partial}{r\partial r}\left(r\frac{\partial (v_z^0 + \varepsilon v_z^1)}{\partial r} \right) - \frac{\partial (p^0 + \varepsilon p^1)}{\partial z} = \mathcal{O}(\varepsilon^2). \tag{5.5}$$

To simplify notation, define

$$p = p^0 + \varepsilon p^1,\ s = s^0 + \varepsilon s^1,\ \eta = \eta^0 + \varepsilon \eta^1,\ v_r = v_r^1 + \varepsilon v_r^2, \quad \text{and} \quad v_z = v_z^0 + \varepsilon v_z^1.$$

The incompressibility condition (5.3) then reads

$$\frac{\partial}{\partial r}(rv_r) + \frac{\partial}{\partial z}(rv_z) = 0, \tag{5.6}$$

and the coefficient in front of ε^{-1} in (5.4) implies

$$p = p(z,t), \tag{5.7}$$

which says that pressure is constant along each fixed cross-section of a vessel. The second momentum equation (5.5) implies

$$r\frac{\partial p}{\partial z} = \frac{\partial}{\partial r}\left(r\frac{\partial v_z}{\partial r}\right). \tag{5.8}$$

Equations (5.6), (5.7), and (5.8) are the standard asymptotic equations obtained from the flow equations before any boundary conditions are taken into account. These are obtained, for example, in [4, 20], using direct methods, in the case when the acceleration terms are not negligible. In that case, additional acceleration terms appear in (5.8). We mention that, typically, the viscous term on the right-hand side of (5.8) contains a factor with viscosity μ. In our reduction, this term is hidden by its inclusion in the asymptotic expansion of the velocity (4.19). What is typically done next in the reduction to a one-dimensional model is the averaging of the flow equations over a fixed cross-sectional area. This is where boundary conditions and modeling of vessel walls are included. In the case frequently studied in literature, the wall (structure) is modeled using the "independent ring model" [3, 4, 20, 10, 11, 22]. In that case the longitudinal displacement s is assumed to be zero, and the shear term $\partial\eta/\partial z$ is assumed negligible. The no-slip boundary condition is used at the lateral boundary.

In this paper, we do not use the no-slip condition, but we use the continuity of forces and velocity at the lateral boundary, which is more realistic when compliant vessel walls are considered. The structure (wall) is modeled by (2.2) and (2.3), whose asymptotic form is given by

$$p(z,t) = \frac{E_0}{R(1-\sigma^2)}\left(\sigma\frac{\partial s}{\partial z} + \frac{\eta}{R}\right) - G_0\frac{\partial^2\eta}{\partial z^2} + \mathcal{O}(\varepsilon^2), \tag{5.9}$$

$$\frac{\partial v_z}{\partial r}|_{r=R} = \frac{E_0}{1-\sigma^2}\frac{\partial}{\partial z}\left(\frac{\partial s}{\partial z} + \frac{\sigma\eta}{R}\right) + \mathcal{O}(\varepsilon^2). \tag{5.10}$$

Here again, with a slight abuse of notation, we used E_0 and G_0 to denote the expressions $E(\varepsilon)h(\varepsilon)/\varepsilon$ and $G(\varepsilon)k(\varepsilon)h(\varepsilon)\varepsilon$, respectively. It turns out that the reduced equations in this case are much easier to write in terms of the effective pressure, p, rather than in terms of the cross-sectional area and the volumetric flow rate which are typically used when the independent ring model is employed. With this in mind, we proceed as follows. We focus on (5.9) and (5.10) to obtain the PDE for the pressure. We eliminate η, v_z, and $\frac{\partial s}{\partial z}$ from (5.9) and (5.10) by using the incompressibility condition (5.6) and the momentum equations (5.7) and (5.8) in the following way. Integrate the leading order term in the second momentum equation (5.8) twice, from r to R, and use

$$v_z(z,R,t) = \frac{\partial s}{\partial t}(z,t) \tag{5.11}$$

to obtain the axial velocity

$$v_z(z,r,t) = \frac{r^2-R^2}{4}\frac{\partial p}{\partial z}(z,t) + \frac{\partial s}{\partial t}(z,t). \tag{5.12}$$

This is a variant of the Biot law for velocity; see [1, 24]. Combined with the incompressibility condition it provides the following relation used in the proof of the convergence theorem (see section 6):

$$v_r = \frac{R}{r}\frac{\partial \eta}{\partial t} - \frac{(R^2-r^2)^2}{16r}\frac{\partial^2 p}{\partial z^2} + \frac{R^2-r^2}{2r}\frac{\partial^2 s}{\partial z \partial t} = \frac{r}{2}\frac{\partial}{\partial z}\left\{\frac{2R^2-r^2}{8}\frac{\partial p}{\partial z} - \frac{\partial s}{\partial t}\right\}.$$

Next use (5.12) in (5.10) to eliminate v_z by observing that

$$\frac{\partial v_z}{\partial r}|_{r=R} = \frac{R}{2}\frac{\partial p}{\partial z}(z,t). \tag{5.13}$$

To eliminate $\frac{\partial s}{\partial z}$, plug expression (5.12) for v_z into the reduced incompressibility condition (5.6) integrated from $r=0$ to $r=R$,

$$\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial z}\left(\frac{1}{R}\int^R v_z(r,z,t)\ r dr\right) = 0, \tag{5.14}$$

to obtain the reduced, averaged, incompressibility condition

$$\frac{\partial \eta}{\partial t} - \frac{R^3}{16}\frac{\partial^2 p}{\partial z^2} + \frac{R}{2}\frac{\partial^2 s}{\partial z \partial t} = 0. \tag{5.15}$$

Integrate with respect to t, and obtain

$$\frac{\partial s}{\partial z} = \frac{R^2}{8}\frac{\partial^2}{\partial z^2}\int_0^t p - \frac{2}{R}\eta. \tag{5.16}$$

Insert (5.13) into (5.10) to obtain

$$\frac{R}{2}p(z,t) - \frac{E_0}{1-\sigma^2}\left(\frac{\partial s}{\partial z} + \frac{\sigma \eta}{R}\right) = C_0(t).$$

We claim that $C_0(t) = C_0 = 0$. To see this, consider even extensions of the axial velocity v_z^ε and the axial displacement s^ε for negative values of z. Analogously, make odd extensions of the radial velocity v_r^ε, radial displacement η^ε, and the pressure p^ε. It is easy to check that all the equations are valid for $z \in (-L, L)$. Consequently, the partial derivatives with respect to z of all the quantities have the same regularity as the unknown functions. Hence

$$\eta(0,t) = 0 = p(0,t) = \frac{\partial s}{\partial z}(0,t). \tag{5.17}$$

Therefore, $C_0(t) = 0$.

From this calculation we obtain the following relation between η and p:

$$\frac{E_0}{R(1-\sigma^2)}\eta = \frac{R}{2-\sigma}\left(-\frac{1}{2}p + \frac{E_0 R}{8(1-\sigma^2)}\frac{\partial^2}{\partial z^2}\int_0^t p\right). \tag{5.18}$$

We use this to eliminate η from (5.9) and (5.10).

Note that at this point we can write (5.9) and (5.10) entirely in terms of the pressure: (5.18) relates η with p, (5.16) relates $\frac{\partial s}{\partial z}$ with η, and therefore with p, and (5.12) implies that the second equation (5.10) reads

$$\frac{\partial p}{\partial z}(z,t) = \frac{2}{R}\left(\frac{E_0}{1-\sigma^2}\frac{\partial}{\partial z}\left(\frac{\partial s}{\partial z} + \frac{\sigma\eta}{R}\right)\right) + \mathcal{O}(\varepsilon^2). \tag{5.19}$$

By differentiating the pressure equation (5.9) with respect to z and by setting it equal to (5.19), we obtain the reduced equation for the pressure

$$\frac{\partial}{\partial t}\left\{\left(\frac{5}{2} - 2\sigma\right)p - (1-\sigma^2)\frac{G_0R^2}{2E_0}\frac{\partial^2 p}{\partial z^2}\right\} = \frac{\partial^2}{\partial z^2}\left\{\frac{E_0R}{8}p - \frac{G_0R^3}{8}\frac{\partial^2 p}{\partial z^2}\right\}. \tag{5.20}$$

Remark 1. *Note that the "correct" time-scale for the pressure is*

$$t_p = \frac{E_0R}{8}\tilde{t} = \left(\lim_{\varepsilon\to 0}\frac{h(\varepsilon)E(\varepsilon)\varepsilon R}{8\mu}\right)t.$$

It is interesting to notice that for the values given in Table 2.1 this is 0.75 *of the time-scale* $\tilde{t}$ *and* 0.35 *of the original physical time-scale* t*. The physical time-scale* t *and the time-scale* $\tilde{t}$ *are related via* $\tilde{t} = (\varepsilon^2/\mu)t = 0.47t$*.*

Depending on the problem, the coefficients containing shear modulus G_0 may or may not be negligible. In the two subsections that follow, we summarize the initial-boundary-value problems corresponding to the two cases.

5.1. The reduced problem with small shear modulus. In this subsection we study the case when the coefficients containing shear modulus G_0 are zero or negligible. After taking into account the obvious regularity of p^ε with respect to z (but not necessarily the regularity of s^ε), we see that $p = A(t)$ for $z = L$. The reduced initial-boundary-value problem for the effective pressure reads

$$\begin{cases} \left(\frac{5}{2} - 2\sigma\right)\frac{\partial p}{\partial t} = \frac{E_0R}{8}\frac{\partial^2 p}{\partial z^2} & \text{in} \quad (0,L)\times(0,T), \\ p(0,t) = 0, \qquad p(L,t) = A(t) & \text{in} \quad (0,T), \\ p(z,0) = 0 \quad \text{in} \quad (0,L), \end{cases} \tag{5.21}$$

and the relationship between $\frac{\partial s}{\partial z}$ and η becomes

$$\frac{\partial s}{\partial z} = \frac{1-2\sigma}{2-\sigma}\frac{\eta}{R}. \tag{5.22}$$

There is a unique smooth solution of problem (5.21). For incompressible materials, $\sigma = 1/2$ [23], in which case (5.22) implies $s = 0$ and

$$p = \frac{4E_0}{3R^2}\eta = \frac{4E_0}{3R}\left(1 - \sqrt{\frac{\mathcal{A}(0)}{\mathcal{A}}} + \mathcal{O}\left(\frac{\eta}{R}\right)\right), \tag{5.23}$$

where $\mathcal{A} = \mathcal{A}(x,t)$ denotes the cross-sectional area at (t,x), and $\mathcal{A}(0)$ is the unstressed cross-sectional area at x (corresponding to the zero pressure). This is the law of Laplace, or the independent ring model, found in [11, 20] using different approaches. In general, for negligible shear modulus, we find that pressure is directly related to the radial displacement via

$$p = \frac{E_0\eta}{(1-\sigma/2)R^2} = \frac{E_0}{(1-\sigma/2)R}\left(1 - \sqrt{\frac{\mathcal{A}(0)}{\mathcal{A}}} + \mathcal{O}\left(\frac{\eta}{R}\right)\right), \tag{5.24}$$

where (5.23) is the first term in the expansion for the pressure in terms of the radial displacement. We see that, for general σ, the diffusion equation for the effective pressure can be easily written in terms of the radial displacement. The resulting equation is again parabolic, reflecting the fact that acceleration terms in the fluid equation have been ignored. More precisely, if the acceleration terms were present (Navier–Stokes equations), the resulting equation would include the second derivative of η with respect to time and give a hyperbolic problem. Hyperbolic problems are typically obtained when the reduced Navier–Stokes equations are coupled with the independent ring model [12, 20, 4].

Remark 2. *If $A \in C_0^\infty(0,+\infty)$, then $p \in C^\infty([0,L]\times[0,T])$.*

5.2. The reduced problem for nonnegligible shear modulus. In the case when the shear modulus coefficients are not small, we need more boundary conditions for (5.20). Furthermore, for $G_0 > 0$ the boundary conditions for the radial displacement are preserved in the limit. By using (5.18) we get the boundary conditions for $\partial_{zz}p$ at $z = 0, L$. The reduced initial-boundary-value problem reads

$$\begin{cases} \dfrac{\partial}{\partial t}\left\{\left(\dfrac{5}{2} - 2\sigma\right)p - (1-\sigma^2)\dfrac{G_0R^2}{2E_0}\dfrac{\partial^2 p}{\partial z^2}\right\} = \dfrac{\partial^2}{\partial z^2}\left\{\dfrac{E_0R}{8}p - \dfrac{G_0R^3}{8}\dfrac{\partial^2 p}{\partial z^2}\right\}, \\ p(0,t) = 0, \qquad p(L,t) = A(t) \quad \text{in} \quad (0,T), \\ \dfrac{\partial^2 p}{\partial z^2}(0,t) = 0, \quad \dfrac{E_0R}{8(1-\sigma^2)}\dfrac{\partial^2 p}{\partial z^2}(L,t) = \dfrac{1}{2}\dfrac{dA}{dt} \quad \text{in} \quad (0,T), \\ p(z,0) = 0 \quad \text{in} \quad (0,L). \end{cases} \tag{5.25}$$

Remark 3. *If $A \in C_0^\infty(0,+\infty)$, then $p \in C^\infty([0,L]\times[0,T])$.*

5.3. The reduced problem in the pressure-velocity form. It is useful to cast the above reduced problem in terms of the leading order velocity and pressure. In fact, in section 6 we will show that the solution of the original problem converges to the solution of the reduced problem written in terms of (v_z, p, η, s). For this purpose we define the reduced problem for (v_z, p, η, s) by using (5.14) and (5.8) to describe conservation of mass and momentum, (5.11), (5.9), and (5.10) to describe the lateral boundary conditions, and conditions corresponding to (2.15), (2.16), and (2.17) as the inlet and outlet boundary data. Summarized, the problem can be written as follows. Find (v_z, p, η, s) such that the following equations describing conservation of mass and momentum hold:

$$\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial z}\left(\frac{1}{R}\int^R v_z(r,z,t)\, r dr\right) = 0,$$

$$r\frac{\partial p}{\partial z} = \frac{\partial}{\partial r}\left(r\frac{\partial v_z}{\partial r}\right),$$

the lateral boundary conditions are

$$v_z(z,R,t) = \frac{\partial s}{\partial t}(z,t),$$
$$p(z,t) = \frac{E_0}{R(1-\sigma^2)}\left(\sigma\frac{\partial s}{\partial z} + \frac{\eta}{R}\right) - G_0\frac{\partial^2\eta}{\partial z^2},$$
$$\frac{\partial v_z}{\partial r}|_{r=R} = \frac{E_0}{1-\sigma^2}\frac{\partial}{\partial z}\left(\frac{\partial s}{\partial z} + \frac{\sigma\eta}{R}\right),$$

and the inlet and outlet boundary data are given by

$$\eta(0,t) = 0 = p(0,t) = \frac{\partial s}{\partial z}(0,t),$$
$$\eta(L,t) = 0 = s(L,t) = 0, \quad p(L,t) = A(t),$$
$$\eta(z,0) = 0 = s(z,0).$$

In the next section we show that the solution of the original problem converges to the solution of the above reduced problem. Furthermore, in section 7 we show that the error between the solution of this reduced problem and the solution of the full Stokes problem coupled with the Navier equations for the membrane is of order ϵ^2 except at the outlet boundary, where, due to the existence of a boundary layer, the error increases to $\mathcal{O}(\epsilon^{3/2})$.

6. Convergence theorem. In this section we study the rescaled problem P(ε) in the limit as $\varepsilon \to 0$. We show that there exists a subsequence of a sequence of solutions of the rescaled problem which converges weakly and that the weak limit (each cluster point) satisfies the weak form of the reduced problem presented in section 5.3. Due to the uniqueness of the solution of the reduced problem, we conclude that every convergent subsequence of the sequence of rescaled solutions converges weakly to the unique solution of the reduced problem.

We start by proving the following weak compactness result for the sequence of solutions of problem P(ε).

Let

$$W = \left\{\varphi \in L^2(\Omega) \mid \frac{\partial\varphi}{\partial r} \in L^2(\Omega)\right\} \quad \text{and} \tag{6.1}$$

$$W_r = \left\{r^{-1}\varphi \in L^2(\Omega) \mid \frac{\partial\varphi}{\partial r} \in L^2(\Omega)\right\}. \tag{6.2}$$

Theorem 6.1. *There exists a subsequence $(v(\varepsilon), p(\varepsilon), \eta^\varepsilon, s^\varepsilon)$ chosen from a sequence of solutions of the rescaled problem* P(ε) *such that $(\mu/\varepsilon^2\, v(\varepsilon), p(\varepsilon), \eta^\varepsilon, s^\varepsilon)$ converges weakly as $\epsilon \to 0$. Denote its limit (cluster point) by (u,p,η,s), where $u = u_r\vec{e}_r + u_z\vec{e}_z$. Then*

$$u(\varepsilon)_z = \frac{\mu}{\varepsilon^2}v(\varepsilon)_z \to u_z \qquad \text{weakly in} \quad L^2(0,T;W), \tag{6.3}$$

$$u(\varepsilon)_r = \frac{\mu}{\varepsilon^2}v(\varepsilon)_r \to u_r \qquad \text{weakly in} \quad L^2(0,T;W_r), \tag{6.4}$$

$$s^\varepsilon \to s \qquad \text{weakly in} \quad H^1((0,T)\times(0,L)), \tag{6.5}$$

$$\left.\begin{array}{rcl}\sqrt{G_0}\varepsilon^{-1}\partial_z\eta^\varepsilon & \to & \sqrt{G_0}\partial_z\eta,\\ \varepsilon^{-1}\eta^\varepsilon & \to & \eta\end{array}\right\} \quad \text{weak}^* \text{ in} \quad L^\infty(0,T;L^2(0,L)), \tag{6.6}$$

$$p(\varepsilon)\to p \qquad \text{weakly in} \quad L^2((0,T)\times\Omega). \tag{6.7}$$

Moreover, $u_r=0$, $\frac{\partial p}{\partial r}=0$, *and*

$$\frac{\partial\eta}{\partial t}+\frac{\partial}{\partial z}\left(\frac{1}{R}\int_0^R u_z(r,z,t)\ rdr\right)=0 \quad \text{in the sense of distributions.} \tag{6.8}$$

Finally,

$$\varepsilon^{-1}\partial_t\eta^\varepsilon\to\partial_t\eta \quad \text{weakly in} \quad L^2(0,T;H^{-1}(0,L)). \tag{6.9}$$

Proof. We need only to prove $u_r=0$ and $\frac{\partial p}{\partial r}=0$ and the conservation of mass equation (6.8).

To show $u_r=0$, multiply the incompressibility condition (4.3) by a test function $\varphi\in C_0^\infty(\Omega)$ to obtain

$$\int_\Omega \varphi\frac{\partial}{\partial r}\big(ru(\varepsilon)_r\big)\ drdz=-\varepsilon\int_\Omega\frac{\partial\varphi}{\partial z}u(\varepsilon)_z\ rdrdz\to 0 \quad \text{as} \quad \varepsilon\to 0.$$

Therefore, $\frac{\partial}{\partial r}(ru_r)=0$, and, since $u_r\in W_r$, it follows that $u_r=0$.

To show that $\frac{\partial p}{\partial r}=0$, observe that (4.18) implies that for every $\xi\in L^2(0,T;H_0^1(\Omega))$ we have

$$\begin{aligned}\left|\int_0^T\left\langle\frac{\partial}{\partial r}p,\xi\right\rangle_\Omega dt\right| &= \left|-\int_0^T\int_\Omega p\frac{\partial\xi}{\partial r}\ rdrdzdt\right|\\ =\left|-\lim_{\varepsilon\to0}\int_0^T\int_\Omega p(\varepsilon)\frac{\partial\xi}{\partial r}\ rdrdzdt\right| &= \left|\lim_{\varepsilon\to0}\int_0^T\left\langle\frac{\partial}{\partial r}p(\varepsilon),\xi\right\rangle_\Omega dt\right|\\ \le\left\{\lim_{\varepsilon\to0}\left\|\frac{\partial}{\partial r}p(\varepsilon)\right\|_{L^2(0,T;H^{-1}(\Omega))}\right\}&\|\xi\|_{L^2(0,T:H_0^1(\Omega))}=0,\end{aligned}$$

and we conclude that p is independent of r, namely $p=p(z,t)$.

To show that conservation of mass (6.8) holds, start with the incompressibility condition (4.3) and multiply (4.3) by the test functions $\varphi\in C_0^\infty(0,L)$ and $\psi\in C_0^\infty(0,T)$. Integrating over Ω implies

$$-R\int_0^T\int_0^L\frac{\eta^\varepsilon}{\varepsilon}\varphi\frac{d\psi}{dt}\ dzdt-\int_0^T\int_\Omega u(\varepsilon)_z\psi(t)\frac{d\varphi}{dz}\ rdrdzdt=0.$$

After passing to the limit as $\varepsilon\to0$, we obtain (6.8) and (6.9). ∎

Corollary 6.2. *Let* $G_0>0$. *Then for any cluster point* $\{u_z,p,\eta,s\}\in L^2(0,T;W)\times L^2(\Omega\times(0,T))\times\big(L^\infty(0,T;H^1(0,L))^2\cap C([0,T];L^2(0,L))^2\big)$ *we have*

$$u_z(R,z,t)=\frac{\partial s}{\partial t}(z,t) \text{ for } (z,t)\in(0,L)\times(0,T), \tag{6.10}$$

$$\eta(0,t)=0=p(0,t)=\frac{\partial s}{\partial z}(0,t) \text{ for } t\in(0,T), \tag{6.11}$$

$$\eta(L,t)=0=s(L,t)=0,\ p(L,t)=A(t) \text{ for } t\in(0,T), \tag{6.12}$$

$$s(z,0)=\eta(z,0)=0 \text{ for } z\in(0,L). \tag{6.13}$$

For $G_0 = 0$, $\eta \in L^\infty(0,T;L^2(0,L))$, and we drop the boundary condition for η at $z = L$.

Proof. Relation (6.10) is a consequence of the preceding theorem. To show the rest, make an even extension of the axial velocity v_z^ε and the axial displacement s^ε for negative values of z and odd extension of the radial velocity v_r^ε, radial displacement η^ε, and pressure p^ε. It is easy to check that all the equations are valid for $z \in (-L, L)$. Consequently, the partial derivatives with respect to z of all the quantities have the same regularity as the unknown functions, and we conclude that (6.11) holds. The boundary condition for the limit pressure in (6.12) is obtained in the same way. The boundary conditions for the displacements follow from the corresponding convergence results proved in Theorem 6.1. The initial conditions (6.13) for s and η follow from (6.5), (6.6), and (6.9). ∎

Motivated by the weak formulation of the rescaled problem P(ε), assumptions (2.5) and (2.6), and the weak compactness result above, we define the following weak formulation of the limiting problem; call it P($\varepsilon \to 0$). Denote

$$\mathcal{E}_\mu(u,\varphi,\psi;0) = \frac{1}{2}\int_0^T \int_\Omega \frac{\partial u_z}{\partial r}\frac{\partial \varphi_z}{\partial r}\psi(t)\, r dr dz dt,$$

$$\mathcal{E}_{\rm el}(\eta,s,\varphi,\psi;0) = R\int_0^T\int_0^L \Big\{ G_0 \frac{\partial \eta}{\partial z}\frac{\partial \varphi_r}{\partial z} + \frac{E_0}{1-\sigma^2}\Big(\frac{\sigma}{R}\frac{\partial s}{\partial z} + \frac{\eta}{R^2}\Big)\varphi_r + \frac{E_0}{1-\sigma^2}\Big(\frac{\partial s}{\partial z}\frac{\partial \varphi_z}{\partial z} - \frac{\sigma}{R}\frac{\partial \eta}{\partial z}\varphi_z\Big)\Big\}\psi(t)\, dz dt,$$

and

$$\mathcal{E}_{\rm src}(A,\varphi,\psi) = \int_0^T\int_0^R A(t)\varphi_z\psi(t)\, r dr dt.$$

Definition 6 (weak formulation of the limiting problem P($\varepsilon \to 0$)). *Let $G_0 > 0$. Vector function $(u_z,\eta,s) \in L^2(0,T;W)\times L^\infty(0,T;H^1(0,L))^2$ such that $\{\partial_t\eta, \partial_t s\} \in L^2(0,T;H^{-1}(0,L))\times L^2((0,T)\times(0,L))$ is called a weak solution of the limiting problem* P($\varepsilon \to 0$) *if*

$$\mathcal{E}_\mu(u,\varphi,\psi;0) + \mathcal{E}_{\rm el}(\eta,s,\varphi,\psi;0) = -\mathcal{E}_{\rm src}(A,\varphi,\psi), \quad \text{in } \mathcal{D}'(\mathbb{R}_+)\ \forall \varphi \in V \tag{6.14}$$

and conditions (6.10)–(6.13) *are satisfied.*

If $G_0 = 0$, then we need $\eta \in L^\infty(0,T;L^2(0,L))$ and $\partial_t\eta \in L^2(0,T;H^{-1}(0,L))$.

It will be convenient to work with the weak formulation written in terms of the pressure and velocity. For this purpose we define the following.

Definition 7 (weak formulation of P($\varepsilon \to 0$) in the pressure-velocity form). *Let $G_0 > 0$. Vector function $\{u_z,\eta,s\} \in L^2(0,T;W)\times L^\infty(0,T;H^1(0,L))^2$, $\{\partial_t\eta,\partial_t s\} \in L^2(0,T;H^{-1}(0,L))\times L^2((0,T)\times(0,L))$, and $p \in L^2((0,T)\times\Omega)$ form a weak solution of the limiting problem* P($\varepsilon \to 0$) *if*

$$\begin{aligned} &\mathcal{E}_\mu(u,\varphi,\psi;0) - \int_0^T\int_\Omega p(\operatorname{div}\varphi)\,\psi\, r dr dz dt \\ &+ \mathcal{E}_{\rm el}(\eta,s,\varphi,\psi;0) = -\mathcal{E}_{\rm src}(A,\varphi,\psi) \quad \forall \varphi \in \mathcal{D}(\mathbb{R}_+) \text{ and } \forall \varphi \in V_{div\neq 0}, \end{aligned} \tag{6.15}$$

and the conditions (6.10)–(6.13) *are satisfied.* $V_{div\neq 0}$ *is defined by* (4.9).

If $G_0 = 0$, *then we require* $\eta \in L^\infty(0,T;L^2(0,L))$ *and* $\partial_t \eta \in L^2(0,T;H^{-1}(0,L))$.

The following theorem shows that solutions of problems P(ε) converge, as $\varepsilon \to 0$, to the solution of the limiting problem.

Theorem 6.3. *Any cluster point* $\{u_z, p, \eta, s\} \in L^2(0,T;W)\times L^2((0,T)\times\Omega)\times L^\infty(0,T;L^2(0,L))$ $\times\, L^\infty(0,T;H^1(0,L))$ *satisfies the weak formulation of the limiting problem* P($\varepsilon \to 0$).

Proof. Consider the test functions $\varphi \in V_{div\neq 0}$ which are not divergence-free. First observe that since

$$\frac{\mu}{\varepsilon^2}\frac{\partial v(\varepsilon)_z}{\partial r} \rightharpoonup \frac{\partial u_z}{\partial r} \qquad \text{weakly in} \quad L^2((0,T)\times\Omega),$$

we have

$$\lim_{\varepsilon\to 0} \mathcal{E}_\mu(v(\varepsilon),\varphi,\psi;\varepsilon) = \int_0^T \int_\Omega \frac{\partial u_z}{\partial r}\frac{\partial \varphi_z}{\partial r}\psi(t)\, r dr dz dt.$$

Next, weak convergence of $p(\varepsilon)$ gives

$$\lim_{\varepsilon\to 0}\int_0^T\int_\Omega p(\varepsilon)\left(\frac{\partial \varphi_z}{\partial z}+\frac{\partial \varphi_r}{\partial r}+\frac{\varphi_r}{r}\right)\psi(t)\, r dr dz dt = \int_0^T\int_\Omega p\ (\operatorname{div}\varphi)\ \psi\ r dr dz dt.$$

Similarly, using $\rho_w R h(\varepsilon)\varepsilon^3/\mu^2 \le C\varepsilon^{3/2}$, we get

$$\begin{aligned}\lim_{\varepsilon\to 0}\mathcal{E}_{\mathrm{el}}(\eta^\varepsilon, s^\varepsilon, \varphi, \psi;\varepsilon) = R\int_0^T\int_0^L \Bigg\{ & G_0\frac{\partial\eta}{\partial z}\frac{\partial\varphi_r}{\partial z} + \frac{E_0}{1-\sigma^2}\left(\frac{\sigma}{R}\frac{\partial s}{\partial z}+\frac{\eta}{R^2}\right)\varphi_r \\ & + \frac{E_0}{1-\sigma^2}\left(\frac{\partial s}{\partial z}\frac{\partial\varphi_z}{\partial z} - \frac{\sigma}{R}\frac{\partial\eta}{\partial z}\varphi_z\right)\Bigg\}\psi(t)\, dz dt. \end{aligned} \tag{6.16}$$

Therefore, $\{u_z, p, \eta, s\}$ satisfies the weak form (6.15). ∎

The proof of Theorem 6.3 implies that any cluster point $\{u_z, p, \eta, s\}$ satisfies the momentum equation

$$\frac{\partial}{r\partial r}\left(r\frac{\partial u_z}{\partial r}\right) - \frac{\partial p}{\partial z} = 0 \qquad \text{in} \quad \Omega\times(0,T), \tag{6.17}$$

the incompressibility condition (6.8), the initial and boundary conditions (6.10)–(6.13), and the following lateral boundary conditions:

$$p(z,t) = \frac{E_0}{R(1-\sigma^2)}\left(\sigma\frac{\partial s}{\partial z}+\frac{\eta}{R}\right) - G_0\frac{\partial^2\eta}{\partial z^2} \quad \text{on} \quad (0,L)\times(0,T), \tag{6.18}$$

$$\frac{\partial u_z}{\partial r}|_{r=R} = \frac{E_0}{1-\sigma^2}\frac{\partial}{\partial z}\left(\frac{\partial s}{\partial z}+\frac{\sigma\eta}{R}\right) \quad \text{on} \quad (0,L)\times(0,T). \tag{6.19}$$

We note that system (6.8), (6.17), with the lateral boundary conditions (6.10), (6.18), and (6.19) and the inlet and outlet boundary conditions (6.11) and (6.12) has a unique solution. This solution coincides with the solution of the reduced problem obtained using asymptotic

reduction, presented in section 5.3. Its weak form is the problem (6.15). Therefore, we have shown the following.

Corollary 6.4. *Every convergent subsequence of solutions of the rescaled problem* $\mathrm{P}(\varepsilon)$ *converges weakly, as* $\epsilon \to 0$*, to the unique solution of the reduced problem obtained using asymptotic reduction, presented in section* 5.3*. The weak formulation of the reduced problem is given by* (6.10)–(6.13) *and* (6.15)*.*

The methods used in this section are standard and can be found in [5, 7, 14, 15]. We mention that the same convergence techniques can be used in the nonlinear case when monotone operators are considered (see, e.g., [18, 2, 17]). Establishing the precise error estimates is nonstandard, and we present the analysis in the next section.

7. Error estimates. In this section we establish the accuracy of the approximation and the relationship between the function v_r and the radial velocity. The literature on higher order approximations for fluid-structure problems in thin domains is sparse because of the various difficulties associated with the boundary conditions that can be prescribed on a finite domain. The closest result on higher order approximations for the Stokes flow can be found in [19], where Stokes flow through *fixed* domains with small thickness was studied with the prescribed velocity field as the lateral boundary condition. This is different from our approach.

There are various difficulties associated with error estimates for the fluid-structure interaction problem considered in the current paper. The main source of difficulties lies in the choice of the appropriate outlet boundary conditions. The minimum requirement for the outlet boundary conditions is that they should lead to a well-posed reduced problem. We found, for example, that periodic boundary conditions do not lead to a well-posed reduced problem. A further requirement is that the outlet boundary conditions for the reduced problem should be compatible with the fluid-structure interaction in the rest of the domain. Many "standard" boundary conditions typically used in fixed domains (rigid wall tubes) do not seem to satisfy this requirement. For example, zero tangential velocity at the outlet boundary, considered in our paper, although it gives rise to a well-posed initial-boundary-value problem, is incompatible with the flow generated in the rest of the domain. This generates a boundary layer at the outlet boundary and provides technical difficulties in obtaining the error estimates.

To get around this difficulty we construct a boundary layer in a manner similar to the one used in [8]. The construction is based on the following approach. We consider the same differential operator which holds for the original problem, but defined on a *fixed* domain, with the appropriate lateral boundary condition which will serve as a correction for the boundary layer effects. The rest of the boundary data is designed so that the boundary layer solution decreases exponentially away from the outlet boundary and therefore does not contaminate the original solution to the leading order accuracy. Nevertheless, due to the fact that the boundary layer solution ignores the interaction between the wall and the fluid, important shear stress effects will be created at the lateral boundary. This will give rise to a lower accuracy in the error estimate. More precisely, we will get an estimate of order $\mathcal{O}(\varepsilon^{3/2})$. This is by $\mathcal{O}(\sqrt{\varepsilon})$ weaker then the analogous estimates in [19]. Getting a better estimate would, in turn, lead to a complicated modification of the effective equations.

We construct the boundary layer explicitly by considering the following abstract problem

on a semi-infinite rigid-wall cylinder $Z^- = S \times \mathbb{R}_-$, where $S = \{r < R\} \times \{y_3 = 0\}$:

$$-\Delta\beta^{bl} + \nabla\pi^{bl} = 0 \qquad \text{in} \quad Z^-, \tag{7.1}$$

$$\operatorname{div}\beta^{bl} = 0 \qquad \text{in} \quad Z^-, \tag{7.2}$$

$$\beta_r^{bl} = v_r(t, r, L) \qquad \text{on} \quad y_3 = 0, \tag{7.3}$$

$$-2\frac{\partial\beta_z^{bl}}{\partial y_3} + \pi^{bl} = -2\frac{\partial v_z}{\partial z}(t, r, L) \qquad \text{on} \quad y_3 = 0, \tag{7.4}$$

$$\beta^{bl} = 0 \qquad \text{on} \quad \partial S \times \mathbb{R}_-. \tag{7.5}$$

Proposition 7.11 from [8] provides the existence of a unique variational solution $\{\beta^{bl}, \pi^{bl}\} \in C^\infty_{loc}(Z^+ \cup Z^-)^3 \times C^\infty_{loc}(Z^+ \cup Z^-)$ for problem (7.1)–(7.5) such that there exist constants $\gamma_0 > 0$ and C_π satisfying

$$\begin{cases} e^{\gamma_0|y_3|} D^\alpha \beta^{bl} \in L^2(Z^-)^9 & \text{for each multi-index } \alpha \in \mathbb{N}^3 \text{ and } \alpha = 0, \\ e^{\gamma_0|y_3|}(\pi^{bl} - C_\pi) \in L^2(Z^-), & \text{and} \\ e^{\gamma_0|y_3|} D^\alpha \pi^{bl} \in L^2(Z^-)^9 & \text{for each multi-index } \alpha \in \mathbb{N}^3. \end{cases} \tag{7.6}$$

In addition, $\int_S \beta_z^{bl} = 0$ and we have $W^{2,q}$-regularity of β^{bl} and $W^{1,q}$-regularity of π^{bl} at $\partial S \times \{0\}$.

Define the boundary layer velocity and the boundary layer pressure to be

$$\beta^\varepsilon(t, r, z) = \varepsilon\beta^{bl}\left(t, r, \frac{z - L}{\varepsilon}\right), \quad \pi^\varepsilon(t, r, z) = \varepsilon^2\left(\pi^{bl}\left(t, r, \frac{z - L}{\varepsilon}\right) - C_\pi\right). \tag{7.7}$$

The boundary layer velocity and pressure satisfy

$$\operatorname{Div}_\varepsilon\left(\pi^\varepsilon I - 2\varepsilon^2 D_\varepsilon(\beta^\varepsilon)\right) = 0 \qquad \text{in } \Omega \times \mathbb{R}_+, \tag{7.8}$$

$$\operatorname{div}_\varepsilon \beta^\varepsilon = 0 \qquad \text{in } \Omega \times \mathbb{R}_+, \tag{7.9}$$

$$\beta_r^\varepsilon = \varepsilon v_r \text{ and } \pi^\varepsilon - 2\varepsilon^2\frac{\partial\beta_z^\varepsilon}{\partial z} = \varepsilon^2\left(-2\frac{\partial v_z}{\partial z} - C_\pi\right) \quad \text{for } z = L, \tag{7.10}$$

and the following estimates for the behavior at the lateral and at the inlet boundary:

$$\left\|\pi^\varepsilon|_{r=R} - 2\varepsilon\frac{\partial\beta_r^\varepsilon}{\partial r}|_{r=R}\right\|_{L^2(0,L)} \le C\varepsilon^{5/2} \quad \forall t > 0, \tag{7.11}$$

$$\left\|\varepsilon^2\left(\frac{\partial\beta_r^\varepsilon}{\partial z} + \frac{1}{\varepsilon}\frac{\partial\beta_z^\varepsilon}{\partial r}\right)|_{r=R}\right\|_{L^2(0,L)} \le C\varepsilon^{5/2} \quad \forall t > 0, \tag{7.12}$$

$$\beta_r^\varepsilon \text{ and } \pi^\varepsilon - 2\varepsilon\frac{\partial\beta_z^\varepsilon}{\partial z} \text{ are exponentially small for } z = 0. \tag{7.13}$$

Furthermore, the following estimate holds on domain Ω:

$$\varepsilon\|\beta^\varepsilon\|_{L^2(\Omega)^3} + \|\pi^\varepsilon\|_{L^2(\Omega)} + \varepsilon\left\|\frac{\partial\beta^\varepsilon}{\partial r}\right\|_{L^2(\Omega)^3} + \left\|\frac{\partial\pi^\varepsilon}{\partial r}\right\|_{L^2(\Omega)} \le C\varepsilon^{5/2}. \tag{7.14}$$

We are now in a position to estimate the difference between the solution of the rescaled problem and the solution of the limiting problem, modified by the boundary layer. We introduce the following notation:

$$\overline{p}(\varepsilon) = p(\varepsilon) - p + \pi^\varepsilon,\ \overline{v}(\varepsilon) = \frac{\mu}{\varepsilon^2} v(\varepsilon) - v_z \vec{e}_z - \varepsilon v_r \vec{e}_r + \beta^\varepsilon,$$
$$\overline{s}(\varepsilon) = s^\varepsilon - s,\ \overline{\eta}(\varepsilon) = \eta^\varepsilon - \varepsilon\eta.$$

The following theorem provides the error estimates.

Theorem 7.1. *Suppose that $A \in C_0^\infty(0,+\infty)$, and let the parameters satisfy*

$$\left| E_0 - \frac{h(\varepsilon)E(\varepsilon)}{\varepsilon} \right| \le C\varepsilon^{3/2},\ |G_0 - h(\varepsilon)G(\varepsilon)k(\varepsilon)\varepsilon| \le C\varepsilon^{3/2},\ \frac{\rho_w h(\varepsilon)\varepsilon^3}{\mu^2} \le C\varepsilon^{3/2}. \tag{7.15}$$

Then for each fixed $\epsilon > 0$ the following estimates hold for (1) *the error in the velocity and pressure:*

$$\left\| \frac{\overline{v}(\varepsilon)_r}{r} \right\|_{L^2} + \|\overline{v}(\varepsilon)_z\|_{L^2} + \left\| \frac{\partial \overline{v}(\varepsilon)_r}{\partial r} \right\|_{L^2} + \left\| \frac{\partial \overline{v}(\varepsilon)_z}{\partial r} \right\|_{L^2} \le C\varepsilon^{3/2}\|A\|_{H^5(0,T)}, \tag{7.16}$$

$$\left\| \frac{\partial \overline{v}(\varepsilon)_r}{\partial z} \right\|_{L^2} + \left\| \frac{\partial \overline{v}(\varepsilon)_z}{\partial z} \right\|_{L^2} \le C\varepsilon^{1/2}\|A\|_{H^5(0,T)}, \tag{7.17}$$

$$\|\overline{p}(\varepsilon)\|_{L^2} \le C\varepsilon^{3/2}\|A\|_{H^5(0,T)}, \tag{7.18}$$

where the L^2-norms are defined on $\Omega \times (0,T)$, and for (2) *the error in the radial and longitudinal displacement:*

$$\sup_{0\le t\le T} \left\{ \sqrt{G_0} \left\| \frac{\partial \overline{\eta}(\varepsilon)}{\varepsilon \partial z}(t) \right\|_{L^2} + \sqrt{E_0} \left\| \frac{\overline{\eta}(\varepsilon)}{\varepsilon}(t) \right\|_{L^2} + \left\| \frac{\partial \overline{s}(\varepsilon)}{\partial z}(t) \right\|_{L^2} \right\}$$
$$\le C\varepsilon^{3/2}\|A\|_{H^5(0,T)}, \tag{7.19}$$

where the L^2-norms are defined on $(0,L)$.

The proof of this theorem is rather technical, and we present it in the appendix. Here we summarize the main steps.

We first calculate the PDE and the boundary conditions that are satisfied by the error $(\overline{v}, \overline{p}, \overline{\eta}, \overline{s})$. Without loss of generality suppose $G_0 > 0$. Using the equations for the rescaled problem (4.1)–(4.6), the reduced equations obtained via asymptotic reduction (5.3), (5.4)–(5.5), (5.9), and (5.10), and the boundary layer problem (7.8)–(7.14), we obtain the PDEs that hold in $\Omega \times \mathbb{R}_+$:

$$\mathrm{Div}_\varepsilon\big(\overline{p}(\varepsilon)I - 2\varepsilon^2 D_\varepsilon(\overline{v}(\varepsilon))\big) = \varepsilon^2\left(2\frac{\partial^2 v_z}{\partial z^2}\vec{e}_z + \varepsilon\frac{\partial^2 v_r}{\partial z^2}\vec{e}_r\right), \tag{7.20}$$

$$\mathrm{div}_\varepsilon\, \overline{v}(\varepsilon) = 0. \tag{7.21}$$

At the lateral boundary $\{R\} \times (0,L) \times \mathbb{R}_+$ the error satisfies

$$\overline{p}(\varepsilon) - 2\varepsilon\frac{\partial \overline{v}(\varepsilon)_r}{\partial r} = F_r(\varepsilon) + \Phi_r^\varepsilon, \tag{7.22}$$

$$-\varepsilon^2\left(\frac{\partial \overline{v}(\varepsilon)_r}{\partial z} + \frac{1}{\varepsilon}\frac{\partial \overline{v}(\varepsilon)_z}{\partial r}\right) = F_z(\varepsilon) + \Phi_z^\varepsilon, \tag{7.23}$$

where $F_r(\varepsilon)$ and $F_z(\varepsilon)$ are the elastic forces generated by the error on the membrane, given by

$$F_r(\varepsilon) = \frac{h(\varepsilon)E(\varepsilon)}{\varepsilon}\frac{1}{R(1-\sigma^2)}\left(\sigma\frac{\partial\overline{s}(\varepsilon)}{\partial z} + \frac{\overline{\eta}(\varepsilon)}{R\varepsilon}\right) + (\omega^\varepsilon)^2\rho_w h(\varepsilon)\frac{\partial^2\overline{\eta}(\varepsilon)}{\partial t^2}$$
$$- h(\varepsilon)G(\varepsilon)k(\varepsilon)\varepsilon\frac{\partial^2}{\partial z^2}\frac{\overline{\eta}(\varepsilon)}{\varepsilon},$$
$$F_z(\varepsilon) = -h(\varepsilon)E(\varepsilon)\frac{1}{1-\sigma^2}\frac{\partial}{\partial z}\left(\frac{\partial\overline{s}(\varepsilon)}{\partial z} + \frac{\sigma\overline{\eta}(\varepsilon)}{R\varepsilon}\right) + (\omega^\varepsilon)^2\rho_w h(\varepsilon)\frac{\partial^2}{\partial t^2}\overline{s}(\varepsilon),$$

and Φ_r^ε and Φ_z^ε are the contributions of the remaining terms, given by

$$\Phi_r^\varepsilon = -\frac{\partial}{\partial z}\Phi_r^{1,\varepsilon} + \Phi_r^{2,\varepsilon},$$
$$\Phi_z^\varepsilon = -\frac{\partial}{\partial z}\Phi_z^{1,\varepsilon} + \Phi_z^{2,\varepsilon},$$

where

$$\Phi_r^{1,\varepsilon} = \varepsilon^2\left(\frac{R^2}{8}\frac{\partial p}{\partial z} + \frac{\partial s}{\partial t}\right)\left(\frac{\partial v_z}{\partial z} + 2\frac{\partial v_r}{\partial r}\right) - \big(G_0 - h(\varepsilon)G(\varepsilon)k(\varepsilon)\varepsilon\big)\frac{\partial\eta}{\partial z},$$
$$\Phi_r^{2,\varepsilon} = (\omega^\varepsilon)^2\varepsilon\rho_w h(\varepsilon)\frac{\partial^2\eta}{\partial t^2} + \pi^\varepsilon - 2\varepsilon\frac{\partial\beta_r^\varepsilon}{\partial r} - \left(E_0 - \frac{h(\varepsilon)E(\varepsilon)}{\varepsilon}\right)\frac{\sigma\frac{\partial s}{\partial z} + \frac{\eta}{R}}{R(1-\sigma^2)}$$

and

$$\Phi_z^{1,\varepsilon} = -\left(E_0 - \frac{h(\varepsilon)E(\varepsilon)}{\varepsilon}\right)\frac{\varepsilon}{1-\sigma^2}\frac{\partial s}{\partial z} - \varepsilon^3\frac{R}{2}\left(\frac{R^2}{8}\frac{\partial^2 p}{\partial z^2} - \frac{\partial^2 s}{\partial t\partial z}\right),$$
$$\Phi_z^{2,\varepsilon} = (\omega^\varepsilon)^2\rho_w h(\varepsilon)\frac{\partial^2 s}{\partial t^2} + \varepsilon\frac{\partial\beta_z^\varepsilon}{\partial r} + \left(E_0 - \frac{h(\varepsilon)E(\varepsilon)}{\varepsilon}\right)\frac{\varepsilon}{1-\sigma^2}\frac{\partial}{\partial z}\frac{\sigma\eta}{R}.$$

Finally, at the inlet and the outlet boundaries we have

$$\overline{v}(\varepsilon)_r = 0 \text{ and } \overline{p}(\varepsilon) - 2\varepsilon^2\frac{\partial\overline{v}(\varepsilon)_z}{\partial z} = -\varepsilon^2 C_\pi \text{ for } z = L, \tag{7.24}$$

$$\overline{v}(\varepsilon)_r \text{ and } \overline{p}(\varepsilon) - 2\varepsilon^2\frac{\partial\overline{v}(\varepsilon)_z}{\partial z} \text{ are exponentially small for } z = 0. \tag{7.25}$$

Next we write the variational formulation of system (7.20)–(7.25) (see (8.1)) and obtain an energy equality. This is (8.2) in the appendix. From the energy equality, L^2-estimates for (7.20)–(7.25) are calculated. Estimate (7.12) is crucial in obtaining the leading order behavior of the error as $\mathcal{O}(\epsilon^{3/2})$. We recall that estimate (7.12) is a consequence of the geometry of the domain and the choice of the appropriate scaling of the boundary layer. The appropriate scaling of the boundary layer is determined from the underlying operator and the time-scale of the problem. Details of the proof are presented in the appendix.

In Figure 7.1 we show numerical simulations obtained for the pressure drop $A(t)$ prescribed on the right boundary $z = L$, given by $A(t) = 950\sin(2\pi\, t)$Pa. The reference pressure

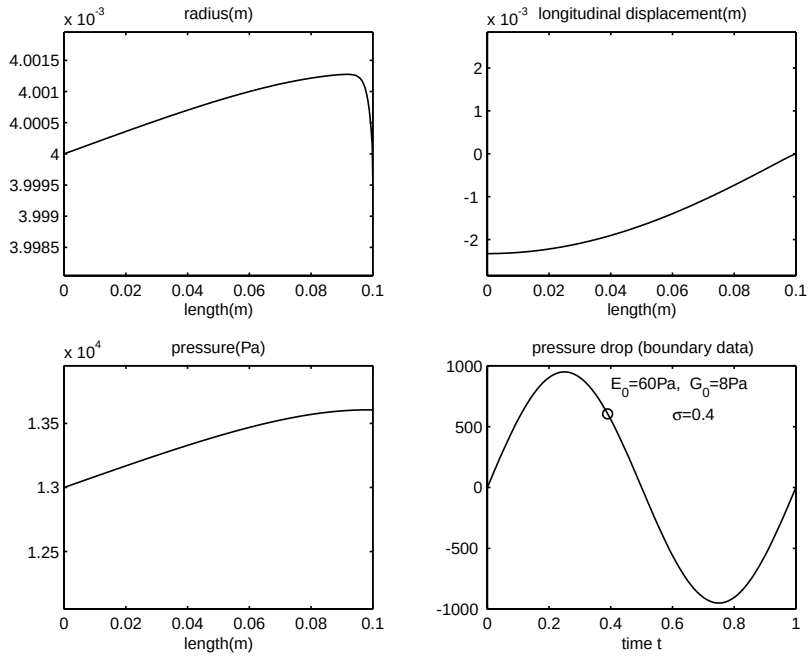


Figure 7.1. *Numerical simulation: Effective displacements with the given pressure drop $A(t)$. The reference pressure is* 1.3×10^4Pa.

$P_0 = 1.3 \times 10^4$Pa. The initial configuration is unperturbed, corresponding to the initial radius $R = 0,0004$m with the initial radial and longitudinal displacement equal to zero. The values of Young's modulus and the shear modulus are $E_0 = 60$Pa, $G_0 = 8$Pa, and the Poisson ratio is $\sigma = 0.4$. Movies showing the dynamics of the vessel wall can be viewed at http://www.math.uh.edu/~canic/smallarteries/menu.html. The figure below is taken at time $t = 0.40$. The dot on the graph of the pressure data $A(t)$ indicates the time in the sinusoidal cycle when the "snap-shot" is taken. We see clearly the formation of the boundary layer in the radial displacement at the boundary $z = L$. The pressure wave coming from the right "pushes" the particles on the membrane to the left, introducing negative longitudinal displacement in the first half of the sinusoidal pulse. This is consistent with the continuity of velocity and forces at the lateral boundary. The radial displacement is positive, and it is decreasing as the points move to the left, away from the right boundary $z = L$. Notice how the radial displacement is by ε order of magnitude smaller than the longitudinal displacement. The movie mentioned above shows how the wall particles move up and to the left in the first half of the sinusoidal cycle and down and to the right in the second half of the pressure cycle.

7.1. The reduced problem and error in physical domain. We conclude this section by writing the reduced equation in the physical (spatial) coordinates, namely, in the domain Ω_ε. Since p is independent of r, the equation for the pressure can be obtained directly from (5.25) by taking into account that the physical inner vessel radius equals εR and that ε is equal to

the inner vessel radius divided by the length of the vessel L. We obtain

$$
\begin{cases}
\dfrac{\partial}{\partial t}\left\{\left(\dfrac{5}{2}-2\sigma\right)p-(1-\sigma^2)\dfrac{G_0L^2}{2E_0}\dfrac{\partial^2 p}{\partial z^2}\right\}=\dfrac{\partial^2}{\partial z^2}\left\{\dfrac{E_0L}{8}p-\dfrac{G_0L^3}{8}\dfrac{\partial^2 p}{\partial z^2}\right\},\\[2ex]
p(0,t)=0, \qquad p(L,t)=A(t) \quad \text{in} \quad (0,T),\\
\dfrac{\partial^2 p}{\partial z^2}(0,t)=0, \quad \dfrac{E_0L}{8(1-\sigma^2)}\dfrac{\partial^2 p}{\partial z^2}(L,t)=\dfrac{1}{2}\dfrac{dA}{dt} \quad \text{in} \quad (0,T),\\
p(z,0)=0 \quad \text{in} \quad (0,L).
\end{cases}
\tag{7.26}
$$

We emphasize that this problem is written in the $\tilde{t}$ time-scale.

The error between the solution of the reduced problem defined on Ω_ε and the solution of the original problem P^ε is obtained as follows. First, recall the notation used in section 5, where the asymptotic equations were written in terms of

$$
p=p^0+\varepsilon p^1,\ s=s^0+\varepsilon s^1,\ \eta=\eta^0+\varepsilon\eta^1,\ v_r=v_r^1+\varepsilon v_r^2, \quad \text{and} \quad v_z=v_z^0+\varepsilon v_z^1.
$$

Then asymptotic expansions of the rescaled quantities (4.19)–(4.22) imply the following asymptotic expansions on Ω_ε:

$$
v^\varepsilon(z,r,t)=\sum_{i\geq 0}\varepsilon^i v^i(z,r,t), \quad p^\varepsilon(z,r,t)=\sum_{i\geq 0}\epsilon^i p^i(z,r,t),
$$
$$
\eta^\varepsilon(z,t)=\varepsilon\sum_{i\geq 0}\epsilon^i\eta^i(z,t), \quad s^\varepsilon(z,t)=\sum_{i\geq 0}\epsilon^i s^i(z,t).
$$

Denote the error in the velocity and pressure, modified by the boundary layer, by

$$
\bar{v}^\varepsilon=v^\varepsilon-v_z(z,r/\varepsilon,t)\vec{e}_z-\varepsilon v_r(z,r/\varepsilon,t)\vec{e}_r+\beta^\varepsilon(z,r/\varepsilon,t),
$$
$$
\bar{p}^\varepsilon=p^\varepsilon-p(z,t)+\pi^{bl}(z,r/\varepsilon,t).
$$

Notice that

$$
v_r^\varepsilon-\varepsilon v_r=\varepsilon^3\sum_{i\geq 3}\varepsilon^{i-3}v^i, \quad v_z^\varepsilon-v_z=\varepsilon^2\sum_{i\geq 2}\varepsilon^{i-2}v^i, \quad p^\varepsilon-p=\varepsilon^2\sum_{i\geq 2}\varepsilon^{i-2}p^i,
$$
$$
\eta^\varepsilon-\varepsilon\eta=\varepsilon^3\sum_{i\geq 3}\varepsilon^{i-3}\eta^i, \quad s^\varepsilon-s=\varepsilon^2\sum_{i\geq 2}\varepsilon^{i-2}s^i.
$$

Theorem 7.1 implies the following error estimates on $\Omega_\varepsilon\times(0,T)$:

$$
\left\|\frac{\bar{v}_r^\varepsilon}{r}\right\|_{L^2(\Omega_\varepsilon\times(0,T))}+\frac{1}{\sqrt{|\Omega_\varepsilon|}}\|\bar{v}_z^\varepsilon\|_{L^2(\Omega_\varepsilon\times(0,T))}
$$
$$
+\left\|\frac{\partial\bar{v}_r^\varepsilon}{\partial r}\right\|_{L^2(\Omega_\varepsilon\times(0,T))}+\left\|\frac{\partial\bar{v}_z^\varepsilon}{\partial r}\right\|_{L^2(\Omega_\varepsilon\times(0,T))}\leq C\varepsilon^{3/2}\|A\|_{H^5(0,T)},
$$
$$
\frac{1}{\sqrt{|\Omega_\varepsilon|}}\left\|\frac{\partial\bar{v}_r^\varepsilon}{\partial z}\right\|_{L^2(\Omega_\varepsilon\times(0,T))}+\frac{1}{\sqrt{|\Omega_\varepsilon|}}\left\|\frac{\partial\bar{v}_z^\varepsilon}{\partial z}\right\|_{L^2(\Omega_\varepsilon\times(0,T))}\leq C\varepsilon^{1/2}\|A\|_{H^5(0,T)},
$$
$$
\frac{1}{\sqrt{|\Omega_\varepsilon|}}\|\bar{p}^\varepsilon\|_{L^2(\Omega_\varepsilon\times(0,T))}\leq C\varepsilon^{3/2}\|A\|_{H^5(0,T)}.
$$

The estimates for the displacement (7.19) remain unchanged.

Finally we mention that if we assume that the approximation of the data in (7.15) is of order ε^2, the resulting error estimates are of order ε^2 in the norms defined on the domain Ω_ε excluding the outlet boundary. Furthermore, we can obtain local estimates of order ε^2 for the derivatives with respect to z of all orders.

8. Appendix. Here we present the proof of Theorem 7.1.

Proof. We start by writing the variational form of system (7.20)–(7.25) satisfied by the error $(\overline{v}, \overline{p}, \overline{\eta}, \overline{s})$:

$$\begin{aligned}
&\mathcal{E}_\mu(\overline{v}, (\varphi_r/\varepsilon, \varphi_z), \psi; \varepsilon) + \mathcal{E}_{\mathrm{el}}(\overline{\eta}, \overline{s}, (\varphi_r/\varepsilon, \varphi_z), \psi; \varepsilon) \\
&\quad = \varepsilon^2 \int_0^t \int_\Omega \left(2\frac{\partial^2 v_z}{\partial z^2}\varphi_z + \varepsilon \frac{\partial^2 v_r}{\partial z^2}\varphi_r \right) \psi(\tau)\, r dr dz d\tau \\
&- \frac{R}{\varepsilon} \int_0^t \int_0^L \left\{ \Phi_r^{2,\varepsilon}\varphi_r(t,R,z) + \Phi_z^{2,\varepsilon}\varphi_z(t,R,z) + \Phi_r^{1,\varepsilon}\frac{\partial \varphi_r}{\partial z} + \Phi_z^{1,\varepsilon}\frac{\partial \varphi_z}{\partial z} \right\} \psi(\tau)\, dz d\tau \psi(\tau)\, r dr d\tau, \\
&-\varepsilon^2 \int_0^t \int_0^R C_\pi \varphi_z(r,L)\psi(\tau)\, r dr d\tau \text{ in } \mathcal{D}'(\mathbb{R}_+)\ \forall \varphi \in V_{div\neq 0},\ \mathrm{div}_\varepsilon\, \varphi = 0.
\end{aligned} \tag{8.1}$$

Consider exponentially small quantities to be equal to zero. Then $\overline{v}(\varepsilon) \in V$ for every t and $\overline{v}(\varepsilon)$ can be used as a test function to obtain the following energy equality:

$$\begin{aligned}
&2\int_0^t \int_\Omega \left(\left(\frac{\partial \overline{v}(\varepsilon)_r}{\partial r} \right)^2 + \left(\frac{\overline{v}(\varepsilon)_r}{\partial r} \right)^2 + \varepsilon^2 \left(\frac{\partial \overline{v}(\varepsilon)_z}{\partial z} \right)^2 \right) + \int_0^t \int_\Omega \left(\varepsilon \frac{\partial \overline{v}(\varepsilon)_r}{\partial z} + \frac{\partial \overline{v}(\varepsilon)_z}{\partial r} \right)^2 \\
&+ \frac{R}{2} \left\{ \frac{(\omega^\varepsilon)^2 \rho_w h(\varepsilon)}{\varepsilon} \int_0^L \left(\left| \frac{\partial \overline{\eta}(\varepsilon)}{\partial t}(t) \right|^2 + \left| \frac{\partial \overline{s}(\varepsilon)}{\partial t}(t) \right|^2 \right) dz + \varepsilon h(\varepsilon) G(\varepsilon) k(\varepsilon) \int_0^L \left| \frac{\partial \overline{\eta}(\varepsilon)}{\varepsilon \partial z}(t) \right|^2 \right. \\
&\left. + \frac{h(\varepsilon) E(\varepsilon)}{\varepsilon(1-\sigma^2)} \left(\sigma \int_0^L \left(\frac{\overline{\eta}(\varepsilon)}{\varepsilon R}(t) - \frac{\partial \overline{s}(\varepsilon)}{\partial z}(t) \right)^2 + (1-\sigma) \int_0^L \left(\left| \frac{\overline{\eta}(\varepsilon)}{\varepsilon R}(t) \right|^2 + \left| \frac{\partial \overline{s}(\varepsilon)}{\partial z}(t) \right|^2 \right) \right) \right\} \\
&= -R \int_0^L \left(\Phi_r^{2,\varepsilon}(t) \frac{\overline{\eta}(\varepsilon)}{\varepsilon}(t) + \Phi_z^{2,\varepsilon}(t) \frac{\overline{s}(\varepsilon)}{\varepsilon}(t) + \Phi_r^{1,\varepsilon}(t) \frac{\partial \overline{\eta}(\varepsilon)}{\varepsilon \partial z}(t) + \Phi_z^{1,\varepsilon}(t) \frac{\partial \overline{s}(\varepsilon)}{\varepsilon \partial z}(t) \right) \\
&+ R \int_0^t \int_0^L \left\{ \frac{\partial}{\partial \tau} \Phi_r^{2,\varepsilon} \frac{\overline{\eta}(\varepsilon)}{\varepsilon} + \frac{\partial}{\varepsilon \partial \tau} \Phi_z^{2,\varepsilon} \overline{s}(\varepsilon) + \frac{\partial}{\partial \tau} \Phi_r^{1,\varepsilon} \frac{\partial \overline{\eta}(\varepsilon)}{\varepsilon \partial z} + \frac{\partial}{\partial \tau} \Phi_z^{1,\varepsilon} \frac{\partial \overline{s}(\varepsilon)}{\varepsilon \partial z} \right\} \\
&- \varepsilon^2 \int_0^t \int_0^R C_\pi \overline{v}(\varepsilon)_z(\tau, r, L)\, r dr d\tau + \varepsilon^2 \int_0^t \int_\Omega \left(2\frac{\partial^2 v_z}{\partial z^2} \overline{v}(\varepsilon)_z + \varepsilon \frac{\partial^2 v_r}{\partial z^2} \overline{v}(\varepsilon)_r \right) r dr dz d\tau.
\end{aligned} \tag{8.2}$$

Now hypothesis (7.15), the regularity of p, and estimates (7.11)–(7.12) allow us to use the Gronwall inequality to estimate all the terms involving $\Phi^{j,\varepsilon}$ by $C\varepsilon^{3/2}$ times the norm of $\overline{v}(\varepsilon)$ at $r = R$. We note that because of (7.12) we cannot get an estimate better than $\mathcal{O}(\varepsilon^{3/2})$.

Next, the outlet term $\varepsilon^2 \int_0^t \int_0^R C_\pi \overline{v}(\varepsilon)_z(\tau, r, L)\, r dr d\tau$ is estimated as in the derivation of

the a priori estimates. Observe that

$$-\varepsilon^2\int_0^t\int_0^R C_\pi\overline{v}(\varepsilon)_z(\tau,r,L)\;rdrd\tau=\varepsilon^2\int_0^t\int_0^R\frac{C_\pi}{L}\overline{v}(\varepsilon)_z(\tau,r,z)\;rdrdzd\tau$$
$$+\frac{R\varepsilon^2C_\pi}{L}\int_0^t\int_0^L z\overline{\eta}(\varepsilon)_z(\tau,z)\;dzd\tau,$$
$$\left|\frac{R\varepsilon^2C_\pi}{L}\int_0^t\int_\Omega\overline{v}(\varepsilon)_z(\tau,r,z)\;rdrdzd\tau-\frac{R^2\varepsilon^2C_\pi}{2L}\int_0^L\overline{s}(\varepsilon)(t)\;dz\right|$$
$$\le\frac{R^2\varepsilon^2|C_\pi|}{2\sqrt{L}}\|\varepsilon D_\varepsilon(\overline{v}(\varepsilon))\|_{L^2((0,t)\times\Omega)}.$$

Therefore, the outlet term is controlled by the sum of two terms—a lateral boundary term, which could be included in $\Phi_r^{2,\varepsilon}$, and a term estimated by $C\varepsilon^2\|\varepsilon D_\varepsilon(\overline{v}(\varepsilon))\|_{L^2((0,t)\times\Omega)}$.

The volume term is more difficult to deal with since we have to use the L^2-norm of the symmetrized gradient. Rewrite the volume term in two parts:

$$I_1=\varepsilon^2\int_0^t\int_\Omega 2\frac{\partial^2v_z}{\partial z^2}(\tau,r,z)\overline{v}(\varepsilon)_z(\tau,r,z)\;rdrdzd\tau$$

and

$$I_2=\varepsilon^3\int_0^t\int_\Omega\frac{\partial^2v_r}{\partial z^2}(\tau,r,z)\overline{v}(\varepsilon)_r(\tau,r,z)\;rdrdzd\tau.$$

Furthermore, note that $\int_0^R 2\xi\frac{\partial^2v_z}{\partial z^2}(\tau,\xi,z)\,d\xi=-2R\frac{\partial^2\eta}{\partial t\partial z}$, and get

$$I_1=-2\varepsilon^2\int_0^t\int_0^L R\frac{\partial^2\eta}{\partial t\partial z}\frac{\partial\overline{s}(\varepsilon)}{\partial\tau}-2\varepsilon^2\int_0^t\int_\Omega\left(\frac{1}{r}\int_0^r\xi\frac{\partial^2v_z}{\partial z^2}\;d\xi\right)\frac{\partial\overline{v}(\varepsilon)_z}{\partial r}\;rdrdzd\tau$$
$$=-2\varepsilon^2\int_0^t\int_0^L R\frac{\partial^2\eta}{\partial t\partial z}\frac{\partial\overline{s}(\varepsilon)}{\partial\tau}-2\varepsilon^2\int_0^t\int_\Omega\left(\frac{1}{r}\int_0^r\xi\frac{\partial^2v_z}{\partial z^2}\;d\xi\right)\left(\frac{\partial\overline{v}(\varepsilon)_z}{\partial r}+\varepsilon\frac{\partial\overline{v}(\varepsilon)_r}{\partial z}\right)$$
$$+2\varepsilon^3\int_0^t\int_\Omega\frac{\partial}{\partial z}\left(\frac{1}{r}\int_0^r\xi\frac{\partial^2v_z}{\partial z^2}\;d\xi\right)\overline{v}(\varepsilon)_r\;rdrdzd\tau. \tag{8.3}$$

Therefore, I_1 is a sum of three terms. The first is controlled in the same way as the terms involving $\Phi^{j,\varepsilon}$, the second is easily estimated using the L^2-norm of the symmetrized gradient, and the third is to be estimated as a part of

$$I_3=I_2+2\varepsilon^3\int_0^t\int_\Omega\frac{\partial}{\partial z}\left(\frac{1}{r}\int_0^r\xi\frac{\partial^2v_z}{\partial z^2}\;d\xi\right)\overline{v}(\varepsilon)_r\;rdrdzd\tau.$$

Since $v_r+\frac{2}{r}\int_0^r\xi\frac{\partial v_z}{\partial z}\,d\xi=-v_r$, we have

$$I_3=-\varepsilon^3\int_0^t\int_\Omega\frac{\partial^2v_r}{\partial z^2}\overline{v}(\varepsilon)_r\;rdrdzd\tau=\varepsilon^3\int_0^t\int_\Omega r\frac{\partial^2v_r}{\partial z^2}\left(\frac{\partial\overline{v}(\varepsilon)_r}{\partial r}+\varepsilon\frac{\partial\overline{v}(\varepsilon)_z}{\partial z}\right)\;rdrdzd\tau$$

and

$$|I_3| \le C\varepsilon^3\left(\left\|\frac{\partial \overline{v}(\varepsilon)_r}{\partial r}\right\|_{L^2((0,t)\times\Omega)} + \varepsilon\left\|\frac{\partial \overline{v}(\varepsilon)_z}{\partial z}\right\|_{L^2((0,t)\times\Omega)}\right). \tag{8.4}$$

Consequently, (7.19) follows and

$$\varepsilon^2\int_0^T \|D_\varepsilon(\overline{v}(\varepsilon))(t)\|^2_{L^2(\Omega)}\,dt \le C\varepsilon^3\|A\|^2.$$

It remains to estimate the term $(D_\varepsilon(\overline{v}(\varepsilon))_{rz}$. Arguing as in the proof of Proposition 3.8, we obtain

$$\left\|\frac{\partial \overline{v}(\varepsilon)_z}{\partial r}\right\|_{L^2((0,T)\times\Omega)} + \varepsilon\left\|\frac{\partial \overline{v}(\varepsilon)_r}{\partial z}\right\|_{L^2((0,T)\times\Omega)} \le C\varepsilon^{3/2}\|A\| + C_1\|\partial_t\overline{s}(\varepsilon)\|_{L^2((0,T)\times(0,L))}.$$

The pressure estimate (7.18) is proved in the same way as in Proposition 4.4. ■

Remark 4. *As in Corollary 3.9, we see that estimates (7.16) and (7.17) are also valid for the time derivative of* $\overline{v}(\varepsilon)$.

Acknowledgments. The authors would like to thank Daniele Lamponi for the discussions about the model and Elise Delavaud and Jérôme Coulon for the numerical simulations shown in Figure 7.1.

REFERENCES

[1] M. A. Biot, *Theory of propagation of elastic waves in a fluid-saturated porous solid.* I. *Lower frequency range* and II. *Higher frequency range*, J. Acoust Soc. Amer., 28 (1956), pp. 168–191.
[2] A. Bourgeat, A. Mikelić, and R. Tapiéro, *Dérivation des équations moyennées décrivant un écoulement non-newtonien dans un domaine de faible épaisseur*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 965–970.
[3] S. Čanić, *Blood flow through compliant vessels after endovascular repair: Wall deformations induced by the discontinuous wall properties*, Comput. Vis. Sci., 4 (2002), pp. 147–155.
[4] S. Čanić and E.-H. Kim, *Mathematical analysis of the quasilinear effects in a hyperbolic model of blood flow through compliant axi-symmetric vessels*, Math. Methods Appl. Sci., 26 (2003), pp. 1–16.
[5] P. G. Ciarlet, *Plates and Junctions in Elastic Multi-structures*, Rech. Math. Appl. 14, Masson, Paris, 1990.
[6] U. Dinar, *Cardiovascular Fluid Dynamics*, CRC Press, Boca Raton, FL, 1981.
[7] H. Dridi, *Comportement asymptotique des équations de Navier-Stokes dans des domaines "applatis,"* Bull. Sci. Math. (2), 106 (1982), pp. 369–385.
[8] W. Jäger and A. Mikelić, *On the effective equations for a viscous incompressible fluid flow through a filter of finite thickness*, Comm. Pure Appl. Math., 51 (1998), pp. 1073–1121.
[9] G. Fichera, *Existence theorems in elasticity*, in Handbook der Physik VIa/2, Springer-Verlag, Berlin, 1972.
[10] L. Formaggia, F. Nobile, and A. Quarteroni, *A one dimensional model for blood flow: Application to vascular prosthesis*, in Mathematical Modeling and Numerical Simulation in Continuum Mechanics, Lect. Notes Comput. Sci. Engrg. 19, I. Babuska, T. Miyoshi, and P. G. Ciarlet, eds., Springer-Verlag, Berlin, 2002, pp. 137–153.
[11] Y. C. Fung, *Biomechanics: Mechanical Properties of Living Tissues*, Springer-Verlag, New York, 1993.
[12] J. Keener and J. Sneyd, *Mathematical Physiology*, Interdiscip. Appl. Math. 8, Springer-Verlag, New York, 1998.

[13] L. D. Landau and E. M. Lifschitz, *Elasticity Theory*, Pergamon Press, Oxford, UK, 1975.
[14] H. Le Dret, *Problemes variationnels dans les multidomaines*, Rech. Math. Appl. 19, Masson, Paris, 1991.
[15] J.-L. Lions, *Perturbations singulières dans les problèmes aux limites et en controle optimal*, Springer-Verlag, Berlin, 1973.
[16] P. Luchini, M. Lupo, and A. Pozzi, *Unsteady Stokes flow in a distensible pipe*, ZAMM Z. Angew. Math. Mech., 71 (1991), pp. 367–378.
[17] A. Mikelić and R. Tapiéro, *Mathematical derivation of the power law describing polymer flow through a thin slab*, M2AN Math. Model. Numer. Anal., 29 (1995), pp. 3–22.
[18] F. Murat and A. Sili, *Problèmes monotones dans des cylindres de faible diamètre*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 567–572.
[19] S. A. Nazarov, *Asymptotic solution of the Navier-Stokes problem on the flow of a thin layer of fluid*, Siberian Math J., 31 (1990), pp. 296–307.
[20] M. S. Olufsen, C. S. Peskin, W. Y. Kim, E. M. Pedersen, A. Nadim, and J. Larsen, *Numerical simulation and experimental validation of blood flow in arteries with structured-tree outflow conditions*, Annals of Biomedical Engineering, 28 (2000), pp. 1281–1299.
[21] M. S. Olufsen, *A structured tree outflow condition for blood flow in the larger systemic arteries*, Amer. J. Physiology, 276 (1999), pp. 257–268.
[22] K. Perktold and G. Rappitsch, *Mathematical modeling of local arterial flow and vessel mechanics*, in Computational Methods for Fluid-Structure Interaction, Pitman Res. Notes Math. Ser. 306, J. Crolet and R. Ohayon, eds., Longman, Harlow, UK, 1994, pp. 230–245.
[23] A. Quarteroni, M. Tuveri, and A. Veneziani, *Computational vascular fluid dynamics: Problems, models and methods. Survey article*, Comput. Vis. Sci., 2 (2000), pp. 163–197.
[24] I. Tolstoy, ed., *Acoustics, Elasticity, and Thermodynamics of Porous Media. Twenty-One Papers by M. A. Biot*, Acoustical Society of America, New York, 1992.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 3, pp. 464–486

Amplitude Equations for Locally Cubic Nonautonomous Nonlinearities*

Dirk Blömker†

Abstract. For systems of partial differential equations (PDEs) with locally cubic nonlinearities, which are perturbed by additive noise, we describe the essential dynamics for small solutions. If the system is near a change of stability, then a natural separation of time-scales occurs and the amplitudes of the dominant modes are given on a long time-scale by a stochastic ordinary differential equation. We consider applications to dynamic pitchfork bifurcation, pattern formation below the threshold of stability, and transient dynamics of stochastic PDEs near this deterministic bifurcations.

Key words. amplitude equation, pattern formation, SPDE, slow modes, separation of time-scales, approximate center-manifold, bifurcation, multiple scale analysis

AMS subject classifications. 60H15, 60H10, 37L55, 37L65

DOI. 10.1137/S1111111103421355

1. Introduction.

Amplitude equations are well known in the physics literature (see, e.g., [H83] or [W97]). They usually describe some order parameter for the system, which evolves on a much slower time-scale. This separation of time-scales occurs, for example, very naturally in a small neighborhood of bifurcations, where a change of stability occurs.

Amplitude equations can be used either for spatially extended systems, where they are stochastic partial differential equations (SPDEs), or for systems on bounded domains, where they are given as stochastic ordinary differential equations (SODEs). This paper will focus on the second case, where an SODE describes the dynamics of the amplitudes of dominant modes evolving on some slow time-scale. On the other hand, all nondominant modes evolve rapidly on a fast time-scale, but they stay much smaller than the dominant ones. The modes in our context are given by the Fourier series expansion with respect to the eigenfunctions of the corresponding linearized operator.

For deterministic systems the theory is rigorously understood even for spatially extended systems (see, e.g., [KSM92, vH91] for the first results). However, there is a lack of results for stochastic systems. The only rigorous example is [BMS01] for a stochastic Swift–Hohenberg equation with periodic boundary conditions on a bounded interval. In this example, a complex-valued SODE was derived describing the amplitude of the dominant mode in a standard complex Fourier series on a very long time-scale.

Our main theorems will extend the results of [BMS01] to a large class of SPDEs and systems of SPDEs. Moreover, our applications will demonstrate the power of this approach, when describing transient dynamics of stochastic equations.

*Received by the editors January 17, 2003; accepted for publication (in revised form) by B. Sandstede June 30, 2003; published electronically September 17, 2003DATEDATE. This work was supported by the DFG Forschungsstipendium BL535-5/1.
http://www.siam.org/journals/siads/2-3/42135.html
†Mathematisches Institut, RWTH Aachen, 52052 Aachen, Germany (bloemker@instmath.rwth-aachen.de). Currently a visitor at Mathematics Research Centre, University of Warwick, Coventry CV4 7AL, UK.

We consider the differential equation

$$\partial_t u(t) = L_\varepsilon u(t) + f(\varepsilon, u(t), t) + \varepsilon^2 \xi(t), \quad u(0) = u_0, \tag{1.1}$$

in a real Banach space X with norm $\|\cdot\|_X$. Equations like this arise, for example, when considering some SPDE (or systems of SPDEs) on bounded domains.

The stochastic perturbation is given by the (generalized) stochastic process ξ, which is the derivative of some cylindrical Q-Wiener process W in X. Therefore, we can treat additive noise that is white in time, but we allow correlations in space.

The main assumptions on the operator L_ε will be that it generates a C_0-semigroup $\{e^{tL_\varepsilon}\}_{t\geq 0}$ on X and that the nullspace $\mathcal{N} := N(L_\varepsilon)$ of L_ε is independent of ε and finite-dimensional. In applications this operator is usually some differential operator equipped with suitable boundary conditions.

The nonlinearity f depends not only on $u(t)$, but it could also involve derivatives of u. The important property is that f contains a small part that is linear in u but no quadratic term in u. For example, $f(\varepsilon, u) = \varepsilon^2 u - u^3$ or $f(\varepsilon, u) = \varepsilon^2 u - u(\partial_x u)^2$.

The fact that the linear part is small will reflect that the unperturbed deterministic system is near a change of stability. Moreover, the reason for neglecting the much more complicated quadratic terms is the following. If we want to separate the dynamics of the dominant modes from the other modes, a cubic nonlinearity helps a lot. In contrast to that, quadratic nonlinearities tend to mix the dynamics of different modes much more strongly. In many examples, quadratic nonlinearities map the dominant modes completely onto nondominant ones, allowing complicated interaction of dominant and nondominant modes. This will be the topic of a forthcoming paper [B03].

One typical example we have in mind is an SPDE such that the unperturbed deterministic PDE exhibits a pitchfork-bifurcation. In a specific example, this was already treated in [BMS01]. To be more precise, f should depend on an additional parameter ν such that in the absence of noise there is a bifurcation at $\nu = 0$ in the deterministic equation. One goal of the presented approach is then to describe the dynamics of the SPDE near this deterministic bifurcation, where the time-scales of the dynamics for stable and unstable modes separate. We sketch briefly some results in section 6.3. We can identify regimes of the bifurcation parameter ν, where the transient behavior of the SPDE is, for instance, almost deterministic or independent of ν.

In contrast to the deterministic setting, the stochastic bifurcation is not that well understood. There are at least two main concepts. The first is a phenomenological bifurcation, where a qualitative change in the unique invariant measure of the corresponding Markov semigroup occurs. The second concept of bifurcation considers changes in the number of invariant measures for the corresponding random dynamical system. Moreover, this second concept is also related to structural changes in random attractors. See, for instance, [Ar98, section 9] for precise definitions and discussions of both concepts. One of the major drawbacks of stochastic bifurcation theory is that, only for one-dimensional SODEs (i.e., state space $\mathbb{R}$), there is a complete picture of possible bifurcations (see [CIS99]).

Note that we call the second concept simply bifurcation instead of dynamical bifurcation as in [Ar98]. The reason is that this terminology also exhibits a completely different meaning (see

section 6.1). We denote by dynamic bifurcation the situation when the bifurcation parameter is slowly moved through a deterministic bifurcation point.

In our example of a deterministic pitchfork-bifurcation perturbed by additive noise, it is, at least for simple examples, known that there is only a phenomenological bifurcation (see, e.g., [CF98]). In the general SPDE case the shape of the of the invariant measure for the Markov semigroup is in general unknown, and the precise structure will be the topic of forthcoming research.

The approach presented does not describe the stochastic bifurcation, as we characterize only typical transient behavior on long time-scales, in contrast to the behavior for time to infinity, which is usually not accessible in real-world experiments. Our results are on one hand related to phenomenological bifurcations, as we could draw conclusions about the structure of invariant measures, but on the other hand the approach reaches far beyond that, as it is capable of describing the evolution in time of typical trajectories on very long transient time-scales.

For multiplicative noise the picture is much more complicated, as in this case bifurcations can occur easily (see, e.g., [CLR01]). We did not try to apply the machinery of amplitude equations to these equations, but it is possible, at least on a nonrigorous level, to reduce the dynamics to SODEs in some of the typical examples.

Our *main results* can be on a formal level described as follows. We make an ansatz of the type

$$u(t) = \varepsilon a(\varepsilon^2 t) \cdot e + \mathcal{O}(\varepsilon^2), \tag{1.2}$$

where $e = (e_1, \dots, e_n)$ is some basis in $\mathcal{N} = N(L_\varepsilon)$. Then in many physical examples a well-known formal calculation shows that plugging ansatz (1.2) into (1.1) yields a system of SODEs for the amplitude a of the dominant modes corresponding to the basis e. This is the *amplitude equation*, and it will be universal in the sense that it is actually independent of ε. The classical example arising in many applications is the so-called Landau equation $a' = \nu a - ca|a|^2 + \dot{\beta}$, where $\dot{\beta}$ is some noise and c, ν are constant coefficient matrices.

The main theorems of this article are first the *attractivity* (cf. Theorem 3.3) justifying the ansatz (1.2) for some initial time, which is not too big. Second, we obtain the *approximation* (cf. Theorem 4.3), showing by rigorous estimates of the error that (1.2) remains true on a very long time-scale of order $\mathcal{O}(\varepsilon^{-2})$, where a is given by the amplitude equation. In the proofs we will follow the strategy to reduce the probability of events giving approximation and attractivity to large deviation estimates for various random fields like the stochastic convolutions and the amplitude a. These are usually easy to derive and treated in the applications.

Another interpretation of the main results is the following. We describe a deterministic approximate center-manifold given by the vector space $\mathcal{N}$. For small solutions of order ε, where ε^2 denotes the noise strength, we show (up to small errors) that $\mathcal{N}$ locally attracts solutions of (1.1) with high probability. Moreover, in an $\mathcal{O}(\varepsilon^2)$-neighborhood of $\mathcal{N}$, most solutions of (1.1) are described (up to small errors) by an SODE on $\mathcal{N}$. This is in contrast to the concept of random invariant manifolds (cf., e.g., [DLS03]). There one can describe the transient dynamics of all solutions as a flow on the manifold, but the manifold is allowed to move in time. Nevertheless, the presented result can be used to estimate probabilities of how the random invariant manifold evolves in time.

Our theory describes solutions with small noise strength, but in contrast to the well-known Freidlin–Wentzell theory (cf. [FW98]), we consider small coefficients in the equation, too. Therefore, we approximate the solutions of (1.1) by the solutions of an SODE, and not by the solutions of the unperturbed PDE. This is motivated by the fact that we want to describe a perturbed deterministic pitchfork-bifurcation when the noise strength is of comparable order to the distance from the bifurcation.

The paper is organized as follows. In section 2 we give the main standing assumptions valid throughout the whole paper and the formulation of the abstract results. Sections 3 and 4 provide the proofs of our main results, first for the local attractivity of $\mathcal{N}$ and then for the approximation of solutions by the amplitude equation.

The last two sections are devoted to applications. Section 5 summarizes simple large deviation results necessary to estimate various probabilities occurring in the application of the main results. Finally, in section 6 we discuss examples. First we consider a dynamic pitchfork-bifurcation, which was discussed for a one-dimensional SODE in [BG02]. Note again that this is not the concept introduced in [Ar98]. In our case dynamic bifurcation means that the bifurcation parameter is time-dependent and is moved slowly through the bifurcation. Problems like this are for SPDEs still the topic of active numerical and experimental research (see [MG99, GM03] and the references therein). In an example we carry over the results of [BG02] to some SPDEs, allowing us to describe the transient behavior of solutions very precisely.

Then we focus on pattern formation below the threshold of instability, where, due to the presence of noise, a pattern appears in an otherwise stable deterministic system. This effect is also well known from experiments (see, e.g., [SR94]) and, for instance, in the context of convection problems still subject to recent experimental investigation (see [SA02]). Nevertheless the problem is not fully understood and there is no rigorous mathematical verification of the pattern being present (see, e.g., [HS92] or [SA02] and the references therein). We present a method giving for an example a first step into that direction.

2. Notation and formulation of the problem.

This section summarizes standing assumptions valid throughout the whole article. For the linear operator L_ε in (1.1), we assume the following.

Assumption 2.1. *For all $\varepsilon > 0$ suppose L_ε is some possibly unbounded linear operator on X. The kernel (or nullspace) of L_ε is denoted by $\mathcal{N} := \{v \in D(L_\varepsilon) : L_\varepsilon v = 0\}$, and it is assumed to be independent of ε. We denote a projection onto $\mathcal{N}$ by P_c and define $n := \dim(\mathcal{N})$.*

Later there will be further restrictions on the choice of the projection P_c. The typical example we have in mind is the spectral projection onto $\mathcal{N}$, and many assumptions would be automatically fulfilled in that case.

One complementary projection to P_c is given by $P_s := I - P_c$. As the dimension of $\mathcal{N}$ is finite, it is well known that both P_c and P_s are bounded linear operators (cf. [W80]).

The second assumption on L_ε and P_c is the following.

Assumption 2.2. *We assume that L_ε from Assumption 2.1 generates a strongly continuous semigroup $\{e^{tL_\varepsilon}\}_{t\geq 0}$ of linear operators on X which is exponentially stable on P_sX. To be more precise, there are constants $\omega > 0$ and $M \geq 1$, independent of ε, such that*

$$\|e^{tL_\varepsilon}P_s x\|_X \leq Me^{-t\omega}\|x\|_X \quad \text{for all } t \geq 0,\ x \in X. \tag{2.1}$$

To deal with the nonlinearity, suppose there is a second Banach space Y, such that X is dense and continuously imbedded into Y. Assume that e^{tL_ε} and P_c can be extended to operators on Y, and for some $\alpha \in [0,1)$ we have

$$\|P_s e^{tL_\varepsilon} y\|_X \le M(1+t^{-\alpha})e^{-t\omega}\|y\|_Y \quad \text{for all } t>0,\ y\in Y. \tag{2.2}$$

Moreover, suppose that P_c, and hence P_s, commutes with e^{tL_ε} on X and Y.

Let us briefly comment on the previous assumption. First, the assumption that P_c commutes with e^{tL_ε} is always true for the spectral projection of L_ε onto $\mathcal{N}$. For self-adjoint operators L_ε in Hilbert spaces, we can, for instance, simply choose the orthogonal projection.

Moreover, under the other assertions of Assumption 2.2 the spectral projection can be extended to a continuous linear operator from Y to $\mathcal{N}$ (i.e., $P_c \in \mathcal{L}(Y,\mathcal{N})$). The main ideas are first to use (2.2) to verify that the residual of L_ε is in $\mathcal{L}(Y,X)$ and second to use the Dunford calculus giving a representation for P_c in terms of the residual (see, e.g., [K95]).

One typical example for Y that we have in mind is an interpolation space between the dual of $D(L)$ and X—for instance, the dual of fractional power spaces in the case when L_ε generates an analytic semigroup.

As $L_\varepsilon \equiv 0$ on $\mathcal{N}$, it is easy to verify that $e^{tL_\varepsilon} = Id$ on $\mathcal{N}$ for all $t \ge 0$. Therefore, we can assume without loss of generality that M is large enough such that

$$\|e^{tL_\varepsilon}x\|_X \le M\|x\|_X \quad \text{for all } t\ge 0,\ x\in X.$$

Moreover, as $\mathcal{N}$ is finite-dimensional, we can also assume that M is sufficiently large such that

$$\|e^{tL_\varepsilon}P_c y\|_X \le M\|P_c y\|_Y \quad \text{for all } t\ge 0,\ y\in Y.$$

For the stochastic perturbation the following assumption is true. For a detailed discussion of Q-Wiener processes and stochastic convolutions, see [dPZ92].

Assumption 2.3. *Suppose that ξ is the generalized derivative of some Q-Wiener process $\{W(t)\}_{t\ge 0}$ on some probability space $(\Omega,\mathcal{F},\mathbb{P})$ such that the stochastic convolution*

$$W_{L_\varepsilon}(t) = \int_0^t e^{(t-\tau)L_\varepsilon} dW(\tau) \tag{2.3}$$

is a well-defined stochastic process with continuous paths in X.

As $P_s e^{tL_\varepsilon} = e^{tL_\varepsilon}P_s$, it is straightforward to verify that

$$P_s[W_{L_\varepsilon}(t)] = \int_0^t e^{(t-\tau)L_\varepsilon} dP_s W(\tau) \quad \text{and} \quad P_c[W_{L_\varepsilon}(t)] = P_c W(t).$$

To give a meaning to (1.1) we will always consider mild solutions.

Assumption 2.4. *We assume that for any (stochastic) initial condition $u_0 \in X$, (1.1) has a mild local solution u. This means we have a stopping time $t^*>0$ and a stochastic process u such that $u:[0,t^*]\to X$ is $\mathbb{P}$-a.s. a solution of*

$$u(t) = e^{tL_\varepsilon}u_0 + \int_0^t e^{(t-\tau)L_\varepsilon} f(\varepsilon,u(\tau),\tau)d\tau + \varepsilon^2 W_{L_\varepsilon}(t) \qquad \text{for } t\le t^*. \tag{2.4}$$

The existence of local solutions is standard if we consider locally Lipschitz-continuous nonlinearities (see, e.g., [dPZ92], and for L^p-theory with application to the Navier–Stokes equation see, e.g., [BP99, BP00]).

We can split the variation of constants formula (2.4) into two parts:

$$P_s u(t) = e^{tL_\varepsilon} P_s u_0 + \int_0^t e^{(t-\tau)L_\varepsilon} P_s f(\varepsilon, u(\tau), \tau) d\tau + \varepsilon^2 \int_0^t e^{(t-\tau)L_\varepsilon} dP_s W(t) \tag{2.5}$$

and

$$P_c u(t) = P_c u_0 + \int_0^t P_c f(\varepsilon, u(\tau), \tau) d\tau + \varepsilon^2 P_c W(t). \tag{2.6}$$

We call $u_s(t) = P_s u(t)$ *fast modes*, as they are subject to an exponential decay on a time-scale of order $\mathcal{O}(1)$. Moreover, $u_c(t) = P_c u(t)$ will be the *slow modes*.

For f we suppose the following.

Assumption 2.5. *There is an $\varepsilon_0 > 0$ such that f defines a family of (nonlinear) operators $f(\varepsilon, \cdot, t) : X \to Y$ for all $\varepsilon \in (0, \varepsilon_0]$ and $t > 0$, where Y was defined in Assumption 2.2.*

Suppose we have the following Taylor expansion for f with respect to u:

$$f(\varepsilon, u, t) = \varepsilon^2 f_1(\varepsilon, t) u + f_3(\varepsilon, t)[u]^3 + g(\varepsilon, u, t), \tag{2.7}$$

where $f_1(\varepsilon, t) \in \mathcal{L}(X, Y)$ is a continuous linear map that leaves $\mathcal{N}$ invariant (i.e., $f_1(\varepsilon, t)\mathcal{N} \subset \mathcal{N}$), and $f_3(\varepsilon, t) \in \mathcal{L}_3(X, Y)$ is a continuous trilinear operator.

Suppose there are constants all denoted by C_f such that

$$\sup_{t \in [0, T_0 \varepsilon^{-2}]} \|g(\varepsilon, u, t)\|_Y \le C_f \|u\|_X^4 \quad \text{for } \|u\|_X \le \delta_0,\ \varepsilon \in (0, \varepsilon_0], \tag{2.8}$$

$$\sup_{t \in [0, T_0 \varepsilon^{-2}]} \|f_1(\varepsilon, t)\|_{\mathcal{L}(X,Y)} \le C_f \quad \text{for } \varepsilon \in (0, \varepsilon_0], \tag{2.9}$$

$$\sup_{t \in [0, T_0 \varepsilon^{-2}]} \|f_3(\varepsilon, t)\|_{\mathcal{L}_3(X,Y)} \le C_f \quad \text{for } \varepsilon \in (0, \varepsilon_0]. \tag{2.10}$$

There are functions $\nu : [0, T_0] \to \mathcal{L}(X, Y)$ and $\mu : [0, T_0] \to \mathcal{L}_3(X, Y)$ and an $\eta > 0$ such that

$$\sup_{t \in [0, T_0 \varepsilon^{-2}]} \|f_1(\varepsilon, t) - \nu(\varepsilon^2 t)\|_{\mathcal{L}(X,Y)} \le C_f \varepsilon^\eta \quad \text{for } \varepsilon \in (0, \varepsilon_0] \tag{2.11}$$

and

$$\sup_{t \in [0, T_0 \varepsilon^{-2}]} \|f_3(\varepsilon, t) - \mu(\varepsilon^2 t)\|_{\mathcal{L}_3(X,Y)} \le C_f \varepsilon^\eta \quad \text{for } \varepsilon \in (0, \varepsilon_0]. \tag{2.12}$$

Example 2.6. *The simplest examples fulfilling Assumption 2.5 are for some given constants $\tilde{\nu} \in \mathbb{R}$ and $\tau_0 > 0$*

$$f(\varepsilon, u, t) = \tilde{\nu} \varepsilon^2 u - u^3 \quad \text{or} \quad f(\varepsilon, u, t) = \tilde{\nu} \varepsilon^4 (t - \tau_0 \varepsilon^{-2}) u - u^3, \tag{2.13}$$

for example, with $X = Y = C^0([a,b])$, which is the space of continuous functions from $[a,b]$ to $\mathbb{R}$. We will use these nonlinearities in our examples in section 6. Note that we can take $\eta > 0$ arbitrarily large and $g \equiv 0$.

Remark 2.7. *We will see later in the proof that the assumption on $f_1(\varepsilon,t)$ to leave $\mathcal{N}$ invariant is important to decouple the dynamics of (2.6) for the slow modes from the dynamics of the fast modes. The assumption is true, for example, if f_1 commutes with L_ε, which in turn is obviously true if $f_1(\varepsilon,t)$ is just multiplication by a scalar.*

The time-dependence of ν and μ (cf. (2.11) and (2.12)) reflects the fact that the slow modes should change on a slow time-scale $T = \varepsilon^2 t$.

2.1. The amplitude equation. The amplitude equation is a (system of) SODE that describes the essential dynamics of mild solutions of (1.1) near 0. Some constants in this equation depend heavily on the choice of some basis e in $\mathcal{N}$. One can try to simplify the structure by changing e.

Consider some basis $e = (e_1, \dots, e_n)$ of $\mathcal{N}$ with $\|e_k\|_X = 1$ for all $k = 1, \dots, n$. For $a \in \mathbb{R}^n$ denote $a \cdot e = \sum_{k=1}^n a_k e_k$. Moreover, define the canonical projection $\Pi : X \to \mathbb{R}^n$ by $\Pi(a \cdot e + z) = a$ for all $a \in \mathbb{R}^n$ and all $z \in \operatorname{kernel}(P_c)$. As the spaces $\mathcal{N}$ and $\mathbb{R}^n$ are finite-dimensional, we easily obtain that Π is continuous; i.e., there is a constant $C_\pi > 0$ such that $|\Pi(x)| \le C_\pi \|x\|$ for all $x \in X$.

We define the cubic nonlinearity $\mu_e(T)[\cdot]^3 : \mathbb{R}^n \to \mathbb{R}^n$ for $T \in [0, T_0]$ by

$$\mu_e(T)[a]^3 = \Pi\{\mu(T)[a \cdot e]^3\} = \sum_{i,j,k=1}^n a_i a_j a_k \Pi\{\mu(T)[e_i, e_j, e_k]\} \tag{2.14}$$

and the linearity $\nu_e(T) : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\nu_e(T) a = \Pi\{\nu(T)(a \cdot e)\} = \sum_{i}^n a_i \Pi\{\nu(T) e_i\}. \tag{2.15}$$

The amplitude equation is now given by

$$a(T) = a_0 + \int_0^T \nu_e(s) a(s) ds + \int_0^T \mu_e(s)[a(s)]^3 ds + \beta(T), \tag{2.16}$$

where $\{\beta(T)\}_{T \ge 0}$ is a Wiener process in $\mathbb{R}^n$ given by $\beta(T) = \varepsilon \Pi(W(\varepsilon^{-2} T))$.

Remark 2.8. *The distribution of β is actually independent of ε due to the scaling properties of a Wiener process. Hence the distribution of solution of (2.16) is independent of ε.*

2.2. Sketch of the results. Our main results are the attractivity (see Theorem 3.3) and the approximation (see Theorem 4.3). In the following we sketch these main results.

For the *attractivity* result assume that the initial condition is of order $\mathcal{O}(\varepsilon)$ and we have large deviation bounds on the stochastic convolution. Then there is a time $t_\varepsilon = \mathcal{O}(\ln(\varepsilon^{-1}))$ such that with high probability for all mild solutions of (1.1) we have

$$u(t_\varepsilon) = \varepsilon a_\varepsilon \cdot e + \varepsilon^2 R_\varepsilon \quad \text{with} \quad |a_\varepsilon|_{\mathbb{R}^n} = \mathcal{O}(1) \text{ and } \|R_\varepsilon\|_X = \mathcal{O}(1).$$

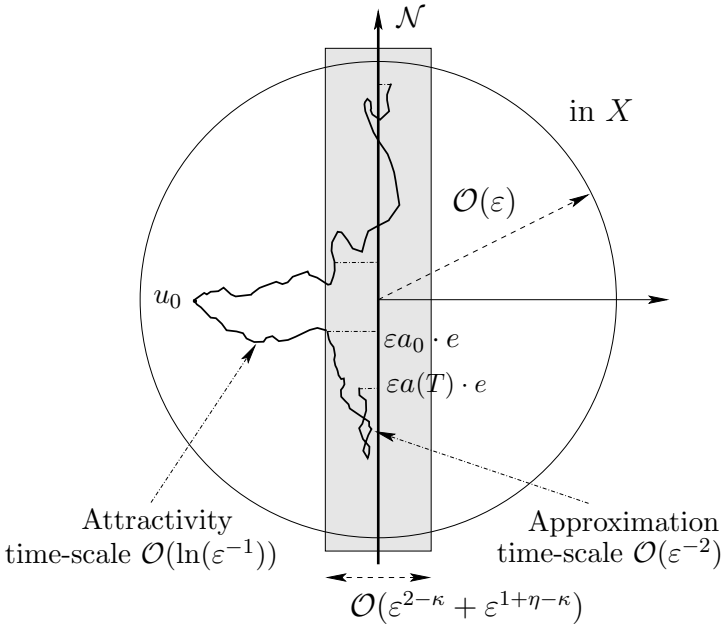


Figure 2.1. *Two typical trajectories of solutions of* (1.1).

For the *approximation* result consider some solution $a(T)$ of the amplitude equation (2.16) and define the approximation $\varepsilon\psi(t) := \varepsilon a(\varepsilon^2 t) \cdot e$. Assume that we have nice large deviation bounds on the stochastic convolution and on the solution $a(T)$, and fix some small $0 < \kappa \ll \eta$. Then we obtain for all mild solutions u of (1.1) that

$$\sup_{t\in[0,T_0\varepsilon^{-2}]} \|u(t) - \varepsilon\psi(t)\|_X < \text{Const} \cdot \left(\varepsilon^{2-\kappa} + \varepsilon^{1+\eta-\kappa}\right) \tag{2.17}$$

with high probability provided $\|u(0) - \varepsilon\psi(0)\|_X = \mathcal{O}(\varepsilon^{2-\kappa} + \varepsilon^{1+\eta-\kappa})$ with high probability. The optimal bound in (2.17) would be $\mathcal{O}(\varepsilon^2)$. Nevertheless, for technical reasons, we are by some κ slightly smaller than that. The constant $\eta > 0$ was defined in (2.11) and (2.12), and we expect $\eta \geq 1$ in many examples.

Combining the attractivity and approximation results, we get a good description of the typical trajectories of (1.1), once we have a good control on various probabilities. A sketch of the typical dynamics is given in Figure 2.1.

Moreover, we can give estimates for the stopping time t^* from Assumption 2.4 like $t^* \geq T_e\varepsilon^{-2}$ with high probability.

The $\mathcal{O}$-notation is used in the following way. A term $G_\varepsilon = \mathcal{O}(g_\varepsilon)$ if and only if there are positive constants ε_0 and C depending only on other constants such that $|G_\varepsilon| \leq C g_\varepsilon$ for all $\varepsilon \in (0, \varepsilon_0]$.

3. The attractivity. First we establish a bound on mild solutions u of (2.4). We show that solutions with initial conditions of order $\mathcal{O}(\varepsilon)$ stay of order $\mathcal{O}(\varepsilon)$ on a large time-scale of order $\mathcal{O}(\varepsilon^{-1})$.

Lemma 3.1. *Suppose all assumptions of section 2 are true. For all times $t_\varepsilon \le \varepsilon^{-1}$ and all constants $0 \le \kappa < 1$, $\delta > 0$, and $D := 2M\delta$, there is an $\varepsilon_0 > 0$ sufficiently small such that for all $\varepsilon \in (0, \varepsilon_0]$ we obtain*

$$\left\{ \sup_{t\in[0,t_\varepsilon]} \|W_{L_\varepsilon}(t)\|_X \le \varepsilon^{-\kappa},\ \|u_0\|_X \le \delta\varepsilon \right\} \Rightarrow \sup_{t\in[0,t_\varepsilon]} \|u(t)\|_X \le D\varepsilon. \tag{3.1}$$

Remark 3.2. *The bound $\varepsilon^{-\kappa}$ on the supremum of the stochastic convolution looks strange at first glance; nevertheless, it is natural. In general we expect (cf. section 5)*

$$\mathbb{P}\left(\sup_{t\in[0,t_\varepsilon]} \|P_s W_{L_\varepsilon}(t)\|_X \le C_w \right) \to 0 \quad \text{for } t_\varepsilon \to \infty.$$

Moreover, for the Brownian motion $P_c W = P_c W_{L_\varepsilon}$

$$\mathbb{P}\left(\sup_{t\in[0,t_\varepsilon]} \|P_c W_{L_\varepsilon}(t)\|_X \le C_w \right) = \mathcal{O}(e^{ct_\varepsilon/C_w^2}) \quad \text{for } t_\varepsilon \text{ and } C_w > 0 \text{ large.}$$

Proof. Using the Assumption 2.5 on f for $\varepsilon D \le \delta_0$, we easily show that there is a constant still denoted by C_f such that for $\|v\|_X \le D\varepsilon$ and $\tau \le t_\varepsilon$

$$\|f(\varepsilon, v, \tau)\|_Y \le C_f(\varepsilon^2\|v\|_X + \|v\|_X^3). \tag{3.2}$$

Define the stopping time $\tau_\varepsilon^* := \inf\{\tau > 0 : \|u(\tau)\|_X > D\varepsilon\}$. Hence, as long as $\tau < \tau_\varepsilon^*$, we obtain

$$\|f(\varepsilon, u(\tau), \tau)\|_Y \le C_f\varepsilon^3(1 + D^2)D. \tag{3.3}$$

Now we derive from (2.4) for $t \le \min\{t_\varepsilon, \tau_\varepsilon^*\}$

$$\begin{aligned}
\|u(t)\|_X &\le M\|u_0\|_X + M\int_0^t (1 + (t-\tau)^{-\alpha})\|f(\varepsilon, u(\tau), \tau)\|_Y d\tau + \varepsilon^2\|W_{L_\varepsilon}(t)\|_X \\
&\le [M\delta + \varepsilon^{1-\kappa}]\varepsilon + MC_f\varepsilon^3(1 + D^2)D\int_0^{t_\varepsilon} (1 + \tau^{-\alpha})d\tau \\
&\le [M\delta + \varepsilon^{1-\kappa}]\varepsilon + MC_f(1 + D^2)D\frac{2-\alpha}{1-\alpha}\cdot\varepsilon^2 \\
&< D\varepsilon
\end{aligned}$$

for ε sufficiently small. This yields immediately $\tau_\varepsilon^* \ge t_\varepsilon$ on the set on interest, which finishes the proof. ∎

Theorem 3.3 (attractivity). *Suppose all assumptions of section 2 are true. Fix the time $t_\varepsilon = \frac{1}{\omega}\ln(\varepsilon^{-2})$ with ω from (2.1) and some $0 \le \kappa < 1$. We can write the mild solution of (1.1) as*

$$u(t_\varepsilon) = \varepsilon a_\varepsilon \cdot e + \varepsilon^2 R_\varepsilon$$

with $a_\varepsilon \in \mathbb{R}^n$ and $R_\varepsilon \in P_s X$ such that for all $\delta > 0$ and $C_w > 0$

$$\left\{ \|u_0\| \le \delta\varepsilon, \quad \sup_{t\in[0,t_\varepsilon]} \|W_{L_\varepsilon}(t)\| \le \varepsilon^{-\kappa}, \quad \|P_s W_{L_\varepsilon}(t_\varepsilon)\| \le C_w \right\}$$
$$\Rightarrow \left\{ |a_\varepsilon|_{\mathbb{R}^n} \le 2C_\pi M\delta, \quad \|R_\varepsilon\| \le 2C_w \right\}$$

for sufficiently small $\varepsilon > 0$.

Proof. Define $\varepsilon a_\varepsilon = \Pi(u(t_\varepsilon))$ and $\varepsilon^2 R_\varepsilon = P_s u(t_\varepsilon)$. By Lemma 3.1 all we need to show is a bound on $P_s u$, as $|\varepsilon a| = |\Pi(u(t_\varepsilon))| \le C_\pi D\varepsilon$ with C_π from subsection 2.1.

Using (2.5) and (2.1), we obtain

$$\|P_s u(t_\varepsilon)\|_X \le Me^{-\omega t_\varepsilon}\|u_0\|_X + \varepsilon^2\|P_s W_{L_\varepsilon}(t_\varepsilon)\|_X + M\int_0^{t_\varepsilon} (1+(t_\varepsilon-\tau)^{-\alpha})e^{-(t_\varepsilon-\tau)\omega}\|f(\varepsilon,u(\tau),\tau)\|_Y d\tau.$$

As $\tau \le \varepsilon^{-1}$ and $\|u(\tau)\| \le D\varepsilon$ by Lemma 3.1, we use (3.3) to finally end up with

$$\|P_s u(t_\varepsilon)\|_X \le M\delta\varepsilon^3 + MC_f\varepsilon^3(D+D^3)\int_0^\infty (1+\tau^{-\alpha})e^{-\tau\omega}d\tau + C_w\varepsilon^2.$$

This implies the result. ∎

4. Approximation. For a solution a of (2.16) we define the approximation $\varepsilon\psi$ depending on a slow time-scale $T = \varepsilon^2 t$ by

$$\varepsilon\psi(t) := \varepsilon a(\varepsilon^2 t)\cdot e.$$

The residual of $\varepsilon\psi$ is given by

$$\mathrm{Res}(\varepsilon\psi(t)) = -\varepsilon\psi(t) + e^{tL_\varepsilon}\varepsilon\psi(0) + \int_0^t e^{(t-\tau)L_\varepsilon} f(\varepsilon,\varepsilon\psi(\tau),\tau)d\tau + \varepsilon^2 W_{L_\varepsilon}(t). \tag{4.1}$$

In order to show that $\varepsilon\psi$ is a good approximation of a solution u of (2.4), we have to control the residual.

Theorem 4.1 (residual). *Suppose all assumptions of section 2 are true. Fix $0 < \kappa < \eta$ and constants $C_a, C_w > 0$. Then there exists a constant $C_{\mathrm{res}} > C_w$ such that for sufficiently small $\varepsilon > 0$ we obtain for all solutions a of (2.16)*

$$\left\{ \sup_{s\in[0,T_0]} |a(s)|_{\mathbb{R}^n} \le C_a\varepsilon^{-\kappa/4}, \quad \sup_{t\in[0,T_0\varepsilon^{-2}]} \|P_s W_{L_\varepsilon}(t)\|_X \le C_w\varepsilon^{-\kappa} \right\}$$
$$\Rightarrow \sup_{t\in[0,T_0\varepsilon^{-2}]} \|\mathrm{Res}(\varepsilon\psi(t))\|_X \le C_{\mathrm{res}}(\varepsilon^{1+\eta-\kappa} + \varepsilon^{2-\kappa}).$$

Remark 4.2. *The results of Theorem 4.1 obviously remain true if we replace $\sup_{s\in[0,T_0]}|a(s)| \le C_a\varepsilon^{-\kappa/4}$ by $\sup_{s\in[0,T_0]}|a(s)| \le C_a$.*

Proof. Let $T = \varepsilon^2 t$ be the slow time. Now (2.7) and $e^{tL_\varepsilon} = Id$ on $\mathcal{N}$ readily imply

$$\text{Res}(\varepsilon\psi(t)) = \varepsilon^2 P_s W_{L_\varepsilon}(t) \tag{4.2}$$

$$- \varepsilon a(T)\cdot e + \varepsilon a(0)\cdot e + \varepsilon^2 P_c W(T\varepsilon^{-2}) \tag{4.3}$$

$$+ \int_0^{T\varepsilon^{-2}} [\varepsilon^2 f_1(\varepsilon,\tau)(\varepsilon a(\varepsilon^2\tau)\cdot e) + P_c f_3(\varepsilon,\tau)[\varepsilon a(\varepsilon^2\tau)\cdot e]^3]d\tau \tag{4.4}$$

$$+ \int_0^t e^{(t-\tau)L_\varepsilon} P_s f_3(\varepsilon,\tau)[\varepsilon a(\varepsilon^2\tau)\cdot e]^3 d\tau \tag{4.5}$$

$$+ \int_0^t e^{(t-\tau)L_\varepsilon} g(\varepsilon,\varepsilon a(\varepsilon^2\tau)\cdot e,\tau)d\tau. \tag{4.6}$$

Now (4.2) is bounded by $C_w\varepsilon^{2-\kappa}$ by assumption, and $C_{\text{res}} > C_w$ is necessary. We choose $C_{\text{res}} > 3C_w$. Using (2.8), we obtain for $t \le T_0\varepsilon^{-2}$

$$\begin{aligned}
\|(4.6)\|_X &\le M\int_0^t (1+(t-\tau)^{-\alpha})\|g(\varepsilon,\varepsilon a(\varepsilon^2\tau)\cdot e,\tau)\|_Y d\tau \\
&\le MC_f\varepsilon^4 \sup_{\tau\in[0,T_0\varepsilon^{-2}]} \|a(\varepsilon^2\tau)\cdot e\|_X^4 \int_0^t (1+(t-\tau)^{-\alpha})d\tau \\
&\le MC_f \sup_{s\in[0,T_0]} |a(s)|^4 n^2 \left(T_0\varepsilon^{-2} + \frac{(T_0\varepsilon^{-2})^{1-\alpha}}{1-\alpha}\right) \\
&\le MC_f C_a^4 n^2 T_0 \left(1 + \frac{T_0^{-\alpha}}{1-\alpha}\right)\cdot \varepsilon^{2-\kappa} < \frac{1}{3}C_{\text{res}}\varepsilon^{2-\kappa},
\end{aligned}$$

which gives us a second condition on C_{res}. Note that for all $a\in\mathbb{R}^n$, $\|a\cdot e\|_X \le \sum_{i=1}^n |a_i| \le n^{1/2}|a|_{\mathbb{R}^n}$ (as $\|e_i\|_X = 1$). Using (2.10) and (2.1), we analogously derive

$$\begin{aligned}
\|(4.5)\|_X &\le MC_f n^{3/2} \sup_{\tau\in[0,t]} |\varepsilon a(\varepsilon^2\tau)|^3 \int_0^t (1+(t-\tau)^{-\alpha})e^{-\tau\omega}d\tau \\
&\le MC_f n^{3/2} C_a^3 \int_0^\infty (1+\tau^{-\alpha})e^{-(t-\tau)\omega}d\tau\cdot\varepsilon^{3-3\kappa/4}.
\end{aligned}$$

This can be also bounded by $\frac{1}{3}C_{\text{res}}\varepsilon^{2-\kappa}$, which implies a third condition on C_{res}.

For (4.4) recall that $\mathcal{N}$ is invariant under $f_1(\varepsilon,\tau)$ by Assumption 2.5. Using the substitution $s = \varepsilon^2\tau$ together with (2.11) and (2.12), we obtain

$$\begin{aligned}
(4.4) &= \int_0^T \left[\varepsilon\Pi\left\{f_1(\varepsilon,\varepsilon^{-2}s)a(s)\cdot e + f_3(\varepsilon,\varepsilon^{-2}s)[a(s)\cdot e]^3\right\}\cdot e\right] ds \\
&= \varepsilon\int_0^T \left[\nu_e(s)(a(s)) + \mu_e(s)[a(s)]^3\right]\cdot e\, ds \\
&\quad + \varepsilon T_0 C_\pi C_f \varepsilon^\eta \cdot \left[n^{1/2}C_a\varepsilon^{-\kappa/4} + n^{3/2}C_a^3\varepsilon^{-3\kappa/4}\right] \\
&= \varepsilon\int_0^T \left[\nu_e(s)(a(s)) + \mu_e(s)[a(s)]^3\right] ds\cdot e + \mathcal{O}(\varepsilon^{1+\eta-\kappa}).
\end{aligned}$$

Now we can use the amplitude equation (2.16) to cancel out the remaining terms in (4.3) and (4.4). This yields a fourth condition on C_{res} if we compute the $\mathcal{O}$-terms explicitly.

We finally derive $\|\text{Res}(\varepsilon\psi(t))\| \le C_{\text{res}}(\varepsilon^{1+\eta-\kappa} + \varepsilon^{2-\kappa})$ for all $t \in [0, T_0\varepsilon^{-2}]$. ■

Theorem 4.3 (approximation). *Suppose all assumptions of section 2 are true. Fix the constants $C_{\text{res}}, T_0, C_a, \delta > 0$, and $\kappa \in (0, \eta)$. Then there is a constant C_{att} such that for sufficiently small $\varepsilon > 0$ we obtain for all solutions u of (2.4) and all solutions a of (2.16)*

$$\left\{ \|u_0 - \varepsilon\psi(0)\|_X \le \delta\varepsilon^2, \quad \sup_{s\in[0,T_0]} |a(s)|_{\mathbb{R}^n} \le C_a n^{-1/2}, \right.$$
$$\left. \sup_{t\in[0,T_0\varepsilon^{-2}]} \|\text{Res}(\varepsilon\psi(t))\|_X \le C_{\text{res}}(\varepsilon^{2-\kappa} + \varepsilon^{1+\eta-\kappa}) \right\}$$
$$\Rightarrow \sup_{t\in[0,T_0\varepsilon^{-2}]} \|u(t) - \varepsilon\psi(t)\|_X \le C_{\text{att}}(\varepsilon^{1+\eta-\kappa} + \varepsilon^{2-\kappa}).$$

Proof. Define $\varepsilon^2 R(t) := u(t) - \varepsilon\psi(t)$. Now (2.4) and (4.1) imply

$$R(t) = e^{tL_\varepsilon}R(0) + \varepsilon^{-2}\text{Res}(\varepsilon\psi(t)) + \varepsilon^{-2}\int_0^t e^{(t-\tau)L_\varepsilon}[f(\varepsilon, u(\tau), \tau) - f(\varepsilon, \varepsilon\psi(\tau), \tau)]d\tau. \tag{4.7}$$

The Taylor expansion of f from (2.7) yields

$$\begin{aligned} &f(\varepsilon, u(\tau), \tau) - f(\varepsilon, \varepsilon\psi(\tau), \tau) \\ &= \varepsilon^4 f_1(\varepsilon, \tau)R(\tau) + f_3(\varepsilon, \tau)[\varepsilon\psi(\tau) + \varepsilon^2 R(\tau)]^3 - f_3(\varepsilon, \tau)[\varepsilon\psi(\tau)]^3 \\ &\quad + g(\varepsilon, +\varepsilon^2 R(\tau), \tau) - g(\varepsilon, \varepsilon\psi(\tau), \tau). \end{aligned}$$

First expand the trilinear form to cancel $[\varepsilon\psi(\tau)]^3$. Then using the bound on $a(s)$, it is easy to derive $\|\psi(\tau)\|_X \le C_a$. Therefore, as long as $\|R(\tau)\|_X \le \varepsilon^{-1}$,

$$\begin{aligned} &\|f(\varepsilon, u(\tau), \tau) - f(\varepsilon, \varepsilon\psi(\tau), \tau)\|_Y \\ &\le \varepsilon^4\|R(\tau)\|_X C_f + \varepsilon^4\|R(\tau)\|_X C_f[3C_a^2 + 3C_a + 1] + C_f\varepsilon^4([C_a + 1]^4 + C_a^4) \\ &\le 3\varepsilon^4\|R(\tau)\|_X C_f[C_a + 1]^2 + 2\varepsilon^4 C_f[C_a + 1]^4. \end{aligned}$$

By assumption $\|R(0)\|_X \le \delta$. In a first step we prove that $\|R(t)\|_X \le \varepsilon^{-1}$ for all $t \in [0, T_0\varepsilon^{-2}]$. Therefore, we further split

$$R = R_c + R_s = P_c R + P_s R.$$

Hence, as long as $\|R(t)\|_X \le \varepsilon^{-1}$ and $t \le T_0\varepsilon^{-2}$, we obtain from (4.7)

$$\begin{aligned} \|R_s(t)\|_X &\le M\|R_s(0)\|_X + C_{\text{res}}(\varepsilon^{-\kappa} + \varepsilon^{-1+\eta-\kappa})\|P_s\|_{\mathcal{L}(X)} \\ &\quad + \varepsilon^2 M\int_0^\infty (1 + \tau^{-\alpha})e^{-\tau\omega}d\tau[3\varepsilon^{-1} + 2]C_f[C_a + 1]^4\|P_s\|_{\mathcal{L}(X)} \\ &\le 2C_{\text{res}}(\varepsilon^{-\kappa} + \varepsilon^{-1+\eta-\kappa})\|P_s\|_{\mathcal{L}(X)} \qquad (4.8) \\ &< \frac{1}{2}\varepsilon^{-1} \qquad \text{for sufficiently small } \varepsilon > 0. \qquad (4.9) \end{aligned}$$

Additionally we obtain

$$\|R_c(t)\|_X \le M\|R_c(0)\|_X + C_{\rm res}(\varepsilon^{-\kappa}+\varepsilon^{-1+\eta-\kappa})\|P_c\|_{\mathcal{L}(X)} + 3C_f[C_a+1]^2\|P_c\|_{\mathcal{L}(X)}\varepsilon^2 M\int_0^t \|R(\tau)\|_X d\tau + \mathcal{O}(1).$$

Now we use $\|R(t)\|_X \le \|R_c(t)\|_X + \mathcal{O}(\varepsilon^{-\kappa}+\varepsilon^{-1+\eta-\kappa})$ by (4.8) and the Gronwall inequality to obtain

$$\|R_c(t)\|_X \le \mathcal{O}(\varepsilon^{-\kappa}+\varepsilon^{-1+\eta-\kappa})\cdot\exp\{3C_f[C_a+1]^2T_0M\|P_c\|_{\mathcal{L}(X)}\} \tag{4.10}$$

$$< \frac{1}{2}\varepsilon^{-1}. \tag{4.11}$$

Hence, for sufficiently small $\varepsilon > 0$ we obtain from (4.9) and (4.11) first that $\|R(t)\|_X < \varepsilon^{-1}$ for all $t \le T_0\varepsilon^{-2}$.

Moreover, (4.8) and (4.10) imply $\sup_{t\in[0,T_0\varepsilon^{-2}]}\|R(t)\|_X = \mathcal{O}(\varepsilon^{-\kappa}+\varepsilon^{-1+\eta-\kappa})$. ■

5. Large deviation results. This section provides large deviation results to control the various probabilities that arise in our application of the abstract result to SPDEs. First we provide estimates for solutions of the amplitude equation (2.16). Then we discuss the stochastic convolution W_L in $C^0([a,b])$, where the operator L is a differential operator.

5.1. Amplitude equation. Consider any solution $a(T)$ of (2.16). Without the cubic nonlinearity or with Lipschitz-continuous nonlinearities there are numerous results, especially for small noise strength (see, e.g., [FW98]). Nevertheless, for our examples we provide an elementary result, which is based only on a priori estimates and large deviation results for Wiener processes. In our cubic case we have to distinguish between the case of stable or unstable cubic nonlinearities.

Theorem 5.1. *Suppose Assumption 2.5 is true, and fix some solution $a(T)$ of (2.16). Then there is a constant $c > 0$ depending only on the covariance matrix of β such that the following are true:*

(I) *The unstable case: For all constants $C_a > 0$ and all $T_1 \in (0, T_0]$ with $T_1 < (2C_f(1+C_a^2))^{-1}$ we obtain*

$$\mathbb{P}\left(\sup_{s\in[0,T_1]}|a(s)| \ge C_a\right) \le \mathbb{P}\Big(|a(0)| \ge C_a/4\Big) + 4ne^{-cC_a^2/T_1n}.$$

(II) *The stable case: Suppose that $\mu_e(T)[b]^3\cdot b \le -C_e|b|^4$ for all $b \in \mathbb{R}^n$. Then there is a constant $C > 0$ depending only on C_f and C_e such that for all constants $T_1 \in (0, T_0]$ and all $C_a > C$ we obtain*

$$\mathbb{P}\left(\sup_{s\in[0,T_1]}|a(s)| \ge C_a\right) \le \mathbb{P}\Big(|a(0)| \ge C_a/2\Big) + 4n\exp\left\{-\frac{c}{T_1n}\sqrt{\frac{C_a^4}{C^4}-1}\right\}.$$

Remark 5.2. *Unfortunately the unstable case has serious drawbacks. High probability is paid by validity of the result only on small time-intervals, although this still gives us for the original equation a time-scale of order $\mathcal{O}(\varepsilon^{-2})$.*

Proof. Using (2.10) and (2.12), we easily obtain that there is a constant also denoted by C_f such that

$$|\mu_e(T)[a]^3| \le C_f|a|^3 \quad \text{for all} \quad a \in \mathbb{R}^n,\ T \in [0, T_0], \tag{5.1}$$

where C_f actually depends only on C_π, n, and the constants in Assumption 2.5. Analogously we obtain

$$|\nu_e(T)[a]| \le C_f|a| \quad \text{for all} \quad a \in \mathbb{R}^n,\ T \in [0, T_0]. \tag{5.2}$$

Using (5.2) and (5.1), we obtain from (2.16)

$$|a(T)| \le |a(0)| + C_f \int_0^T (|a(s)| + |a(s)|^3)ds + |\beta(T)|.$$

Suppose $|a(0)| < C_a/4$ and $\sup_{T\in[0,T_1]} |\beta(T)| < C_a/4$. As long as $|a(T)| < C_a$ we obtain

$$|a(T)| \le C_a/2 + C_f T(C_a + C_a^3) < C_a,$$

as long as $T \le 1/(2C_f(1 + C_a^2))$. Hence

$$\mathbb{P}\left(\sup_{s\in[0,T_1]} |a(s)| < C_a\right) \ge \mathbb{P}\left(|a(0)| < C_a/4, \quad \sup_{s\in[0,T_1]} |\beta(s)| < C_a/4\right).$$

To finish the proof of the unstable case, use, e.g., [DZ98, section 5.2]. As β is a Brownian motion in $\mathbb{R}^d$, we easily obtain the existence of a constant c depending only on the covariance matrix of β such that

$$\mathbb{P}\left(\sup_{s\in[0,T_1]} |\beta(s)| > C_a/4\right) \le 4ne^{-cC_a^2/T_1 n}. \tag{5.3}$$

In the *stable case* define $b = a - \beta$. Hence

$$b \in C^1 \quad \text{with} \quad \partial_T b = \nu_e[b + \beta] + \mu_e[b + \beta]^3.$$

In the following we denote all constants depending only on C_f or C_e simply by C. Using (5.2), the assumption on μ_e, and Young's inequality, we obtain

$$\begin{aligned}\frac{1}{2}\partial_T|b|^2 &\le C_f|b + \beta||b| + \mu_e[b + \beta]^3 \cdot b \\ &\le -\frac{C_e}{2}|b|^4 + C|\beta|^4 + C.\end{aligned}$$

Suppose $|\beta|^4 < R$ and $|a(0)| < \delta$. Then Lemma 5.3 implies

$$|b(T)|^2 < \max\left\{\delta^2,\ C\frac{\sqrt{R^4 + 1}}{C_e^{1/2}}\right\}.$$

Hence

$$|a(T)| < \frac{1}{2}C(R^4+1)^{1/4} + \delta$$

for all $T > 0$.

Define $R^4 = (C_a/C)^4 - 1$ and $\delta = C_a/2$; then for all T_1 and all $C_a > C$

$$\mathbb{P}\left(\sup_{s\in[0,T_1]} |a(s)| \le C_a\right) \ge \mathbb{P}\left(|a(0)| < C_a/2, \quad \sup_{s\in[0,T_2]} |\beta(s)| < \sqrt[4]{(C_a/C)^4 - 1}\right).$$

By (5.3) we easily finish the proof. ∎

Lemma 5.3. *Suppose for some constant $d, c > 0$ we have a real-valued function y such that $y(0) \ge -c/d$ and $y' \le -d^2y^2 + c^2$. Then*

$$y \le \max\{y(0),\ c/d\} \quad \textit{for all} \quad t > 0.$$

Proof. A comparison principle (see, e.g., [Ha80]) yields $y \le x$ with $x(0) = y(0)$ and $x' = -d^2x^2 + c^2$. For the proof we just have to use first that $\pm c/d$ are the only stationary solutions for x and then that x is growing if and only if $|x| \le c/d$. ∎

5.2. Stochastic convolution. There are many general results for exponential tail estimates for stochastic convolutions in Banach or Hilbert spaces. One of the first results is [CM90]; for recent results, see, for instance, [BP00b] and the references therein. A recent new approach relying on Zygmund's interpolation inequality is [SS03].

For our applications we need estimates for sectorial differential operators in the space of continuous functions. We need especially the dependence of the constants on the time-interval $[0, T]$ which is frequently not covered. For simplicity we will basically apply the results of [P92]. This is not optimal, but it is sufficient for our examples.

Assumption 5.4. *Suppose Assumption 2.2 is true, and let L be some nonpositive self-adjoint differential operator of order $2m$ subject to suitable boundary conditions on some sufficiently smooth bounded domain $G \subset \mathbb{R}^d$ (e.g., $L = p(\Delta)$ for some polynomial p of degree m). Suppose that Assumption 2.1 is true with P_c as the L^2-orthogonal projection onto $\mathcal{N} = N(L)$ and that L generates an analytic semigroup $\{e^{tL}\}_{t\ge0}$ on $L^2(G)$ and therefore also on $H = P_sL^2(G)$, where $P_s = I - P_c$ as before.*

Define $E = P_sC^0(G)$, where $C^0(G)$ is the standard space of continuous functions with sup-norm. Then it is easy to verify that the assumptions (E.1) and (E.2) of [P92] are fulfilled (see, e.g., [L94, section 3]). Also, Assumption 2.2 is true with $X = E = Y$. Note that the L^2-orthogonal projection P_c is still the spectral projection for L defined on E.

Suppose W is a Q-Wiener process as in Assumption 2.3. Then we can write $W(t) = Q^{1/2}\tilde{W}(t)$ with $\tilde{W}(t) = \sum_{k=1}^{\infty} \beta_k(t) f_k$, where $\{\beta_k\}_{k\in\mathbb{N}}$ is some family of independently and identically distributed real-valued Brownian motions, and $\{f_k\}_{k\in\mathbb{N}}$ is an orthonormal basis in $L^2(G)$ of eigenfunctions of Q.

For all $\gamma_0 \in (0, 1/2)$ and $p_0 > 1$ define as in [P92]

$$\kappa_T^{p_0} := \int_0^T t^{(\gamma_0-1)p_0} \|e^{tL}\|^{p_0}_{\mathcal{L}(H,E)} dt, \tag{5.4}$$

$$\eta_T := \sup_{t\in[0,T]} \int_0^t \tau^{-2\gamma_0} \mathrm{tr}_{L^2}\left(Q^{1/2} P_s e^{2\tau L} Q^{1/2}\right) d\tau. \tag{5.5}$$

Provided κ_T and η_T are finite, then [P92, Theorem 1.1] implies

$$\mathbb{P}\left(\sup_{t\in[0,T]} \|P_s W_L(t)\|_{C_0} \geq \delta\right) \leq C(T, p_0) \exp\left\{\frac{-\delta^2}{\kappa_T^2 \eta_T}\right\},$$

with $C(T, p_0) = 4T \exp\{(4Tn_0!)^{1/n_0}\}$ and $n_0 = [p_0(2p_0 - 2)^{-1}] + 1$.

For $s > d/2$ we easily check using Sobolev imbedding of E into $H^s(G)$ that Assumption 2.2 is also true with $X = E$, $Y = H$, $\alpha = s/2m$, and some $\omega > 0$ which is some spectral gap to the first nonzero eigenvalue of L. Now for $u \in H$

$$\|e^{tL}u\|_E = \|P_s e^{tL} u\|_{C^0} \leq M(t^{-s/2m} + 1)e^{-\omega t}\|u\|_{L^2}.$$

We easily obtain $\kappa_T < C_\kappa$ for all $T > 0$ with some constant $C_\kappa > 0$ (depending only on s, m, p_0, γ_0, and ω) if and only if $(1/p_0 - 1 + \gamma_0)2m > s$. If we choose p_0 near 1 and γ_0 near $\frac{1}{2}$, then we will always find such an $s > \frac{d}{2}$ provided $2m > d$. Moreover, if $\gamma_0 = \frac{1}{4}$, we will always find s provided $m > d$.

If $\mathrm{tr}(Q) < \infty$, then $\mathrm{tr}(Q^{1/2} P_s e^{2tL} Q^{1/2}) \leq \mathrm{tr}(Q) M e^{-2\omega\tau}$. This implies the existence of some constant C_η depending only on α_0 and ω such that $\eta_T < C_\eta M \mathrm{tr}(Q)$.

For $Q = I$ we obtain

$$\eta_T = \sup_{t\in[0,T]} \int_0^t \tau^{-2\gamma_0} \sum_{\lambda_k \neq 0} e^{2\tau\lambda_k} d\tau \leq \sum_{\lambda_k \neq 0} \frac{1}{|\lambda_k|^{1-2\gamma_0}} \int_0^\infty s^{-2\gamma_0} e^{-2s} ds. \tag{5.6}$$

As λ_k is proportional to $-k^{2m/d}$ (cf., e.g., [EE87]), we obtain that (5.6) is finite if and only if $2m(1 - 2\gamma_0) > d$, which in turn is true for $\gamma_0 = \frac{1}{4}$ and $m > d$.

Applying the results of [P92], we immediately prove the following.

Theorem 5.5. *Let Assumption 5.4 be true, and let W be a Q-Wiener process as in Assumption 2.3. Fix some arbitrary small $\zeta > 0$, and suppose either $\mathrm{tr}(Q) < \infty$ and $2m > d$ or $Q = I$ and $m > d$.*

Then there are constants $c_1, c_2 > 0$ such that for all $T, \delta > 0$ we have

$$\mathbb{P}\left(\sup_{t\in[0,T]} \|P_s W_L(t)\|_{C^0} > \delta\right) \leq \exp\{c_1 T^\zeta - c_2 \delta^2\}.$$

We remark without proof that the condition $m > d$ is not optimal for $Q = I$; here $2m > d$ should be also true.

Using a result similar to (5.3) for P_cW, we obtain the following.

Corollary 5.6. *Let the assumptions of the previous theorem be true. Then there are constants $c_i > 0$ such that for all $T, C_w > 0$, we obtain*

$$\mathbb{P}\left(\sup_{t\in[0,T]} \|W_L(t)\|_{C^0} > C_w \right) \le c_3 \exp\{-c_4 C_w^2/T\} + \exp\{c_1 T^\zeta - c_2 C_w^2\}.$$

For sufficiently large $t > 0$ and all $\delta > 0$ it is well known that

$$\mathbb{P}\Big(\|P_s W_{L_\varepsilon}(t)\|_{C^0} > C_w \Big) \le c_6 \exp\{-c_5 C_w^2\}. \tag{5.7}$$

This is, for example, easily proven by using the imbedding of C^0 into H^s for $s > d/2$ and by calculating that for some small $h > 0$ the exponential moment $\mathbb{E}\exp\{h\|P_sW_L(t)\|_{H^s}^2\}$ is uniformly bounded with respect to t.

6. Applications. In our applications we consider for simplicity of presentation only examples which are restricted to scalar SPDEs in one space dimension $d = 1$. Moreover, we will consider only examples where the amplitude equation is only an SODE in $\mathbb{R}$. We could treat systems of SPDEs in higher dimension or higher-dimensional SODEs as amplitude equations, but we want to keep the notation as simple as possible in order to demonstrate the main ideas of our applications.

Let us fix some notation needed in what follows. Suppose W is some Q-Wiener process with either $\mathrm{tr}(Q) < \infty$ or $Q = I$. Define $L = -(1 + \partial_x^2)^2$, which is a self-adjoint operator on $L^2([0, \pi])$ subject to zero Dirichlet boundary conditions for u and $\partial_x^2 u$.

It is well known that the fourth order differential operator L generates a bounded semigroup on $Y = X = C^0([0, \pi])$, which fulfills Assumptions 2.1 and 2.2 with $\mathcal{N} = \mathrm{span}\{\sin\}$ and $\alpha = 0$ in (2.2). Therefore, we fix P_c to be the L^2-orthogonal projection onto $\mathcal{N}$, which coincides in this case with the spectral projection of L defined on X. It is also easy to check that Assumption 2.3 is true.

Moreover, we have the L^2-orthogonal basis of eigenfunctions $e_k(x) = \sin(kx)$ in X with $\|e_k\|_X = 1$. Note that Assumption 5.4 is true with $m = 2$ and $p(z) = -(1 + z^2)^2$. Therefore, we can apply all large deviation results of the previous subsection.

6.1. Dynamic pitchfork-bifurcation. The dynamic pitchfork-bifurcation is a well-studied experimental effect (see, e.g., [BK99, ME87, GM03] and the references therein). It refers to a system which is moved slowly though some deterministic bifurcation point by slowly changing the bifurcation parameter with time. This leads to hysteresis-type effects, where the solution stays near the unstable equilibrium after passing through the deterministic bifurcation point. This result was studied rigorously for the deterministic and the stochastic equations (see [BG02] and the references therein), both in an ODE setting. We will describe how the latter result can be immediately carried over to SPDEs by the results presented in the previous sections.

Consider as an example the scalar SPDE

$$\partial_t u(t) = Lu(t) + \tilde{\nu}\varepsilon^4(t - \tau_0\varepsilon^{-2})u(t) - u^3(t) + \sigma\varepsilon^2\partial_t W(t) \quad \text{for} \quad t > 0, \tag{6.1}$$

subject to zero Dirichlet boundary conditions on $[0,\pi]$ and initial condition $u(0)=u_0$. Here $\tilde{\nu}$ and τ_0 are some constants fixed later on. It was already discussed in Example 2.6 that Assumption 2.5 is true with $X=Y=C^0([0,\pi])$.

As obviously $\mathcal{N}=\operatorname{span}\{\sin\}$, the corresponding amplitude equation describes the amplitude $a\in\mathbb{R}$ of the sine. It is given by an easy calculation (cf. section 2.1):

$$a'(T)=(s-\tau_0)\tilde{\nu}a(T)-\frac{3\pi}{8}a^3(T)+\sigma\beta'(T)\quad\text{for}\quad T>0, \tag{6.2}$$

where $\beta(T)=\langle \varepsilon W(T\varepsilon^{-2}),\sin\rangle_{L^2}$ is a real-valued Brownian motion.

We will first state rigorous results that verify $u(t)\approx a(\varepsilon^2 t)\sin$. After that we briefly comment on the dynamic pitchfork-bifurcation. Note that (6.2) exhibits exactly the same structure as the equations discussed in [BG02]. Therefore, an analogue of their result for (6.1) follows immediately.

Theorem 6.1 (attractivity). *There are constants $c_i>0$ such that for any choice of $\tau_0>0$ and $\tilde{\nu},\sigma\in[-1,1]$, all mild solutions u of (6.1), and all solutions a of (6.2), we obtain the following.*

For $t_\varepsilon=\ln(\varepsilon^{-1})$ we can write $u(t_\varepsilon)=\varepsilon a_\varepsilon\sin+\varepsilon^2 R_\varepsilon$ with

$$\begin{aligned}&\mathbb{P}\Big(|a_\varepsilon|\le 2C_\pi M\delta,\quad \|R_\varepsilon\|_{C^0}\le 2C_w\Big)\\&\ge \mathbb{P}\Big(\|u_0\|_{C^0}\le\delta\varepsilon\Big)-c_1e^{-c_2\varepsilon^{-1}}-c_3e^{-c_4C_w^2}\end{aligned}$$

for all $C_w,\delta>0$, and sufficiently small $\varepsilon>0$.

Note that the probability bound on the right-hand side is in general only positive for ε small and C_w large.

Proof. The proof is by straightforward application of Theorem 3.3 together with Corollary 5.6 and (5.7) to bound the probabilities. ∎

Theorem 6.2 (approximation). *For all $T_0>\tau_0>0$ and $0<\kappa<1$ there are constants $c_i>0$ such that the following is true.*

Given $\delta>0$, $C_a>0$, $C_w>0$, there is a constant $C_{\text{att}}>0$ such that for any choice of $\tilde{\nu},\sigma\in[-1,1]$, all mild solutions u of (6.1), and all solutions a of (6.2),

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in[0,T_0\varepsilon^{-2}]}\|u(t)-\varepsilon a(\varepsilon^2t)\sin\|_{C^0}\le C_{\text{att}}\varepsilon^{2-\kappa}\right)\\&\ge 1-\mathbb{P}\Big(\|u_0-\varepsilon a(0)\sin\|_{C^0}>C_w\varepsilon\Big)\\&\qquad\qquad-\mathbb{P}\Big(|a(0)|>C_a\Big)-c_1e^{-c_2\varepsilon^{-\kappa}}-c_3e^{-c_4C_a^2/T_0}\end{aligned} \tag{6.3}$$

for sufficiently small $\varepsilon>0$.

Proof. The proof is by application of Theorems 4.3 and 4.1 together with Theorems 5.5 and 5.1 to bound probabilities. ∎

We can further bound (6.3) in Theorem 6.2 by Theorem 6.1. Summarizing both theorems, we can write with high probability all mild solutions u of (6.1) with $\|u(0)\|=\mathcal{O}(\varepsilon)$ as

$$u(t)=\varepsilon a(\varepsilon^2t)\sin+\mathcal{O}(\varepsilon^{2-\kappa}) \tag{6.4}$$

for all $t \in \ln(\varepsilon^{-1}) + [0, T_0\varepsilon^{-2}]$, where a is a solution of the amplitude equation (6.2) with initial condition $a_0 = \varepsilon^{-1}\Pi u(t_\varepsilon)$, where the projection Π was defined in section 2.1.

Now we can immediately apply the results of [BG02] first to (6.2) and then via (6.4) to solutions of (6.1). Hence, for ε sufficiently small, (6.1) exhibits the same dynamics as the one-dimensional dynamic pitchfork-bifurcation, but on a much slower time-scale. To keep the presentation short, we refrain from restating the elaborate description of the transient dynamics from [BG02]. Note, finally, that it is essential to derive a one-dimensional kernel $\mathcal{N}$, as the results of [BG02] are only valid for one-dimensional amplitude equations.

6.2. Pattern formation. The formation of a pattern below the threshold of a change of stability is a well-known experimental phenomenon. See, for example, [SA02] or [SR94] for noise-induced convection rolls below the onset of convection in Bénard's problem. In this case the system is slightly below a change of stability in the unperturbed (deterministic) system, which undergoes a pitchfork-bifurcation.

Unfortunately, this problem is out of reach for the present approach, as it consists of a three-dimensional Navier–Stokes equation which is coupled to a heat equation. We could treat systems of SPDEs on three-dimensional domains, but we cannot treat the quadratic nonlinearity with the method presented in this article. This will be a topic of further research (see [B03]).

We sketch a simple problem, which exhibits pattern formation below threshold of stability. Let us consider the well-known Swift–Hohenberg equation, which is in the theory of convection frequently used as a simplified model for the first convective instability. It is similar to the equations of the previous section,

$$\partial_t u(t) = Lu(t) + \nu\varepsilon^2 u(t) - u^3(t) + \varepsilon^2 \partial_t W(t) \quad \text{for} \quad t > 0, \tag{6.5}$$

subject to zero Dirichlet-type boundary conditions on $[0, \pi]$ and initial condition $u_0 = 0$, where L and W are given in the beginning of section 6 and ν is some parameter.

The pattern in our toy-model us just the sine representing the convection roll in the full problem. Due to the special size of the domain, we have only one period of the sine. If we would consider large domains, we would get several periods of the pattern depending on the size of the domain. Nevertheless, for sake of simplicity we stay with this very simple model.

For the unperturbed deterministic equation it is well known that it undergoes a pitchfork-bifurcation at $\nu = 0$. There $\sin \in N(L)$ becomes unstable. For $\nu < 0$ the homogeneous solution $u = 0$ is the only stable solution, and for $\nu > 0$ we end up with a stable pattern that is a small deformation of the sine. To verify this result is a lot of work but is standard, using, for instance, the celebrated theory of Crandel and Rabinowitz. In contrast to that, we will see that also in the case of $\nu < 0$ due to additive noise the pattern appears and sustains for long times, although it should decay due to the stability of the homogeneous solution $u = 0$. In the following we will verify a result that the probability $\mathbb{P}$(pattern visible for "most" $t \in [0, T_0\varepsilon^{-2}]$) is near 1, where T_0 is just some arbitrary constant.

Let us first apply the main results of this paper to (6.5). Due to $u(0) = 0$ it is obvious that the assertion of the attractivity result readily holds with $t_\varepsilon = 0$, $a_\varepsilon = 0$, and $R_\varepsilon = 0$.

Moreover, an easy calculation (see section 2.1) establishes that the corresponding ampli-

tude equation is

$$a' = \nu a - \frac{3\pi}{8}a^3 + \beta', \tag{6.6}$$

where $\beta(T) = \langle \varepsilon W(T\varepsilon^{-2}), \sin\rangle_{L^2}$ as before in section 6.1. Now we can easily verify an approximation result as in Theorem 6.2 to obtain

$$u(t) = \varepsilon a(\varepsilon^2 t)\sin + \mathcal{O}(\varepsilon^{2-\kappa}) \qquad \text{uniformly for all} \quad t \in [0, T_e\varepsilon^{-2}] \tag{6.7}$$

with high probability for all $\varepsilon \in [0, \varepsilon_0]$. We are free to choose any $T_0 > 0$ we want, but we have to pay for that with $\varepsilon_0 > 0$ small. We refrain from restating the precise result, as it is completely analogous to Theorem 6.2.

To prove a pattern result, we can, for example, verify that $|a(T)| \geq C\varepsilon^{1/2}$ for "a lot of" times $T \in [0, T_0]$. In what follows we give a short argument for this.

First define

$$l_\varepsilon(T) := |\{s \in [0, T] \ : \ |a(s)| \leq \varepsilon^{1/2}\}|.$$

This is the "bad" set of times, where we possibly do not see the pattern. However, we will definitely see the pattern for all times in $[0, T_0\varepsilon^{-2}] - l_\varepsilon(T_0)$. The following remark summarizes the result, which is now possible to verify. We refrain from stating an abstract theorem.

Remark 6.3. *For a main result on pattern formation, we will verify*

$$\mathbb{P}(l_\varepsilon(T_0) \geq \varepsilon^{1/4}) \leq C\varepsilon^{1/4}T_0 \tag{6.8}$$

for T_0 large and ε small enough.

Hence the probability is high to see the pattern on a set of times with measure $(T_0 - C\varepsilon^{1/4})\varepsilon^{-2}$ for any choice of the bifurcation parameter $\varepsilon^2\nu$, provided, for example, $|\nu| \leq 1$.

Note finally that there is nothing special about dimension one. Similar results will apply in case $n = \dim(\mathcal{N}) > 1$, where $\mathcal{N}$ is then some space of pattern. The only thing we rely on are some technical assumptions for the existence of an invariant Markov measure for the amplitude equation.

To establish (6.8) first recall that the distribution of a is independent of ε by Remark 2.8. Moreover, it is well known under certain assumptions on the noise (e.g., full rank of the covariance matrix) and the stability of the cubic nonlinearity that there exists a unique invariant Markov measure $\mathbb{P}^*$ for the amplitude equation. Moreover, the Lebesgue-density $p^* := d\mathbb{P}^*/d\lambda$ of this measure is continuous.

By the definition of l_ε

$$\mathbb{E}l_\varepsilon(T) = \int_0^T \mathbb{P}(|a(s)| \leq \varepsilon^{1/2})ds.$$

Now the celebrated Birkhoff ergodic theorem (cf., e.g., [dPZ96]) for invariant Markov measures implies

$$\frac{1}{T}\mathbb{E}l_\varepsilon(T) \to \mathbb{P}^*([-\varepsilon^{1/2}, \varepsilon^{1/2}]) \quad \text{for} \quad T \to \infty.$$

Moreover, we obtain $\frac{1}{2}\varepsilon^{-1/2}\mathbb{P}^*([-\varepsilon^{1/2},\varepsilon^{1/2}]) \to p^*(0)$ by the continuity of the density. Furthermore, by the Chebyshev inequality,

$$\mathbb{P}(l_\varepsilon(T) \geq \varepsilon^{1/4}) \leq \varepsilon^{-1/4}\mathbb{E}l_\varepsilon(T) \approx 2\varepsilon^{1/4}Tp^*(0),$$

and it is straightforward to establish (6.8).

6.3. Bifurcation. In this section we briefly sketch transient stochastic dynamics near a deterministic pitchfork-bifurcation. As mentioned in the introduction, we will not describe the whole bifurcation but rather will focus on examples of parameter regimes, where we can establish the amplitude equation describing finite time behavior of solutions.

As we will see, there are different scenarios depending on the ratio between noise-strength and bifurcation parameter. This will give another indication to the well-known fact that a stochastically perturbed bifurcation leads to a soft transition of the transient dynamics (cf., e.g., [Ar98]), in contrast to the sharp separation in case of a deterministic bifurcation.

Consider the same type of equation as in the previous section, which is for $\sigma = 0$ a classical example of a pitchfork-bifurcation in a PDE:

$$\partial_t u = Lu + \mu u - u^3 + \sigma^2\partial_t W, \tag{6.9}$$

subject to zero Dirichlet-type boundary conditions on $[0,\pi]$.

We can distinguish between three different regimes, and in the following we just give the corresponding amplitude equation and discuss the transient dynamics that we expect. Moreover, in the end we sketch the basic ideas of how to modify our results to derive the amplitude equations for the different cases. Note that we do not specify the constant c appearing in the equations, as it depends on the normalization of the function e spanning $\mathcal{N}$.

1. Case $|\mu| \approx \sigma^2 \approx \varepsilon^2$.
   Fix $\sigma^2 = \varepsilon^2$ and $\mu = \nu\varepsilon^2$. The amplitude equation is $a' = \nu a - ca^3 + \beta'$.
2. Case $|\mu| \ll \sigma^2 \approx \varepsilon^2$.
   Fix $\sigma^2 = \varepsilon^2$ and $|\mu| \leq \varepsilon^3$. The amplitude equation is $a' = -ca^3 + \beta'$.
3. Case $\sigma^2 \ll |\mu| \approx \varepsilon^2$.
   Fix $\varepsilon^2 = |\mu|$ and $\sigma \leq \varepsilon^3$. The amplitude equation is $a' = \operatorname{sgn}(\mu)a - ca^3$.

Case 2 corresponds to the case when we are very near to the deterministic bifurcation point. The amplitude equation and the dynamics are independent of the bifurcation parameter in this case. Hence the deterministic bifurcation point is widened to a longer interval.

Case 3 corresponds to the case when we are far away from the bifurcation. Here the dynamics is essentially given by the stable equilibria of the deterministic equation. The stochastic nature of the original SPDE is only seen in small fluctuations around these fixed points.

In Case 1, we have an intermediate regime, when we are of order noise-strength away from the bifurcation. Here the amplitude equation is stochastic, and the dynamical behavior interpolates between the deterministic behavior (Case 3) and the bifurcation regime (Case 2).

Let us finally comment on how to rigorously derive the amplitude equations in the previous statements. Case 1 was already discussed in the previous sections. The second case is rather easy. We can, for instance, follow our proofs of the approximation result to see that all ν-dependent terms are now of lower order. Hence they disappear into the error terms, and there

will be no contribution to the amplitude equation. For Case 3 we can also follow our approach. This would give an amplitude equation of the type $a' = \operatorname{sgn}(\mu)a - ca^3 + \sigma^2\varepsilon^{-2}\beta'$. Then we can add a Freidlin–Wentzell-type argument to eliminate the noise term, as $\sigma^2\varepsilon^{-2} \to 0$.

Acknowledgments. The author would like to thank the Mathematical Research Centre of the University of Warwick for its hospitality.

Moreover, the author would like to thank the referees and especially the editor Björn Sandstede for many helpful remarks improving the presentation of the manuscript.

REFERENCES

[Ar98] L. Arnold, *Random Dynamical Systems*, Springer Monogr. Math., Springer-Verlag, Berlin, 1998.

[BG02] N. Berglund and B. Gentz, *Pathwise description of dynamic pitchfork bifurcations with additive noise*, Probab. Theory Related Fields, 122 (2002), pp. 341–388.

[BK99] N. Berglund and H. Kunz, *Memory effects and scaling laws in slowly driven systems*, J. Phys. A, 32 (1999), pp. 15–39.

[BMS01] D. Blömker, S. Maier-Paape, and G. Schneider, *The stochastic Landau equation as an amplitude equation*, Discrete Contin. Dyn. Syst. Ser. B, 1 (2001), pp. 527–541.

[B03] D. Blömker, *Approximation of the Stochastic Rayleigh-Bénard Problem near the Onset of Convection and Related Problems*, preprint, 2003.

[BP99] Z. Brzeźniak and S. Peszat, *Space-time continuous solutions to SPDE's driven by a homogeneous Wiener process*, Studia Math., 137 (1999), pp. 261–299.

[BP00] Z. Brzeźniak and S. Peszat, *Strong local and global solutions for stochastic Navier-Stokes equations*, in Infinite Dimensional Stochastic Analysis (Amsterdam, 1999), Verh. Afd. Natuurkd. 1. Reeks. K. Ned. Akad. Wet. 52, R. Neth. Acad. Arts Sci., Amsterdam, 2000, pp. 85–98.

[BP00b] Z. Brzeźniak and S. Peszat, *Maximal inequalities and exponential estimates for stochastic convolutions in Banach spaces*, in Stochastic Processes, Physics and Geometry: New Interplays, I (Leipzig, 1999), CMS Conf. Proc. 28, AMS, Providence, RI, 2000, pp. 55–64.

[CLR01] T. Caraballo, J. A. Langa, and J. C. Robinson, *A stochastic pitchfork bifurcation in a reaction-diffusion equation*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 2041–2061.

[CIS99] H. Crauel, P. Imkeller, and M. Steinkamp, *Bifurcations of one-dimensional stochastic differential equations*, in Stochastic Dynamics (Bremen, 1997), Springer-Verlag, New York, 1999, pp. 27–47.

[CF98] H. Crauel and F. Flandoli, *Additive noise destroys a pitchfork bifurcation*, J. Dynam. Differential Equations, 10 (1998), pp. 259–274.

[CM90] P. L. Chow and J.-L. Menaldi, *Exponential estimates in exit probability for some diffusion processes in Hilbert spaces*, Stochastics Stochastics Rep., 29 (1990), pp. 377–393.

[dPZ92] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[dPZ96] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite-Dimensional Systems*, London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.

[DZ98] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, 2nd ed., Springer-Verlag, New York, 1998.

[DLS03] J. Duan, K. Lu, and B. Schmalfuss, *Invariant manifolds for stochastic partial differential equations*, Ann. Probab., to appear.

[EE87] D. E. Edmunds and W. D. Evans, *Spectral Theory and Differential Operators*, Oxford Math. Monogr., Oxford Science Publications, The Clarendon Press, Oxford University Press, New York, 1987.

[FW98] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, 2nd ed., Springer-Verlag, New York, 1998.

[GM03] T. Galla and E. Moro, *Defect formation in the Swift-Hohenberg equation*, Phys. Rev. E, 67 (2003), 035101.

[H83] H. Haken, *Synergetics. An Introduction. Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology*, Springer Ser. Synergetics 1, Springer-Verlag, Berlin, 1983.

[Ha80] J. Hale, *Ordinary Differential Equations*, Robert E. Krieger Publishing, Huntington, NY, 1980.

[HS92] P. C. Hohenberg and J. B. Swift, *Effects of additive noise at the onset of Rayleigh-Bénard convection*, Phys. Rev. A, 46 (1992), pp. 4773–4785.

[K95] T. Kato, *Perturbation Theory for Linear Operators. Reprint of the* 1980 *Edition*, Classics Math., Springer-Verlag, Berlin, 1995.

[KSM92] P. Kirrmann, G. Schneider, and A. Mielke, *The validity of modulation equations for extended systems with cubic nonlinearities*, Proc. Roy. Soc. Edinburgh Sect. A, 122 (1992), pp. 85–91.

[L94] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Progr. Nonlinear Differential Equations Appl. 16, Birkhäuser-Verlag, Basel, 1995.

[ME87] P. Mandel and T. Erneux, *The slow passage through a steady bifurcation: Delay and memory effects*, J. Statist. Phys., 48 (1987), pp. 1059–1070.

[MG99] E. Moro and L. Grant, *Dynamics of defect formation*, Phys. Rev. E, 59 (1999), pp. R1303–R1306.

[P92] S. Peszat, *Exponential tail estimates for infinite-dimensional stochastic convolutions*, Bull. Polish Acad. Sci. Math., 40 (1992), pp. 323–333.

[SA02] M. A. Scherer and G. Ahlers, *Temporal and spatial properties of fluctuations below a supercritical primary bifurcation to traveling oblique-roll electroconvection*, Phys. Rev. E, 65 (2002), 051101.

[SR94] W. Schöpf and I. Rehberg, *The influence of thermal noise on the onset of travelling-wave convection in binary fluid mixtures: An experimental investigation*, J. Fluid Mech., 271 (1994), pp. 235–265.

[SS03] J. Seidler and T. Sobukawa, *Exponential integrability of stochastic convolutions*, J. London Math. Soc., 67 (2003), pp. 245–258.

[vH91] A. van Harten, *On the validity of Ginzburg–Landau's equation*, J. Nonlinear Sci., 1 (1991), pp. 397–422.

[W97] D. Walgraef, *Spatio-temporal Pattern Formation*, Partially Ordered Systems, Springer-Verlag, New York, 1997.

[W80] J. Weidmann, *Linear Operators in Hilbert Spaces*, Springer-Verlag, New York, 1980.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 3, pp. 487–516

PDE Methods for Nonlocal Models*

Carlo R. Laing† and William C. Troy‡

Abstract. We develop partial differential equation (PDE) methods to study the dynamics of pattern formation in partial integro-differential equations (PIDEs) defined on a spatially extended domain. Our primary focus is on scalar equations in two spatial dimensions. These models arise in a variety of neuronal modeling problems and also occur in material science. We first derive a PDE which is equivalent to the PIDE. We then find circularly symmetric solutions of the resultant PDE; the linearization of the PDE around these solutions provides a criterion for their stability. When a solution is unstable, our analysis predicts the exact number of peaks that form to comprise a multipeak solution of the full PDE. We illustrate our results with specific numerical examples and discuss other systems for which this technique can be used.

Key words. pattern formation, integro-differential equation, PDE, nonlocal

AMS subject classifications. 34B15, 34C23, 93C15, 34C11

DOI. 10.1137/030600040

1. Introduction.

Pattern formation in neuronal networks is an area of ongoing interest [10, 11, 12, 13, 19, 26, 31, 35, 37, 40, 42, 43, 53, 54, 59, 60, 65]. In this paper, we investigate spatially localized regions of high activity, often referred to as "bumps." These are of interest in modeling working memory, the ability to remember information over a time-scale of a few seconds [18, 40, 63, 66]. Experiments on primates show that there exist regions of neurons that have elevated firing rates during the period that the animal is "remembering" some aspect of an object or event [17, 29, 45]. These regions are spatially localized in a location determined by the relevant aspect of the object or event being remembered. Further applications of pattern formation in neural systems include head-direction systems [58, 67], where a constantly updated bump of activity represents the current heading of an animal, and feature selectivity in the visual cortex [12, 13, 37], where bump formation may be related to the "tuning" of a particular neuron's response. Similar models to those studied here have been used to model the "look, plan, reach, remember" dynamics in the perseverative reaching of infants and their longer term cognitive development [60]. Also, in a recent book, Giese [31] uses systems of the form (1.1) to study problems related to visual perception of motion, the planning of eye movements, and robot navigation.

Realistic models for these types of activity involve spatially extended systems of coupled neural elements and the study of localized areas of high activity in these systems. Previous studies have focused on nonlocal rate models [1, 2, 3, 10, 11, 12, 13, 31, 37, 42, 53, 54, 66],

*Received by the editors February 12, 2003; accepted for publication (in revised form) by T. Kaper May 25, 2003; published electronically September 17, 2003.
http://www.siam.org/journals/siads/2-3/60004.html
†Institute of Information and Mathematical Sciences, Massey University, Private Bag 102-904, North Shore Mail Centre, Auckland, New Zealand (c.r.laing@massey.ac.nz).
‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (troy@math.pitt.edu).

in which a neural element is described by a scalar variable (e.g., a firing rate), and more complicated "spiking neuron" models [18, 35, 40, 63], which take into account the intrinsic dynamics of single neurons. We concentrate here on rate models exclusively.

In this paper, our goal is to develop PDE methods to analyze the dynamics of localized pattern formation in rate models of the form

$$\frac{\partial u(x,y,t)}{\partial t} = -u(x,y,t) + \iint_\Omega w(x-q,y-s) f(u(q,s,t)-th)\,dq\,ds, \tag{1.1}$$

where $\Omega \subset \mathrm{R}^2$. In the context of the neuronal models discussed above, $u(x,y,t)$ represents the synaptic input to a neuron at position $(x,y) \in \Omega$ at time t, w denotes the connectivity, or coupling, between neural elements, $f(u(x,y,t)-th)$ is the firing rate of the neuron at the position (x,y), and th is the threshold of excitation. Equations of the form (1.1) and its one-dimensional analogue have been used extensively in neural modeling [11, 12, 13, 26, 31, 37, 40, 43, 59, 65]. However, most of the previous attention has been focused on the one-dimensional analogue of (1.1):

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_\Omega w(x-y) f(u(y,t)-th)\,dy. \tag{1.2}$$

Because of the lower dimensionality of this system, it is easier to study than (1.1), and for applications such as the head-direction system and feature selectivity in the visual system this one-dimensional system may be appropriate, since the independent variable x corresponds to a one-dimensional quantity, an angle. However, the cortex is an essentially two-dimensional sheet, and thus (1.1) is a more realistic model for pattern formation in a neuronal system.

Little analytical work has been done on (1.1). Specifically, there are results relating to circularly symmetric solutions [59, 65], for which the two-dimensional problem is effectively reduced to a one-dimensional one. There has also been some very recent work on a "two layer" analogue of (1.1), where the activity of two populations of neurons are modeled [25]. However, these authors primarily studied target patterns, which also have circular symmetry. Also, Bressloff et al. have studied (1.1) on a sphere [12] and have also considered this equation when there is a lattice of inhomogeneities in the domain, using the symmetry of this lattice to determine the types of solution that can occur [11, 13].

Equations similar to (1.1) and (1.2) also occur in material science [5, 6, 7, 8, 16, 27, 28]. An often-studied equation is

$$\frac{\partial u(x,t)}{\partial t} = -ju + f(u) + \int_{-\infty}^{\infty} w(x-y)u(y,t)\,dy, \tag{1.3}$$

where $j = \int_{-\infty}^{\infty} w(x)\,dx$, w represents nonlocal interactions, and f typically has three zeros. Physically important solutions include heteroclinic, homoclinic, and periodic orbits. Solutions remain continuous when the function $g(u) \equiv ju - f(u)$ is monotonic. In this case, the equation $v = g(u)$ can be inverted to give $u = h(v) \equiv g^{-1}(v)$, and (1.3) becomes

$$h'(v)\frac{\partial v(x,t)}{\partial t} = -v(x,t) + \int_{-\infty}^{\infty} w(x-y)h(v(y,t))\,dy. \tag{1.4}$$

The similarities between (1.4) and its two-dimensional generalization, and (1.2) and (1.1), suggest that many of the results presented here will also be applicable to those models.

Our main focus will be on the two-dimensional model (1.1). We will develop techniques which describe how families of multibump solutions form when the coupling function $w(x, y)$ is a function of distance in R^2 only, i.e.,

$$w(x, y) = w\left(\sqrt{x^2 + y^2}\right). \tag{1.5}$$

Solutions of (1.1) are translationally invariant when (1.5) holds.

Our approach is twofold: first, we develop a method to derive a fourth order PDE which is equivalent to the integral equation. To our knowledge an equivalent PDE for the two-dimensional problem has not previously been derived. Second, we will analyze the PDE and develop techniques which show how families of peaks form when circularly symmetric steady states of (1.1) are unstable.

We hope that the insights obtained by considering coupling functions satisfying (1.5) will provide a basis for extensions to more complicated settings. For example, in the cortex it is important to consider connectivity functions that incorporate the "patchy" nature of neural connections [9, 34, 44], which break the translational invariance of the system. As mentioned above, Bressloff et al. have studied this effect [11, 13].

One assumption in studying (1.1) and (1.2) is that the behavior of neurons can be characterized by their firing rate alone and, more importantly, that excitatory and inhibitory neurons can be represented by a single population with a connectivity function that takes both positive and negative values. A more realistic approach would involve two variables, representing the activities of the excitatory and inhibitory populations, and coupling functions between and within the populations that are nonnegative. Although our models are less realistic in the sense that we use only one population, lumping the excitatory and inhibitory neurons together, they have the advantage of involving only one variable.

Note that we are not addressing the processes involved in the formation of the connectivities represented by $w(x)$ but are instead interested in the possible patterns of neural activity that can exist in the system once these connectivities are in place.

Overview. In the next section, we summarize previous results for one-dimensional models. The insights obtained in the study of the one-dimensional case play an important role in section 3, where we study two-dimensional models. Section 3 begins with numerical results which show the types of multibump solutions that can arise for specific examples. We then proceed with our analytical approach in which we (i) derive a PDE which approximates the integral equation and (ii) show how an appropriate linearization of the PDE will lead to a prediction of the exact number of peaks that form when a circularly symmetric solution is unstable. At the end of section 3, we show how families of 3-bump, 7-bump, and 12-bump solutions form for a specific example. Section 4 gives a summary of our results and some directions for future study.

2. Background: One-dimensional models.

Much of the present research into rate models stems from the early work of Wilson and Cowan [66] and subsequent studies by Amari [1, 2, 3] and Kishimoto and Amari [39]. These authors model the dynamics of a single layer of neurons

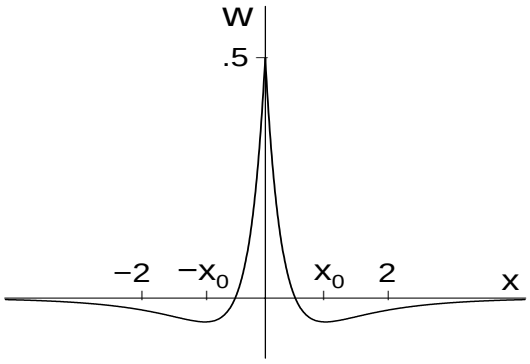


Figure 2.1. *An example of a coupling function giving "lateral inhibition" (positive for small $|x|$ and negative for large $|x|$).*

with rate equations of the form

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-y) f(u(y,t))\, dy + s(x,t) + h. \tag{2.1}$$

Here $u(x,t)$ is the synaptic input to a neural element at time $t \geq 0$ and position $x \in (-\infty, \infty)$, $w(x)$ determines the coupling between elements, and $f(u)$ gives the firing rate of a neuron with input u. Neurons are said to be "active" if $f(u(x,t)) > 0$. The function $s(x,t)$ denotes a variable external stimulus, and h is a constant stimulus applied to the entire field.

In [2], Amari set $f(u) = H(u)$, where $H(u)$ is the Heaviside function, and assumed the following:

(H_1) $w > 0$ and $w' < 0$ on an interval $(0, \bar{x})$, $w(-\bar{x}) = w(\bar{x}) = 0$, and $w(x) < 0$ if $|x| > \bar{x}$.

(H_2) w is a continuous even function, and $\int_{-\infty}^{\infty} w(y)\, dy$ is finite.

An example of this "lateral inhibition"–type coupling is given in Figure 2.1. He then analyzes stationary solutions of (2.1) when $s(x,t) = 0$. Setting $\partial u(x,t)/\partial t = 0$ reduces (2.1) to

$$u(x) = \int_{-\infty}^{\infty} w(x-y) H(u(y))\, dy + h. \tag{2.2}$$

The "region of excitation" of a stationary solution is the set $R(u) = \{x | u(x) > 0\}$. A 1-bump solution is a solution whose region of excitation is a finite interval. If the region of excitation consists of $N \geq 1$ disjoint finite intervals, then $u(x)$ is an N-bump solution. In [1, 2], Amari analyzes the existence, multiplicity, and stability of 1-bump solutions of (2.2).

In [43], we extended Amari's work and chose a specific $w(x)$, which changes sign infinitely often. For simplicity, we set $s(x,t) \equiv 0$ and studied

$$\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} w(x-y) f(u(y,t) - th)\, dy, \tag{2.3}$$

where

$$w(x) = e^{-b|x|}(b \sin|x| + \cos x) \quad \text{and} \quad f(u) = 2e^{-\tau/u^2} H(u), \tag{2.4}$$

and b, th, and τ are positive constants. The parameter b controls the rate at which oscillations in w decay with x, th is the threshold (effectively replacing h in (2.1)), and τ controls the slope of $f(u)$. Note that $f(u)$ is a C^∞ extension of (twice) the Heaviside function when $\tau > 0$, and $f(u - th) = 0$ if $u \le th$. An example of w and f are shown in Figure 2.2.

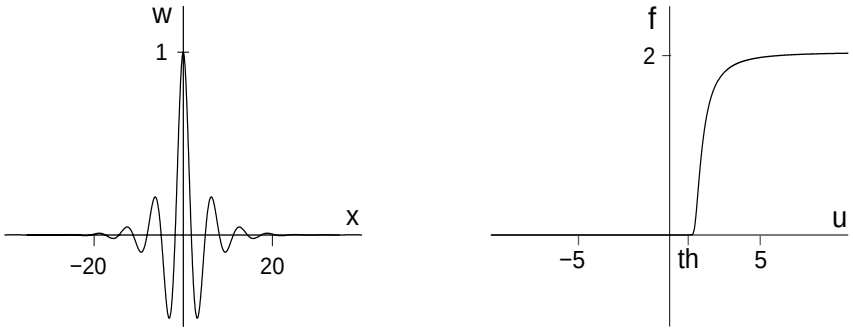


Figure 2.2. *Examples of $w(x)$ (left) and $f(u)$ (right) (2.4). Parameters are $b = 0.25$, $th = 1.5$, and $\tau = 0.005$.*

It is thought that the oscillatory form of coupling, (2.4), better represents the connectivity known to exist in the prefrontal cortex, where labeling studies have shown that coupled groups of neurons have "patchy" distributions, with a characteristic distance between patches [9, 34, 44].

As before, we define a stationary solution of (2.3)–(2.4) to be a solution of

$$u(x) = \int_{-\infty}^{\infty} w(x - y) f(u(y) - th)\, dy. \tag{2.5}$$

To compare the two-dimensional results in the next section with the one-dimensional case, we give a brief derivation of the PDE and ODE which are equivalent to (2.3) and (2.5). We use the Fourier transform, defined by $\widehat{\mathrm{F}}(g) = \int_{-\infty}^{\infty} e^{-i\alpha\eta} g(\eta)\, d\eta$, where $g \in L^1(\mathrm{R})$ and $\alpha \in \mathrm{R}$. Assume that u satisfies (2.5) and that u and u_t are continuous and integrable for $x \in \mathrm{R}$. Applying the Fourier transform to (2.3) and using its convolution property, we obtain

$$\widehat{\mathrm{F}}(u + u_t) = \widehat{\mathrm{F}}(w)\widehat{\mathrm{F}}(f(u - th)). \tag{2.6}$$

Evaluating $\widehat{\mathrm{F}}(w)$ results in

$$\widehat{\mathrm{F}}(u + u_t) = \frac{4b(b^2 + 1)}{\alpha^4 + 2\alpha^2(b^2 - 1) + (b^2 + 1)^2} \widehat{\mathrm{F}}(f(u - th)). \tag{2.7}$$

Next, multiplying (2.7) by $\alpha^4 + 2\alpha^2(b^2 - 1) + (b^2 + 1)^2$ and taking the inverse Fourier transform of both sides, we obtain the PDE

$$(u + u_t)_{xxxx} - 2(b^2 - 1)(u + u_t)_{xx} + (b^2 + 1)^2(u + u_t) = 4b(b^2 + 1) f(u - th). \tag{2.8}$$

This PDE is exactly equivalent to the partial integro-differential equation (PIDE) (2.3). Using this equivalence and setting $u_t = 0$ in (2.8), we see that N-bump stationary solutions of (2.3)

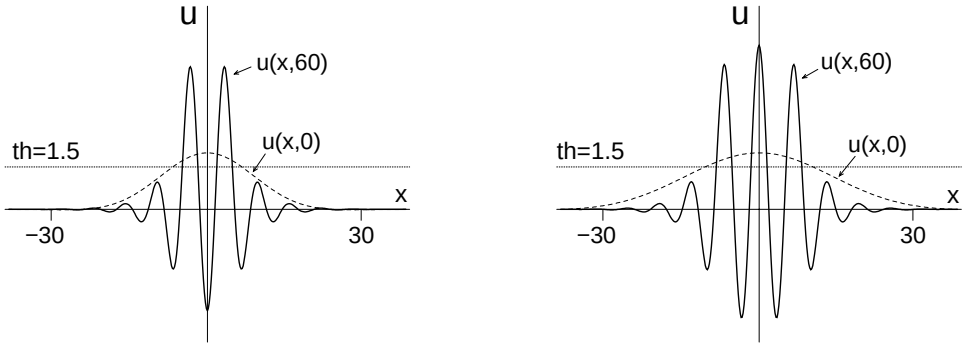


Figure 2.3. *Stable 2-bump (left) and 3-bump (right) solutions of (2.3)–(2.4). Parameters are $\tau = 0.1$, $th = 1.5$, and $b = 0.25$.*

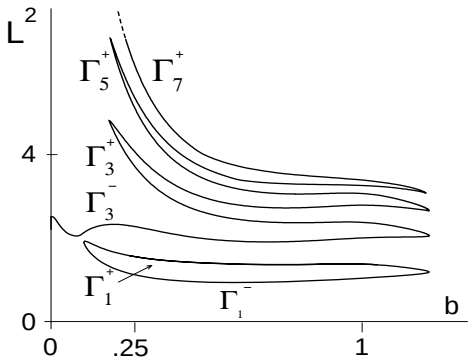


Figure 2.4. *Bifurcation diagram for N-bump solutions of (2.9) when N is odd. The vertical axis is the L^2 norm of the solution. Parameters are $\tau = 0.1$, $th = 1.5$.*

must satisfy the ODE problem

$$\begin{cases} u'''' - 2(b^2-1)u'' + (b^2+1)^2 u = 4b(b^2+1)f(u-th), \\ \lim_{x\to\pm\infty}(u,u',u'',u''') = (0,0,0,0). \end{cases} \tag{2.9}$$

In summary, by a judicious choice of coupling function w, we have exactly transformed the PIDE (2.3) into the PDE (2.8), whose stationary solutions satisfy the ODE problem (2.9). The new work in this paper consists of applying similar ideas to the two-dimensional system (1.1).

Numerical results. In [43], we solved (2.3) with initial conditions of the form

$$u(x,0) = \cos\left(\frac{Lx}{12.5\pi}\right)\exp\left(-\left(\frac{Lx}{12.5\pi}\right)^2\right), \quad -12.5\pi < x < 12.5\pi, \tag{2.10}$$

for different values of L (dashed curves in Figure 2.3). For appropriately chosen L, the initial condition evolves into a stationary N-bump solution which satisfies (2.9). To understand the global structure of solutions, we then used AUTO97 [22] to continue these N-bump solutions as b varied. This resulted in the bifurcation diagram shown in Figure 2.4, where Γ_N^+ and Γ_N^- are branches of stable and unstable N-bump solutions. (N is odd in this diagram—a similar

diagram showing families of N-bump solutions exists when N is even.) For $N \geq 3$, solutions arise at $b = 0$ through a bifurcation from a periodic orbit. Figure 2.4 suggests that a "snaking" phenomenon occurs in the branches of the bifurcation curve; solutions gain more bumps as the L^2 norm of the solution increases, with branches of stable solutions separated by branches of unstable solutions. See [43] for more details. Similar phenomena occur in other higher order ODE models [19, 38, 52]. For some applications, it is important to find ways to steer a system from one stable N-bump configuration to another. For example, switching from an N-bump to a 1-bump solution is of particular interest in the work of Thelen et al. [60], as it is proposed that this represents the process of decision-making by infants in the face of multiple choices.

3. The two-dimensional model. In this section, we analyze the formation of N-bump solutions in the two-dimensional model

$$\frac{\partial u(x,y,t)}{\partial t} = -u(x,y,t) + \iint_\Omega w(x-q, y-s) f(u(q,s,t) - th)\, dq\, ds, \tag{3.1}$$

where $f(u)$ is a positive multiple of the firing function defined in (2.4) and $w(x,y)$ satisfies (1.5). Stationary solutions of (3.1) satisfy the associated equation

$$u(x,y) = \iint_\Omega w(x-q, y-s) f(u(q,s) - th)\, dq\, ds. \tag{3.2}$$

For a given solution u of (3.2), we define its region of excitation to be

$$R(u) = \{(x,y) |\ u(x,y) > th\}. \tag{3.3}$$

A solution of (3.2) is an N-bump solution if its region of excitation consists of N finite disjoint components.

We will address the following basic questions:

(i) Is there a correspondence between families of N-bump solutions in one dimension and those in two dimensions? Do solutions exist in two dimensions that do not have one-dimensional counterparts?

(ii) What are the dynamics responsible for the formation of N-bump solutions? How can we derive a PDE that is equivalent to (3.1)?

3.1. Numerical examples. Thus far we have numerical results for three specific systems [42, 43]. The first is (3.1) with $f(u) = H(u)$ and

$$w(x,y) = Ke^{-k\sqrt{x^2+y^2}} - Me^{-m\sqrt{x^2+y^2}}. \tag{3.4}$$

In polar coordinates, (3.4) becomes $w(r) = Ke^{-kr} - Me^{-mr}$. If $K > M$ and $k > m$, then $w(r)$ has one positive zero. Figure 3.1 (upper left panel) shows a numerically stable 2-bump solution on a 10×10 square domain. Here $K = 3.5$, $k = 1.8$, $M = 2.8$, $m = 1.52$, and $th = 0$. In our study of the one-dimensional case, we found that stable 2-bump solutions could not exist for these parameter values, yet this computation suggests that stable 2-bump solutions do exist in two dimensions. We conjecture that this property can be explained using the analytical approach described in section 3.3.

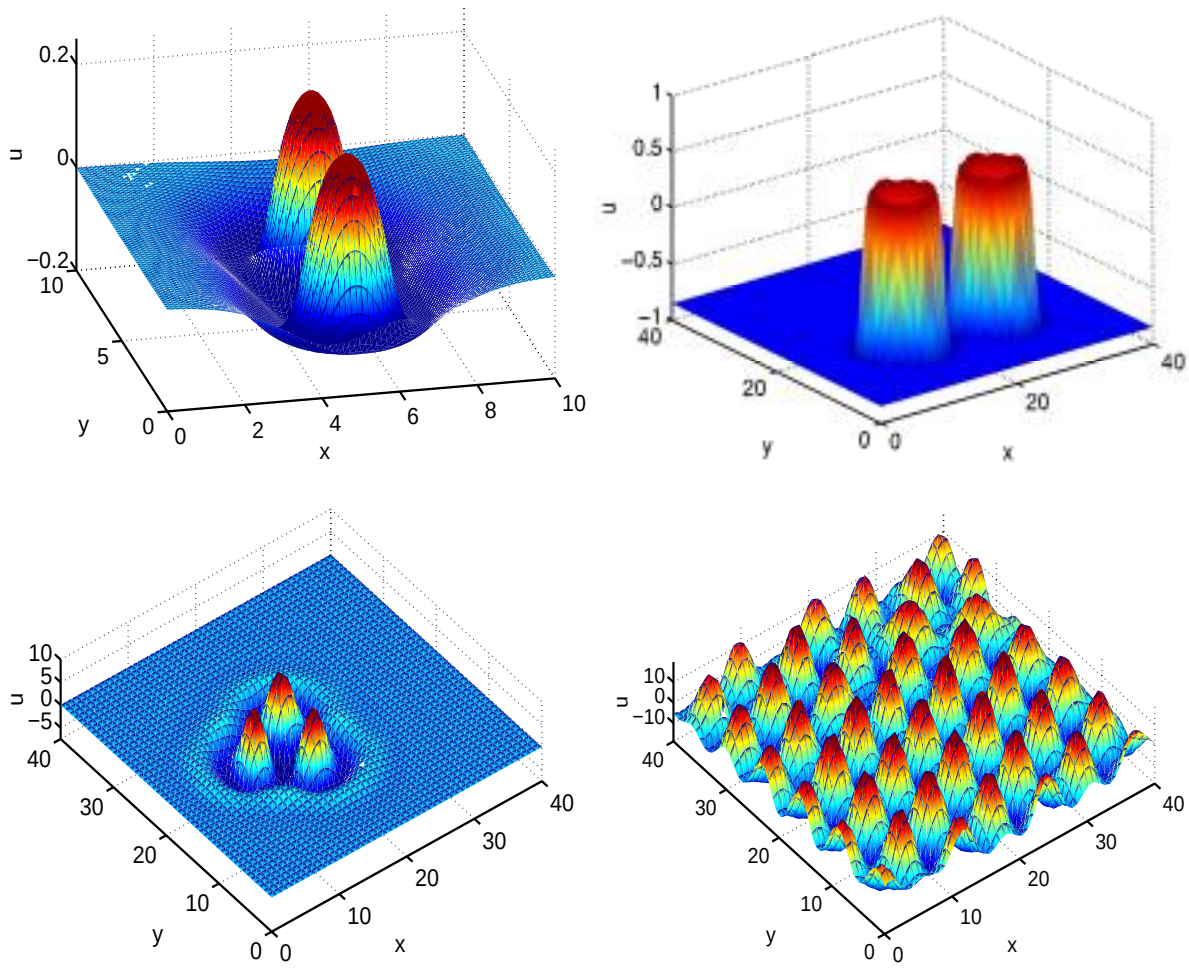


Figure 3.1. *Examples of multibump solutions for couplings* (3.4), (3.5), *and* (3.6). *See text.*

In our second example, we solve the two-dimensional analogue of (2.1) on a 40×40 grid with $f(u) = H(u)$, $s(x,t) \equiv 0$, and $h = -1$, and with the following coupling (studied in [42]):

$$w(x,y) = 2e^{-k\sqrt{x^2+y^2}} \left[1 - d_1(x^2+y^2) + d_2(x^2+y^2)^2 - d_3(x^2+y^2)^3\right]. \tag{3.5}$$

In polar coordinates, (3.5) becomes $w(r) = 2e^{-kr}(1 - d_1 r^2 + d_2 r^4 - d_3 r^6)$. For the choice of parameters $(k, d_1, d_2, d_3) = (1, \frac{2}{3}, \frac{1}{18}, \frac{1}{1200})$ the function $w(r)$ has three positive zeros. Figure 3.1 (upper right) shows a numerically stable 2-bump solution for this case.

The third problem we have studied consists of (3.1), with the firing rate coupling (2.4) and with the coupling function

$$w(x,y) = e^{-b\sqrt{x^2+y^2}} \left(b \sin\left(\sqrt{x^2+y^2}\right) + \cos\left(\sqrt{x^2+y^2}\right)\right), \tag{3.6}$$

where $b > 0$. The coupling in (3.6) is the two-dimensional analogue of (2.4). In polar coordinates, (3.6) becomes $w(r) = e^{-br}(b \sin r + \cos r)$, and $w(r)$ has infinitely many zeros. Figure 3.1 (lower panels) shows two numerically stable solutions computed on a square domain of size 40×40. For $b = 0.3$, the solution fills the entire domain with bumps (lower right panel). A similar "progressive recruitment" phenomenon is found by Gutkin, Ermentrout, and O'Sullivan in a one-dimensional model [34]. Usher, Stemmler, and Olami [62] found similar patterns in a stochastic model of spiking neurons that had short-range excitation and long-range inhibition. Raising b to $b = 0.4$, we find a 3-bump solution (lower left panel). As in the one-dimensional case, other N-bump solutions coexist.

3.2. Circularly symmetric solutions. Our numerical study shows that (3.1) has a rich structure of stable N-bump solutions for a wide range of coupling functions. In section 3.3, we will describe the fundamental role of circularly symmetric solutions in the formation of these solutions. To set the stage for section 3.3, our goal here is to summarize the important properties of the circularly symmetric solutions.

In polar coordinates, writing $x = r\cos\theta$, $y = r\sin\theta$, $q = \eta\cos\alpha$, $s = \eta\sin\alpha$, (3.1) becomes

$$\frac{\partial u(r,\theta,t)}{\partial t} = -u + \int_0^\infty \int_0^{2\pi} w\left(\sqrt{r^2+\eta^2-2r\eta\cos(\theta-\alpha)}\right) f(u(\eta,\alpha,t)-th)\eta\, d\alpha\, d\eta. \tag{3.7}$$

Stationary solutions of (3.7) satisfy

$$u(r,\theta) = \int_0^\infty \int_0^{2\pi} w\left(\sqrt{r^2+\eta^2-2r\eta\cos(\theta-\alpha)}\right) f(u(\eta,\alpha)-th)\eta\, d\alpha\, d\eta. \tag{3.8}$$

A solution is a circularly symmetric 1-bump solution if u is independent of θ and there is an $R_0 > 0$ such that

$$u(r) > th \quad \text{for} \quad 0 < r < R_0, \quad u(R_0) = 0, \quad \text{and} \quad u(r) < th \quad \text{for} \quad r > R_0. \tag{3.9}$$

Since u is assumed to be independent of θ, we set $\theta = 0$, and (3.8) reduces to

$$u(r) = \int_0^{R_0} \int_0^{2\pi} w\left(\sqrt{r^2+\eta^2-2r\eta\cos\alpha}\right) f(u(\eta,\alpha)-th)\eta\, d\alpha\, d\eta. \tag{3.10}$$

Thus a 1-bump circularly symmetric solution satisfies (3.9)–(3.10). When $N > 1$, circularly symmetric N-bump solutions are similarly defined.

Thus far, the only analytical results for circularly symmetric solutions are those given by Taylor [59] and Werner and Richter [65]. Taylor [59] discusses the case $f(u) = H(u)$, in which case $u(r)$ satisfies

$$u(r) = \int_0^{R_0} \int_0^{2\pi} w\left(\sqrt{r^2+\eta^2-2r\eta\cos\alpha}\right) \eta\, d\alpha\, d\eta \tag{3.11}$$

for couplings of the form

$$w(r) = Ke^{-kr^2} - Me^{-mr^2}, \tag{3.12}$$

where $K > M > 0$ and $k > m > 0$. He classifies some of the solutions and discusses their stability with respect to perturbations that vary only the radius of the solution. Werner and Richter [65] also discuss solutions of (3.11), in particular, circular and ring solutions, some of which were not found by Taylor. They also discuss the stability of these solutions with respect to perturbations that vary only the radius of the solution. In the next section, we will show how circularly symmetric N-bump solutions play an important role in the formation of asymmetric N-bump patterns.

We end this section by stating an open problem. Suppose that the region of excitation of a solution of (3.8) is a disc and that the firing rate function is a continuous increasing function of u. Then is it the case that u must be independent of θ? Or is it possible that there are coupling functions for which solutions exist which are not circularly symmetric? For elliptic PDEs the analogous problem of classifying positive solutions on a disc is very important and has been extensively studied [30, 57].

3.3. Noncircularly symmetric solutions: The PDE approach. There have been few attempts to analyze solutions of (3.1) that do not have circular symmetry. Recently we have made progress on this problem by successfully deriving a PDE that is equivalent to the PIDE (3.1). We have also developed a method of analysis of the PDE which explains the formation of N-bump solutions similar to those in Figures 3.7, 3.8, and 3.12. Our approach is described below.

The first step is to apply the two-dimensional Fourier transform, defined by $F(g) \equiv (2\pi)^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(\alpha x+\beta y)}g(x,y)dx\,dy$ to (3.1). Note that $F(g)$ is a function of α and β. We obtain

$$F(u+u_t) = F(w)F(f(u-th)). \tag{3.13}$$

For functions $w(x,y)$ that depend only on $\sqrt{x^2+y^2}$, it is known that $F(w)$ is a function of $\sqrt{\alpha^2+\beta^2}$ only. See Appendix A for a short proof. The coupling functions given in (3.4), (3.5), and (3.6) satisfy these properties. However, in each case $F(w)$ has a complicated form which prevents the use of (3.13) to derive a PDE. To circumvent this problem, we approximate $F(w)$ by a rational function, G, of $\sqrt{\alpha^2+\beta^2}$ containing only even powers of its argument. The rationale behind this is the same as for the one-dimensional case: we are using the observation that $F(\nabla^2 f) = -(\alpha^2+\beta^2)F(f)$.

We begin by choosing functions of the form

$$G\left(\sqrt{\alpha^2+\beta^2}\right) = \frac{A}{B+(\alpha^2+\beta^2-M)^2}, \tag{3.14}$$

where A, B, and M are parameters. Once G is known, the approximate coupling function $\widehat{w}(x,y)$ is given by the inverse two-dimensional Fourier transform of it. Because of the symmetry of G, this reduces to a Hankel transform of order 0:

$$\widehat{w}(r) = \int_0^{\infty} sG(s)J_0(rs)\,ds, \tag{3.15}$$

where J_0 is the Bessel function of the first kind of order zero (see Appendix A). In Figure 3.2, we illustrate an example of a coupling function $\widehat{w}$ when G is of the form (3.14). Note the similarity between this coupling function and the coupling function (3.6).

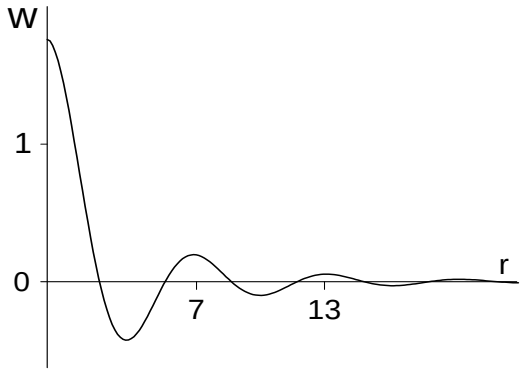


Figure 3.2. $\widehat{w}(r)$ (3.15) *for parameter values* $M = 1$, $A = 0.4$, *and* $B = 0.1$.

We now derive the PDE. First, replace $F(w)$ in (3.13) with (3.14). Next, multiply both sides of (3.13) by $B + (\alpha^2 + \beta^2 - M)^2$ and take the inverse two-dimensional Fourier transform to obtain

$$\nabla^4(u + u_t) + 2M\nabla^2(u + u_t) + (B + M^2)(u + u_t) = Af(u(x, y, t) - th). \tag{3.16}$$

This equation is exactly equivalent to the PIDE (3.1) if w is given by (3.15), where $F(w)$ is given by (3.14). It is interesting to note that the derivative of u with respect to time cannot be separated from (3.16).

One can view this process of derivation of a PDE in two different ways. One is that, given a coupling $w(r)$, we can find its Fourier transform and then approximate that by an appropriate rational function of $\sqrt{\alpha^2 + \beta^2}$. This rational function can then be used to derive a PDE whose dynamics will in some way approximate the dynamics of the original PIDE. The other way to view it is that we *define* $w(r)$ through (3.15) and, by varying the parameters A, B, and M in (3.14) (or in another appropriate rational function of $\sqrt{\alpha^2 + \beta^2}$), move through the space of possible coupling functions that can be treated this way. From this second point of view, there are no approximations made, but the tradeoff is that, by restricting G to be only rational functions of $\sqrt{\alpha^2 + \beta^2}$ with even powers of its argument, we may not be able to investigate all coupling functions $w(r)$ of interest. See section 3.5 for more discussion.

We now seek circularly symmetric solutions of (3.16). Under the assumption that u is not a function of θ, (3.16) becomes

$$\begin{aligned}\left[\frac{\partial^4}{\partial r^4} + \frac{2}{r}\frac{\partial^3}{\partial r^3} - \frac{1}{r^2}\frac{\partial^2}{\partial r^2} + \frac{1}{r^3}\frac{\partial}{\partial r} + 2M\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right)\right.\\ \left. + B + M^2\right]\left(u + \frac{\partial u}{\partial t}\right) = Af(u - th).\end{aligned} \tag{3.17}$$

Stationary solutions of (3.17) satisfy the ODE boundary value problem

$$\begin{cases} u'''' + \frac{2}{r}u''' - \frac{1}{r^2}u'' + \frac{1}{r^3}u' + 2M\left(u'' + \frac{1}{r}u'\right) + (B + M^2)u = Af(u(r) - th), \\ u'(0) = u'''(0) = 0, \text{ and } \lim_{r\to\infty}(u, u', u'', u''') = (0, 0, 0, 0). \end{cases} \tag{3.18}$$

In order to determine the stability of a stationary solution $\widetilde{u}(r)$ of (3.17), we linearize the full PDE (3.16) around it. To do this, we follow [20] and write

$$u(r,\theta,t) = \widetilde{u}(r) + \mu\nu(r,t)\cos(m\theta), \tag{3.19}$$

where μ is a small parameter, $\nu(r,t)$ is a perturbation function, and $m \geq 0$, an integer, is the azimuthal index. We choose this form of solution in order to investigate solutions that break the circular symmetry of the system. Substituting (3.19) into (3.16) and linearizing in μ, we obtain a PDE for ν:

$$\begin{aligned}\left[\frac{\partial^4}{\partial r^4} + \frac{2}{r}\frac{\partial^3}{\partial r^3} + \left(\frac{2Mr^2 - 2m^2 - 1}{r^2}\right)\frac{\partial^2}{\partial r^2} + \left(\frac{2m^2 + 1 + 2Mr^2}{r^3}\right)\frac{\partial}{\partial r}\right.\\ \left. + \frac{m^4 - 4m^2 + (B+M^2)r^4 - 2Mm^2r^2}{r^4}\right]\left(\nu + \frac{\partial\nu}{\partial t}\right) = Af'(\widetilde{u} - th)\nu.\end{aligned} \tag{3.20}$$

Since this is a linear equation in ν, we expect the solution to be of the form $\nu(r,t) \sim \overline{\nu}(r)e^{\lambda t}$ as $t \to \infty$, where λ is the most positive (real) eigenvalue and $\overline{\nu}(r)$ is the corresponding eigenfunction. In order to determine the stability of a particular circularly symmetric solution with radial profile $\widetilde{u}(r)$, we substitute $\widetilde{u}(r)$ into (3.20), and for each integer $m \geq 0$ we find the largest value of $\lambda(m)$. Then we determine the positive integer N at which λ is the greatest. If $\lambda(N) > 0$, then the solution with radial profile $\widetilde{u}(r)$ is unstable. Our analysis predicts that N bumps will form if the initial condition for (3.16) consists of a small random perturbation of the circularly symmetric solution whose radial profile is $\widetilde{u}(r)$. The distance from the origin at which these N bumps appear is determined by the shape of the eigenfunction $\overline{\nu}(r)$ (see examples below).

To numerically determine λ and $\overline{\nu}(r)$, we integrated (3.20) with a randomly chosen initial condition $\nu(r,0)$. In general this is composed of many eigenfunctions, but due to the exponential growth or decay in time, for large t, $\nu(r,t)$ is dominated by the eigenmode with the most positive corresponding λ. The quantity λ can thus be determined by plotting the log of the norm of $\nu(r,t)$ as a function of time and measuring the slope of the corresponding graph after transients have died away. The eigenfunction $\overline{\nu}(r)$ is simply $\nu(r,t)$ when t is large, suitably scaled in amplitude if necessary. This process was repeated with a number of different random initial conditions to verify that they did not affect the determination of λ and $\overline{\nu}$. In all of our experiments, we found that the eigenfunctions either grew or decayed monotonically in time as $t \to \infty$. This reinforces our assumption that the dominant eigenvalue is real. It would be interesting to investigate this further and provide a proof to rigorously determine the nature of the dominant eigenvalue.

Below we use the procedures described above to compute specific multibump solutions. To solve the PDEs (3.17) and (3.20), we used a finite difference scheme with 100 equally spaced r values in the interval $(0, 30]$ and an Euler step in time of length $\mathrm{dt} = 0.5$. Boundary conditions were $u = \partial u/\partial r = 0$ at $r = 30$. We solved the full PDE (3.16) on the disc $\Omega = \{(r,\theta)|0 \leq r \leq 30,\ 0 \leq \theta \leq 2\pi\}$ with boundary conditions $u = \partial u/\partial r = 0$ at $r = 30$ for all $\theta \in [0, 2\pi]$. Here we also used a finite difference scheme, discretizing the disc Ω into a 100×90 grid of (r,θ) values. In the t direction we again used an Euler step of length $\mathrm{dt} = 0.5$. The results shown were insensitive to changes in the time-step size, the number of points used

in the spatial discretization, and whether finite-difference or spectral methods [61] were used to approximate the spatial derivatives.

3.4. Examples of multibump formation. In this section, we demonstrate how three different families of multibump solutions form. The first is a 3-bump solution, the second is a 12-bump solution, and the third is a 7-bump solution. Throughout we use the parameter values $M = 1$, $A = 0.4$, and $B = 0.1$ in (3.14). For the firing rate we use

$$f(u) = e^{-\tau/u^2} H(u). \tag{3.21}$$

This function is a scalar multiple of the function in (2.4) and is a natural extension of the Heaviside function $H(u)$ (i.e., (3.21) reduces to the Heaviside function when $\tau = 0$). We fix $\tau = 0.1$ in (3.21). The first step is to find a circularly symmetric solution. For this we set $th = 0.25$ and solve (3.17) to obtain the solution shown in Figure 3.3 (left, solid curve). Note that this is only one of a number of stable solutions; it was selected by letting the initial condition $u(r, 0)$ be an appropriately chosen Gaussian.

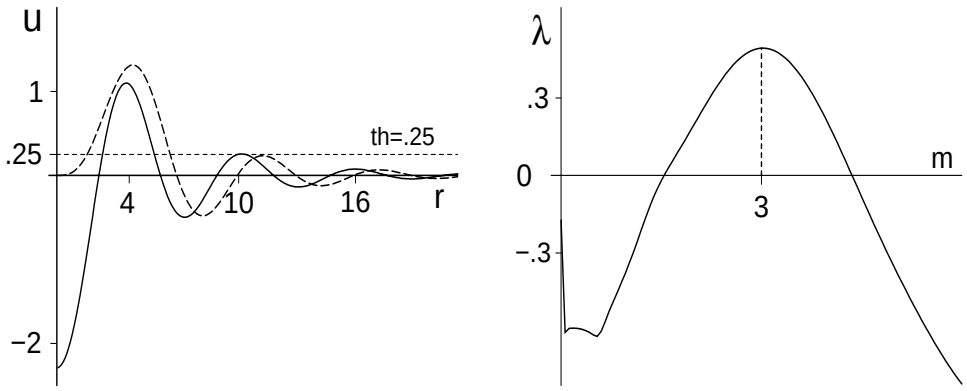


Figure 3.3. *Left: Stable solution of (3.17) (solid curve, defined as γ_1 in the text), and the eigenfunction corresponding to $m = 3$ (dashed). The solution is the one at P_1 in Figure 3.4. Right: λ as a function of m for the solution in the left panel.*

Next, we use AUTO97 [22] (applied to the system (3.18)) to continue this solution as the parameter th varies. Figure 3.4 shows the resulting bifurcation curve. The vertical axis is the maximum value of a solution, and the horizontal axis denotes the parameter th. The bifurcation curve has multiple folds (compare with Figure 2.4), and solutions gain more bumps as u_{max} increases. For example, at $th = 0.25$ there are several coexisting solutions, three of which are denoted by the points P_1, P_2, and P_3.

The solutions corresponding to these points are shown in Figures 3.3, 3.5, and 3.6, respectively. We have studied the stability of these solutions using (3.20), and for each we have computed the corresponding function $\lambda(m)$, shown in the right panels of Figures 3.3, 3.5, and 3.6.

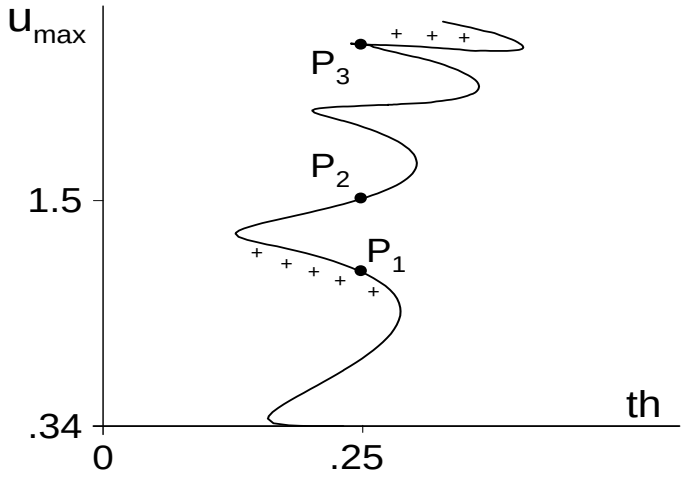


Figure 3.4. *Bifurcation diagram for solutions of* (3.18) *satisfying* $u(0) < 0$ *and* $u''(0) > 0$. *The solutions at* P_1, P_2, *and* P_3 *are shown in Figures* 3.3, 3.5, *and* 3.6, *respectively. Branches marked with a "+" sign are stable solutions of* (3.17).

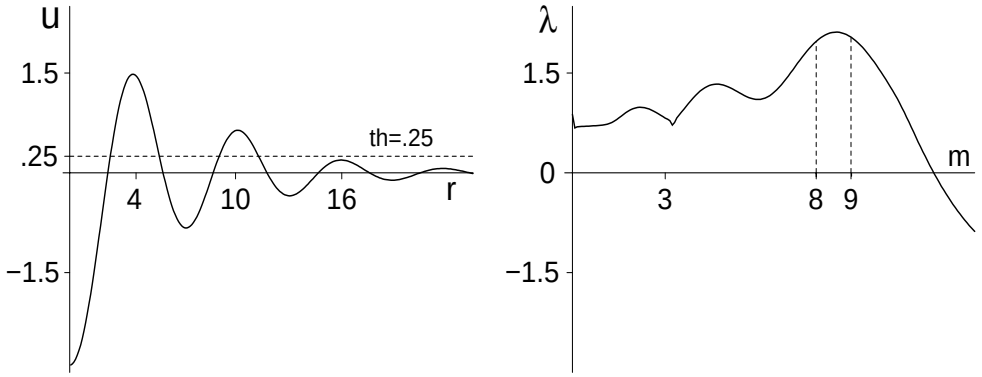


Figure 3.5. *Left: Unstable stationary solution of* (3.17) *corresponding to the point* P_2 *in Figure* 3.4. *The dashed line is* $u = th$. *Right:* λ *vs.* m *for the solution in the left panel. Note that* $\lambda(0) > 0$, *and hence we see the instability of the solution.*

3.4.1. The formation of a 3-bump solution. Define γ_1 to be the solution of (3.18) shown with a solid curve in the left panel of Figure 3.3 and Γ_1 to be the surface obtained by rotating γ_1 through a full circle about the line $r = 0$. Note that the positive parts of Γ_1 form a set of concentric annuli (see Figure 3.7, top left). Figure 3.3 (right panel) shows that for γ_1, $\lambda < 0$ when $m = 0$. This implies that γ_1 is actually a stable solution of (3.17). (Indeed it must be, as we found it by numerically integrating (3.17).) We also see that $m = 3$ is the integer with the largest value of λ and that $\lambda(3)$ is positive. The eigenfunction corresponding to $m = 3$ is shown in the left panel of Figure 3.3 (dashed curve). Its largest peak is centered over the

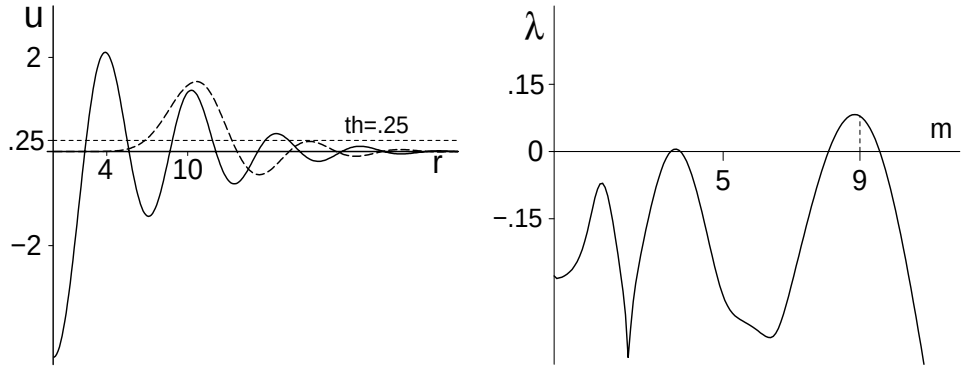


Figure 3.6. *Left: Stable stationary solution of* (3.17) *(solid curve) corresponding to* P_3 *in Figure* 3.4 *(defined to be* γ_3 *in the text), and the eigenfunction corresponding to* $m = 9$ *(dashed). Right:* λ *vs.* m *for the solution in the left panel.*

first positive bump of γ_1. Therefore, we expect any instability of Γ_1 in the full PDE (3.16) to have three-fold rotational symmetry and to appear at the annular part of Γ_1 corresponding to the first positive bump of γ_1. That is, if the initial condition of (3.16) is a small perturbation from Γ_1, we predict that the resulting solution will evolve into a 3-bump solution. Figure 3.7 illustrates that this is what happens. The upper left panel shows the initial condition, a small random perturbation of Γ_1 (the circularly symmetric solution generated by rotating γ_1 through a full circle). The next two panels illustrate the formation of a 3-bump solution as t increases from $t = 0$ to $t = 30$. If the integration is continued past $t = 30$, the 3-bump solution stimulates nearby regions and more bumps form, eventually filling the entire region (not shown). The resulting pattern is similar to that seen in the lower right panel of Figure 3.1. However, if we raise the threshold from $th = 0.25$ to $th = 0.3$ at $t = 30$, the nearby regions are *not* sufficiently stimulated, and the (now stable) 3-bump pattern persists (compare with Figure 3.1, lower left).

We now briefly describe properties of the solution corresponding to P_2, shown in the left panel of Figure 3.5. This solution has two intervals on which $u > th$ and is an *unstable* solution of (3.17) since $\lambda(0) > 0$. Note that we had to use AUTO97 to find this solution due to its instability. (Note also that it is not a stable solution of (3.16)).

3.4.2. The formation of a 12-bump solution. We now focus on the stationary solution of (3.17), which we define as γ_3, corresponding to the point P_3 in Figure 3.4 (see Figure 3.6). Upon rotation through a full circle about the line $r = 0$, γ_3 generates the multiring annular solution which we define to be Γ_3, similar to that shown in the upper left panel of Figure 3.8. γ_3 is a stable solution of (3.17) since $\lambda(0) < 0$. However, it is an unstable solution of (3.16) since for some m, $\lambda(m) > 0$. (Indeed, $m = 9$ is the integer with the largest positive λ.) The eigenfunction corresponding to $m = 9$ is shown dashed in Figure 3.6 (left). The largest peak of the eigenfunction is centered over the second positive bump of γ_3. Thus, if a small perturbation of Γ_3 was used as the initial condition for (3.16), we predict that the solution

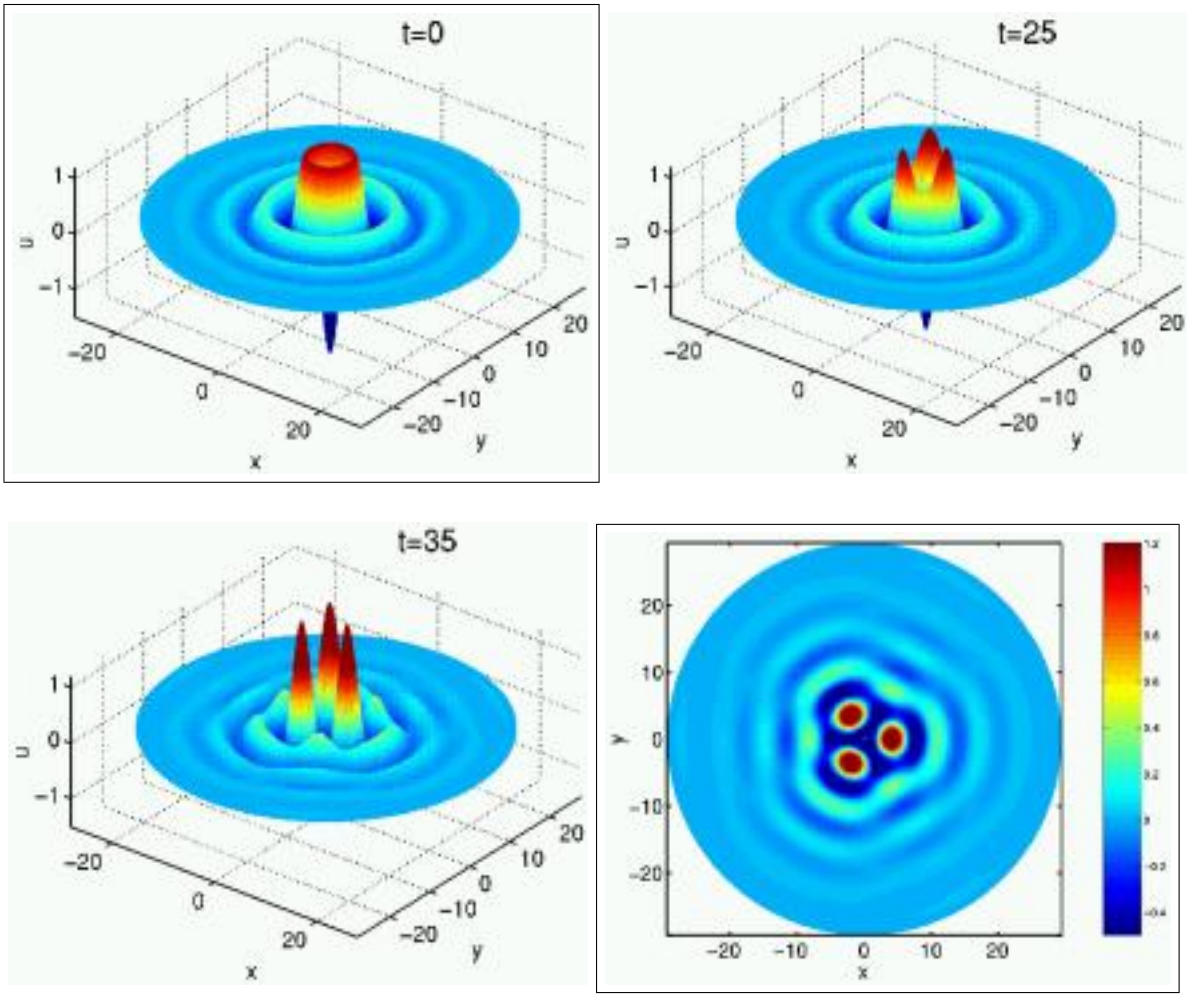


Figure 3.7. *The 3-peak solution resulting from the instability of the circularly symmetric solution whose radial profile is given by the solid curve in Figure 3.3 (left). Top left: Initial condition (a small random perturbation of Γ_1). Top right: At $t = 25$. Bottom left: At $t = 35$. Bottom right: Level curve diagram at $t = 35$. Clicking on the top left or bottom right panels will show movies of the development of the solution from different viewpoints.*

would develop 9-fold rotational symmetry, with the nine new bumps appearing in place of the annulus corresponding to the second positive bump of γ_3. That is, we predict that the second ring will break into nine bumps. In Figure 3.8 we see that this is what happens. As t increases from $t = 0$ to $t = 25$, the inner ring retains its circularly symmetric structure.

Although it is not easily seen in Figure 3.8, there is a subtle two-step process that happens next. First, as t increases from $t = 25$, the amplitude of the inner ring shrinks until the inner ring (taken in isolation) is the same size as the single ring shown in the upper left panel of Figure 3.7. (Recall that this solution corresponds to the point P_1 in Figure 3.4.) After this point, the inner ring begins its evolution into a 3-bump structure, while the outer nine bumps remain (lower panels of Figure 3.8). To understand the "shrinking" phenomenon, we have a

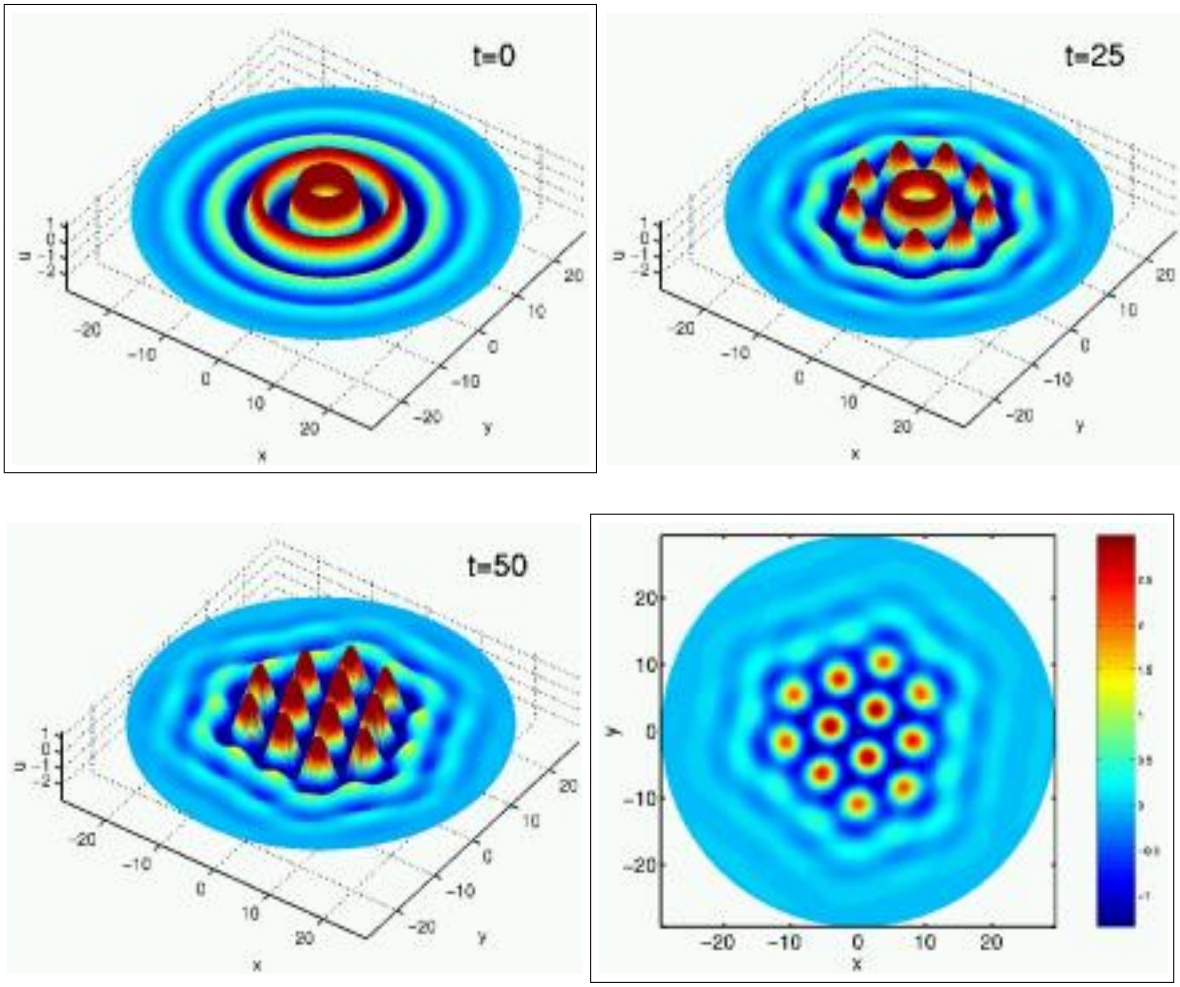


Figure 3.8. *The formation of a* 12*-peak pattern resulting from the instability of the circularly symmetric solution whose radial profile is given by the solid curve in Figure* 3.6 *(left), the axially symmetric solution at* P_3*. Top left: Initial condition (a small random perturbation of* Γ_3*). Top right:* $t = 25$*. Lower panels:* $t = 50$ *(left), and corresponding level curve diagram (right). Clicking on the top left or bottom right panels will show movies of the development of the solution from different viewpoints.*

plausible explanation based on the following calculation: First, we let the initial condition for a solution of (3.17) consist of the function which is equal to γ_3 until its first negative-going zero crossing and which is zero otherwise. (This initial condition is shown dashed in Figure 3.9.) Next, we solved (3.17) with this initial condition and found that the solution quickly shrank in amplitude and evolved into γ_1 (shown in Figure 3.3), which corresponds to the point P_1 in Figure 3.4. Since we know that Γ_1 breaks into a 3-peak structure under the dynamics of (3.16), we expect the "shrunken" inner ring to begin breaking into a 3-peak structure as t increases further. Figure 3.8 shows that this is indeed what happens.

3.4.3. The formation of a 7-bump solution. For our third example, we consider the class of circularly symmetric solutions consisting of a central peak surrounded by one or more

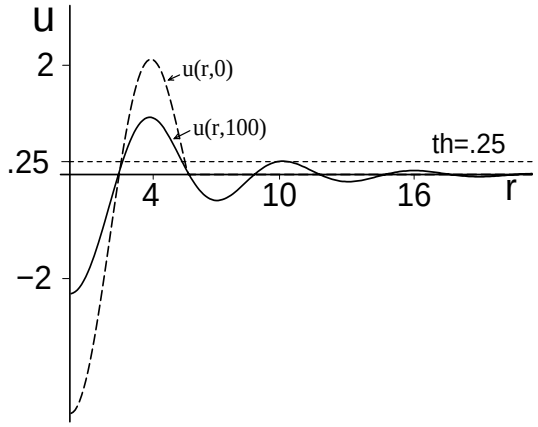


Figure 3.9. *The dashed curve shows the initial condition for* (3.17) *(its construction is given in the text), and the solid line shows the stable stationary state,* $u(r, 100)$*. The line* $u = th$ *is also shown.*

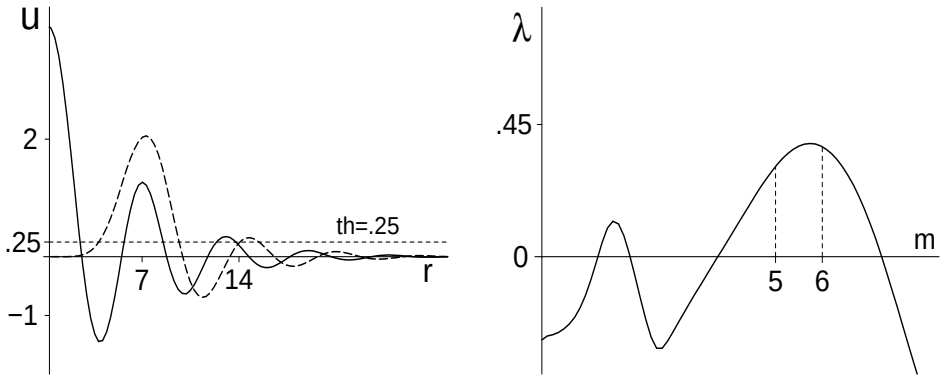


Figure 3.10. *Left: Stable solution of* (3.17) *(solid, defined as* γ_4 *in the text) corresponding to the point* Q_1 *in Figure* 3.11*, and the eigenfunction (dashed) corresponding to* $m = 6$*. Right:* λ *as a function of* m *for the solution in the left panel. Parameters are* $th = 0.25$*,* $M = 1$*,* $A = 0.4$*,* $B = 0.1$*,* $r = 0.1$*.*

annular rings. To find one such solution we again set $th = 0.25$ and solved (3.17). The resulting curve is shown (solid line) in Figure 3.10 (left). We denote it by γ_4 and denote by Γ_4 the surface produced by rotating γ_4 through a full circle about the line $r = 0$ (Figure 3.12, upper left). The initial condition was chosen so that the system (3.17) approached a solution with $u(0) > 0$, in contrast with the solutions previously studied in this section.

The solution γ_4 satisfies the ODE (3.18) with initial conditions of the form

$$u(0) > 0,\ u'(0) = 0,\ u''(0) < 0,\ u'''(0) = 0. \tag{3.22}$$

As in the previous examples, we use AUTO97 to continue the solutions of (3.18) as the

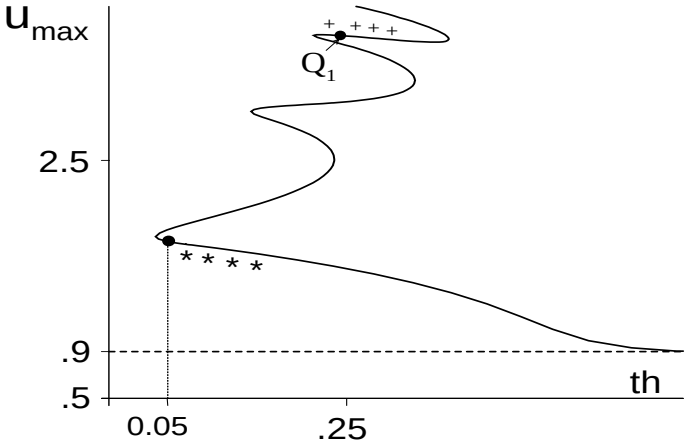


Figure 3.11. *Bifurcation diagram for solutions of* (3.18) *satisfying* $u(0) > 0$ *and* $u''(0) < 0$. *The solution at* Q_1 *is shown in Figure* 3.10 *(left). The "+" signs indicate that a particular branch is a stable solution of* (3.17). *The "∗" signs indicate stable solutions of* (3.16), *i.e.,* $\lambda < 0$ *for all* m. *At the critical value* $th \approx 0.05$, *there is a change in stability. See text.*

parameter *th* varies. Figure 3.11 gives the resultant bifurcation curve. Again, the curve has multiple folds, and solutions gain more bumps as u_{max} increases (compare with Figure 3.4). The point Q_1 in Figure 3.11 represents γ_4. We numerically solve (3.20) with $\widetilde{u} = \gamma_4$ to determine its stability. The corresponding plot of λ as a function of m is shown in Figure 3.10 (right). We see that γ_4 is a stable solution of (3.17) since $\lambda < 0$ when $m = 0$. However, Γ_4 is an unstable solution of (3.16) since $\lambda > 0$ for some $m > 0$. ($m = 6$ is the integer with the largest positive λ.) The eigenfunction corresponding to $m = 6$ is shown dashed in Figure 3.10 (left). Its largest component is concentrated near the second positive bump of γ_4, corresponding to the innermost annular ring of Γ_4. Thus, if the initial condition of (3.16) is a small random perturbation of Γ_4, we expect the innermost annular ring of the solution to break into six bumps, which will surround the central peak, resulting in a total of seven bumps. Figure 3.12 shows that this is indeed what happens.

3.4.4. Discussion. In this section, we have investigated the formation of three particular multibump patterns. These are not the only three, as there are many more solutions of (3.18), with more superthreshold oscillations before their decay to zero at $r = \infty$. These can presumably be analyzed in the same way as we have done here and will all lead to different patterns being formed.

We have presented one particular way of forming multibumps here, namely, finding circularly symmetric solutions that are unstable with respect to perturbations that break the circular symmetry and using small perturbations of these circularly symmetric solution as initial conditions for fixed parameter values. Another, perhaps more realistic, way of causing these patterns to form is to find circularly symmetric solutions of (3.16) that are stable with respect to perturbations that break the circular symmetry (i.e., have $\lambda < 0$ for all m) and

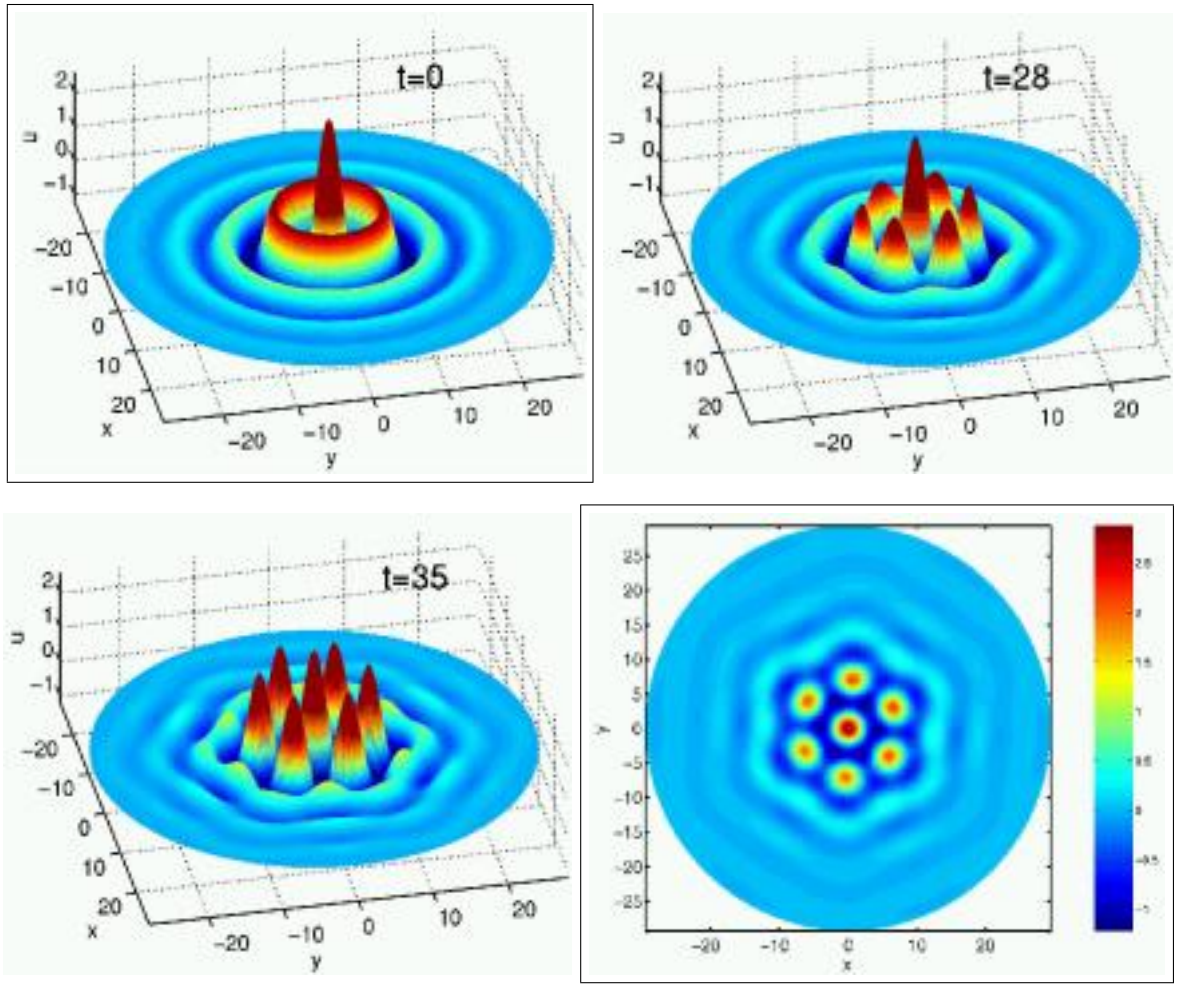


Figure 3.12. *The 7-peak solution resulting from the instability of the circularly symmetric solution whose radial profile is given by the solid curve in Figure 3.10 (left), the axially symmetric solution at* Q_1. *Top left: Initial condition (a small random perturbation of* Γ_4*). Top right:* $t = 28$. *Lower panels:* $t = 35$ *(left), and corresponding level curve diagram (right). Clicking on the top left or bottom right panels will show movies of the development of the solution from different viewpoints.*

then vary the parameters in such a way as to make the solution unstable with respect to these perturbations. This can be thought of as mimicking the change in the bulk properties of the neural tissue that would result from, for example, the action of neuromodulators (see, for example, [32].)

As an example, in Figure 3.13 we show a solution of (3.17) with $th = 0.08$ corresponding to the lowest branch in Figure 3.11. The corresponding plot of λ as a function of m is shown in Figure 3.14 (squares). We see that $\lambda < 0$ for all m, and thus the circularly symmetric solution formed by rotating the curve in Figure 3.13 through a full circle about the line $r = 0$ is a stable solution of (3.16). We now decrease th. The solution shown in Figure 3.13 changes very little, but its stability changes markedly. In Figure 3.14, we show plots of λ as a function

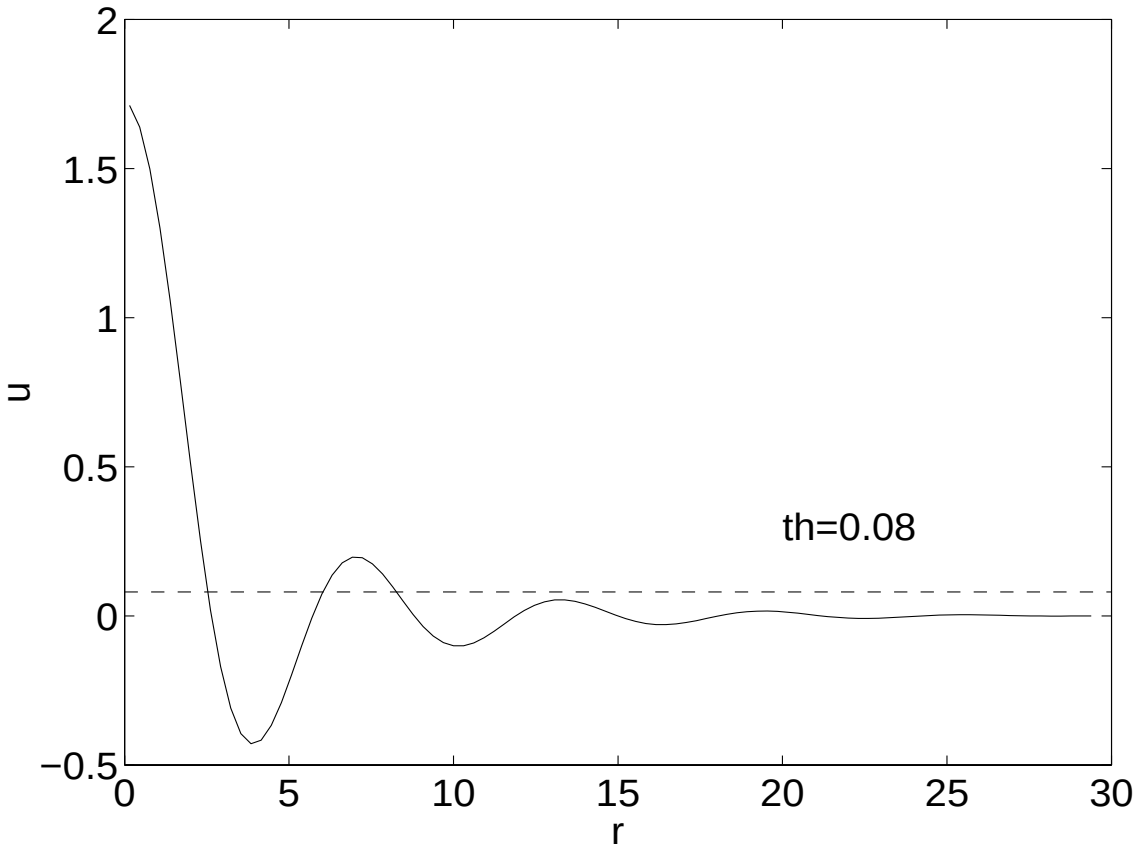


Figure 3.13. *A stable solution of* (3.17). *Parameter values are* $A = 0.4$, $B = 0.1$, $M = 1$, $r = 0.1$, $th = 0.08$.

of m for $th = 0.0536$ and $th = 0.046$. As th is decreased, the solution becomes unstable at $th \approx 0.05$ (see Figure 3.11), with $m = 6$ being the integer with the most positive value of λ. The eigenfunction corresponding to $m = 6$ has its largest peak near the bump of the solution in Figure 3.13 between $r = 5$ and $r = 10$ (not shown).

Thus we expect that if the initial condition of (3.16) is a small perturbation of the circularly symmetric surface with radial profile given by the solution in Figure 3.13 and at some point in the simulation th is reduced sufficiently from $th = 0.08$, we should see a breakup of the solution in a way similar to that shown in Figure 3.7. This is indeed what is seen (not shown).

3.5. General couplings. In this paper we have investigated one particular family of coupling functions, defined through their Fourier transform (3.14). However, a number of other types of coupling functions have been studied in the past, notably "Mexican hat"–type coupling [26, 31, 50]. (Examples are a difference of Gaussians [40, 55] and a difference of exponentials [6, 16, 33, 42].) Gaussian functions have also been used [35, 53, 54, 64] as have sinusoidal functions on a periodic domain [37, 41] and general nonnegative even functions [5, 7, 8, 27]. Specific nonneural applications whose models involve these types of coupling function include martinsitic phase transitions in steel, the behavior of diblock copolymers, and population

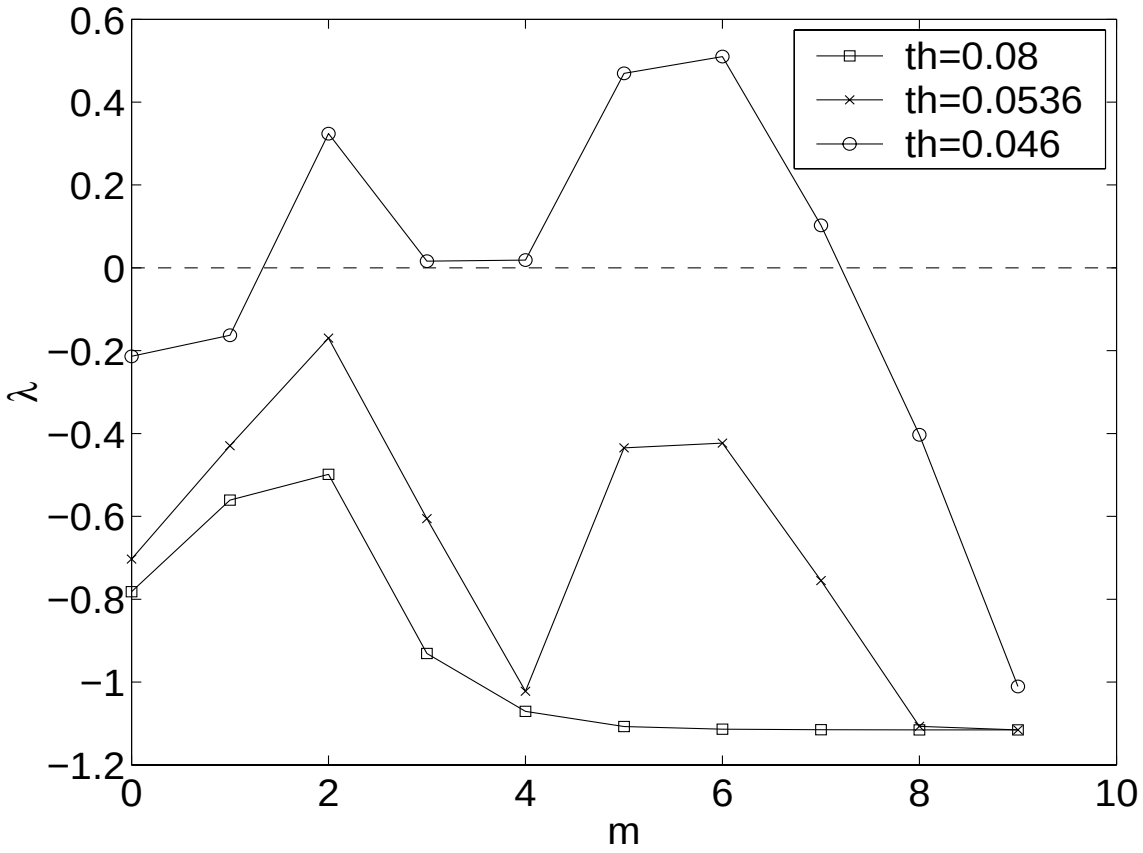


Figure 3.14. *λ as a function of m for three different values of th.*

dynamics. We now show that our technique of approximating the Fourier transform of the coupling function so that a PDE can be derived can also be applied to these types of coupling.

First consider a two-dimensional coupling function whose radial dependence is Gaussian. Since the two-dimensional Fourier transform of such a function is also circularly symmetric with a Gaussian dependence on distance in Fourier space, the problem reduces to approximating a one-dimensional Gaussian in η by a rational function of η^2. For concreteness, choose $F(\eta) = \exp(-\eta^2)$. Figure 3.15 shows the Gaussian for $0 < \eta < 4$ and two approximations to it. One approximation is a Padé approximant, a generalization of a Taylor series that matches F and as many derivatives at $\eta = 0$ as possible. This approximant is of degree $(0, 4)$, as the numerator is a polynomial of degree 0 and the denominator is of degree 4. (Note that due to the evenness of the Gaussian, only even powers of η will appear in the approximant, automatically satisfying the general condition that only even powers appear in the approximation of the Fourier transform of the coupling function.) The other function shown is the least squares fit of a function of the form $a_1/(a_2 + a_3\eta^2 + a_4\eta^4)$ to 100 evenly spaced points on the Gaussian curve. Note that these coefficients will change if the domain over which the Gaussian is considered is changed. Both of the approximations presented here are good and can be used to derive a fourth order PDE. Since both approximations have numerators of or-

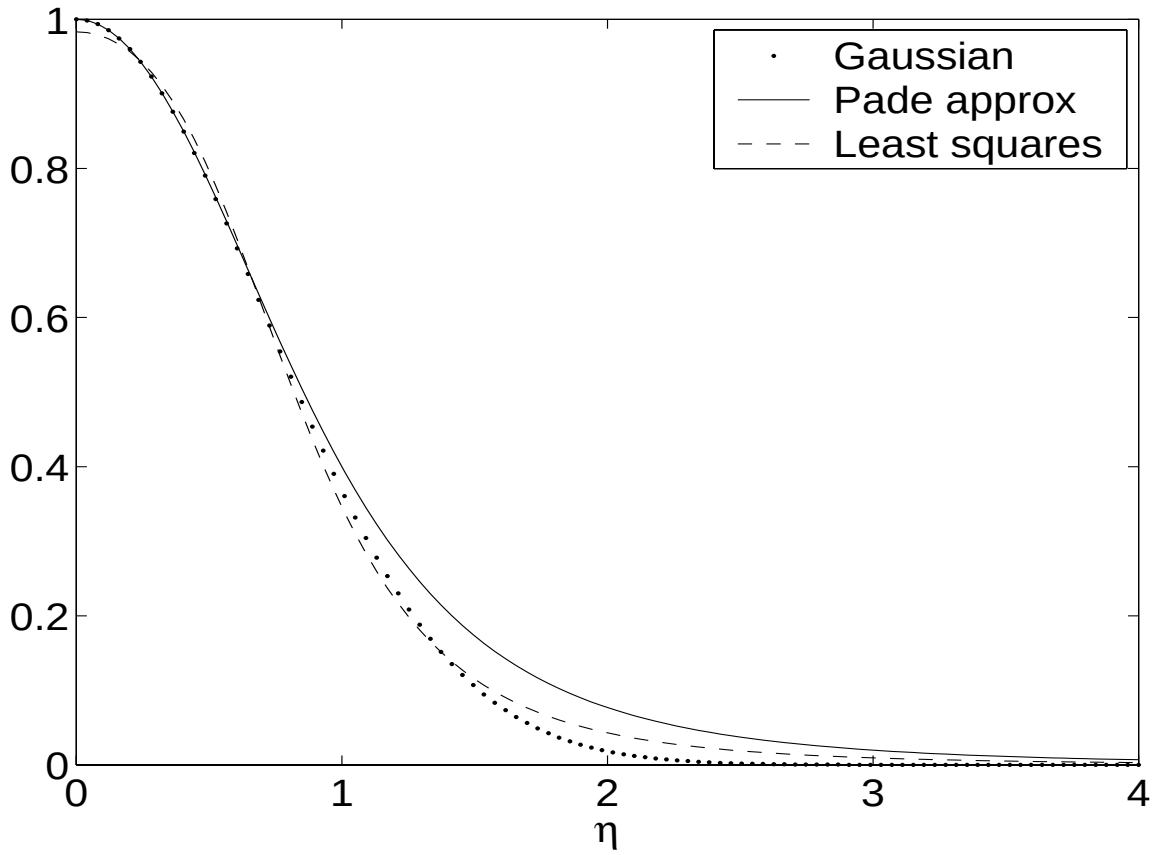


Figure 3.15. *Dotted: The Gaussian* $\exp(-\eta^2)$. *Solid: The Padé approximant* $1/(1+\eta^2+\eta^4/2)$. *Dashed: The least squares approximation* $2.7287/(2.7758+1.7504\eta^2+3.3574\eta^4)$.

der zero, the right-hand side of the corresponding equation of the form (3.16) will not contain spatial derivatives. Clearly, as either the degree of the Padé approximant or the degree of the rational function that is fit to the Fourier transform of the coupling function is increased, the approximation will become better and the order of the resulting PDE will rise.

Since taking the Fourier transform is a linear operation, coupling functions formed from the difference of Gaussians can be dealt with using the ideas just presented.

As a second example, we consider a difference of exponentials of the same form as (3.4). We need to find the two-dimensional Fourier transform of the function

$$w(r) = Ke^{-kr} - Me^{-mr}, \tag{3.23}$$

where $M < K$, $m < k$, and r is the radius. From the result in Appendix A, we have

$$F(w)(\eta) = \int_0^\infty \left[Ke^{-kr} - Me^{-mr}\right] J_0(r\eta) r\, dr, \tag{3.24}$$

where J_0 is the Bessel function of first kind of order 0. Using the results of Appendix B, we

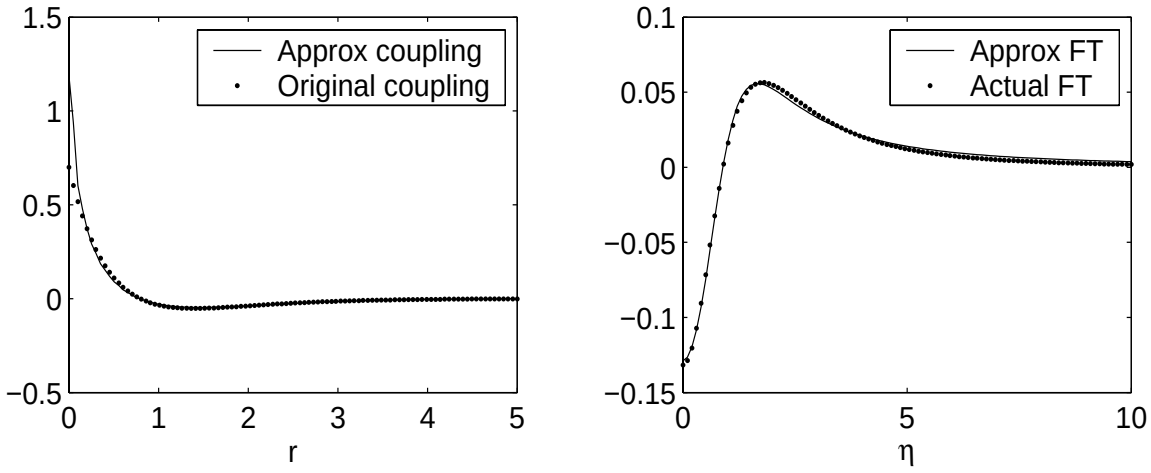


Figure 3.16. *Left: The coupling function* (3.23) *(dotted), and the inverse Fourier transform of the function shown solid in the right panel (solid). Right: The exact Fourier transform of* w*, given by* (3.25) *(dotted), and its approximation* (3.26) *(solid).*

have

$$F(w)(\eta) = \frac{Kk}{(k^2+\eta^2)^{3/2}} - \frac{Mm}{(m^2+\eta^2)^{3/2}}. \tag{3.25}$$

We set $K = 3.5$, $M = 2.8$, $k = 1.8$, and $m = 1.52$, the values used in section 3.1. In Figure 3.16 (left, dotted) we show the coupling function $w(r)$, and in the right panel (dotted) we show its exact Fourier transform (3.25). We have approximated (3.25) by an appropriate rational function of η, minimizing the least squares error for the data shown in Figure 3.16, right. The result is

$$G(\eta) = \frac{1.2636\eta^2 - 1}{7.7592 + 4.1991\eta^2 + 3.3163\eta^4}. \tag{3.26}$$

This function is shown with a solid curve in Figure 3.16, right. In the left panel of Figure 3.16, we show the coupling function resulting from taking the inverse Fourier transform of $G(\eta)$, given by (3.15) (solid line). We chose a rational function of the form (3.26) as it gave a good approximation. Using higher order polynomials (with even powers of η) in the numerator and denominator of the approximation would result in a better approximation of w by $\widehat{w}$ but would also result in higher order PDEs. Clearly, the coupling in (3.26) can be used to derive a fourth order PDE for u, and the dynamics of this equation will be equivalent to the integral equation (3.1) for the coupling shown with the solid curve in Figure 3.16 (left).

4. Summary. In this paper we have studied a class of PIDEs which have been used extensively in neuronal modeling. Our goal throughout has been to develop methods which help us understand the dynamics of multibump formation in two space dimensions.

Section 2 summarizes results for the model in one space dimension. These include the existence, multiplicity, and stability of N-bump solutions. In section 3, we focus on the two-dimensional model. This part of our investigation has led to the following results:

(i) the development of a method to approximate a PIDE with a PDE;
(ii) a description of the important properties of circularly symmetric solutions of the PDE;
(iii) the development of a method to analyze the PDE and determine the stability of circularly symmetric solutions. For unstable solutions, our methods predict the exact number of bumps that form as the unstable solution evolves.

We then applied these techniques to a specific equation and illustrated the dynamic formation of multibump solutions in three different scenarios. Finally, in section 3.5, we discuss the feasibility, both numerical and theoretical, of extending our methods to models with other couplings.

Similar results regarding the breakup of annular rings include [20], in which the stability of higher-bound states in self-focusing optical media is studied, and [46], in which the breakup of concentric rings in a reaction-diffusion system is studied. Both of these examples involve PDEs, and our results appear to be the first for integro-differential equations.

A first step in making our methods mathematically rigorous is to prove the existence of the fundamentally important solutions of the ODE problem (3.18). These describe circularly symmetric solutions of the PDE. This problem is especially challenging since (3.18) is nonautonomous and is neither reversible nor Hamiltonian. One approach is to cast (3.18) as a two-dimensional shooting problem. Here the two free parameters are the values of $u(0)$ and $u''(0)$. Note that while the relationship between homoclinic orbits and spatially localized patterns in one dimension is well known [14, 38, 43], we use such orbits here to find patterns in two spatial dimensions.

Another issue to be addressed is the correspondence between solutions of the PDE (3.16) and the integral equation (3.1), when the coupling function is given by (3.15). Formally, the equations are equivalent, but it remains to be proven that solutions of one are also solutions of the other, and if so, whether stability of a solution of one equation implies stability of that solution from the point of view of the other equation. We have not attempted to numerically solve the integral equation (3.1).

There are several ways to extend the techniques developed in this paper. A more general extension could involve combining the methods introduced here with the ideas of Bressloff [10] regarding pattern formation on inhomogeneous domains. It would also be interesting to see if the results found here could be extended to a two layer system using one population of excitatory neurons and one of inhibitory neurons with appropriate nonnegative coupling weights [25, 35, 54].

We have concentrated only on the instability of circularly symmetric bumps with respect to perturbations that break that symmetry. There are many other pattern-forming mechanisms that can potentially be studied using the ideas presented here. One example is spiral wave formation [4, 50], a phenomenon that cannot occur in one-dimensional domains. We have observed these patterns in a system of the form (3.1) with purely positive (excitatory) coupling and a simple form of adaptation like that used in [41] to prevent the whole domain from becoming active (not shown).

Another extension would be to use the ideas presented here to study a network of spiking neurons to see whether the appearance of the sorts of patterns investigated here could be predicted in such a network. The firing rate function f would have to be the appropriate function for the neurons used, but provided the neurons do not synchronize, the profiles of

firing frequency as a function of space should be the same for the rate and spiking models [40], and the techniques presented here should be applicable.

In conclusion, the main result presented here is the link between PIDEs and PDEs. The techniques introduced here enable one to apply the results for pattern formation in PDEs (of which there are many [15, 21, 23, 24, 36, 48, 49, 51]) to systems involving spatial integrals [5, 6, 7, 8, 10, 12, 13, 16, 27, 28, 47, 53, 54, 55] for which there are far fewer results, but which are of great interest.

Appendix A. Symmetry properties of the two-dimensional Fourier transform. We define the two-dimensional Fourier transform of a function $g(x, y)$ to be

$$F(g) \equiv \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i(\alpha x+\beta y)} g(x, y) dx\, dy. \tag{A.1}$$

Move to polar coordinates with $x = r\cos\theta$, $y = r\sin\theta$, $\alpha = \eta\cos\psi$, and $\beta = \eta\sin\psi$, and assume that g is a function of r only. Then (A.1) becomes

$$F(g) = \frac{1}{2\pi} \int_0^{\infty} g(r) r \left[\int_0^{2\pi} e^{-ir\eta\cos(\theta-\psi)} d\theta \right] dr. \tag{A.2}$$

It is clear that the inner integral (and therefore $F(g)$) is independent of ψ, so we set $\psi = \pi/2$, and the inner integral in (A.2) becomes

$$\int_0^{2\pi} e^{-ir\eta\sin\theta} d\theta. \tag{A.3}$$

Letting $z = e^{i\theta}$ and moving to the complex plane, we have

$$\int_0^{2\pi} e^{-ir\eta\sin\theta} d\theta = \int_C \frac{e^{-r\eta(z-1/z)/2}}{iz} dz, \tag{A.4}$$

where C is the unit circle in the complex plane. In [56, p. 161], it is shown that

$$e^{\tau(z-1/z)/2} = \sum_{N=-\infty}^{\infty} J_N(\tau) z^N, \tag{A.5}$$

where $J_N(\cdot)$ is the Bessel function of the first kind of order N. Thus, setting $\tau = -r\eta$, we have

$$\int_0^{2\pi} e^{-ir\eta\sin\theta} d\theta = \sum_{N=-\infty}^{\infty} J_0(-r\eta) \int_C \frac{z^{N-1}}{i} dz = 2\pi J_0(-r\eta), \tag{A.6}$$

and using the evenness of J_0, (A.2) becomes

$$F(g) = \int_0^{\infty} g(r) J_0(r\eta) r\, dr. \tag{A.7}$$

This is clearly a function of η $(= \sqrt{\alpha^2 + \beta^2})$ only.

Appendix B. A particular Fourier transform. We now show that

$$I \equiv \int_0^\infty e^{-kr} J_0(r\eta) r \, dr = \frac{k}{(k^2+\eta^2)^{3/2}}, \tag{B.1}$$

which, in combination with (A.7), gives (3.25). We start with the series expansion of the Bessel function

$$J_0(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{2^{2n}(n!)^2}. \tag{B.2}$$

Using this,

$$I = \sum_{n=0}^{\infty} \frac{(-1)^n \eta^{2n}}{2^{2n}(n!)^2} \int_0^\infty e^{-kr} r^{2n+1} dr = \sum_{n=0}^{\infty} \frac{(-1)^n \eta^{2n}(2n+1)!}{2^{2n}(n!)^2 k^{2n+2}}, \tag{B.3}$$

where the integral has been evaluated using the Gamma function. Now, using the notation $(a)_n = a(a+1)(a+2)\cdots(a+n-1)$ and the identity $(2n+1)! = 2^{2n} n! (3/2)_n$, we have

$$I = \sum_{n=0}^{\infty} \frac{(-1)^n \eta^{2n}(3/2)_n}{n!\, k^{2n+2}} = \frac{1}{k^2} \sum_{n=0}^{\infty} \frac{(3/2)_n}{n!} \left(\frac{-\eta^2}{k^2}\right)^n = \frac{(1+\eta^2/k^2)^{-3/2}}{k^2} = \frac{k}{(k^2+\eta^2)^{3/2}} \tag{B.4}$$

as was claimed in (B.1).

REFERENCES

[1] S. Amari, *Homogeneous nets of neuron-like elements*, Biol. Cybernet., 17 (1975), pp. 211–220.

[2] S. Amari, *Dynamics of pattern formation in lateral-inhibition type neural fields*, Biol. Cybernet., 27 (1977), pp. 77–87.

[3] S. Amari, *Mathematical Theory of Neural Networks*, Sangyo-Tosho Publishers, Tokyo, 1978.

[4] D. Barkley, *Euclidian symmetry and the dynamics of rotating spiral waves*, Phys. Rev. Lett., 72 (1994), pp. 164–167.

[5] P. Bates and F. Chen, *Spectral analysis and multidimensional stability of traveling waves for non–local Allen–Cahn equation*, J. Math. Anal. Appl., 273 (2002), pp. 45–57.

[6] P. Bates, X. Chen, and A. Chmaj, *Traveling Waves for Bistable Equations with Nonlocalities*, preprint, 2002.

[7] P. Bates, P. Fife, X. Ren, and X. Wang, *Travelling waves in a convolution model for phase transitions*, Arch. Ration. Mech. Anal., 138 (1997), pp. 105–136.

[8] P. Bates and X. Ren, *Heteroclinic orbits for a higher order phase transition problem*, European J. Appl. Math., 8 (1997), pp. 149–163.

[9] 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.

[10] P. C. Bressloff, *Traveling fronts and wave propagation failure in an inhomogeneous neural network*, Phys. D, 155 (2001), pp. 83–100.

[11] P. C. Bressloff, *Bloch waves, periodic feature maps, and cortical pattern formation*, Phys. Rev. Lett., 89 (2002), 088101.

[12] P. C. Bressloff and J. D. Cowan, *The visual cortex as a crystal*, Phys. D, 173 (2002), pp. 226–258.

[13] P. C. Bressloff, J. D. Cowan, M. Golubitsky, P. J. Thomas, and M. Wiener, *Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex*, Phil. Trans. Roy. Soc. B, 40 (2001), pp. 299–330.

[14] A. R. Champneys, *Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics*, Phys. D, 112 (1998), pp. 158–186.

[15] X. Chen and M. Kowalczyk, *Dynamics of an interior spike in the Gierer–Meinhardt system*, SIAM J. Math. Anal., 33 (2001), pp. 172–193.

[16] A. J. J. Chmaj and X. Ren, *Pattern formation in the nonlocal bistable equation*, Methods Appl. Anal., 8 (2001), pp. 369–386.

[17] C. L. Colby, J. R. Duhamel, and M. E. Goldberg, *Oculocentric spatial representation in parietal cortex*, Cereb. Cortex, 5 (1995), pp. 470–481.

[18] A. Compte, N. Brunel, P. Goldman-Rakic, and X.-J. Wang, *Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model*, Cereb. Cortex, 10 (2000), pp. 910–923.

[19] S. Coombes, G. Lord, and M. Owen, *Waves and bumps in neuronal networks with axo-dendritic synaptic interactions*, Phys. D, 178 (2003), pp. 219–241.

[20] J. M. Soto-Crespo, D. R. Heatley, E. M. Wright, and N. N. Akhmediev, *Stability of higher-bound states in a saturable self-focusing medium*, Phys. Rev. A, 44 (1991), pp. 636–644.

[21] J. Dockery and R. J. Field, *Numerical evidence of stationary and breathing concentration patterns in the Oregonator with equal diffusivities*, Phys. Rev. E, 58 (1998), pp. 823–832.

[22] E. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kutznetsov, B. Sandstede, and X. Wang, *AUTO97: Continuation and Bifurcation Software for Ordinary Differential Equations (with HOMCONT)*, Tech. report, Concordia University, Montreal, Canada, 1997.

[23] A. Doelman, T. Kaper, and P. Zegeling, *Pattern formation in the one-dimensional Gray-Scott model*, Nonlinearity, 10 (1997), pp. 523–563.

[24] A. Doelman and H. van der Ploeg, *Homoclinic stripe patterns*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 65–104.

[25] M. Enculescu and M. Bestehorn, *Activity dynamics in nonlocal interacting neural fields*, Phys. Rev. E, 67 (2003), 041904.

[26] G. B. Ermentrout, *Neural networks as spatio-temporal pattern forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.

[27] P. Fife, *Clines and material interfaces with nonlocal interaction*, in Nonlinear Problems in Applied Mathematics, T. S. Angell, L. P. Cook, R. E. Kleinman, and W. E. Olmstead, eds., SIAM, Philadelphia, 1996, pp. 134–149.

[28] P. C. Fife, *Pattern formation in gradient systems*, in Handbook for Dynamical Systems, Vol. 2, Applications, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 677–722.

[29] S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic, *Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex*, J. Neurophysiol., 61 (1989), pp. 331–349.

[30] B. Gidas, W. M. Ni, and L. Nirenberg, *Symmetry of positive solutions of elliptic equations in* R^N, in Mathematical Analysis and Applications, Part A, Adv. in Math. Suppl. Stud. 7a, Academic Press, New York, 1981, pp. 369–402.

[31] M. A. Giese, *Dynamic Neural Field Theory for Motion Perception*, Kluwer Academic Publishers, Boston, 1998.

[32] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiol., 78 (1997), pp. 1199–1211.

[33] Y. Guo, *Existence and Stability of Standing Pulses in Neural Networks*, Ph.D. thesis, University of Pittsburgh, Pittsburgh, PA, 2003.

[34] B. Gutkin, G. B. Ermentrout, and J. O'Sullivan, *Layer 3 patchy recurrent connections may determine the spatial organization of sustained activity in the primate frontal cortex*, Neurocomputing, 32–33 (2000), pp. 391–400.

[35] B. S. Gutkin, C. R. Laing, C. C. Chow, G. B. Ermentrout, and C. L. Colby, *Turning on and off with excitation: The role of spike–timing asynchrony and synchrony in sustained neural activity*, J. Comput. Neurosci., 11 (2001), pp. 121–134.

[36] J. K. Hale, L. A. Peletier, and W. C. Troy, *Exact homoclinic and heteroclinic solutions of the Gray–Scott model for autocatalysis*, SIAM J. Appl. Math., 61 (2000), pp. 102–130.
[37] D. Hansel and H. Sompolinsky, *Modeling feature selectivity in local cortical circuits*, in Methods in Neuronal Modeling, 2nd ed., C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1998.
[38] G. W. Hunt, M. A. Peletier, A. R. Champneys, P. D. Woods, M. Ahmer Wadee, C. J. Budd, and G. L. Lord, *Cellular buckling in long structures*, Nonlinear Dynam., 21 (2000), pp. 3–29.
[39] K. Kishimoto and S. Amari, *Existence and stability of local excitations in homogeneous fields*, J. Math. Biol., 7 (1979), pp. 303–318.
[40] C. R. Laing and C. C. Chow, *Stationary bumps in networks of spiking neurons*, Neural Comp., 13 (2001), pp. 1473–1494.
[41] C. R. Laing and A. Longtin, *Noise–induced stabilization of bumps in systems with long–range spatial coupling*, Phys. D, 160 (2001), pp. 149–172.
[42] C. R. Laing and W. C. Troy, *Two bump solutions of Amari–type models of working memory*, Phys. D, 178 (2003), pp. 190–218.
[43] 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.
[44] J. B. Levitt, D. A. Lewis, T. Yoshioka, and J. S. Lund, *Topography of pyramidal neuron intrinsic connections in Macaque monkey prefrontal cortex (areas* 9 *and* 46*)*, J. Comp. Neurol., 338 (1993), pp. 360–376.
[45] E. K. Miller, C. A. Erickson, and R. Desimone, *Neural mechanisms of visual working memory in prefrontal cortex of the Macaque*, J. Neurosci., 16 (1996), pp. 5154–5167.
[46] D. Morgan and T. Kaper, *Axisymmetric ring solutions of the* 2*-D Gray Scott model and their destabilization into spots*, Phys. D, submitted.
[47] C. B. Muratov, *Theory of domain patterns in systems with long-range interactions of Coulomb type*, Phys. Rev. E, 66 (2002), 066108.
[48] C. B. Muratov and V. V. Osipov, *General theory of instabilities for patterns with sharp interfaces in reaction-diffusion systems*, Phys. Rev. E, 53 (1996), pp. 3101–3116.
[49] C. B. Muratov and V. V. Osipov, *Stability of the static spike autosolitons in the Gray–Scott model*, SIAM J. Appl. Math., 62 (2002), pp. 1463–1487.
[50] J. D. Murray, *Mathematical Biology*, 2nd ed., Springer–Verlag, Berlin, 1993.
[51] Y. Nishiura and D. Ueyama, *A skeleton structure for self-replicating patterns*, Phys. D, 130 (1999), pp. 73–104.
[52] L. A. Peletier and W. C. Troy, *Patterns: Higher Order Models in Physics and Chemistry*, Birkhäuser Boston, Boston, 2001.
[53] 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.
[54] D. J. Pinto and G. 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.
[55] S. Ruuth, B. Merriman, and S. Osher, *Convolution generated motion as a link between cellular automata and continuum pattern dynamics*, J. Comput. Phys., 151 (1999), pp. 836–861.
[56] M. R. Spiegel, *Complex Variables with an Introduction to Conformal Mapping*, Schaum's Outline Series, McGraw-Hill, New York, 1998.
[57] J. Serrin, *A symmetry problem in potential theory*, Arch. Ration. Mech. Anal., 43 (1971), pp. 304–318.
[58] S. M. Stringer, T. P. Trappenberg, E. T. Rolls, and I. E. T. de Araujo, *Self-organizing continuous attractor networks and path integration: One-dimensional models of head direction cells*, Network-Comp. Neural, 13 (2002), pp. 217–242.
[59] J. G. Taylor, *Neural "bubble" dynamics in two dimensions: Foundations*, Biol. Cybernet., 80 (1999), pp. 393–409.
[60] E. Thelen, G. Schoner, C. Scheier, and L. Smith, *The dynamics of embodiment: A field theory of infant perseverative reaching*, Behavioral and Brain Sciences, 24 (2001), pp. 1–34.
[61] L. N. Trefethen, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.
[62] M. Usher, M. Stemmler, and Z. Olami, *Dynamic pattern formation leads to* $1/f$ *noise in neural populations*, Phys. Rev. Lett., 74 (1995), pp. 326–329.

[63] X. J. Wang, *Synaptic reverberation underlying mnemonic persistent activity*, Trends Neurosci., 24 (2001), pp. 455–463.
[64] T. Wennekers, *Dynamic approximation of spatio–temporal receptive fields in nonlinear neural field models*, Neural Computation, 14 (2002), pp. 1801–1825.
[65] H. Werner and T. Richter, *Circular stationary solutions in two-dimensional neural fields*, Biol. Cybernet., 85 (2001), pp. 211–217.
[66] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetic, 13 (1973) pp. 55–80.
[67] K. Zhang, *Representation of spatial orientation by the intrinsic dynamics of the head–direction cell ensemble: A theory*, J. Neurosci., 16 (1996), pp. 2112–2126.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 4, pp. 517–545

Numerical Analysis of the Novikov Problem of a Normal Metal in a Strong Magnetic Field*

Roberto De Leo†

Abstract. We present the results of our numerical exploration of the fractal structure found by S.P. Novikov in an elementary multivalued Poisson dynamical system on the 3-torus coming from the problem of the dependence of magnetoresistance on the direction of the magnetic field in a normal metal.

Key words. low-dimension topology, Poisson geometry, multivalued functions, magnetoresistance in normal metals

AMS subject classifications. 57M50, 53D17, 37E35, 65D18, 82D35

DOI. 10.1137/S1111111102406646

1. Introduction.

It is somewhat surprising that, in spite of the huge amount of literature covering symplectic and Poissonian geometry, until recent years very little effort has been made to study Hamiltonian dynamical systems with multivalued first integrals.

While no such systems arise from classical mechanics, the main source of inspiration for symplectic and Poissonian geometry, they do arise from quantum mechanical models of very concrete physical phenomena, making their study much more than some "abstract" mathematical generalization.

After extending Morse theory to multivalued functions [Nov82], Novikov became interested in such systems and rediscovered a very promising multivalued Hamiltonian dynamical system coming from solid state physics, whose geometrical study had been started almost half a century ago by a group of theoretical physicists led by Lifshitz [LAK57] and had been popular for about ten years. This study was eventually abandoned before any appropriate tool able to unveil the beautiful geometrical structure that lies within it was found.

Indeed, the topological structure of this dynamical system turned out to be much more complex than was foreseeable, and it ultimately led to the construction of a fractal on the projective plane for a class of triply periodic functions on $\mathbb{R}^3$.

In this paper, we describe the algorithm we constructed to numerically generate a picture of these fractals, and we show the results we obtained in the simplest nontrivial case, i.e., when the level surface of the Hamiltonian has genus equal to 3. An evaluation of the fractal dimension of the pictures we build allows us a rough check of the not yet proven Novikov

*Received by the editors April 30, 2002; accepted for publication (in revised form) by L. Young May 1, 2003; published electronically October 2, 2003. The author thanks Indam and the Cagliari University for financial support during his Ph.D. at the University of Maryland and the Math Department at the University of Roma "La Sapienza" for its financial support in the period January–August 2002 (project Cofin 2000 "Proprietà geometriche delle varietà reali e complesse").

http://www.siam.org/journals/siads/2-4/40664.html

†Department of Mathematics, University of Cagliari, Cagliari, Italy (deleo@unica.it) and Department of Mathematics, University of Maryland at College Park, College Park, MD (rdl@math.umd.edu) and INFN, Cagliari, Italy (roberto.deleo@ca.infn.it).

conjecture claiming that such a dimension is strictly between 1 and 2 (and, in particular, that the Lebesgue measure of the fractal is 0).

2. A short introduction to the model. It is a very basic fact of the quantum theory of solids that, in the semiclassical approximation, valence electrons can be considered as classical particles with a "periodic" momentum, called a "quasi momentum," and the interaction between these electrons and the underlying lattice of ions is totally encoded in a Hamiltonian H, called the "Fermi function," that depends only on quasi momenta [AM76].

In other words, in this model, an electron is a point (q, p) of the symplectic phase space $\mathbb{R}^3 \times \mathbb{T}^3$, and its dynamic is determined by a Hamiltonian $H = H(p)$ so that its equations of motion are the standard

$$\dot{q} = \partial H/\partial p \,, \qquad \dot{p} = 0 \,.$$

As often happens, topology comes into play as soon as a magnetic field is applied; in fact, in this case, setting as usual $-e = c = \hbar = 1$ in the equations yields the canonical

$$\dot{q} = \partial H/\partial p \,, \qquad \dot{p} = \partial H/\partial p \times B(q) = \{p, H\}_B \,.$$

The case we are interested in, widely used in the experiments of solid state physics, is the case when B is constant, in which case the second equation decouples from the first so that the system reduces to the easier equation $\dot{p} = \{p, H\}_B$ in the "reduced" phase space $\mathbb{T}^3$.

This new phase space is not symplectic, because its dimension is odd, but it has the structure of a Poisson space with Poisson bracket $\{p_a, p_b\}_B = \epsilon_{abc} B^c$, called the "magnetic bracket." This bracket is degenerate, because we are in odd dimension, and so it must have a Casimir, i.e., a first integral of algebraic nature, that is in fact the function $I(p) = p_a B^a$.

It is in the compactness of the phase space that topology makes its appearance; in fact, the coordinates p_a are angle coordinates and in particular are not globally well-defined functions on $\mathbb{T}^3$, but they are well-defined *multivalued* functions. Hence the function I is not really a function on $\mathbb{T}^3$ but just a multivalued function, and it is a globally well-defined function only on the universal covering $\mathbb{R}^3$.

So finally we have all the ingredients of our dynamical system: its phase space is the "magnetic torus" $(\mathbb{T}^3, \{,\}_B)$, and its first integrals are the (single-valued) Hamiltonian H and the multivalued Casimir I. Two integrals of motion on a three-dimensional space are enough to allow us to get the orbits by algebraic means as intersections of their level sets, so from the analytical point of view the system is trivial.

Despite the extreme simplicity of the system, no one studied even such elementary multivalued analogues of Poisson and symplectic dynamics until Novikov noticed them in 1982. And although the problem may seem trivial as the analogous single-valued system actually is, it turned out to be extremely rich and highly nontrivial from the topological point of view.

In fact, if I were single-valued, then all of its level sets would be compact, as happens with H, and so in turn all orbits would be compact. No interesting topology shows up under those conditions. When I is multivalued, however, its level sets can be open, and therefore open orbits can exist. In particular, I depends linearly on p_a, and therefore it defines a (flat) foliation of $\mathbb{T}^3$ for every nonzero magnetic field.

Let us point out that the foliation of the level sets of I is of course generated also by its differential $dI = B$, taking B as a constant 1-form, so that another way of thinking about

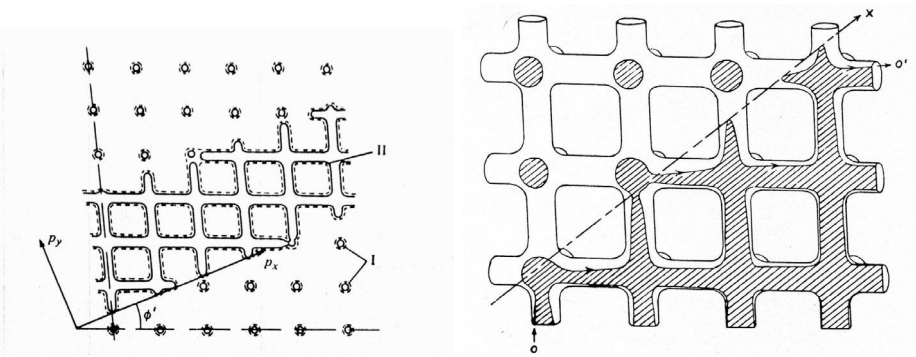


Figure 1. *These pictures, from* [LP59] *and* [Cha60]*, testify to early attempts to understand the behavior of asymptotic direction of orbits as a function of the magnetic field direction in the simpler nontrivial case, i.e., the* 3*-dimensional "prison bars" that we study numerically in this work. In both of those papers, analytical-geometrical rather than topological methods are used. The picture on the left was reprinted with permission from the Journal of Experimental and Theoretical Physics. The picture on the right was reprinted with permission from General Electric.*

the system is as the problem of understanding the topology of foliations of surfaces embedded in $\mathbb{T}^3$ induced on them by the class of 1-forms that we get through the pull-back of constant 1-forms in $\mathbb{T}^3$.

The topological property of orbits, i.e., whether they are homotopic to 0 in $\mathbb{T}^3$ (or, equivalently, whether they are closed in $\mathbb{R}^3$), has strong consequences on the physical quantities in play. As shown by Lifshitz, Azbel, and Kaganov [LAK57], the magnetoresistance behavior depends on the topology of the quasi momenta orbits; if all orbits are closed, then the magnetoresistance tensor on the plane perpendicular to the magnetic field tends to zero as the magnetic field intensity grow, while if there are open orbits with an asymptotic direction d, then it tends to a constant tensor of rank 1 whose eigendirection corresponding to the nonzero eigenvalue is exactly the direction d.

After a subtle topological analysis of the system behavior for magnetic fields close to rational [Zor84] and in "generic position" [Dyn93a, Dyn97], the following picture has been extracted by S. P. Novikov (see [NM98] for a thorough review and extra bibliography): once a Fermi function, i.e., a Hamiltonian, has been fixed on the space of directions of the magnetic field (i.e., the projective plane $\mathbb{R}\mathrm{P}^2$), it defines a fractal consisting of smooth polygons whose union contains all rational directions (i.e., the directions parallel to a direction with three integer entries), and that therefore are dense on the plane. These polygons generically have a finite number of points in common, and every one of them is labeled by an integer plane l (i.e., equivalently, a Miller index in the language of solid state physics or an integer 2-cycle in a more topological language). Moreover, to every point of the plane are associated two values of the energy $e_{1,2}$.

The meaning of these data is the following: let H and E_F be the Fermi function and Fermi energy of a metal, and suppose that we want to know the asymptotic behavior of trajectories of quasi momenta for some magnetic field B lying within a polygon labeled by $l = (i, j, k) \in H_2(\mathbb{T}^3, \mathbb{Z})$. If e_1 and e_2 are two energies associated to B, the answer is that if $e_1 \leq E_F \leq e_2$, there are open orbits, and they are a finite deformation of the straight line of

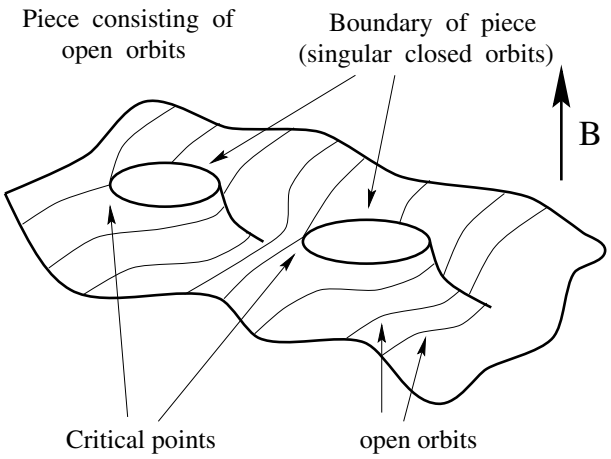


Figure 2. *A genus-2 component filled with open orbits ("warped plane"), from* [NM98]. *This figure was reprinted with permission from Uspekhi Fizicheskikh Nauk.*

direction $d = B \times l$, while if $E_F < e_1$ or $e_2 < E_F$, all orbits are closed. If B does not belong to any polygon, the picture is completely different: in this case, $e_1 = e_2 = e$ so if $E_F \neq e$, all orbits are closed, and if $E_F = e$, then open orbits exist but have no asymptotic directions.

The fact that the polygons are dense on $\mathbb{R}\mathrm{P}^2$ and that all orbits are closed for all but one value of the energy shows that this last case is nongeneric. It is known that at fixed energy the set of directions that gives rise to such behavior for orbits is of measure zero [DL99], but it is still unknown whether the union of these sets over all possible values of the energies is still a measure zero set. Novikov conjectured that the measure of this set is always zero and that its fractal dimension is between 1 and 2, and in this paper we give a numerical confirmation of this fact for the two Hamiltonians we explore.

Let us point out that when we write about the "generic" case we mean generic inside the class of those particular kinds of systems and not among the class of all dynamical systems. Indeed, if we look at this system from the point of view of dynamical systems on surfaces, the system we are studying appears highly nongeneric; every open orbit of a generic dynamical system on a surface is ergodic and therefore dense on it, while in our case the fact that open orbits are strongly asymptotic to a straight line means that their closure fills only a genus-2 component of the surface (see Figure 2) no matter how large its genus is!

Curiously enough, loosely speaking, in our particular dynamical system class we meet the "truly generic" behavior exactly in the nongeneric case, i.e., when B is such that $E_F = e_1 = e_2$. Indeed, for those directions no label is defined in general, unless they belong to the boundary of some stability zone, and open trajectories fill components of higher genus, up to the whole surface.

From these facts it is clear that knowing the geometry of the zones, their labels, and the functions e_1 and e_2 gives us a complete knowledge of the asymptotic behavior of trajectories and provides information about the fractal measure of its complement, the set of "ergodic-like" directions. Even in the most elementary cases, though it is impossible to get analytical expression for functions e_i and l, a numerical analysis of the problem is necessary.

It is important here to point out that without this picture in mind the task of describing the behavior of open orbits as a function of the magnetic field direction seems hopeless, mainly because there is no simple way to get the asymptotic direction from the analytical expression of the orbit and numerical calculations become highly unreliable after the simpler cases.

The discovery of this underlying beautiful topological structure has made it possible to write down an algorithm for the numerical exploration of the system. Indeed, from the analysis of Zorich and Dynnikov, it turns out that to study the behavior of open orbits relative to 2- and 3-irrational directions, it is enough to study it for close enough 1-rational directions, because those directions belong to the same polygon and therefore correspond to the same label, whose knowledge is enough to describe the asymptotic behavior of open orbits coming from directions belonging to the same polygon. This is a huge improvement because open orbits coming from 1-rational directions are periodic and therefore compact in $\mathbb{T}^3$; i.e., it is possible to numerically build the whole orbit.

Moreover, because the 2-cycle has only integer coordinates, an approximate computation, if performed with an error smaller than .5, can lead to an exact result.

Our numerical study aims mainly at finding the first numerical evidence of the existence of this fractal structure and evaluating its Minkowski fractal dimension [Fal97] in the simplest smooth nontrivial case, namely, the function

$$f(x) = \cos(x) + \cos(y) + \cos(z) \ .$$

We also repeat the calculations for a piecewise polynomial function with the same symmetries, which allows us to go further from the analytical point of view.

3. The algorithm. Our ingredients are a triply periodic function H (Fermi function), smooth or piecewise smooth, a noncritical value of it c (Fermi energy), the corresponding smooth level surface $M_c^2 = H^{-1}(c) \in \mathbb{T}^3$ that we will always assume connected and of nonzero genus in the following to avoid trivial cases, and a rational direction $B \in \mathbb{R}\mathrm{P}^2$ (magnetic field). We disregard any effect concerning magnetic breakdown and assume that our magnetic field is strong enough to give rise to the phenomenon; i.e., it is at least of the order of $\sim 10^4$ Gauss, but not strong enough to deform the Fermi surface, so the only free parameter left is its direction.

The goal is to get the Miller index associated with B, i.e., the homology class of the 2-tori (if they exist) on which lie the open orbits generated by B (see [NM98] for details). In other words, we must find the three integer numbers $h = (l, m, n) \in H_2(\mathbb{T}^3, \mathbb{Z})$ that represent the integer irreducible homology class of the 2-torus embedded in $\mathbb{T}^3$ on which the open orbits lie.

Let us remind the reader that all rational directions are "generic," i.e., they belong to some polygon, so that in our setting if there are open orbits, they must lie on a genus-2 component of M_c^2. In other words, we know a priori that in all cases we are going to explore numerically we will never meet any "ergodic-like" behavior.

The topological picture that corresponds to the case of open orbits for rational B (see [Dyn99] for a more thorough discussion with proofs) goes as follows. Let us consider the system from the point of view of foliations: as the system is conservative, the 1-form $\omega \in \Omega^1(M_c^2)$ induced by B can have only two kinds of critical points—saddles (that are always of "figure eight" type because all leaves are compact in $\mathbb{T}^3$) and foci. Moreover, it is easy to imagine

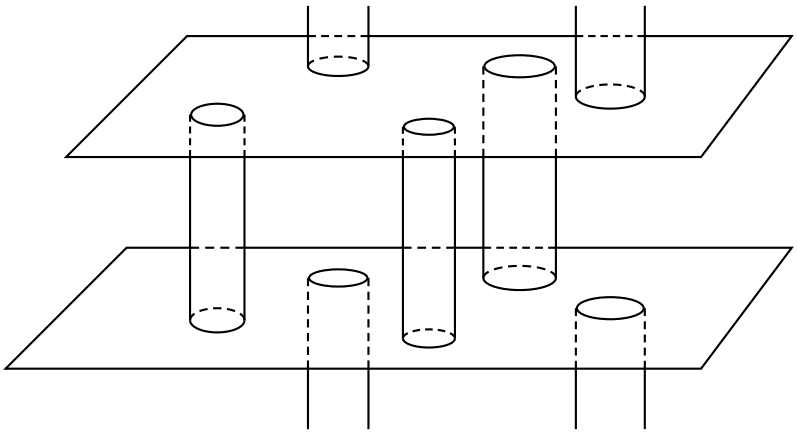


Figure 3. *An example of the splitting in* 2*-tori and cylinders of a surface of genus* 7 *under the action of a constant magnetic field.*

how every focus is associated to a saddle in such a way that one of the two singular loops of the saddle (or their union) is homotopic to zero through the very same foliation in circles that surround the focus. As we are interested only in open orbits, we can totally disregard these pairs that in some sense "cancel" each other so that at the end we are left finally only with saddles.

The saddles left can be of three kinds, depending on the number of singular loops that are closed: if both are closed, then the saddle is internal to the set of closed loops and can be disregarded; if one is open and the other is closed, then it is clear that this saddle lies at the boundary between a cylinder of closed orbits and a genus-2 component of open orbits; if both are open, then their union must be closed and again it lies at the boundary as before. We point out that we totally disregard the case of saddle connections in part because it is very difficult to find them numerically and in part because they are not generic anyway and disregarding them should not result in a big loss.

In the nicest situation, in which all saddles left are half-open, the situation will be similar to the one shown in Figure 3: if g is the genus of M_c^2, then elementary topology tells us that the number of these saddles is $2g-2$, and therefore we will have exactly $g-1$ cylinders that in turn can separate at most the same of number of genus-2 components if $g-1$ is even or one less otherwise.

Let us describe the idea of the algorithm in this setting: every 2-cycle in $H_2(\mathbb{T}^3,\mathbb{Z})$ is essentially an (oriented) integer plane, i.e., a plane orthogonal to an integer direction, so the 2-cycle homology class h of the "warped planes" is clearly in one-to-one correspondence with the sublattice $\Lambda \subset H_1(\mathbb{T}^3,\mathbb{Z})$ that it contains, and therefore to find h it is enough to evaluate the homology class of any two nontrivial independent loops lying on the warped plane. To spot the homology class of such loops, it is enough to look for those 1-cycles of M_c^2 that have zero intersection number with the ones foliating the cylinders and finally send them in $\mathbb{T}^3$ with i_*.

So these are the tasks the we must perform: spot all critical points, find those such that at least one of their singular loops is open, get the homology class (in M_c^2) of the corresponding cylinders, find the symplectic orthogonal to this (i.e., the set of classes with zero intersection with those), and send them in $\mathbb{T}^3$ with i_*. The result of all this will be a pair of independent irreducible loops in $\mathbb{T}^3$ that span a two-dimensional lattice contained in a unique (modulo sign) 2-cycle $h \in H_2(\mathbb{T}^3, \mathbb{Z})$ that is the result we look for.

4. The NTC library. After the previous discussion, it is clear which capabilities we expect from the software we are going to use to perform the numerical analysis. It must be able to deal with the topology of a surface, i.e., it must have the possibility of dealing with simplexes of dimension 0, 1, 2, and 3, and it must be able to perform topological operations like getting a simplicial decomposition of the level set of a function of three variables (to get the Fermi surface M_c^2), intersecting two simplicial complexes (to get the one-dimensional leaves on M_c^2), identifying closed curves in $\mathbb{T}^3$ (i.e., it must be able to deal with the periodic boundary conditions that identify $[0,1]^3$ with $\mathbb{T}^3$), evaluating intersection numbers between 2-cycles on a surface, and finding the homology class of loops in $\mathbb{T}^3$.

When we started this project, after a thorough search on the Internet we found several C++ libraries able to deal with the topology of three- and lower-dimensional objects through simplicial decompositions. None of them directly implements the specific functions we needed, so we decided to write a C++ library on top of one of the preexisting ones to implement the complex topological functions we needed and tried to make the code as reusable as possible as it seems that such a library could be useful in the future for different numerical topological problems. We called our library the Novikov torus conjecture (NTC) library.

After an accurate examination of all libraries available, we chose to use the library Visualization ToolKit (VTK) [SML98] (http://www.vtk.org/). The main reasons for our choice, apart from the fact that it is free of charge and open source, are the fast rate at which it is developing and the existence of a very active mailing list about VTK-related problems and relative solutions. Moreover, as VTK was intended primarily as a visualization tool based on the standard C library OpenGL by SGI (http://www.opengl.org/, http://www.mesa3d.org/), it easily allows us to visualize our surfaces and cycles, making the debugging process much easier.

VTK already contains all functions needed to generate isolines, isosurfaces, and intersections between them, so we just had to add two kinds of functionality for dealing with $\mathbb{T}^3$ topology and with singular slices.

The first method we implemented to deal with $\mathbb{T}^3$ is purely two-dimensional. After normalizing the period of the Hamiltonian to 1 in the three coordinate directions, to get the full picture of the intersection we restrict our sampling to an opportunely chosen parallelogram spanned by a $\mathbb{Z}^2$ basis of the two-dimensional lattice given by the intersection of $\mathbb{Z}^3$ with the plane perpendicular to B, so that we get a picture that glues nicely on the boundary as shown in Figure 4 (left). This procedure unfortunately tends to need too much RAM when the components of B get big, say, around 400, because of the increase of the area of the basic parallelogram.

Therefore, we implemented a second way that uses only the simplicial decomposition of the Fermi surface in the cube $[0,1]^3$. The critical leaf now is obtained by starting on the plane

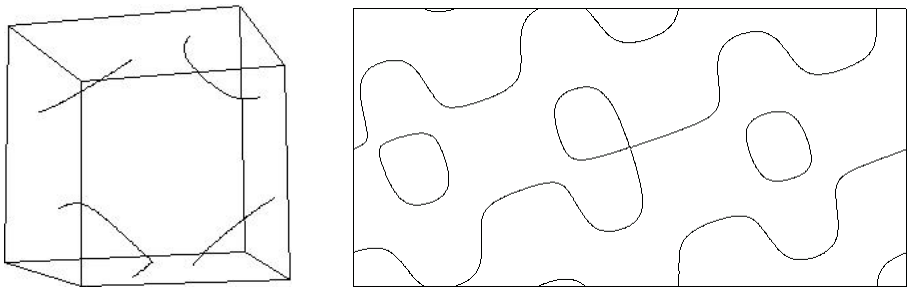


Figure 4. *There are two ways to find orbits: (left) cutting the 3D triangulation of the Fermi surface with periodic boundary conditions (only the closed loop of a saddle is shown); (right) restricting the plane perpendicular to B to a basic cell and evaluating there the isolines of the Hamiltonian restricted to the plane (here we show the complete leaf).*

passing through the critical point and then following the loop. When the trajectory reaches the boundary of the cube, the coordinates of the equivalent point are evaluated, and a new plane is taken passing through that point, as shown in Figure 4 (right). The process stops when the loop comes back for the second time to the critical point (as there are two critical loops for every critical point). Analogously, methods have developed for the leaves to be able to travel along it following the trajectory even through the points that touch the boundaries of the cube.

The singular points, on the other hand, are a problem because the numerical approximations make it impossible to cut exactly through the right point, and the section will be always either slightly above or slightly below the critical level. On the other side it seems very desirable to retrieve the singular leaf with the right topology, because to evaluate the homology class of the singular loops we need to walk over the loop starting from and ending with the singular point, so we implemented a function that for every saddle looks at a neighborhood of the singular point and performs within it a surgery to glue back the singular point to the two critical loops.

Once the critical leaf has been found, then its homology class in the Fermi surface is evaluated by counting (with the right signs) its intersections with planes that cut on the surface the loops chosen to be the basis loops for $H_1(M_c^2, \mathbb{Z})$. After all homology classes are evaluated, simple algebraic steps retrieve the homology class corresponding to B. The present implementation of the NTC library allows calculations only for surfaces of genus 3 and 4. We hope to improve the library to make it able to deal with a surface of any genus. The complete documentation for the NTC library together with the source code is available at the Internet address http://ntc.sourceforge.net/.

5. Study of the trigonometric function. The function

$$f(x, y, z) = \cos(2\pi x) + \cos(2\pi y) + \cos(2\pi z)$$

is the simplest trigonometric function that gives rise to a nontrivial (i.e., rank 3) embedding of a surface in $\mathbb{T}^3$ and the only one that had been studied so far.

The only critical values of f are ± 3 and ± 1, so all level sets $M_E = f^{-1}(E)$ are homeomorphic to spheres for $E \in (-3, -1) \cup (1, 3)$. The level set M_0 shown in Figure 5 shows that

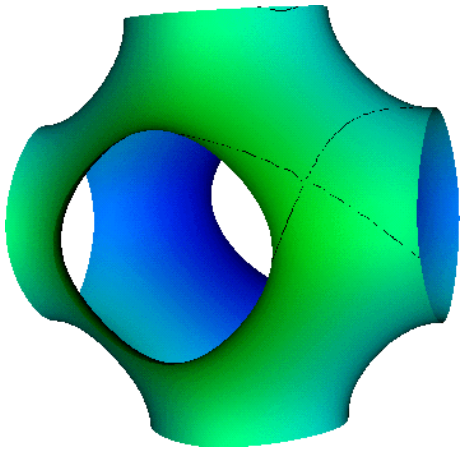


Figure 5. *The surface* $\cos(2\pi x) + \cos(2\pi y) + \cos(2\pi z) = 0$ *restricted to the unit cube.*

for $E \in (-1, 1)$ all level sets are genus-3 surfaces embedded with rank 3 in $\mathbb{T}^3$.

In particular, this means that every generic foliation of M_E induced by a magnetic field B will have at least four saddle points and that all saddles but four will be associated to some center and hence will be homotopic to 0 in the surface. These critical points, which we call "topological" as their origin is due to the topology of the surface and not to the particular embedding, are at the bases of two cylinders that separate two genus-2 components of open leaves.

Each level surface of this function is invariant under the symmetry group of the cube; this action in turn induces an action on $\mathbb{R}\mathrm{P}^2$ under which the fractal picture is invariant, so it is enough for us to analyze its structure in one of the 48 domains in which the action subdivides $\mathbb{R}\mathrm{P}^2$.

In the projective chart of $\mathbb{R}\mathrm{P}^2$ corresponding to the plane $z = 1$, one of these domains is the triangle $x \leq y \subset [0, 1]^2$, so we will refer just to the square $[0, 1]^2$ as our "phase space" from now on. On this square the picture of stability zones will be symmetric with respect to the diagonal, a fact that will be used to check the consistency of our algorithm.

Another symmetry, due to the fact that the cosine is an even function, implies that all level sets are symmetric with respect to the origin. It follows that the four topological saddles are divided into two symmetrical pairs that define one cylinder each. In fact, we can assign to every closed (in $\mathbb{R}^3$) orbit a plus or minus sign, according to whether it bounds a region where f assumes, respectively, values smaller or bigger than the one assumed on the loop. (These two different kinds of loops are called "electrons" and "holes" in physics literature.)

This sign is invariant by homotopy, so the same sign is associated to the whole cylinder and is shared by the two critical loops at the two bases. As the symmetries with respect to the center do not switch their sign, it is clear that every pair of symmetric critical points defines one of the two cylinders.

Finally, the identity $\cos(2\pi x) = -\cos[2\pi(1/2 - x)]$ induces a symmetry between different level surfaces; namely, the surface M_c is obtained from M_{-c} through a translation and a reflection with respect to the origin. As the foliation $p_a B^a = \text{const}$ is invariant by these two

operations, it is clear that the existence of open orbits at energy c implies the existence of open orbits at energy $-c$ so that the interval for which any direction gives rise to open orbits (that is closed, connected, and nonempty by [Dyn97]) has the form $[-E, E]$.

The surface M_0 hence plays a very special role, as at energy $c = 0$ every direction gives rise to open orbits and so every "stability zone" reaches here its biggest size. This means that to study the fractal on $\mathbb{R}\mathrm{P}^2$ corresponding to this function it is enough to study the level $c = 0$, while in general it would be needed to check several different energies for every direction of B to find which homology class, if any, is associated to it.

Moreover, this means that at every energy different from 0 there is no common point between boundaries of different zones, as every zone gets strictly smaller at every change of energy. In the limit for the energy that goes to -1 or 1, all zones tend to disappear as above 1 or below -1 the level surface of f is a sphere.

Let us now examine in more detail the case of 0 energy. It is easy to verify that this surface has curvature everywhere negative except in the eight points $(\pm.5, \pm.5, \pm.5)$ in which it is 0. This means that for every direction different from $(\pm1, \pm1, \pm1)$ we will have exactly four critical points, all of saddle type because of the topological constraints.

The analytical expression of the critical points for a generic $E \in (-1, 1)$ is very complicated, but it gets much simpler in the most interesting case, namely, $E = 0$. Their expression in Cartesian coordinates $(a, b) \in [0, 1]^2$ is

$$
\begin{aligned}
x_1(a,b) &= \frac{1}{2\pi}\sin^{-1}(a\alpha(a,b))\,,\\
y_1(a,b) &= \frac{1}{2\pi}\sin^{-1}(b\alpha(a,b))\,,\\
z_1(a,b) &= \begin{cases} \dfrac{1}{2} - \dfrac{1}{2\pi}\sin^{-1}(\alpha(a,b))\,, & a \le b\,,\\[2ex] \dfrac{1}{2\pi}\sin^{-1}(\alpha(a,b))\,, & a \ge b\,,\end{cases}\\
(x_2, y_2, z_2) &= \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) - (x_1, y_1, z_1)\,,\\
(x_3, y_3, z_3) &= \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right) + (x_1, y_1, z_1)\,,\\
(x_4, y_4, z_4) &= (1, 1, 1) - (x_1, y_1, z_1)\,,
\end{aligned}
$$

$$
\alpha(a,b) = \sqrt{\frac{2\sqrt{a^4 + b^4 + 1 - a^2b^2 - a^2 - b^2} - (a^2 + b^2 + 1)}{a^4 + b^4 + 1 - 2a^2b^2 - 2a^2 - 2b^2}}\,.
$$

It is easy to realize that the averaged Euler characteristic $\chi_{_B}(c) = \langle B, \gamma(c)\rangle$, where $\gamma(c) = \sum w_i x_i(c)$ is the sum over all critical points weighted by the Dynnikov index w_i equal to the "Hamiltonian" index of the critical point (as 0 of the 1-form—see [Dyn97]) times $\langle \nabla_{x_i} f, B\rangle$, is identically 0 for $c = 0$.

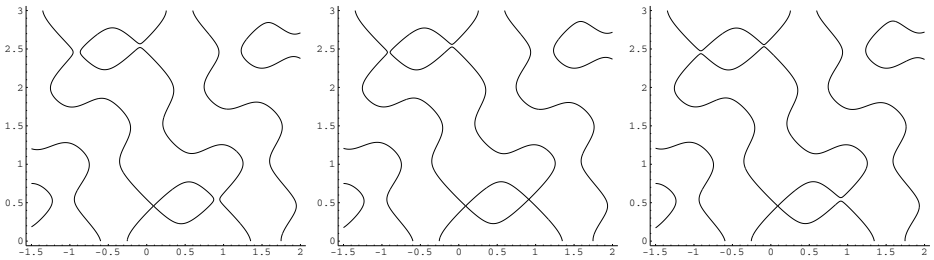


Figure 6. *The disappearance of a cylinder at the boundary of the stability zone* (0,0,1)*: On the left the cylinder has nonzero height, in the central picture the two bases collapse one over the other, and in the third one the cylinder has disappeared, substituted by a new one relative to the stability zone* (1,2,4).

This fact is also clear from the relation $\chi_B(c) = \sum h_+ - \sum h_-$; i.e., the averaged Euler characteristic is equal to the sum of the height of cylinders of "positive" closed leaves (the ones on which the gradient points to the exterior of the loop) minus the height of cylinders of "negative loops."

By the symmetry at $c = 0$ that exchanges "electrons" with "holes," i.e., positive cylinders with negative ones, it is clear that the sum is zero, while it is negative for $c > 0$ and positive for $c < 0$. This corresponds to the fact that all "ergodic" or "nongeneric" directions appear just at energy 0, as the nullity of the averaged Euler characteristic is a necessary condition for the appearance of these directions.

That there could be no "ergodic regime" for energies different from 0 was also clear from the fact that all energy intervals $[e_1(B), e_2(B)]$ for which open orbits exist are of the form $[-e, e]$; "ergodic" directions correspond to the case of length 0 of this interval, which in this case implies $e_1 = e_2 = 0$.

Now let us see what is possible to do "by hand" about stability zones at energy 0. As we have the explicit analytical expression for all critical points, we can use the following procedure: first, we make sure that a direction (a, b) is "generic," i.e., that it is inside some stability zone, for example, by looking at the plane section generated by the NTC library or by any computer algebra program like Mathematica and verifying that just one of the loops is homotopic to 0. (At energy 0 it is enough to examine just one of the critical points because of the symmetry.)

Then we choose one critical point, say, $p_1 = (x_1, y_1, z_1)$, inside the cube $[0, 1]^3$ and follow "vertically" the cylinder of closed orbits until we reach the second base point, as shown in Figure 6. As we observed before, the second base point p must be symmetrical with respect to the origin, namely, the one we called p_4, so in the covering its coordinates will be of the form $p_4 + (l, m, n)$. Equivalently, going from p_1 to p_4 inside the cylinder and coming back to p_1 through the segment that joins them inside the cube will produce a loop of homology class (l, m, n) in $\mathbb{T}^3$. As at the boundary of a zone both cylinders have height 0, i.e., the two bases belong to the same leaf, it follows that the boundary of any zone is a subset of the curves $\{\langle B, p_1 - p_4 - (l, m, n)\rangle = 0\}_{(l,m,n)\in\mathbb{Z}^3}$.

By the topological stability of curves homotopic to 0, this triple of integers depends con-

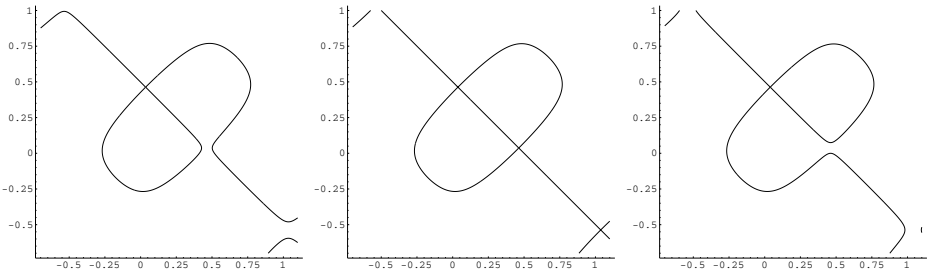


Figure 7. *A change of cylinder inside a stability zone. On the left is shown a critical leaf at the base of a cylinder; the critical point is* $p_1 \simeq (0.035, 0.463, 0.25)$. *At the opposite base lies the critical point* $p = p_4 + (0, 0, 1)$. *The middle picture shows what happens at the boundary between the two stability zones of cylinders; namely, the point* p_1 *has a saddle connection with* p_2. *The picture on the right shows the base of the new cylinder. At one base still lies the point* p_1, *but at the opposite one now lies* $p_4 + (1, 1, 0)$.

tinuously on the magnetic field, so it is locally constant. The number of different triples inside a single stability zone determines the number of sides of the zone as shown in Figure 8.

The cylinder identified by p_1 and p_4 will disappear either when its height goes to 0 or when it gets substituted by a new one; in the first case it means that we reached the boundary of the stability zone.

In the second case it happens that either two different cylinders collide and mutually exchange one of their bases or a single cylinder collides with itself and the base point is exchanged with one equivalent to it but in a different position (see Figure 7 and Tables 1(1)–1(9)).

In Figure 8 we show what happens in the case of the zone $(2, 4, 5)$. There are three different kinds of cylinder, labeled by $(-3, 3, -2)$, $(0, 0, 1)$, and $(-4, 2, -1)$, so the zone is a triangle divided inside in three subzones. At the boundary between the first and the second subzone, the change is determined by the appearance of a saddle connection between p_1 and $p_2 + (-1, 2, -1)$. At the boundary between the first and the third, we have an analogous situation between p_1 and $p_2 + (-3, 3, -1)$, and at the boundary between second and third we have instead the appearance of saddle connection of p_1 with itself, precisely with $p_1 + (2, -1, 0)$.

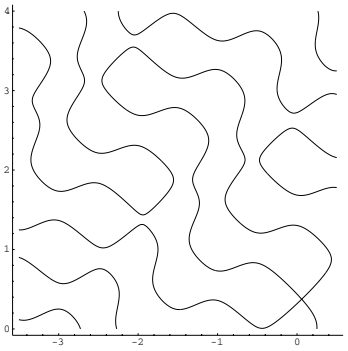
In Table 3 and also in the other pictures with smaller resolution, it is possible to recognize in many stability zones the boundaries between subzones in which p_1 has a saddle connection with itself, as in these points the 2-tori filled by open orbits have rank 1 and so these points are not included in the data and the stability zone is cut by a segment of straight line. It is easy to check that the same straight line, whose equation is $la + mb + n = 0$ for p_1 having a saddle connection with $p_1 + (l, m, n)$, cuts several (possibly infinite) zones.

These rank-1 genus-2 components survive energy changes longer than the rank-2 ones; i.e., we still find them when the rest of the zone has disappeared, but they disappear for any generic perturbation of B. A further confirmation of the accuracy of our algorithm is given by the perfect agreement between the segment found analytically for zone $(2, 4, 5)$ shown above and the one that is possible to see in Table 3.

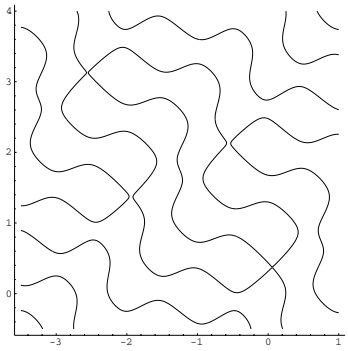
All techniques described above allow us in principle to find analytically all boundaries of stability zones and the boundaries of their subzones, even though they do nothing to help

Table 1

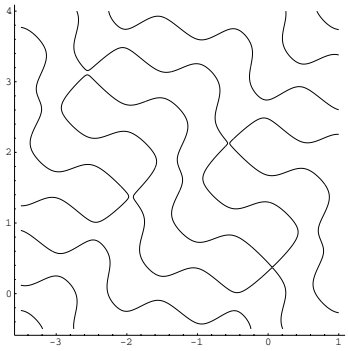
To illustrate the phenomenon of the change of cylinder type inside a stability zone, we show what happens in case of the zone $(2,4,5)$ *shown in Figure* 8. 1–3: *In the first row we move the direction of* B *from subzone* I *to subzone* III. *In the central picture we reach the boundary between the subzones, at which point* p_1 *has a saddle connection with point* $p_2+(-3,3,-2)$. 4–6: *In the second row we move* B *from subzone* III *to subzone* II. *This time on the boundary* p_1 *has a saddle connection with a copy of itself separated by a* 1*-rational vector* $(2,-1,0)$. 7–9: *In the last row we move* B *from subzone* I *to subzone* II. *At the boundary,* p_1 *has a saddle connection with* $p_2+(1,-2,1)$.



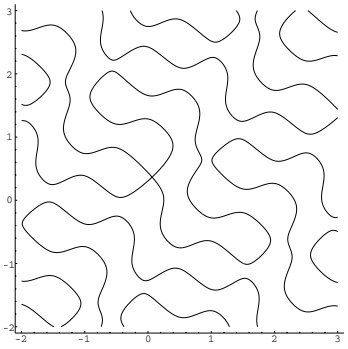
1. $B=(.37,.73,1)$



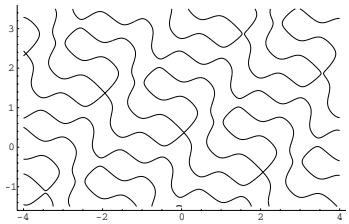
2. $B\simeq(.37,.742,1)$



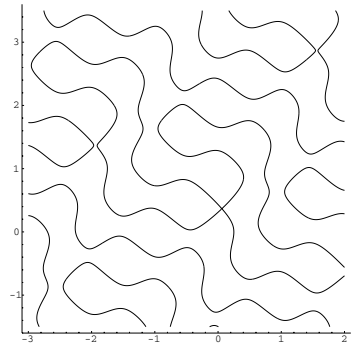
3. $B=(.37,.743,1)$



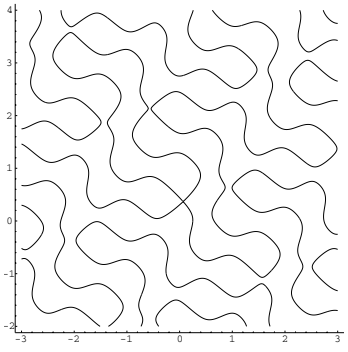
4. $B=(.41,.77,1)$



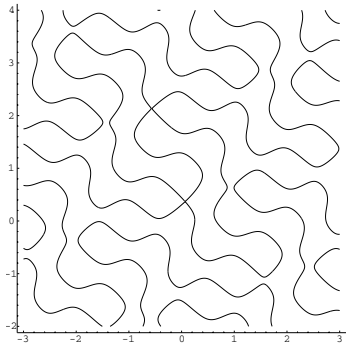
5. $B=(.385,.77,1)$



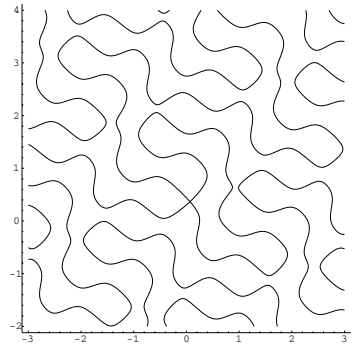
6. $B=(.384,.77,1)$



7. $B=(.42,.76,1)$



8. $B\simeq(.42,.761,1)$



9. $B=(.42,.77,1)$

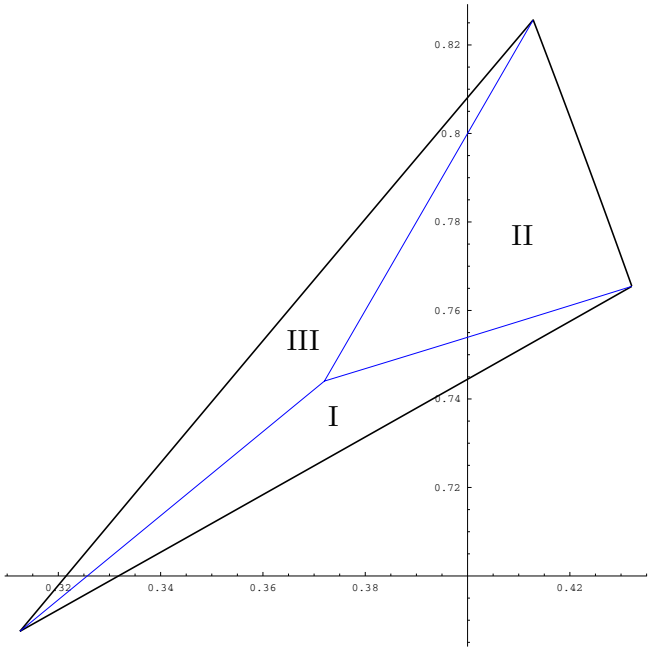


Figure 8. *The "cylinder" structure of the zone* $(2,4,5)$. *Keeping fixed the critical point of one of the bases of the cylinder, say,* p_1, *in subzone* I *in the second base we find the critical point* $p_4+(-3,3,-2)$, *in subzone* II *the critical point* $p_4+(0,0,-1)$, *and in subzone* III *the critical point* $p_4+(-4,2,-1)$. *All three cylinders are formed by closed loops that have the same homology class in* M_0, *namely,* $(2,4,5)$ *(using coordinates with respect to the natural base in* $\ker i_*$*). In Tables* 1(1)–1(9) *are shown the three kinds of cylinder corresponding to the three internal subzones and the way they transform when the magnetic field direction crosses the internal boundaries.*

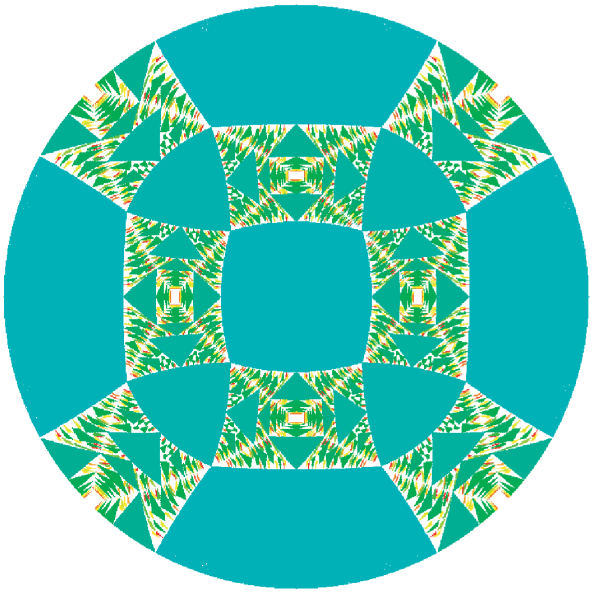
us determine which homology class is associated with them; this quantity is easily obtained through our library. The main problem is that we did not find any way to put these procedures in a simple algorithm for letting a computer do the job, so it has to be done "by hand."

Anyway, to be able to get this analytical expression does not seem to be crucial in itself. With our NTC library we can obtain a good approximation of the interior of any stability zone by sampling the square $[0,1]^2$ with step $1/N$ in both directions. In that way we will get for every point $(m/N, n/N)$, $0<m,n\le N$, the homology class of the stability zone it belongs to (if any). It is good though to have such analytical expressions as they provide a way to double check the accuracy of our algorithm by comparing the interior of the zone found with the NTC library with its analytical boundary.

We initially ran our program with resolution $N=100$ for energies $E=0,-.1,-.2,-.3,-.5,-.7,-.9$ and found the pictures we show in Tables 5–10. In Table 4 are shown the labels associated to the biggest zones together with their boundaries found analytically. The boundaries are also drawn in Table 5 to show the very good agreement with them of numerical data found using the NTC library. After trying several different machines with different operating systems, it turned out that the fastest machines available to us were Pentium II Linux machines, so we ran all our simulations on them. Every sampling with $N=100$ resolution takes around 12 hours of CPU time.

Table 2

The fractal picture in the whole projective space $\mathbb{R}P^2$ *obtained at a resolution* $N = 10^3$. *Of the* $\sim 3 \cdot 10^4$ *zones found, just the ones with at least* 10 *points* (~ 1000) *are shown. The square has been obtained by just symmetrizing the triangular picture obtained. To get this picture, we used five Linux machines with Pentium* II *CPUs for* $\sim$ *three weeks. It is possible to get the homology class corresponding to the biggest zones comparing this picture with Table* 4. *From these data has been extrapolated a fractal dimension of* $d \simeq 1.77$ *for the set of "ergodic" directions.*



In Table 3 are shown the data found with the NTC library at resolution $N = 1000$. The calculation explored just the upper triangle $b \geq a \subset [0,1]^2$; it ran $\sim$ 3 weeks on 5 Linux machines with Pentium II CPUs and found $\sim 3 \cdot 10^4$ distinct stability zones. In Table 3 we show just the 1066 zones containing at least 10 points and then we extend the picture by symmetry to the whole square. In Table 2, we extended the picture to the upper half of the sphere by symmetry to show the global pattern of the fractal.

5.1. Evaluation of the fractal dimension. One of the most standard procedures to get the fractal dimension of a set is to evaluate its "box counting" dimension [ASY96, Fal97]. To double check our results we used two different methods to get this estimate.

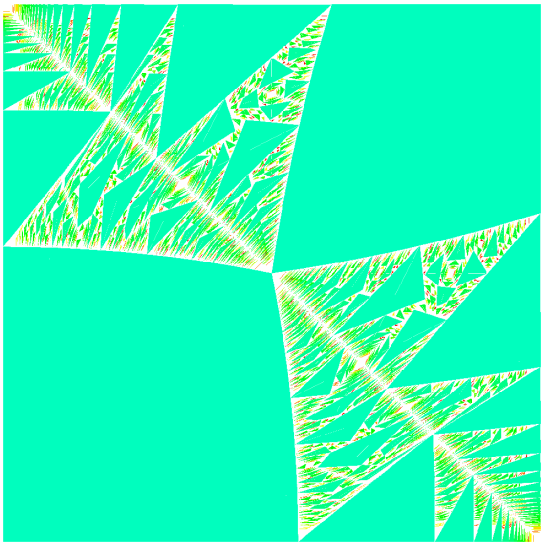
The first method comes directly from the definition; namely, we divide the square into 2^{2n} squares of area $1/2^{2n}$ and count how many of them we need to cover the fractal (i.e., the white spots in Table 3). In Figure 10 we show the data for $n = 1, \ldots, 10$.

After we discard the last two terms, which we probably cannot evaluate well enough because of the finiteness of our resolution, we find that the slope that minimizes the *rms* in a linear fitting of the above plot is $d \simeq 1.78$.

The other method we used is the following: after having fixed a number $r > 1$, we count for every n how many zones have area between r^{-n} and r^{-n-1}. Let us call this number N_n. Then as $n \to \infty$, the ratio between $\log_r(N_n)$ and n converges for fractals for which that dimension is well defined to the box counting dimension divided by the dimension of the ambient space

Table 3

The fractal picture in the square $[0,1]^2$ *obtained at a resolution* $N = 10^3$. *Of the* $\sim 3 \cdot 10^4$ *zones found, just the ones with at least* 10 *points* (~ 1000) *are shown. The square has been obtained by just symmetrizing the triangular picture obtained. To get this picture, we used five Linux machines with Pentium* II *CPUs for* $\sim$ *three weeks. It is possible to get the homology class corresponding to the biggest zones comparing this picture with Table* 4. *From these data has been extrapolated a fractal dimension of* $d \simeq 1.77$ *for the set of "ergodic" directions.*



[Fal97]. The plot in case $r = 2$ is shown in Figure 11.

In this case, the global behavior is much less linear, but it is clear that the first points have no real meaning because there the scale is still too big, and it is safe also to discard the last ones as there we are probably at a scale too small for the resolution of our picture. After discarding these boundary points, we are left with a plot which can be well approximated by a linear function with slope $\alpha \simeq .91$.

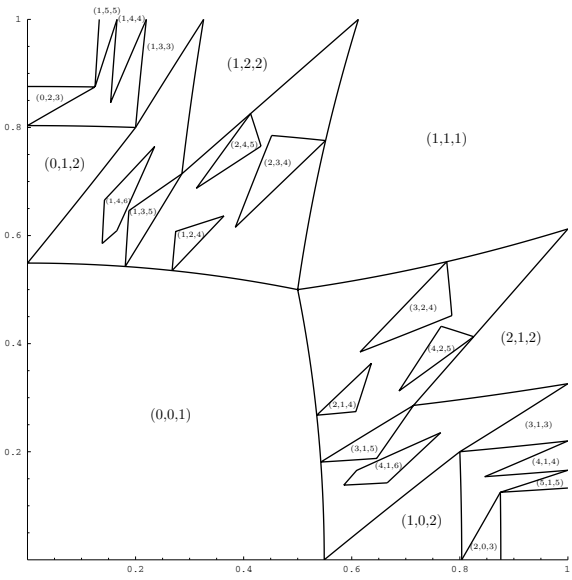
This suggests that $d \simeq 1.82$, which is in very good agreement with the previous estimate for the fractal box counting dimension of the set of "ergodic" directions.

6. Study of the piecewise quadratic function. Using degree-2 polynomials, we can build a function that has the same properties of the previous one but is much easier to deal with analytically. The function will not be any more globally smooth as the second derivatives will not glue smoothly, but it still will be globally $\mathcal{C}^1$ and piecewise quadratic.

In particular, it is possible to build a piecewise polynomial function of degree 2 that allows us to evaluate the expression of all critical points at every energy so that we will be able in principle to verify the agreement of our algorithm with every zone at energies different from zero and to find analytical expressions for topological quantities that depend on them like the averaged Euler characteristic.

Table 4

Boundaries of a few stability zones together with their homology class. All these boundaries have been obtained with the technique explained in section 5, while the associated homology class has been obtained numerically, except in the trivial case of (0,0,1). *Below are listed the zones with biggest sizes, and their area forms the data found at* $N = 1000$.



The function we used is the following:

$$f(x,y,z) = F(x) + F(y) + F(z)\ , \quad F(x) = \begin{cases} 8(2[x]-1)[x]\ , & [x] \in [0,.5]\ , \\ -8(2[x]-1)([x]-1)\ , & [x] \in [.5,1]\ , \end{cases}$$

where $[x]$ is the fractional part of x for $x \geq 0$ and F is extended to $(-\infty, 0)$ by $F(-x) = -F(x)$.

Its level sets are very similar to the ones of the previous function. Figure 9 is a picture of the level $N_0 = f^{-1}(0)$, which has the same peculiarity of the level set M_0 studied in the previous section.

As before, this function in the range of energies $(-1, 1)$ gives rise to genus-3 surfaces embedded in $\mathbb{T}^3$ with rank 3, so just four saddles of the foliation contribute to the topology of our system. All other saddles (if any) will be linked to a center and hence will be homotopic to 0 in the surface and easily eliminated from the surface through a homotopy naturally generated by the center itself.

Table 5

Map of the stability zones at energy 0 *in the square* $[0,1]^2$ *sampled at a resolution* $N = 100$. *Nearly* 700 *zones are found at this resolution; in this picture we show just the* 74 *that contain at least* 5 *points. The boundary found analytically is also shown for a few zones to show the perfect agreement with the numerical results.*

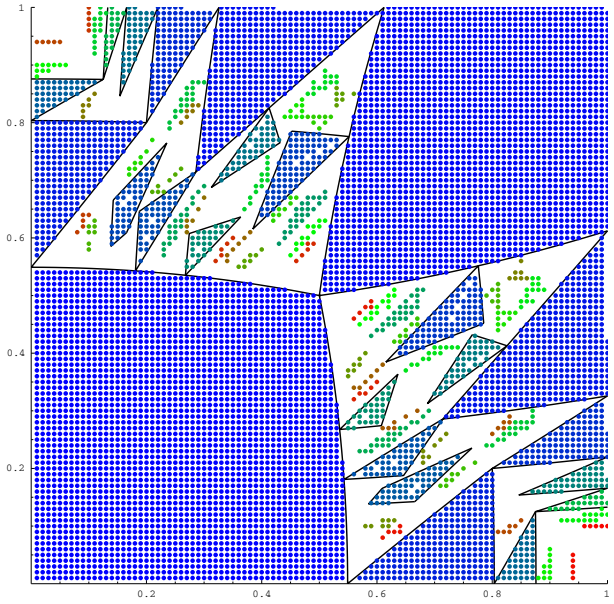


Table 6

Map of the stability zones at energy $E = -.1$ *in the square* $[0,1]^2$ *sampled at a resolution* $N = 100$. *Just* 48 *zones remain at this energy, and here we plotted just the* 34 *with more than* 1 *point.*

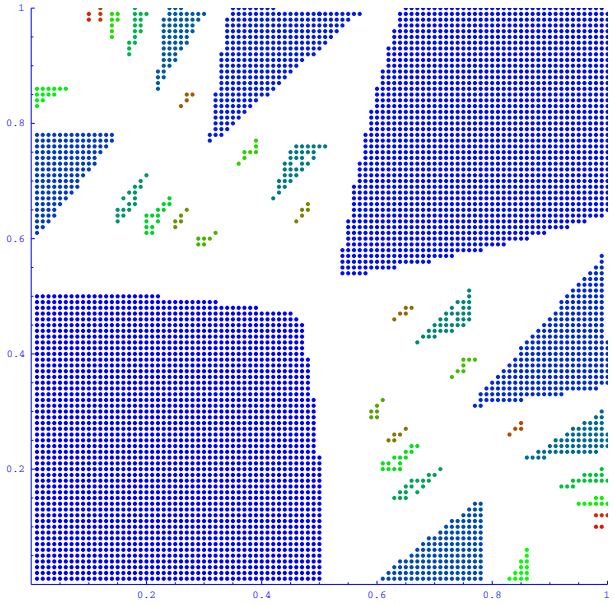


Table 7

Map of the stability zones at energy $E = -.2$ in the square $[0,1]^2$ sampled at a resolution $N = 100$. Just 12 zones remain, and we show all of them in this picture.

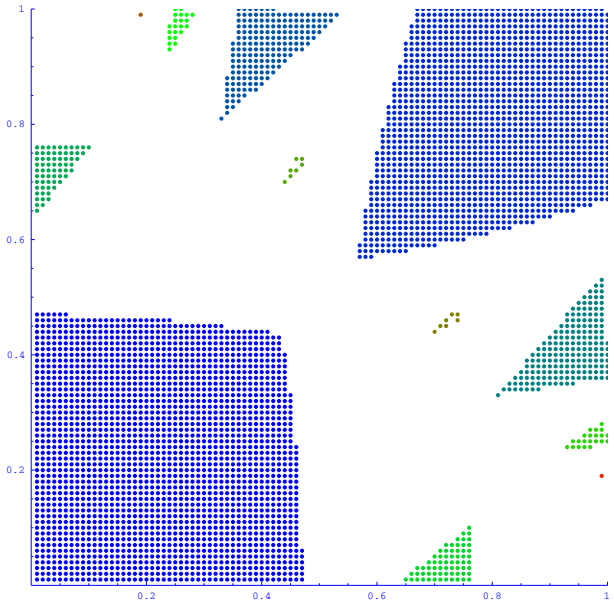


Table 8

Map of the stability zones at energy $E = -.3$ in the square $[0,1]^2$ sampled at a resolution $N = 100$. All 8 zones found are shown.

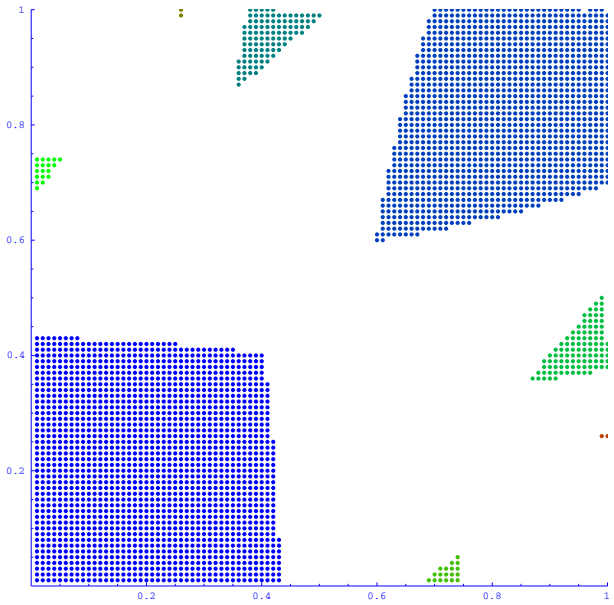


Table 9

Map of the stability zones at energy $E = -.5$ in the square $[0,1]^2$ sampled at a resolution $N = 100$. At this energy just the four zones shown remain.

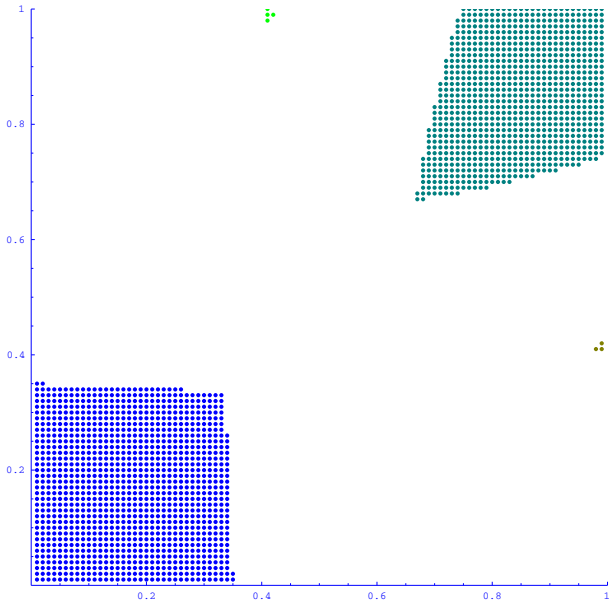


Table 10

Map of the stability zones at energy $E = -.9$ in the square $[0,1]^2$ sampled at a resolution $N = 100$. Just the two biggest zones are now visible.

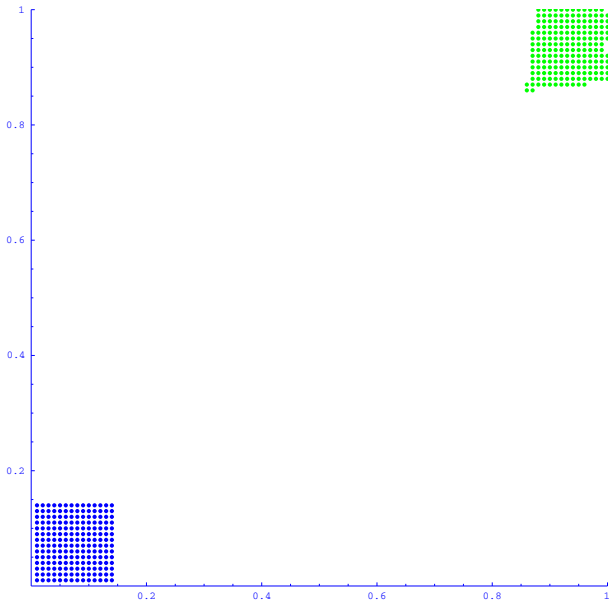


Table 11

Picture analogous to the one in Table 4 in the case of the piecewise quadratic function.

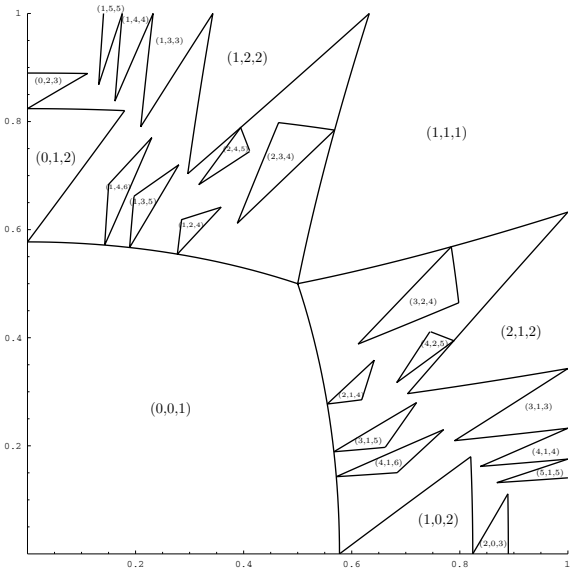


Table 12

Map of the stability zones for the piecewise quadratic function in the square $[0,1]^2$ *sampled at* $E = 0$ *with resolution* $N = 100$. *Of the* 708 *zones found, just the* 74 *with more than* 5 *points are shown. The boundary found analytically as explained in section 5 is also shown for a few zones to show the perfect agreement with the numerical results. They are very close to the boundaries of trigonometric function shown in Table 4, and the homology zones that label them are exactly the same as in the trigonometric case.*

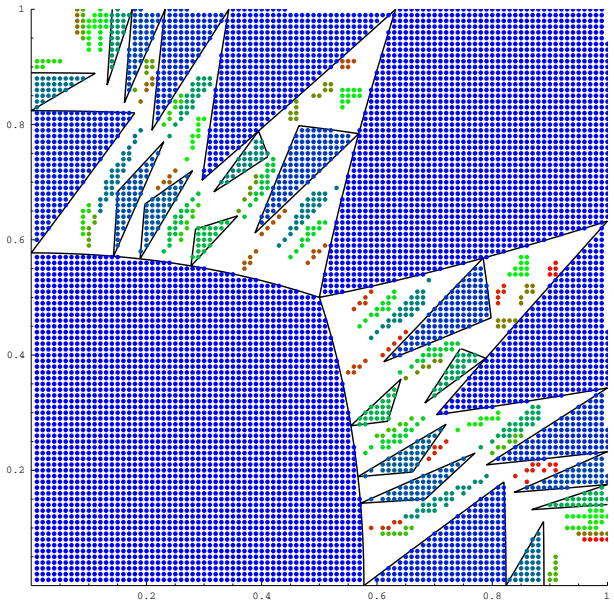


Table 13

Map of the stability zones for the piecewise quadratic function in the square $[0,1]^2$ *sampled at* $E = -.1$ *with resolution* $N = 100$. *We show here all* 42 *zones found together with the boundaries of the biggest ones.*

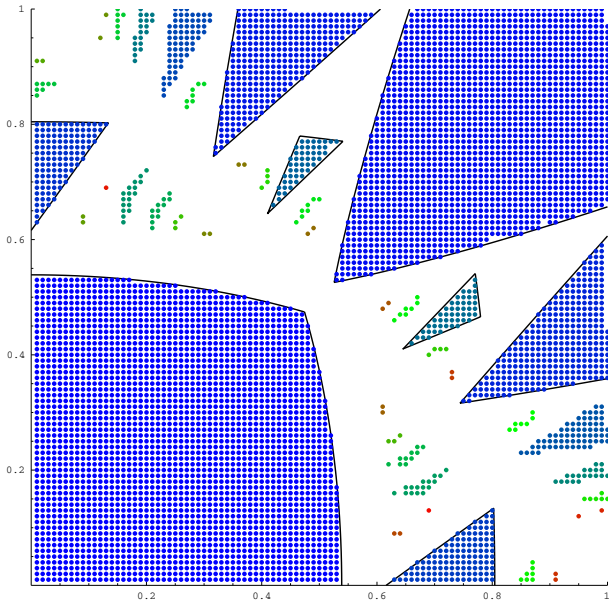
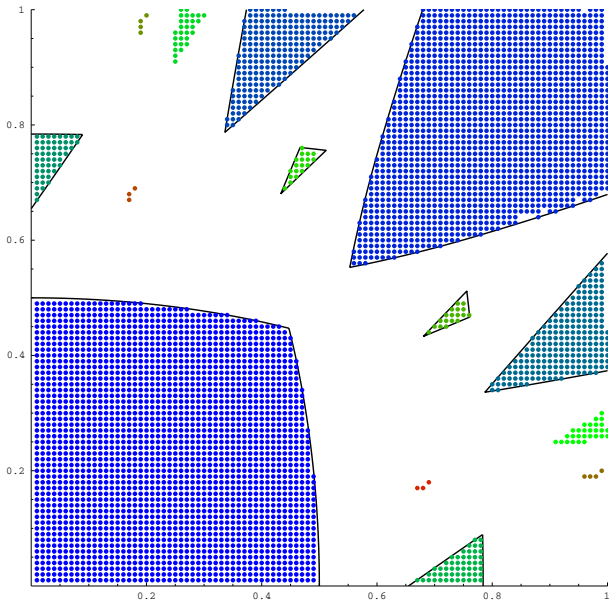


Table 14

Map of the stability zones for the piecewise quadratic function in the square $[0,1]^2$ *sampled at* $E = -.2$ *with resolution* $N = 100$. *All* 14 *zones found are shown together with boundaries of the biggest ones.*



The analytical expression for the critical points for $E \in [-1, 0]$ are the following:

$$x_1(a,b,E) = \begin{cases} \frac{a\sqrt{(1+E)}}{4\sqrt{1-a^2+b^2}}\,, & b^2-a^2 \geq E\,, \\ \frac{a\sqrt{(1-E)}}{4\sqrt{1+a^2-b^2}}\,, & b^2-a^2 \leq E\,, \end{cases} \qquad x_4 = 1 - x_1\,,$$

$$y_1(a,b,E) = \begin{cases} \frac{1}{2} - \frac{b\sqrt{(1+E)}}{4\sqrt{1-a^2+b^2}}\,, & b^2-a^2 \geq E\,, \\ \frac{1}{2} - \frac{b\sqrt{(1-E)}}{4\sqrt{1+a^2-b^2}}\,, & b^2-a^2 \leq E\,, \end{cases} \qquad y_4 = 1 - y_1\,,$$

$$z_1(a,b,E) = \begin{cases} \frac{1}{2} - \frac{\sqrt{(1+E)}}{4\sqrt{1-a^2+b^2}}\,, & b^2-a^2 \geq E\,, \\ \frac{1}{2} - \frac{\sqrt{(1-E)}}{4\sqrt{1+a^2-b^2}}\,, & a^2-b^2 \leq E\,, \end{cases} \qquad z_4 = 1 - z_1\,,$$

$$x_2(a,b,E) = \begin{cases} \frac{1}{2} - \frac{a\sqrt{(1-E)}}{4\sqrt{1-a^2+b^2}}\,, & a^2-b^2 \leq E\,, \\ \frac{1}{2} - \frac{a\sqrt{(1+E)}}{4\sqrt{1+a^2-b^2}}\,, & a^2-b^2 \geq E\,, \end{cases} \qquad x_3 = 1 - x_2\,,$$

$$y_2(a,b,E) = \begin{cases} \frac{b\sqrt{(1-E)}}{4\sqrt{1-a^2+b^2}}\,, & a^2-b^2 \leq E\,, \\ \frac{b\sqrt{(1+E)}}{4\sqrt{1+a^2-b^2}}\,, & a^2-b^2 \geq E\,, \end{cases} \qquad y_3 = 1 - y_2\,,$$

$$z_2(a,b,E) = \begin{cases} \frac{\sqrt{(1-E)}}{4\sqrt{1-a^2+b^2}}\,, & a^2-b^2 \leq E\,, \\ \frac{1}{2} - \frac{\sqrt{(1+E)}}{4\sqrt{1+a^2-b^2}}\,, & a^2-b^2 \geq E\,, \end{cases} \qquad z_3 = 1 - z_2\,.$$

The expression of boundaries of all zones in this case is very simple. For example, the boundary of the zone labeled by $(0,0,1)$ at energy E is the union of the segments of the ellipse $(1+E)a^2 + (3-E)b^2 = 1+E$ for $b \geq a$ and $(1+E)b^2 + (3-E)a^2 = 1+E$ for $b \leq a$, and the boundary of the zone corresponding to $(1,1,1)$ is the union of the segments $8a - (3-E)a^2 - (1+E)b^2 = 3-E$ for $b \geq a$ and $8b - (3-E)b^2 - (1+E)a^2 = 3-E$ for $b \leq a$.

Using the same triples of integers used for boundaries in Table 4, we have been able to find with a few modifications the corresponding zones for this function. As shown in Table 11, to these zones correspond exactly the same homology classes of the previous picture, as we expected given the similarity between the two functions.

We analyzed numerically the stability zones in the square $[0,1]^2$ for the same energies, finding the data reported in Tables 12–18. At every energy we included in the picture also the boundary of a few zones to show the very good agreement of numerical data with the analytical results.

Table 15

Map of the stability zones for the piecewise quadratic function in the square $[0,1]^2$ *sampled at* $E = -.3$ *with resolution* $N = 100$. *All* 10 *zones found are shown together with boundaries of the biggest ones.*

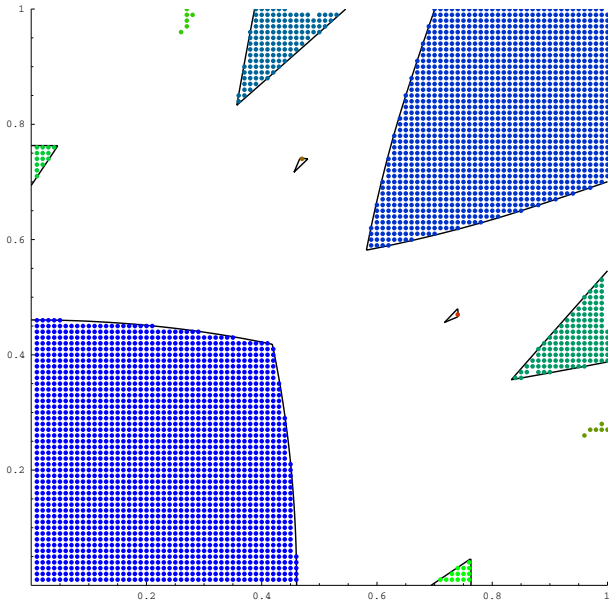


Table 16

Map of the stability zones for the piecewise quadratic function in the square $[0,1]^2$ *sampled at* $E = -.5$ *with resolution* $N = 100$. *All* 4 *zones found are shown together with their boundaries.*

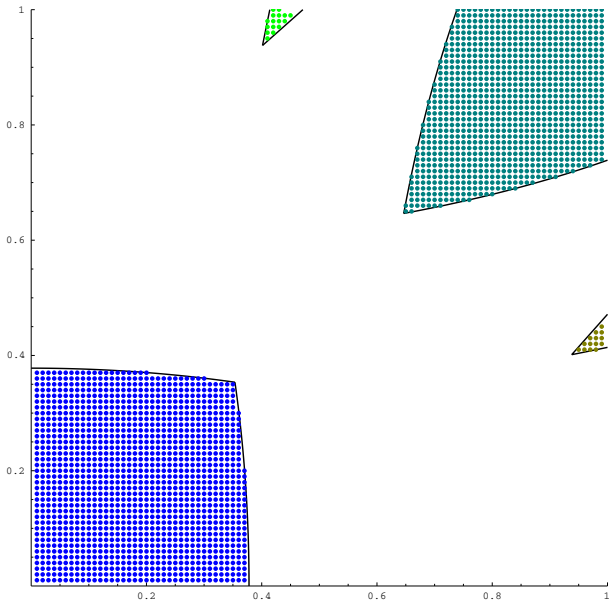


Table 17

Map of the stability zones for the piecewise quadratic function in the square $[0,1]^2$ *sampled at* $E = -.7$ *with resolution* $N = 100$. *Just the two main zones survive at this energy.*

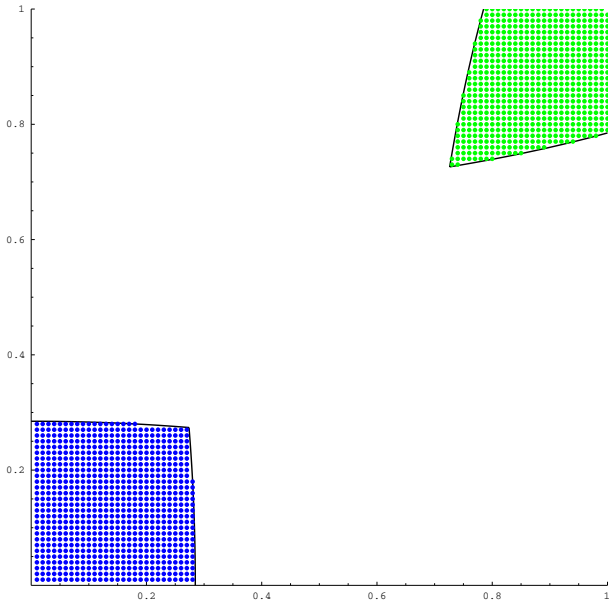


Table 18

Map of the stability zones for the piecewise quadratic function in the square $[0,1]^2$ *sampled at* $E = -.9$ *with resolution* $N = 100$. *Just the two main zones survive at this energy.*

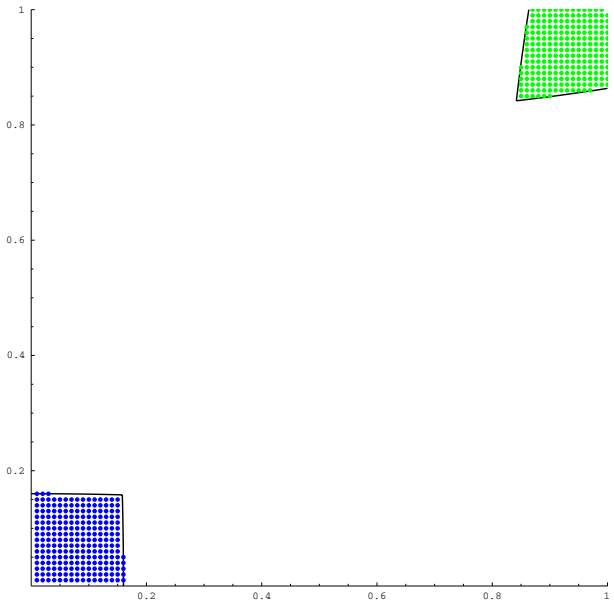
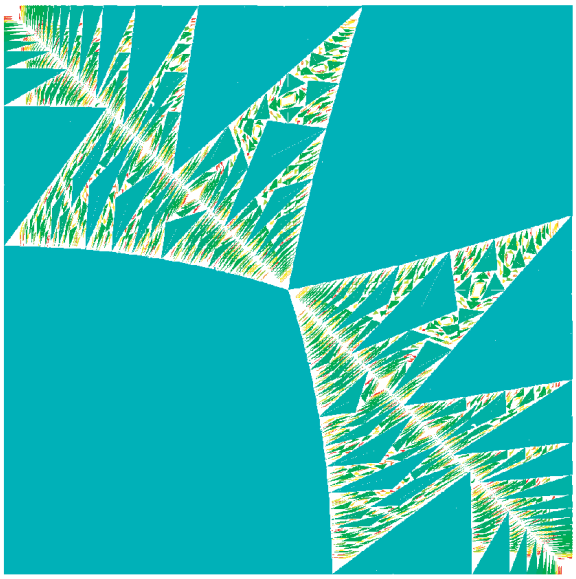


Table 19

The fractal picture for the piecewise quadratic function in the square $[0,1]^2$ *obtained at a resolution* $N = 10^3$. *Of the* $\sim 3 \cdot 10^4$ *zones found, just the ones with at least* 10 *points* (~ 1000) *are shown. The square has been obtained by just symmetrizing the triangular picture obtained.*



Using the data found at resolution 1000 (Table 19) we evaluated again the fractal dimension of the set of ergodic directions with the two methods used for the trigonometric case, finding very similar results: the box counting method gives us an estimate of $d \simeq 1.77$, and from the growth rate of the sizes of the stability zones we get $d/2 \simeq .9$. Therefore, the two different estimates are in very good agreement also in this case and suggest a fractal dimension around $d = 1.8$.

7. Conclusions. We produced a C++ library that implements all functions needed to numerically analyze the topological behavior of orbits of the Poissonian multivalued dynamical system coming from the motion of quasi electrons in a normal metal under a strong magnetic field.

We checked our code on two "toy functions" that produce genus-3 surfaces embedded in $\mathbb{T}^3$ with rank 3 and verified its correctness by comparing numerical data with the analytical data that it was possible to get for the two simple functions chosen, finding a very good agreement between the two.

Finally, we used the library to find a picture of the two fractals corresponding to each function and evaluated numerically their fractal dimension to verify that, according to Novikov's conjecture, its value lies strictly between 1 and 2.

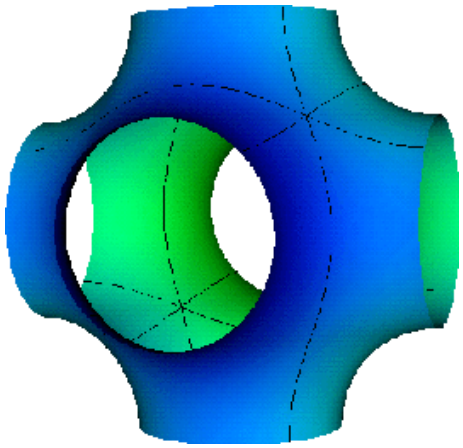


Figure 9. *The surface* $N_0 = f^{-1}(0)$ *in* $\mathbb{T}^3 \simeq [0,1]^3/\sim$. *Two copies of three basic cycles nonhomotopic to* 0 *in* $\mathbb{T}^3$ *are shown, obtained through intersection with planes passing through the center of the unit cube.*

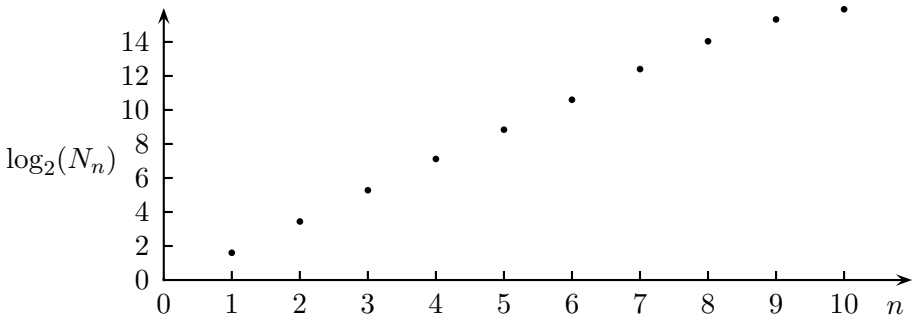


Figure 10. *Plot of the* log *in base* 2 *of number of squares needed to cover the fractal with squares of size* 2^{2n} *versus the size scale* n.

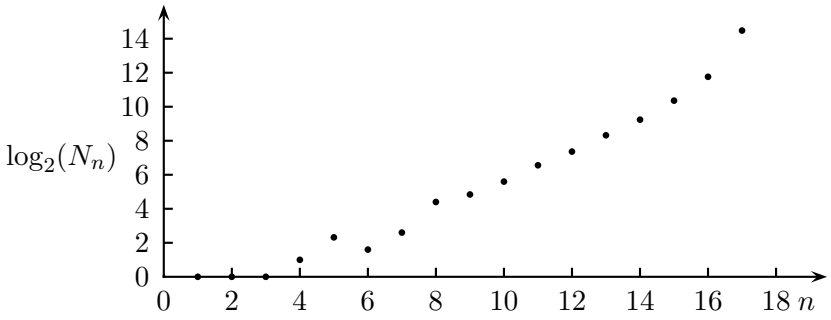


Figure 11. *Plot of the* log *in base* 2 *of number of zones of area between* 2^{-n} *and* 2^{-n-1} *versus the size scale* n.

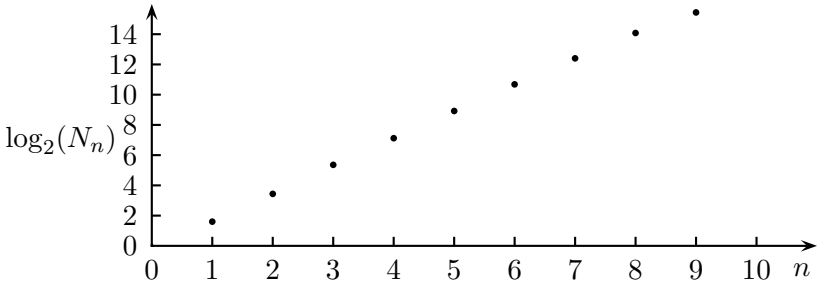


Figure 12. *Plot of the* log *in base* 2 *of number of squares needed to cover the fractal with squares of size* 2^{2n} *versus the size scale* n.

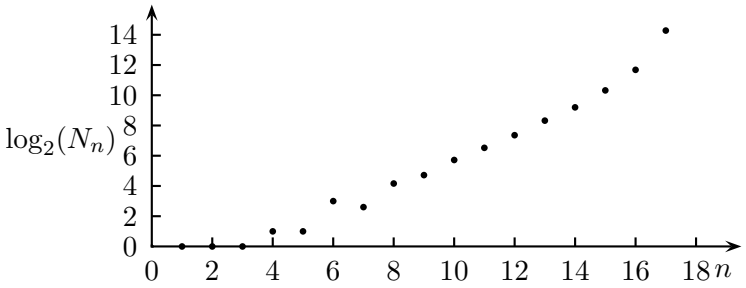


Figure 13. *Plot of the* log *in base* 2 *of number of zones of area between* 2^{-n} *and* 2^{-n-1} *versus the size scale* n.

Acknowledgments. The author gratefully thanks his doctoral advisor S.P. Novikov for his advice and for helpful discussions about the subject. The author also acknowledges many fruitful discussions with I.A. Dynnikov, A. Giacobbe, B. Hunt, D.J. Patil, and K. Snitz.

Finally, the author acknowledges the IPST (University of Maryland), the Cagliari section of INFN, the high school L.B. Alberti (Cagliari), and the CRS4 (Cagliari) for providing the several Alpha Digital and Pentium II/III Linux Machines on which all numerical simulations have been run.

REFERENCES

[AM76] N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Saunders College Publishing, Philadelphia, 1976.

[ASY96] K. T. Alligood, T. D. Sauer, and J. A. Yorke, *Chaos: An Introduction to Dynamical Systems*, Springer-Verlag, New York, 1996.

[Cha60] R. G. Chambers, *Magnetoresistance*, in The Fermi Surface, W. A. Harrison and M. B. Webb, eds., Wiley and Sons, New York, 1960.

[DL99] R. De Leo, *Existence and measure of ergodic leaves in Novikov problem on the semiclassical motion of an electron*, Uspekhi Mat. Nauk, 54 (1999), math-ph/0005031 (in Russian).

[Dyn93a] I. A. Dynnikov, *Proof of S. P. Novikov's conjecture on the semiclassical motion of an electron*, Mat. Zametki, 53 (1993), pp. 57–68.

[Dyn97] I. A. Dynnikov, *Semiclassical motion of the electron. A proof of the Novikov conjecture in general position and counterexamples*, in Solitons, Geometry, and Topology: On the Crossroad, Amer. Math. Soc. Transl. Ser. (2) 179, AMS, Providence, RI, 1997, pp. 45–73.

[Dyn99] I. A. Dynnikov, *Geometry of stability zones in the Novikov problem on the semiclassical motion of an electron*, Russian Math. Surveys, 54 (1999), pp. 21–60.

[Fal97] K. Falconer, *Techniques in Fractal Geometry*, Wiley, New York, 1997.

[LAK57] I. M. Lifschitz, M. Ya. Azbel, and M. I. Kaganov, *The theory of galvanometric effects in metals*, J. Exp. Theor. Phys., 4 (1957), pp. 41–53.

[LP59] I. M. Lifschitz and V. G. Peschanskii, *Galvanometric characteristics of metals with open Fermi surfaces* I, J. Exp. Theor. Phys., 8 (1959), pp. 875–883.

[Nov82] S. P. Novikov, *Hamiltonian formalism and a multivalued analog of Morse theory*, Uspekhi Mat. Nauk, 37 (1982), pp. 3–49 (in Russian).

[NM98] S. P. Novikov and A. Ya. Maltsev, *Topological phenomena in normal metals*, Uspekhi Fiz. Nauk, 41 (1998), pp. 231–239, cond-mat/9709007 (in Russian).

[SML98] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization ToolKit*, Prentice–Hall, Upper Saddle River, NJ, 1998.

[Zor84] A. V. Zorich, *A problem of Novikov on the semiclassical motion of electrons in a uniform almost rational magnetic field*, Uspekhi Mat. Nauk, 39 (1984), pp. 235–236 (in Russian).

SIAM J. Applied Dynamical Systems
Vol. 2, No. 4, pp. 546–569

Computing Geodesic Level Sets on Global (Un)stable Manifolds of Vector Fields*

Bernd Krauskopf† and Hinke M. Osinga†

Abstract. Many applications give rise to dynamical systems in the form of a vector field with a phase space of moderate dimension. Examples are the Lorenz equations, mechanical and other oscillators, and models of spiking neurons. The global dynamics of such a system is organized by the stable and unstable manifolds of the saddle points, of the saddle periodic orbits, and, more generally, of all compact invariant manifolds of saddle type. Except in very special circumstances the (un)stable manifolds are global objects that cannot be found analytically but need to be computed numerically. This is a nontrivial task when the dimension of the manifold is larger than one.

In this paper we present an algorithm to compute the k-dimensional unstable manifold of an equilibrium or periodic orbit (or a more general normally hyperbolic invariant manifold) of a vector field with an n-dimensional phase space, where $1 < k < n$. Stable manifolds are computed by considering the flow for negative time. The key idea is to view the unstable manifold as a purely geometric object, hence disregarding the dynamics on the manifold, and compute it as a list of approximate geodesic level sets, which are (topological) $(k-1)$-spheres. Starting from a $(k-1)$-sphere in the linear eigenspace of the equilibrium or periodic orbit, the next geodesic level set is found in a local (and changing) coordinate system given by hyperplanes perpendicular to the last geodesic level set. In this setup the mesh points defining the approximation of the next geodesic level set can be found by solving boundary value problems. By appropriately adding or removing mesh points it is ensured that the mesh that represents the computed manifold is of a prescribed quality.

The algorithm is presently implemented to compute two-dimensional manifolds in a phase space of arbitrary dimension. In this case the geodesic level sets are topological circles and the manifold is represented as a list of bands between consecutive level sets. We use color to distinguish between consecutive bands or to indicate geodesic distance from the equilibrium or periodic orbit, and we also show how geodesic level sets change with increasing geodesic distance. This is very helpful when one wants to understand the often very complicated embeddings of two-dimensional (un)stable manifolds in phase space.

The properties and performance of our method are illustrated with several examples, including the stable manifold of the origin of the Lorenz system, a two-dimensional stable manifold in a four-dimensional phase space arising in a problem in optimal control, and a stable manifold of a periodic orbit that is a Möbius strip. Each illustration is accompanied by an animation (supplied with this paper).

Key words. global stable and unstable manifolds, numerical method, geodesic parametrization

AMS subject classifications. 37C10, 34K19, 37M20

DOI. 10.1137/030600180

1. Introduction.

Many dynamical systems arising in applications can be written as a system of ordinary differential equations. Well-known examples are the Lorenz system, Chua's

*Received by the editors May 5, 2003; accepted for publication (in revised form) by M. Golubitsky October 27, 2003; published electronically December 22, 2003.
http://www.siam.org/journals/siads/2-4/60018.html

†Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (B.Krauskopf@bristol.ac.uk, H.M.Osinga@bristol.ac.uk). Bernd Krauskopf is supported by an EPSRC Advanced Research Fellowship.

circuit, and any number of periodically forced systems; see, for example, the text books [8, 24, 30], further references therein, and also the examples in section 4. In order to understand the global dynamics of such a system it is necessary to compute its equilibria, periodic orbits, and possibly other normally hyperbolic invariant sets, such as invariant tori. If these objects are of saddle type, then their stable and unstable manifolds are also important. These global manifolds organize the dynamics: stable manifolds (of codimension one) may form boundaries of basins of attraction, and intersections of stable and unstable manifolds are responsible for complicated dynamics and chaos. For example, the sensitivity of the system to perturbations depends largely on the size of the basins of attraction. Hence, the knowledge of how the global stable manifolds that are basin boundaries are embedded in phase space provides information that cannot be obtained from local inspection alone. These global objects can generally be found by numerical methods only.

To fix notation we consider a vector field in its general form

$$\frac{dx}{dt} = f(x), \tag{1.1}$$

where x is from an n-dimensional phase space, which we take to be $\mathbb{R}^n$ for simplicity, and $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is sufficiently smooth. This means in particular that a stable manifold of (1.1) can be computed as an unstable manifold by reversing time. Note that we consider all possible parameters of the vector field to be constant. In order to see what happens to the manifolds under variation of parameters, one will have to compute them for individual fixed values of the parameters.

To keep this introduction simple, we explain the basic idea with the example of an unstable manifold of a saddle point $x_0 \in \mathbb{R}^n$. Suppose that the Jacobian $Df(x_0)$ of (1.1) has $1 < k < n$ eigenvalues with positive real parts and $(n-k)$ eigenvalues with negative real parts (counted with multiplicity). The unstable invariant manifold theorem (see, for example, [24, 27]) guarantees that in a neighborhood of x_0 there exists the local unstable manifold $W^u_{\mathrm{loc}}(x_0)$, which is tangent to the unstable (generalized) eigenspace $E^u(x_0)$ of $Df(x_0)$ at x_0. Furthermore, $W^u_{\mathrm{loc}}(x_0)$ is as smooth as f. The local unstable manifold $W^u_{\mathrm{loc}}(x_0)$ can be globalized by applying the flow ϕ^t of (1.1) to obtain the k-dimensional (immersed) unstable manifold

$$\begin{aligned} W^u(x_0) &= \{x \in \mathbb{R}^n \mid \lim_{t\to-\infty} \phi^t(x) = x_0\} \\ &= \lim_{t\to\infty} \phi^t(W^u_{\mathrm{loc}}(x_0)). \end{aligned}$$

This automatically implies that $W^u(x_0)$ is also as smooth as f.

It is a natural idea to start close to x_0 with a small topological $(k-1)$-sphere S_δ (a topological circle for $k = 2$) approximately in $W^u_{\mathrm{loc}}(x_0)$ at distance δ from x_0 and then "grow" this sphere to obtain $W^u(x_0)$ further away from x_0. In practice, we take $S_\delta \subset E^u(x_0)$, but one could also start with a higher-order approximation of $W^u(x_0)$ close to x_0.

In the special case $k = 1$ of computing a one-dimensional manifold, the initial $(k-1)$-sphere S_δ consists of two points (approximately) on $W^u_{\mathrm{loc}}(x_0)$ at distance δ from x_0, one on each side of $W^s_{\mathrm{loc}}(x_0)$. These two points can be grown efficiently by integration, that is, by evolving them with the flow ϕ^t of (1.1). In other words, computing one-dimensional unstable manifolds simply amounts to integrating from two suitable initial conditions.

The situation becomes much more complicated for $k \geq 2$. One may be tempted to simply evolve S_δ under the flow ϕ^t of (1.1) also in this case. The idea behind this is that $W^u(x_0)$ can be represented as the family $\{\phi^t(S_\delta)\}_{t\in\mathbb{R}}$ of all t-images of S_δ, parametrized by the (integration) time t. However, S_δ deforms very rapidly under the flow ϕ^t due to the dynamics on $W^u(x_0)$. For example, it stretches out along the strong unstable direction (if present) and develops very large aspect ratios. Even when all eigenvalues have about the same strength, the initial circle S_δ generally gets distorted a lot further away from x_0 due to nonlinear effects; see the example in section 4.4. As a consequence, any initial mesh representing S_δ generally deteriorates so rapidly that simply evolving it under the flow does not result in a good mesh representation of $W^u(x_0)$.

Computing higher-dimensional stable and unstable manifolds of vector fields is quite challenging and an active field of research. A number of algorithms have been devised for this task [4, 5, 7, 9, 10, 12, 16, 17, 20, 22, 25, 26]. The method in [4, 5] is special as it computes a box covering of $W^u(x_0)$; it has been implemented for arbitrary n and is independent of k. All other algorithms are implemented specifically for the case of two-dimensional manifolds and produce a triangulation of the manifold that is built up by starting near the saddle point. These methods differ in the way they deal with the problem of mesh deterioration; see section 2 for a more detailed review.

In this paper we present an algorithm for computing a k-dimensional unstable manifold of a vector field with an n-dimensional phase space. Its formulation in terms of geodesic level sets further develops previous work in [22] for the specific case when $k = 2$ and $n = 3$. The method is now implemented for the case $k = 2$ and any n. This includes the case of two-dimensional stable and unstable manifolds of periodic orbits.

The key idea is to step completely away from evolving an existing mesh under the flow. To this end we consider $W^u(x_0)$ purely as a *geometric object*, namely, as a family $\{S_\eta\}_{\eta\in\mathbb{R}^+}$ of geodesic level sets parametrized by their *geodesic distance η along the manifold* from x_0; see section 2 for the precise definition. The manifold is grown by computing more and more (approximate) geodesic level sets on it. We prove in section 5 for the first time that this leads to a mesh whose quality can be guaranteed irrespective of the dynamics on the manifold.

Another goal of this paper is to illustrate and discuss the performance of our algorithm with several examples and to highlight its usefulness for the study of systems arising in applications. The interested reader is encouraged to look ahead to section 4 where we show

1. the two-dimensional stable manifold of the origin of the Lorenz system with parameters chosen such that the attractor is a figure-eight shaped periodic orbit,
2. the two-dimensional stable manifold in a four-dimensional phase space arising in an optimal control problem,
3. the two-dimensional stable manifold, in fact a Möbius strip, of a periodic orbit in the ζ^3-model,
4. a new concocted example that illustrates a possible geometric limitation of our method, and
5. the two-dimensional stable manifold of the origin of the Lorenz system for the standard choice of parameters, where we show how the information of the geodesic distance can be used to understand how the manifold interacts with the well-known Lorenz attractor.

All examples are accompanied by animations.

2. Background and concepts. As was already mentioned in the introduction, there are several possibilities for parametrizing a k-dimensional unstable manifold $W^u(x_0)$ of a saddle point x_0. By definition, $W^u(x_0)$ consists of orbits that can be parametrized by their intersection points with a suitable small $(k-1)$-sphere S_δ in $W^u_{\text{loc}}(x_0)$. The dual parametrization is given by the t-images of S_δ, that is,

$$W^u(x_0) = \{\phi^t(S_\delta)\}_{t\in\mathbb{R}}, \tag{2.1}$$

which has the advantage that its parameter space is one-dimensional. Because it is given by the dynamics, this parametrization has a number of nice properties. All elements of this parametrization are smooth topological $(k-1)$-spheres, just as S_δ. This property holds no matter what the manifold looks like or whether it converges to an attractor or not. (In the special case $k=2$ this means that the two-dimensional manifold $W^u(x_0)$ is a one-parameter family of smooth simple closed curves.) However, as discussed in the introduction, the problem is that the smooth topological $(k-1)$-spheres of this parametrization generally deform rapidly under the dynamics when t is increased. This is why simply computing (approximations to) $\phi^t(S_\delta)$ for a discrete set of integration times is completely impractical as a means of obtaining a mesh approximation of $W^u(x_0)$. Indeed, a strategy is needed to counteract this problem when a topological $(k-1)$-sphere is evolved under the flow.

Appropriate methods have been designed by Guckenheimer and Worfolk [10], Worfolk [34], and Johnson, Jolly, and Kevrekidis [16] and implemented for the case when $k=2$ and $n=3$ of growing topological circles to cover a two-dimensional manifold, starting with S_δ. Guckenheimer and Worfolk [10] and Worfolk [34] rescale the vector field so that the tangential component to the last circle is practically zero and the circle is grown in the radial direction by integration. The idea is to move from one approximate geodesic level set to the next with the help of the flow. This requires that the rescaled vector field points radially outward everywhere along the last circle, which is why this approach has difficulties when $Df(x_0)$ has complex conjugate eigenvalues. Also, the first circle must be chosen such that this condition is satisfied, which means that one may need to start with an ellipse rather than S_δ. Johnson, Jolly, and Kevrekidis [16] use a parametrization by arclength of the trajectories in the product of time and phase space. At each step of their computation the mesh points on the furthest circle are integrated up to a specified arclength, leading to a new circle, on which a uniform mesh is selected by interpolation between the integration points. The lack of control on the interpolation error makes it difficult to assess the accuracy of the computation, a problem that may be even more pronounced for $k \geq 3$.

A different approach is to view the manifold as a family of orbits. Doedel [7] computes two-dimensional manifolds by following orbits of, for example, prescribed arclength by continuation with the package AUTO [6]. The angle of the orbit with a reference direction near the equilibrium is free in the continuation process. The initial condition of each orbit lies again on a small circle $S_\delta \subset E^u(x_0)$. His method is very accurate and is particularly suited for the case $k=2$, where the manifold can be represented as a one-parameter family of orbits. However, it does seem to be more difficult to generalize this to $k \geq 3$. The method leads to an uneven distribution of mesh points on the manifold, because many mesh points may be needed near

the equilibrium to maintain the accuracy of the computation.

The method of Henderson [12] is similar in spirit but tries to better control the mesh by using only the local information of the orbits. His method starts with integrating a particular orbit together with higher-order manifold information. This results in a string of polyhedral patches along the orbit, called a fattened trajectory. Then the new boundary of the computed part of the manifold is found and the next fattened trajectory is added by starting from a suitable point along this boundary. The computation stops when all fattened trajectories have been computed up to a specified arclength from S_δ.

In the method by Guckenheimer and Vladimirsky [9] the unstable manifold is grown by adding new mesh points locally. Their method is fast by keeping the integration of the vector field to a minimum. It is implemented for the case $k = 2$ and adds a simplex locally to the manifold at each step. All but one point of the simplex are known mesh points, and the position of the new point is found by solving a PDE in an Eulerian framework, which formalizes the invariance condition, with an upwind method. Where the next triangle is added is determined in a clever way by the upwinding (essentially by the local direction of the vector field). This allows one to compute the manifold, for example, up to a predetermined arclength of the orbits on the manifold (which are approximated by the mesh structure). Due to the upwinding, the algorithm does not produce fattened trajectories but tends to add triangles in a way similar to a growth method.

A complementary approach is given by Dellnitz and Hohmann [4, 5]. Rather than growing the manifold, they compute an outer approximation of the manifold inside a specified compact region A. This method does not use any specific form of parametrization. They consider the time-τ map of the flow of (1.1) for some fixed $\tau > 0$ and first cover $W^u_{\rm loc}(x_0)$ with n-dimensional boxes. This local box covering is then evolved in A to obtain a box covering of the connected piece of $W^u(x_0)$ in A that contains x_0. This covering method is nonuniform and depends on the dynamics on the manifold. The implementation is independent of the dimension k of the manifold. The practical implementation of reliably detecting when the image of one box intersects another box (for example, by using test points) remains a challenge already for $n = 3$. If a priori bounds on the local growth rate of the vector field are available, a rigorous (slightly larger) covering can be computed [18].

All above algorithms (and this includes our own) can, in principle, be used for higher-dimensional manifold computations but present their own specific challenges of increased computational complexity for $k \geq 3$. Furthermore, visualizing higher-dimensional manifolds is a problem, and no images of three-dimensional (un)stable manifolds have been published.

The conceptual idea behind the algorithm presented here is to use the parametrization that is ideal from the geometrical point of view, irrespective of the dynamics given by the flow on $W^u(x_0)$. By this we mean the one-parameter parametrization whose elements are "the best possible topological $(k-1)$-spheres." These are given by the spheres that consist of points with constant geodesic distance d_g inside $W^u(x_0)$ from x_0. The geodesic distance $d_g(x, y)$ is the arclength of the shortest path in $W^u(x_0)$, called a *geodesic*, connecting x and y, that is,

$$d_g(x,y) := \min_{\gamma} \left\{ \int_\gamma ds \;\middle|\; \gamma \subset W^u(x_0) \text{ is a path connecting } x \text{ with } y \right\}. \tag{2.2}$$

This gives rise to the geodesic parametrization

$$\{S_\eta\}_{\eta>0}, \quad \text{where} \quad S_\eta := \{x \in W^u(x_0) \mid d_g(x, x_0) = \eta\} \tag{2.3}$$

mentioned before. This parametrization does not depend on the dynamics on $W^u(x_0)$ but only on its geometry.

Since $W^u(x_0)$ is a smooth manifold tangent to $E^u(x_0)$ at x_0, there exists $0 < \eta_{\max} \leq \infty$ such that $\mathcal{S}_\eta$ is a single smooth topological $(k-1)$-sphere without self-intersections for all $0 < \eta < \eta_{\max}$ [32]. Our method (see section 3) can compute $W^u(x_0)$ up to $\eta_{\max}$.

It is possible that $\eta_{\max}$ is finite. A trivial case is that the manifold converges to a regular attractor, such as an equilibrium or periodic orbit. Our implementation deals with this case by growing the manifold with different speeds in different directions, as is explained in [22].

More interesting is the situation when $\eta_{\max} < \infty$ even when the manifold does not converge to an attractor. In the case $k = 2$ this occurs when for some η_c the circle S_{η_c} self-intersects, and S_η for $\eta > \eta_c$ consists of two disjoint smooth closed curves. This means that there is a point $y \in W^u(x_0)$ with two different geodesics along which the geodesic distance η_c is realized. We do not know of any model exhibiting this phenomenon, so we constructed an example of a manifold with $\eta_{\max} < \infty$. Indeed, our algorithm stops when $\eta_{\max}$ is reached; see section 4.4 and Figure 6. However, this geometric obstruction does not occur in nonconcocted problems we examined. Moreover, it should be possible to restart the computation from the two disjoint circles; see again section 4.4.

In fact, all methods that try to grow the manifold from an initial $(k-1)$-sphere outward in (approximately) constant steps, including those in [10, 16, 34], encounter this geometric obstruction. It is the price for abandoning the parametrization $\{\phi^t(\mathcal{S}_\delta)\}_{t\in\mathbb{R}}$, whose elements are smooth topological $(k-1)$-spheres in $W^u(x_0)$, in favor of the geodesic parametrization $\{S_\eta\}_{\eta>0}$, whose elements naturally induce a regular discretization of $W^u(x_0)$ with a well-controlled error. We make use of this property of the geodesic parametrization in section 5.

3. The general algorithm.

The setup of the algorithm in [22] for the special case when $k = 2$ and $n = 3$ generalizes to arbitrary k and n. In particular, as we will explain now, it is still possible to find a new mesh point by continuation of a one-parameter family of well-posed boundary value problems.

As before, we start from an initial level set S_δ in $E^u(x_0)$ at some prescribed distance δ from x_0 and then approximate a sequence of level sets $\{S_{\eta_i}\}_{0\leq i\leq l}$ until a prespecified fixed geodesic distance D from x_0 is reached. In other words, $\eta_0 = \delta$ and $\eta_l \geq D$. Each level set S_{η_i} is approximated by a mesh M_i from which we form the simplicial complex C_i, which consists of $(k-1)$-simplices [29] (lines for a 1-sphere, triangles for a 2-sphere, etc.). Hence, C_i is a piecewise linear continuous approximation of the $(k-1)$-sphere S_{η_i} at step i. A first piece of the manifold $W^u(x_0)$ (up to a prescribed geodesic distance $D \leq \eta_l$) is approximated by the k-dimensional simplicial complex $\mathcal{C}$ that is formed from the total mesh $\mathcal{M} = \cup_{0\leq i\leq l} M_i$ in such a way that the $(k-1)$-simplices in C_i appear as faces of the k-simplices of $\mathcal{C}$.

As start data M_0 and C_0 we choose a mesh representation of S_δ. The algorithm now proceeds by adding new $(k-1)$-spheres in steps. Suppose that we have computed M_i and C_i so that we are at step i, where we want to find M_{i+1} and C_{i+1} that represent $S_{\eta_{i+1}}$ with $\eta_{i+1} = \eta_i + \Delta_i$.

3.1. Finding a new point in M_{i+1}. For every point $r \in M_i$ we define an $(n-k+1)$-dimensional hyperplane $\mathcal{F}_r$ with the property that it is "most perpendicular" to C_i at r. By this we mean that the vectors normal to $\mathcal{F}_r$ can be expressed as averages of the vectors between r and its direct neighbors in M_i. Independent of the dimension k of $W^u(x_0)$, the intersection of $W^u(x_0)$ with $\mathcal{F}_r$ locally near r is a well-defined one-dimensional curve. (Globally, this intersection may be a set of disjoint one-dimensional curves.)

Consider now the following one-parameter family, parametrized by the integration time τ, of two-point boundary value problems of finding the orbits $\{\phi^t(q_r(\tau)) \mid t \in [0,\tau]\}$ satisfying

$$\phi^0(q_r(\tau)) = q_r(\tau) \in C_i, \tag{3.1}$$

$$\phi^\tau(q_r(\tau)) = b_r(\tau) \in \mathcal{F}_r. \tag{3.2}$$

We denote the initial condition in C_i by $q_r(\tau)$ and the final point in $\mathcal{F}_r$ by $b_r(\tau)$. Then the intersection curve $W^u(x_0) \cap \mathcal{F}_r$ can be represented in good approximation (and locally near r) as the one-parameter family $b_r(\tau)$, where $b_r(0) = r$.

What we need to find is the point $b_r = b_r(\tau_r)$ defined uniquely by the property that τ_r is the smallest integration time τ for which $\| b_r(\tau) - r \| = \Delta_i$; that is, b_r is the first point on $W^u(x_0) \cap \mathcal{F}_r$ that lies at distance Δ_i from r. To find b_r we start from the trivial solution $q_r(0) = b_r(0) = r$ for $\tau = 0$, which clearly satisfies the boundary conditions (3.1) and (3.2). We then continue this solution in the continuation parameter τ while monitoring the test function

$$\Delta_i - \| b_r(\tau) - r \| . \tag{3.3}$$

When we find the first zero of (3.3) for some $\tau = \tau_r$, the continuation stops and we set $b_r = b_r(\tau_r)$.

3.2. Checking the guess for Δ_i. In order to decide whether Δ_i was appropriate, we consider for each $r \in M_i$ the curvature of the one-dimensional curve through r and the corresponding points $b_r \in M_{i+1}$ and $p_r \in M_{i-1}$. Here, p_r is the mesh point that was used to find r, that is, $r \in \mathcal{F}_{p_r}$. This is motivated by what is done for one-dimensional manifolds in [21] and goes back to an idea of Hobson [13]. Let α_r denote the angle between the line through p_r and r and the line through r and b_r. The step Δ_i was acceptable if both

$$\alpha_r < \alpha_{\max}, \tag{3.4}$$

$$\Delta_i \cdot \alpha_r < (\Delta\alpha)_{\max} \tag{3.5}$$

hold for all $r \in M_i$. In this case we accept M_{i+1} and form the simplicial complex C_{i+1} as the next $(k-1)$-sphere. On the other hand, if there is some $r \in M_i$ such that either (3.4) or (3.5) is not satisfied, then Δ_i was too big. We discard M_{i+1}, decrease Δ_i (in practice we halve it), and compute a new M_{i+1} at this smaller distance from M_i. The algorithm accepts M_{i+1} after all (with a warning message) when Δ_i has already been decreased to a prespecified minimal value $\Delta_{\min}$.

If Δ_i was acceptable, we usually set $\Delta_{i+1} = \Delta_i$. However, if for every $r \in M_i$ both α_r and $\Delta \cdot \alpha_r$ are well below the respective upper bounds in (3.4) or (3.5), say, less than $\alpha_{\min}$ and $(\Delta\alpha)_{\min}$, respectively, then we try a larger Δ_{i+1} in the next step (in practice we set $\Delta_{i+1} = 2\Delta_i$). The parameters $\alpha_{\min}$, $\alpha_{\max}$, $(\Delta\alpha)_{\min}$, and $(\Delta\alpha)_{\max}$ need to be specified before a computation.

3.3. Adding and removing mesh points. It is very important to maintain an accurate approximation of $S_{\eta_{i+1}}$ by the simplicial complex C_{i+1}, because this allows us to control the interpolation error; see section 5. With our method for controlling the growth step size Δ_i we ensure that the distance between C_i and C_{i+1} is within the accuracy bounds of the computation. However, we also need to make sure that, when neighboring points of M_{i+1} are too far away from each other inside this level set, a new point is added between them. To do this we choose a suitable point in C_i between the two respective points in M_i and then use the method in section 3.1 to find a new point in C_{i+1}. In this way, we ensure that no interpolation is performed between points that are further away from each other than a maximally allowed and prespecified distance $\Delta_{\mathcal{F}}$.

Similarily, when two points in M_{i+1} come within a prespecified distance $\delta_{\mathcal{F}}$, we remove one of them to ensure proper order relations between directly neighboring points.

3.4. Implementation details for $k = 2$. When the manifold $W^u(x_0)$ is two-dimensional it is parametrized by a family of (topological) circles S_η. The 1-simplices that form the continuous objects C_i are line segments between neighboring points in M_i. The algorithm grows $W^u(x_0)$ by adding a new circle C_{i+1} at each step, which means that a new band of width Δ_i of triangles is added to $\mathcal{C}$. The simplicial complex $\mathcal{C}$ representing $W^u(x_0)$ consists of triangles that constitute the bands between consecutive circles C_{i-1} and C_i.

The initial circle M_0 is chosen in $E^u(x_0)$ at distance δ. As was explained in section 3.2, the width of each band depends on the curvature of the manifold locally near the band, and it is governed by $\alpha_{\min}$, $\alpha_{\max}$, $(\Delta\alpha)_{\min}$, and $(\Delta\alpha)_{\max}$. Extra points are added to M_i when neighboring mesh points are further apart than $\Delta_{\mathcal{F}}$ and removed if they are closer than $\delta_{\mathcal{F}}$. The implementation of this is quite straightforward for $k = 2$ because of the order structure of the meshes M_i.

The boundary value problem (3.1)–(3.2) is presently solved with a shooting approach, for which we use a fourth-order fixed time-step Runge–Kutta integration routine.

4. Examples. We now present a number of examples to discuss the properties of our algorithm. In particular, we explain how the parametrization of the manifolds by geodesic level sets can be used to gain insight into their often quite complicated geometry. All figures have been rendered with the program Geomview [28], and each figure is accompanied by an animation that is supplied with this paper.

4.1. The Lorenz system with an attracting orbit. We consider the well-known Lorenz system [19]

$$\left\{\begin{array}{rcl} \dot{x} &=& \sigma(y-x), \\ \dot{y} &=& \varrho x - y - xz, \\ \dot{z} &=& xy - \beta z \end{array}\right. \tag{4.1}$$

for $\sigma = 10$, $\varrho = 28$, and the nonstandard value $\beta = 0.4$, for which there is an attracting periodic orbit that encircles two saddle points. Figure 1 (a) shows the two-dimensional stable manifold $W^s(0)$ of the origin.

The manifold $W^s(0)$ was computed with the following accuracy. The computation started with 20 points on a circle in $E^s(0)$ of radius $\delta = 1.0$ around the origin. Then new circles

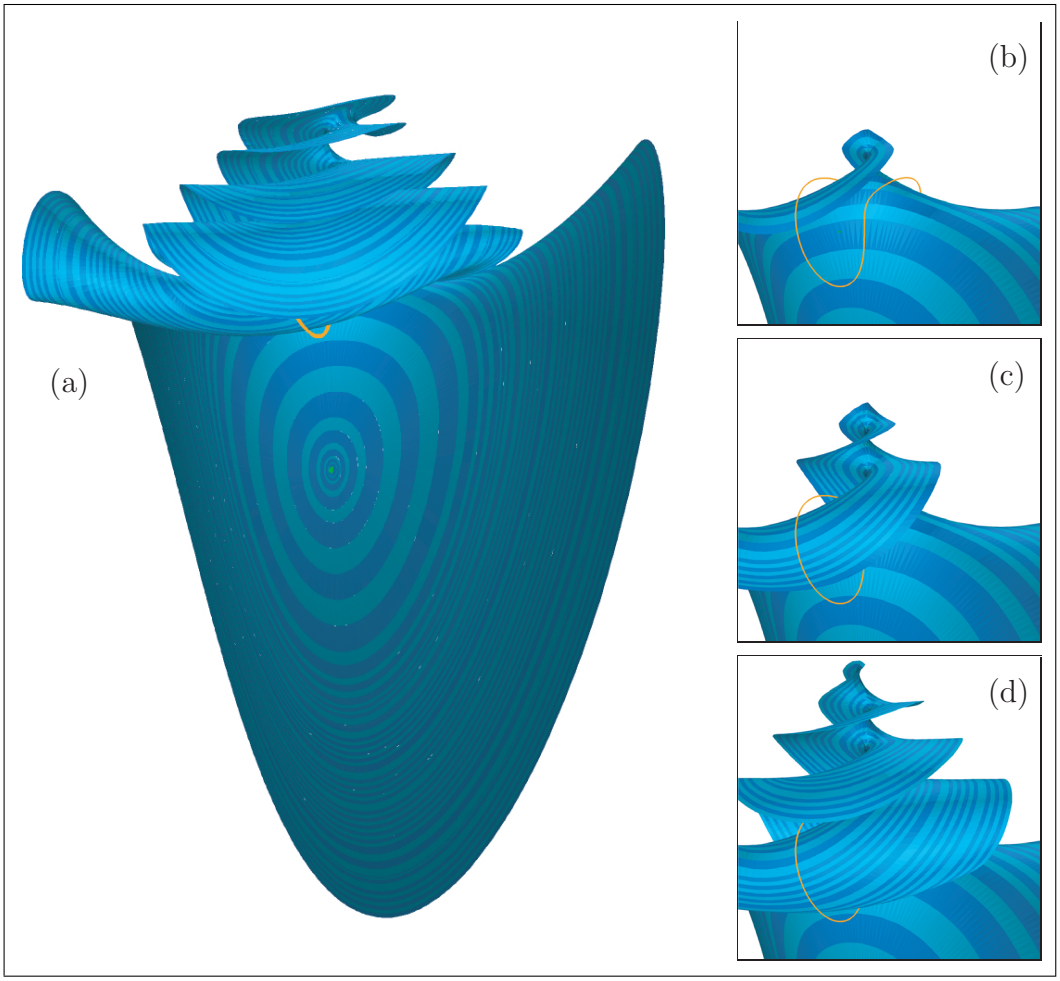


Figure 1. *The stable manifold $W^s(0)$ of the Lorenz system for $\sigma = 10$, $\varrho = 28$, and $\beta = 0.4$, computed up to geodesic distance* 66.28 (a). *Close-ups near the attracting periodic orbit show the growth process:* 24 *bands or geodesic distance* 40.25 (b), 34 *bands or geodesic distance* 46.25 (c), *and* 49 *bands or geodesic distance* 54.75 (d).

were added at distances Δ_i controlled by $\alpha_{\min} = 0.3$, $\alpha_{\max} = 0.4$, $(\Delta\alpha)_{\min} = 0.1$, and $(\Delta\alpha)_{\max} = 1.0$; see sections 3.1 and 3.2. (For practical reasons a new circle was always accepted if $\Delta_i \leq 0.01$.) The mesh points on a circle are never more than $\Delta_{\mathcal{F}} = 1.0$ or less than $\delta_{\mathcal{F}} = 0.25$ apart. In total 75 circles were computed on $W^s(0)$; the last circle is approximately at geodesic distance 66.28 from the origin, and it consists of 1522 mesh points. How $W^s(0)$ is grown during the computation is shown in Figure 1 (b)–(d) with three close-up views near the attracting periodic orbit; see also the accompanying movie. We remark that the manifold in Figure 1 (a) is featured on the advertising poster for SIADS.

The color changes in Figure 1 indicate the positions of the bands that are added during the course of the computation. They illustrate the geodesic level sets on $W^u(0)$ that were computed. To highlight this, Figure 2 shows individual small segments of bands at different

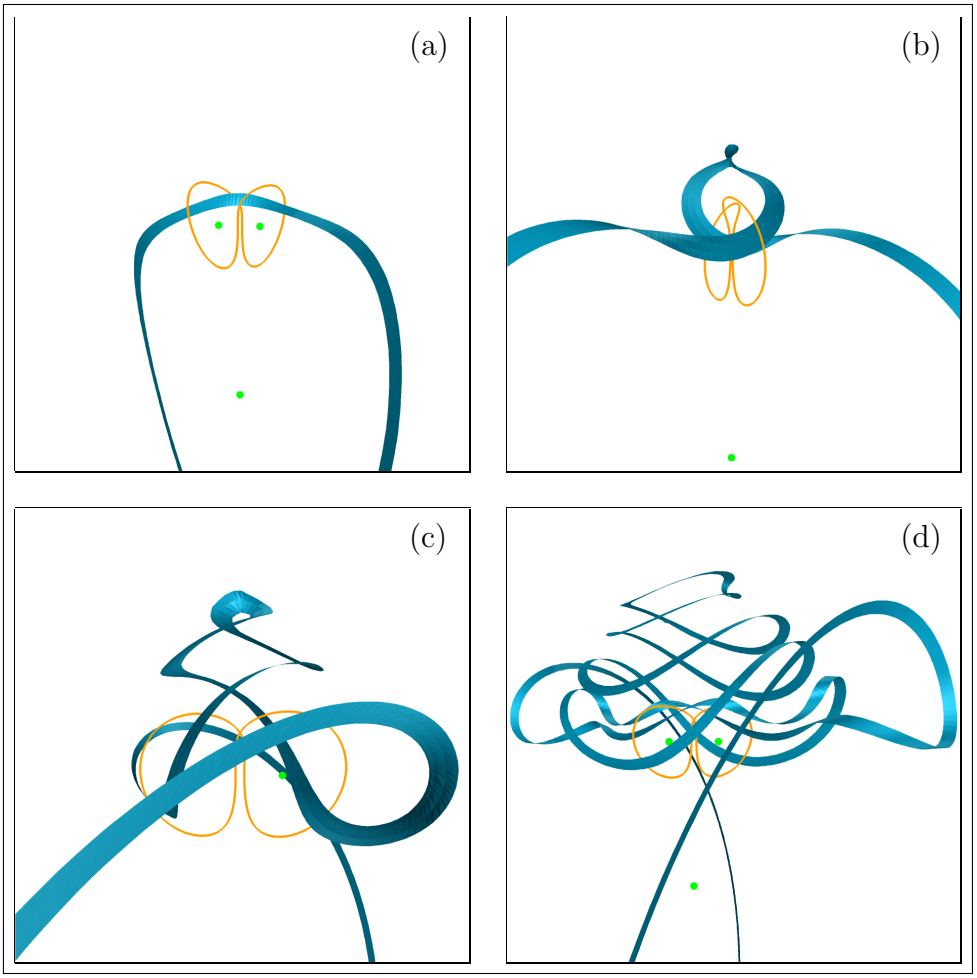


Figure 2. *Growing bands are approximations of the level sets S_η of the stable manifold of the Lorenz system in Figure 1. From* (a) *to* (d) *are shown band* 14 *with* $\eta \in [30.25, 32.25]$, *bands* 23–25 *with* $\eta \in [38.25, 41.25]$, *bands* 33–35 *with* $\eta \in [44.75, 47.25]$, *and bands* 68–71 *with* $\eta \in [63.75, 65.75]$, *respectively.*

(approximately) constant geodesic distance from the origin. Each boundary consists of two topological circles that are both indeed smooth, simple, and closed. Notice that the circles have quite complicated embeddings in the phase space due to the overall shape of the manifold but are nevertheless topologically trivial, that is, contractible to the fixed point from which they were grown. For an animation illustrating the change in shape of the bands and circles during the growth process, see the accompanying movie.

4.2. An optimal control system. The following example from optimal control theory is taken from [11, 15] and demonstrates that our implementation for $k = 2$ can indeed be used in ambient spaces of arbitrary dimension. Furthermore, it is an illustration of how invariant manifold computations can be used to understand the dynamics of systems arising in applications.

Consider an inverted planar pendulum balancing on a cart. The cart moves in the plane

of the pendulum with an applied horizontal force u constituting a control. The mass of the cart is M, the mass of the pendulum is m, and its center of mass is at distance l from the pivot. Disregarding the model equations associated with the cart gives the two-dimensional vector field

$$\left\{\begin{array}{rcl} \dot{x}_1 &=& x_2, \\ \dot{x}_2 &=& f(x_1,x_2)+c(x_1,x_2)u \\ &:=& \dfrac{\frac{g}{l}\sin(x_1)-\frac{1}{2}m_r x_2^2\sin(2x_1)-\frac{m_r}{ml}cos(x_1)u}{\frac{4}{3}-m_r\cos^2(x_1)}. \end{array}\right. \tag{4.2}$$

Here $x_1 \in \mathbb{R}$ is the angle measured from the upright position (not taken modulo 2π), $m_r = m/(m+M)$ is the mass ratio, and g is the gravitational constant. The origin is an unstable equilibrium corresponding to the upright position.

We wish to find an optimal control u as a function of x_1 and x_2 that drives the system to the origin, while minimizing the cost function

$$Q(x_1,x_2,u) = \mu_1 x_1^2 + \mu_2 x_2^2 + \mu_3 u^2 \tag{4.3}$$

for positive parameters μ_1, μ_2, and μ_3. (It is because of this nonperiodic cost function that x_1 is not taken modulo 2π.) Pontryagin's maximum principle [33] ensures that an optimal solution exists and is represented by special solutions of the four-dimensional Hamiltonian system given by the Hamiltonian

$$\begin{aligned} H(x_1,x_2,p_1,p_2) &= Q(x_1,x_2,u^*(x_1,x_2,p_1,p_2)) + p_1 x_2 \\ &\quad + p_2\, f(x_1,x_2) + p_2\, c(x_1,x_2)\, u^*(x_1,x_2,p_1,p_2), \end{aligned} \tag{4.4}$$

where $u^*(x_1,x_2,p_1,p_2) = -\frac{1}{2\mu_3}\, c(x_1,x_2)p_2$. Namely, the optimal solution to reach the origin from an initial condition $(x_1(0),x_2(0))$ is found by lifting $(x_1(0),x_2(0))$ to a point $(x_1(0),x_2(0),p_1(0),p_2(0)) \in \mathbb{R}^4$ that lies on the two-dimensional stable manifold $W^s(0)$ of the origin of the system described by the Hamiltonian (4.4). The trajectory $(x_1(t),x_2(t),p_1(t),p_2(t))$ through this point directly corresponds to a (locally) optimal trajectory $(x_1(t),x_2(t))$ of (4.2) by using the optimal controller $u(t) = u^*(x_1(t),x_2(t),p_1(t),p_2(t))$. If there is more than one point on $W^s(0)$ for given $(x_1(0),x_2(0))$, then there exists more than one choice for u^*. This means that either the solution is only locally optimal or the globally optimal cost can be achieved using two different control actions.

We computed $W^s(0)$ for the parameter values as in [11, 15], namely, $m = 2\,\mathrm{kg}$, $M = 8\,\mathrm{kg}$, $l = 0.5\,\mathrm{m}$, $g = 9.8\,\mathrm{m/s^2}$, $\mu_1 = 0.1$, $\mu_2 = 0.05$, and $\mu_3 = 0.01$. We used the following accuracy parameters for the computation. The starting data was a set of 20 points on a circle in $E^s(0)$ of radius $\delta = 1.0$ around the origin. The distances Δ_i at which new circles were added were controlled by $\alpha_{\min} = 0.3$, $\alpha_{\max} = 0.4$, $(\Delta\alpha)_{\min} = 0.05$, $(\Delta\alpha)_{\max} = 0.2$, where we always accepted the circle if $\Delta_i \le 0.05$; see sections 3.1 and 3.2. The distances between mesh points on a circle were always between $\delta_{\mathcal{F}} = 0.125$ and $\Delta_{\mathcal{F}} = 0.5$.

The result is shown in Figure 3, where $W^s(0)$ was computed up to a geodesic distance of approximately 26.25, corresponding to 111 circles. The last circle has 1229 mesh points. Since the ambient space is four-dimensional it is difficult to view the result. Therefore, Figure 3

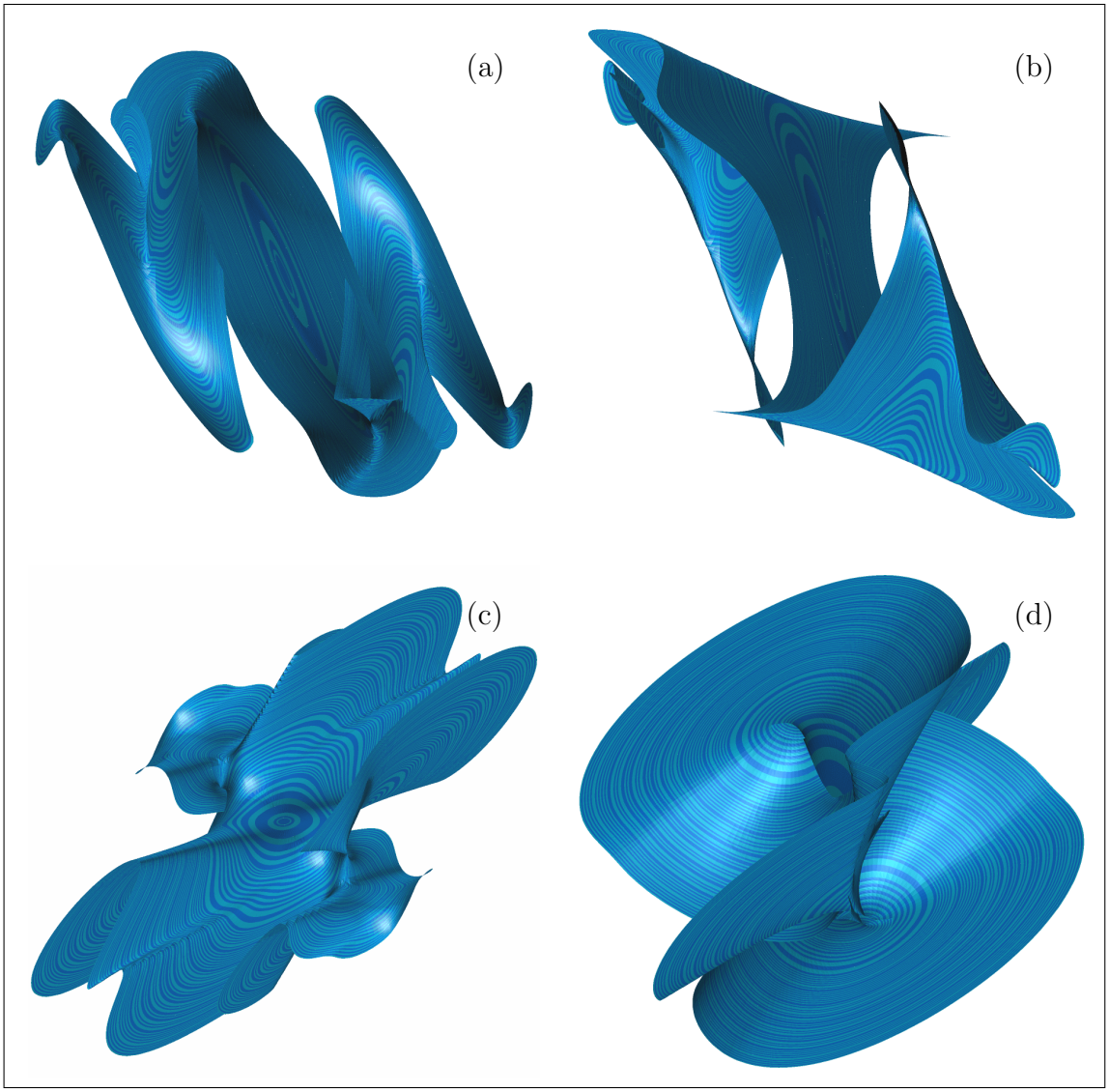


Figure 3. *Four views of the two-dimensional stable manifold $W^s(0)$ up to geodesic distance 26.25 of the Hamiltonian system (4.4) projected onto three-dimensional spaces, namely onto $\{p_2 = 0\}$ (a), onto $\{p_1 = 0\}$ (b), onto $\{x_2 = 0\}$ (c), and onto $\{x_1 = 0\}$ (d). All self-intersections are due to projection.*

shows four views in four different three-dimensional projections. The growth process of the manifold is shown in the same projections in the accompanying animation. Note that the optimal control $u^*(x_1, x_2, p_1, p_2) = -\frac{1}{2\mu_u} c(x_1, x_2)\, p_2$ does not depend on p_1. Therefore, the folds shown in the projection onto $\{p_1 = 0\}$ in Figure 3 (b) indicate that indeed more than one optimal control exists for several choices of the initial condition $(x_1(0), x_2(0))$.

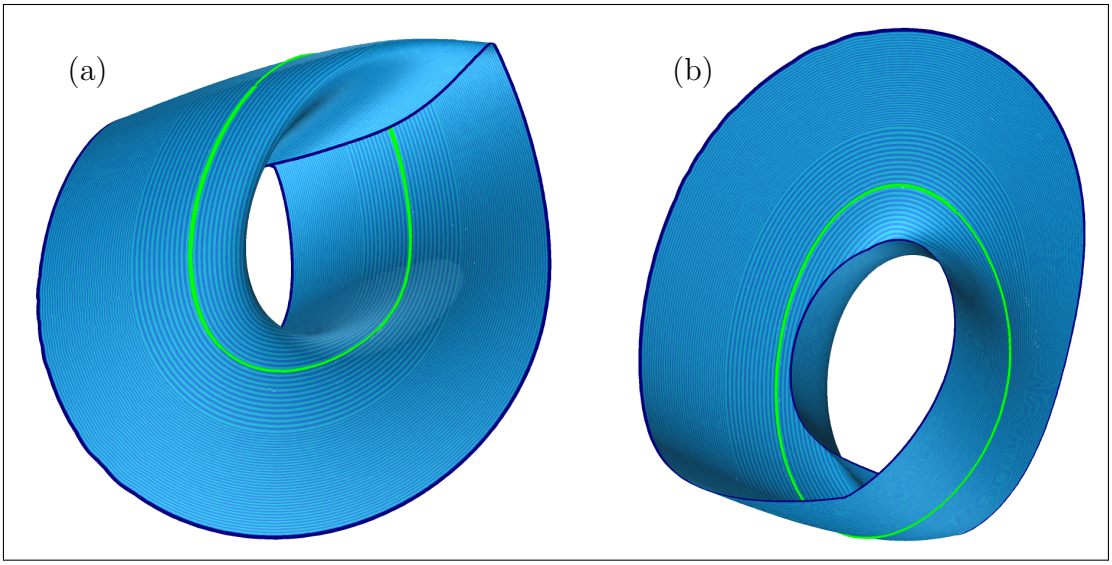


Figure 4. *Two views of the two-dimensional stable manifold up to geodesic distance* 3.1 *of the periodic orbit of the* ζ^3*-model* (4.5) *for* $\alpha = 3.2$ *and* $\beta = 2$*. The manifold is in fact a Möbius strip; to show this, its edge is highlighted.*

4.3. A Möbius strip in the ζ^3-model. To illustrate that the algorithm is able to compute stable or unstable manifolds of periodic orbits of saddle type, we consider the vector field

$$\left\{ \begin{array}{rcl} \dot{x} &=& y, \\ \dot{y} &=& z, \\ \dot{z} &=& \alpha\, x - x^2 - \beta\, y - z. \end{array} \right. \tag{4.5}$$

It is known as the ζ^3-model [2], and it arises as the asymptotic normal form of a system near the simultaneous onset of up to three instabilities; see also [16, 17, 22]. For $\alpha = 3.2$ and $\beta = 2$ the system has a saddle periodic orbit Γ with Floquet multipliers 1, -0.022, and -1.15, approximately. This means that both the stable and unstable manifolds are nonorientable.

We computed $W^s(\Gamma)$, where the accuracy parameters were $\Delta_{\mathcal{F}} = 0.1$, $\delta_{\mathcal{F}} = 0.025$, $\Delta_{\min} = 0.01$, $0.2 < \alpha < 0.3$, and $0.001 < \Delta\alpha < 0.1$. We took 100 mesh points on Γ, and the initial circle in the linear unstable eigenspace $N^s(\Gamma)$ of Γ was a double cover with 200 points at distance $\delta = 0.1$ from Γ. More specific details on how to compute nonorientable manifolds can be found in [25]. The result is shown in Figure 4 for a total computed geodesic distance of 3.1, corresponding to 100 circles. To help with depth perception we colored the last circle differently. It forms the boundary of the computed piece of the manifold, which is indeed a Möbius strip; see also the accompanying animation of the growth process. A more in-depth discussion on when and why the stable manifold of the ζ^3-model is nonorientable is presented in [26].

4.4. A manifold with $\eta_{\max} < \infty$. As was mentioned in section 2, an unstable manifold $W^u(x_0) = \{S_\eta)\}_{\eta>0}$ may be such that $\eta_{\max} < \infty$, even though the manifold does not converge to an attractor. This means that $S_{\eta_{\max}}$ is the first element of the parametrization that ceases

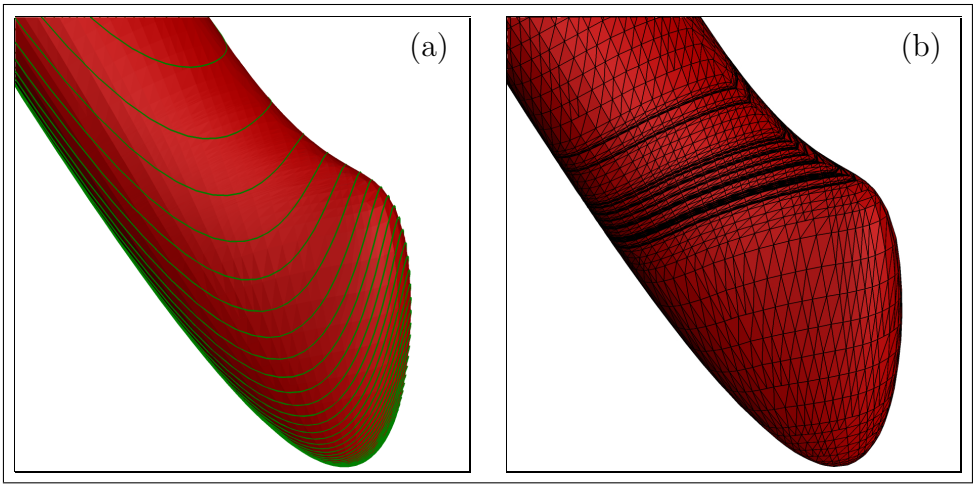


Figure 5. *The computed unstable manifold of the origin of the concocted vector field of section 4.4 for $a = 5$. Panel* (a) *shows a typical orbit on it (in green); panel* (b) *shows the mesh resulting from the computation.*

to be a smooth topological $(k-1)$-sphere without self-intersections. We are not aware of any example of this phenomenon in a dynamical system, and so we constructed an example of this situation for $k = 2$ as follows. Consider the vector field

$$\left\{ \begin{array}{rcl} \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} & = & A \begin{pmatrix} u \\ v \end{pmatrix}, \\ \dot{w} & = & -w + u^2 + v^2, \end{array} \right. \tag{4.6}$$

where

$$A = \begin{pmatrix} \alpha & \omega \\ -\omega & \alpha \end{pmatrix}.$$

For $\alpha > 0$ the origin of this vector field has a parabola-shaped unstable manifold tangent to the (x, y)-plane. The dynamics on this unstable manifold is simply given by the constant rotation $\omega > 0$.

We now apply the coordinate transformation

$$\left\{ \begin{array}{rcl} x & = & u - w, \\ y & = & v/3, \\ z & = & w + a\,(u - w)^2 \end{array} \right. \tag{4.7}$$

to (4.6), where $a > 0$ is a parameter. It deforms the parabola-shaped unstable manifold of (4.6) first in the x-direction so that only a finite piece of it lies to the right of the (y, z)-plane. Furthermore, for sufficiently large a a "bump" is created on the manifold in the region of positive x. The shape of the manifold is independent of ω, the amount of spiraling on it. We took $\omega = 0.2$ in the computations of $W^u(x_0)$.

Figure 5 shows the resulting manifolds for fixed $\alpha = 0.1$ and $a = 5$. For sufficiently large ω one can get a good impression of the shape of the manifold simply by plotting a single orbit

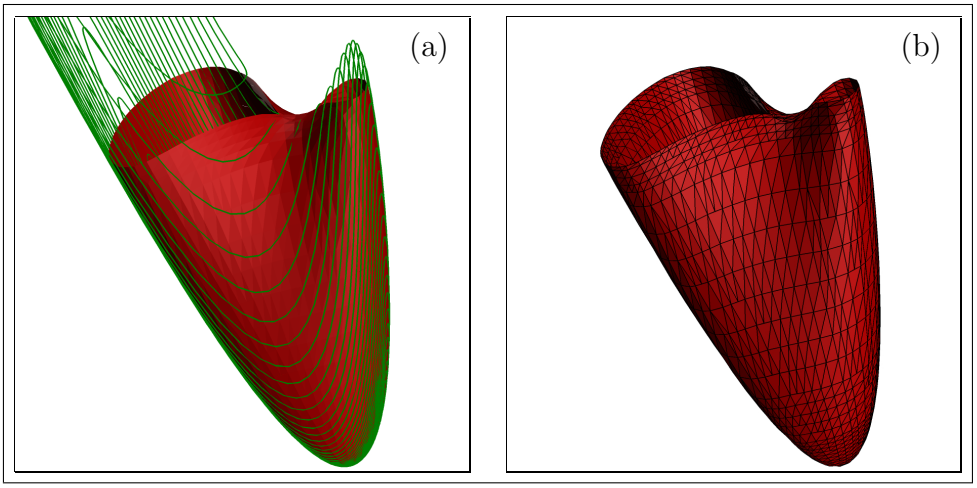


Figure 6. *The computed unstable manifold of the origin of the concocted vector field of section 4.4 for $a = 10$. Panel* (a) *shows a typical orbit on it (in green); panel* (b) *shows the mesh resulting from the computation. The algorithm stops when a geodesic level set self-intersects.*

on $W^u(0)$; this is illustrated in Figure 5 (a) for an initial condition $(0.005, 0, 0)$ in $E^u(x_0)$ with $\omega = 10$. Our algorithm is able to compute $W^u(0)$ for $a = 5$, but it makes quite small Δ_i steps near and just after the emerging bump. Notice that quite a number of mesh points need to be removed before the mesh is restored such that the manifold can be grown further with larger steps, as is further illustrated in the accompanying animation.

On the other hand, for $a = 10$ the bump is fully developed. The trajectory in Figure 6 (a) was computed for $a = 10$ with $\omega = 10$ and the same initial condition as for $a = 5$. There exists a point at the bottom of the "valley" between the infinite part and the bump on the manifold that has two geodesics. This valley point has geodesic distance $\eta_{\max}$ as defined above. In other words, it is a singularity of the geodesic level set on which it lies, which forms a figure eight. Level sets for geodesic distance larger than $\eta_{\max}$ consist of two disjoint circles, one around the bump and one around the "neck" of the manifold. Indeed, the algorithm stops when the valley floor is reached; see also the accompanying animation.

In both computations the starting data was a circle in $E^u(0)$ of radius $\delta = 0.01$ around the origin. Furthermore, we set $\alpha_{\min} = 0.3$, $\alpha_{\max} = 0.4$, $(\Delta\alpha)_{\min} = 0.002$, $(\Delta\alpha)_{\max} = 0.015$, $\Delta_{\min} = 0.001$, $\delta_{\mathcal{F}} = 0.005$, and $\Delta_{\mathcal{F}} = 0.04$.

We remark that a finite $\eta_{\max}$ is a problem for all algorithms that use the approach of calculating approximate geodesic level sets, that is, also for the algorithms in [10, 16, 34]. Actually, this geometric obstruction to growing $W^u(0)$ is, in a sense, an artificial problem. It can be overcome by changing to a different metric; in this specific example one can rescale y to ensure that the geodesic distance in the valley is realized by the path exactly over the top. However, changing the metric as one encounters this obstruction during a computation is not practical. Instead one could split up the figure-eight level set into two separate circles and start two individual computations. Since we did not encounter this geometric obstruction in any other example, this has not been implemented.

4.5. The Lorenz system with a strange attractor. To demonstrate how one can make use of the parametrization by geodesic distance to understand complicated manifolds, we consider again the Lorenz system (4.1), but now for the standard values of the parameters $\sigma = 10$, $\varrho = 28$, and $\beta = 8/3$, which gives the well-known Lorenz attractor. This example has emerged as a test case for the computation of two-dimensional global manifolds, and the task is to compute the stable manifold of the origin as it interacts with the chaotic attractor. How exactly does this happen and what does the manifold look like? A first attempt at visualizing this manifold was a series of three sketches in [1]. Computed images of differing sophistication can be found in [9, 10, 12, 20, 22] and in [23], which features a direct comparison with the sketches in [1].

The manifold $W^s(0)$ in Figure 7 (a) was computed with the accuracy parameters in section 4.1. In total 72 circles were computed on $W^s(0)$; the last circle is approximately at geodesic distance 151.75 from the origin, and it consists of 2300 mesh points.

How $W^s(0)$ is grown during the computation is shown in Figure 7 (b)–(d) with three close-ups showing how a helix develops about the positive z-axis; see also the accompanying movie. The coloring indicates geodesic distance from the origin, as given by the color bar. Also shown is the one-dimensional unstable manifold of the origin in red, whose closure forms the well-known Lorenz attractor, an orbit which is shown in yellow.

The geometry of the manifold is illustrated in Figure 8 and the accompanying movie by following the development of a small set of bands as they grow. This gives an idea of how $W^s(0)$ is parametrized by geodesic level sets. In spite of the complicated geometry, each of the level sets is still an unknotted smooth topological circle, even though it is not very "round" and certainly not planar. The later bands nicely show how the manifold has to cross over and back between the two scrolls of the chaotic attractor. Clearly visible in Figure 8 (e) and (f) is the development of two further helixes of $W^s(0)$ near the z-axis.

5. Convergence of the algorithm. Suppose that we computed an approximation of the manifold $W^u(x_0)$ up to a prescribed and fixed geodesic distance D. Hence, we have a simplicial complex $\mathcal{C}(\delta, \Delta)$ with the mesh points $\mathcal{M}(D, \delta, \Delta) = \cup_{i=0}^{l} M_i$ approximating the compact manifold $\mathcal{S}_D = \{S_\eta\}_{0 \le \eta \le D} \subset W^u(x_0)$. In the discussion of the convergence of the algorithm we assume that $D < \eta_{\max}$. The dependence on δ and Δ indicates that the approximation was computed with these prespecified accuracy parameters, where for simplicity we chose $\Delta_{\mathcal{F}} = \Delta$ and $\Delta_i = \Delta$ constant so that Δ specifies the overall mesh quality.

We want to show that $\mathcal{C}(\delta, \Delta)$ converges to $\mathcal{S}_D$ as the initial step δ from x_0 and the mesh quality parameter Δ go to zero. The fundamental difficulty with any such proof is that when δ and Δ are changed, the entire approximation needs to be recomputed, meaning that the refined set of mesh points $\mathcal{M}(\tilde{\delta}, \tilde{\Delta})$ generally does not have any mesh points in common with $\mathcal{M}(\delta, \Delta)$. This is why one cannot use strategies of proof as they are used when proving the correctness of computations of implicitly defined manifolds (given as the zero set of a function) where it is possible to refine and improve the mesh locally; see, for example, [29]. In fact, the problem here is quite similar to proving the convergence of numerical approximations of attractors to the actual attractor of the underlying continuous dynamical system; see [31].

The key problem is to control the total (interpolation) error between mesh points uniformly over the mesh. We use here the approach in [20] for two-dimensional unstable manifolds of

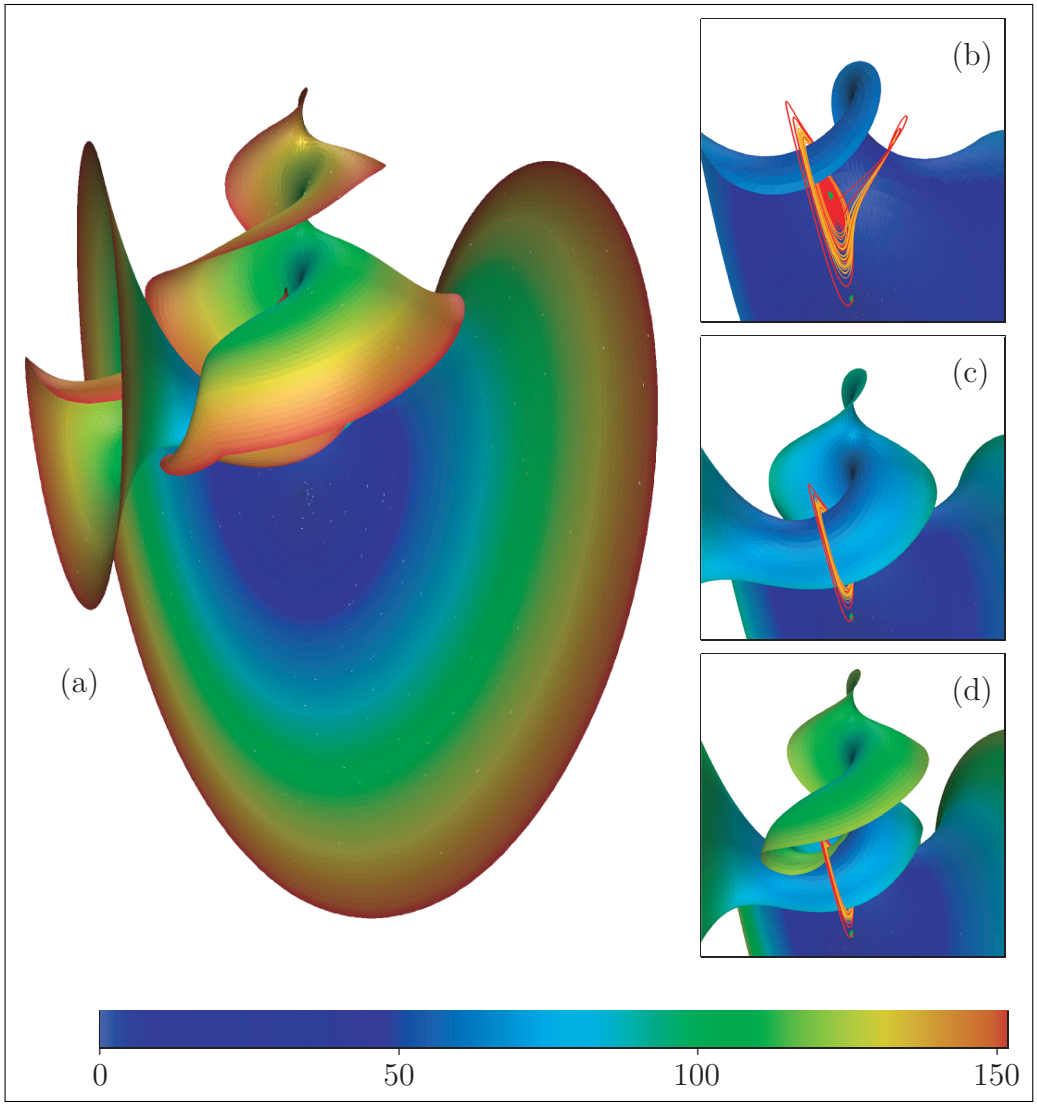


Figure 7. *The stable manifold $W^s(0)$ of the Lorenz system for $\sigma = 10$, $\varrho = 28$, and $\beta = 8/3$, computed up to geodesic distance* 151.75 (a). *Close-ups near the z-axis show the growth process:* 25 *bands or geodesic distance* 62.75 (b), 40 *bands or geodesic distance* 92.75 (c), *and* 55 *bands or geodesic distance* 122.75 (d). *Color indicates geodesic distance from the origin as given by the color bar.*

maps.

Our goal is to prove that for any given $\varepsilon > 0$ the approximation $\mathcal{C}(\delta, \Delta)$ is in an ε-neighborhood of $W^u(x_0)$, provided δ and Δ are chosen small enough. This is known as upper semicontinuity. Our proof essentially tracks how different errors grow along the manifold, which is described by certain Lipschitz bounds on the local growth rate of the vector field. In many situations it is not possible to determine these Lipschitz bounds, so that *a priori* error bounds for the computation of global invariant manifolds are not available. The only practical way of checking a computation is to repeat it with increased accuracy and compare the results.

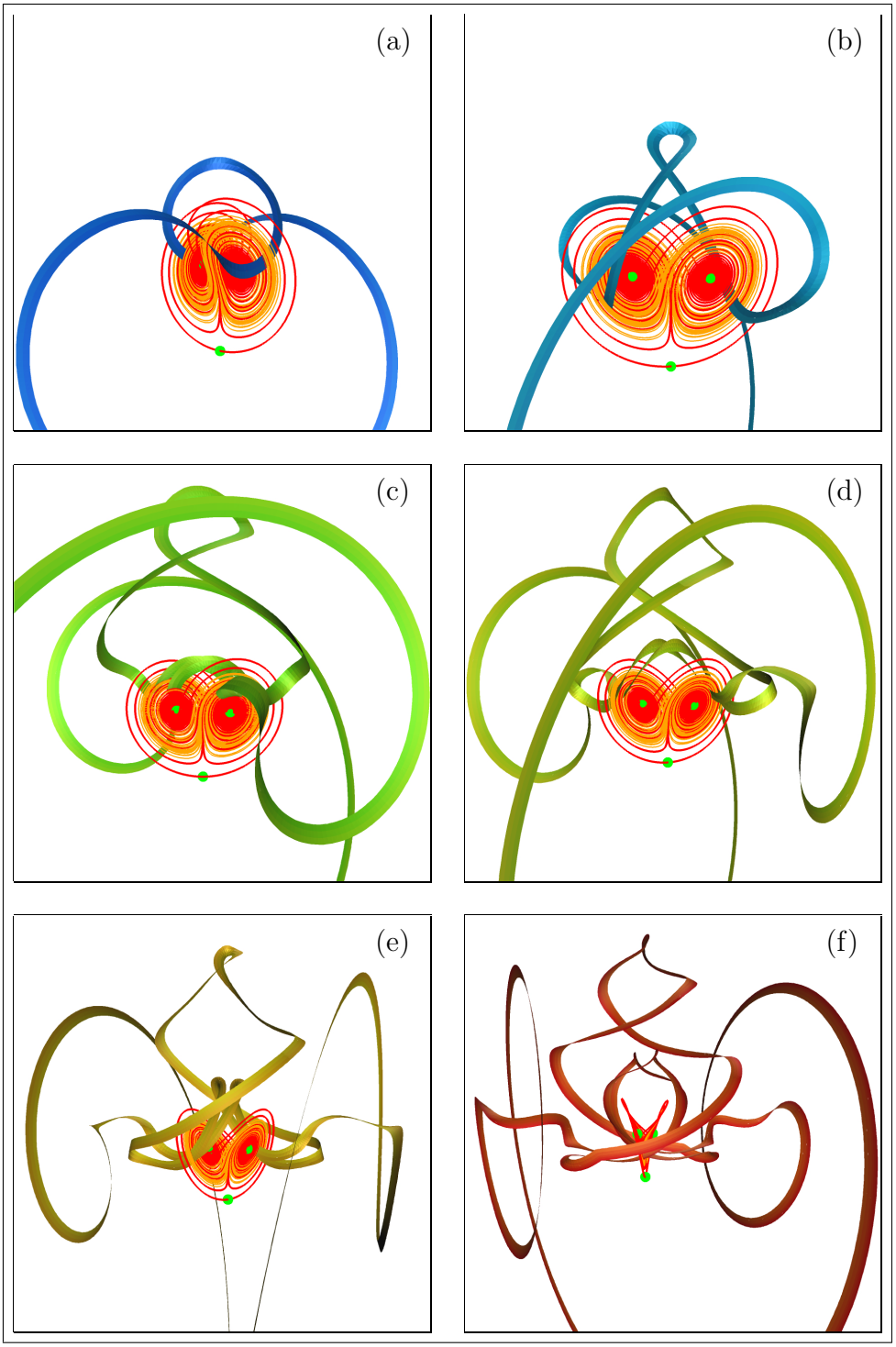


Figure 8. *Growing bands are approximations of the level sets S_η of the stable manifold of the Lorenz system in Figure 7. From* (a) *to* (f) *are shown bands* 24–25 *with* $\eta \in [58.75, 62.75]$, *bands* 30–31 *with* $\eta \in [70.75, 74.75]$, *bands* 52–54 *with* $\eta \in [114.75, 120.75]$, *bands* 56–58 *with* $\eta \in [122.75, 128.75]$, *bands* 60–62 *with* $\eta \in [130.75, 136.75]$, *and bands* 67–71 *with* $\eta \in [144.75, 151.75]$, *respectively.*

5.1. Upper semicontinuity. Recall (see, for example, [14]) that the semidistance between two sets $A, B \subset \mathbb{R}^n$ is

$$d(A,B) = \sup_{x\in A}\{d(x,B)\}, \text{ where } d(x,B) = \inf_{y\in B}\{\|x-y\|\}.$$

We use the notation $\mathcal{N}_\varepsilon(X)$ for the ε-neighborhood of a set $X \subset \mathbb{R}^n$.

Theorem 1 (upper semicontinuity). *Let $D > 0$ be given and fixed, and let $\mathcal{C}(\delta,\Delta)$ denote the simplicial complex approximation of $\mathcal{S}_D = \{S_\eta\}_{0\le\eta\le D}$. Then for any $\varepsilon > 0$*

$$\mathcal{C}(\delta,\Delta) \subset \mathcal{N}_\varepsilon(W^u(x_0))$$

for all sufficiently small δ and Δ.

Proof. In order to show that $\mathcal{C}(\delta,\Delta) \subset \mathcal{N}_\varepsilon(W^u(x_0))$, we will show that $\mathcal{C}(\delta,\Delta) \subset \mathcal{N}_\varepsilon(\mathcal{S}_{\tilde{D}})$ for some $\tilde{D} > D$. Since $\mathcal{S}_{\tilde{D}} \subset W^u(x_0)$, this will automatically imply that $\mathcal{C}(\delta,\Delta) \subset \mathcal{N}_\varepsilon(W^u(x_0))$. The reason for using $\mathcal{S}_{\tilde{D}}$ instead of $W^u(x_0)$ is that $\mathcal{S}_{\tilde{D}}$ is compact; this will be used to get Lipschitz bounds.

The total error $\varepsilon_{\mathcal{C}}$ is defined as

$$\varepsilon_{\mathcal{C}} = d(\mathcal{C}(\delta,\Delta), \mathcal{S}_{\tilde{D}}). \tag{5.1}$$

This means that $\mathcal{C}(\delta,\Delta)$ is contained in $\mathcal{N}_\varepsilon(\mathcal{S}_{\tilde{D}})$ for any $\varepsilon > \varepsilon_{\mathcal{C}}$. We obtain upper semicontinuity by showing that $\varepsilon_{\mathcal{C}} \to 0$ as $\delta, \Delta \to 0$.

There are two contributions to the error $\varepsilon_{\mathcal{C}}$. First, the mesh points in $\mathcal{M}(\delta,\Delta)$ do not lie exactly on $\mathcal{S}_{\tilde{D}}$, which gives rise to the mesh error

$$\varepsilon_{\mathcal{M}} = d(\mathcal{M}(\delta,\Delta), \mathcal{S}_{\tilde{D}}).$$

Second, there is the global interpolation error $\varepsilon_I(\Delta)$ due to the fact that points in $\mathcal{C}(\delta,\Delta) \setminus \mathcal{M}(\delta,\Delta)$ are interpolated linearly between the mesh points. Consequently, we have for the total error

$$\varepsilon_{\mathcal{C}} \le \varepsilon_{\mathcal{M}} + \varepsilon_I(\Delta). \tag{5.2}$$

Note that, unlike in the error analysis of the approximation of an implicitly defined invariant manifold [29], these two errors are not independent. The interpolation error at step i enters into the mesh error at step $i+1$.

Since the manifold is C^2 and interpolation is always done between points that are at most Δ apart, the interpolation error $\varepsilon_I(\Delta)$ is $O(\Delta^2)$. Hence, in light of (5.2), it suffices to show that $\varepsilon_{\mathcal{M}}$ goes to zero with Δ and δ. To this end, we define the mesh error at step i

$$\varepsilon_{\mathcal{M}}(i) = d(M(\delta,\Delta)_i, \mathcal{S}_{\tilde{D}}).$$

The inital mesh error $\varepsilon_{\mathcal{M}}(0)$ is due to the fact that $M(\delta,\Delta)_0$ is chosen in the linear subspace $E^u(x_0)$ at distance δ from x_0 instead of on $W^u(x_0)$. By taking δ small, $\varepsilon_{\mathcal{M}}(0)$ can be made arbitrarily small. Under the assumption that $W^u(x_0)$ is C^2, the initial error $\varepsilon_{\mathcal{M}}(0)$ is of order $O(\delta^2)$.

The mesh error $\varepsilon_{\mathcal{M}}(i+1)$ at step $i+1$ can be estimated in terms of $\varepsilon_{\mathcal{M}}(i)$ as follows. Points in $M(\delta,\Delta)_{i+1}$ are found by continuing the boundary value problem (3.1)–(3.2). Since the accuracy of the boundary value solver is independent of the other computational errors, we assume that solving the boundary value problem does not introduce an extra error. (In practice, it is sufficient to ensure that the boundary value solver has an error of order less than that of the interpolation error.) Recall that, for any $r \in M(\delta,\Delta)_i$, the starting condition (3.1) of the boundary value problem leads to points $q_r(\tau_r) \in C(\delta,\Delta)_i$ and the end condition (3.2) to points $b_r(\tau_r) \in \mathcal{F}_r$, where τ_r is the integration time such that $\| b_r(\tau_r) - r \| = \Delta$. The error at $C(\delta,\Delta)_{i+1}$ comes from the fact that $C(\delta,\Delta)_i$ is only approximately on $\mathcal{S}_{\tilde{D}}$, so that we are solving the "wrong" boundary value problem. Therefore,

$$\begin{aligned} \varepsilon_{\mathcal{M}}(i+1) &= \max_{r\in M(\delta,\Delta)_i} d(b_r(\tau_r), \mathcal{S}_{\tilde{D}}) \\ &= \max_{r\in M(\delta,\Delta)_i} d(\phi^{\tau_r}(q_r(\tau_r)), \mathcal{S}_{\tilde{D}}). \end{aligned} \tag{5.3}$$

Since $\mathcal{S}_{\tilde{D}}$ is compact and the vector field is C^2, a Lipschitz constant $\kappa_U > 0$ and a neighborhood U of $\mathcal{S}_{\tilde{D}}$ exist such that

$$d(\phi^t(p), \mathcal{S}_{\tilde{D}}) \le e^{\kappa_U t}\, d(p, \mathcal{S}_{\tilde{D}}), \tag{5.4}$$

provided the orbit of p up to $\phi^t(p)$ is in U; see also [31]. Combining this with (5.3) and defining τ_i such that $e^{\kappa_U \tau_i} = \max_{r\in M(\delta,\Delta)_i} (e^{\kappa_U \tau_r})$, we get

$$\varepsilon_{\mathcal{M}}(i+1) \le \max_{r\in M(\delta,\Delta)_i} e^{\kappa_U \tau_r} d(q_r(\tau_r), \mathcal{S}_{\tilde{D}}) \le e^{\kappa_U \tau_i} (\varepsilon_{\mathcal{M}}(i) + \varepsilon_I(\Delta)). \tag{5.5}$$

In total, there are $l = \lceil \frac{D-\delta}{\Delta} \rceil$ steps so that

$$\varepsilon_{\mathcal{M}} = \max_{0\le i\le l} \{\varepsilon_{\mathcal{M}}(i)\} \le e^{\kappa_U \tau(\delta,\Delta)} (\varepsilon_{\mathcal{M}}(0) + l\, \varepsilon_I(\Delta)),$$

where $\tau(\delta,\Delta) = \sum_{i=0}^{l} \tau_i$. One should think of $\tau(\delta,\Delta)$ as the maximal integration time needed to reach $C(\delta,\Delta)_l$ starting from a point in $C(\delta,\Delta)_0$. Note that $l\, \varepsilon_I(\Delta)$ is $O(\Delta)$ because $W^u(x_0)$ is C^2. To deal with the fact that $\tau(\delta,\Delta) \to \infty$ as $\delta \to 0$, we use the fact that the local manifold is always an attractor [27]. This means that there exists a fixed neighborhood $V \subset U$ of x_0 on which we find a *negative* Lipschitz constant $-\kappa_V < 0$. In other words, the error decreases from one step to the next as long as one remains in V. Suppose that $S_\eta \subset V$ for $0 \le \eta \le \eta_V$ and $\delta < \eta_V$. Let δ and Δ be small enough so that there exists $J = J(\delta,\Delta) > 0$ ($\eta_J \approx \eta_V$) with $C(\delta,\Delta)_i \subset V$ for $0 \le i \le J$. Then for all $0 \le i < J$

$$\begin{aligned} \varepsilon_{\mathcal{M}}(i+1) &\le e^{-\kappa_V \tau_i} (\varepsilon_{\mathcal{M}}(i) + \varepsilon_I(\Delta)) \\ &\le e^{-\kappa_V \tau_V(\delta,\Delta)} (\varepsilon_{\mathcal{M}}(0) + J\, \varepsilon_I(\Delta)) \le \varepsilon_{\mathcal{M}}(0), \end{aligned} \tag{5.6}$$

provided Δ is small enough. Here, $\tau_V(\delta,\Delta) = \sum_{i=0}^{J} \tau_i$. The important point is that (5.6) holds also in the limit $\delta \to 0$ as long as $\Delta \to 0$ accordingly.

One can now find a fixed upper bound $\bar{\tau} < \infty$ on the total integration time that is needed to cover $U \setminus V$, independent of δ and Δ, that is,

$$\tau(\delta, \Delta) \leq \tau_V(\delta, \Delta) + \bar{\tau}.$$

In other words, $\tau(\delta, \Delta) \to \infty$ as $\delta \to 0$ only because $\tau_V(\delta, \Delta) \to \infty$ with decreasing δ.

By combining (5.6) and (5.5), we get for all $J \leq i < l$

$$\begin{aligned} \varepsilon_{\mathcal{M}}(i) &\leq e^{\kappa_U \bar{\tau}} \left(\varepsilon_{\mathcal{M}}(J) + (l - J)\, \varepsilon_I(\Delta)\right) \\ &\leq e^{\kappa_U \bar{\tau}} \left(\varepsilon_{\mathcal{M}}(0) + O(\Delta)\right). \end{aligned} \tag{5.7}$$

Hence, using (5.6) and (5.7), we have

$$\begin{aligned} \varepsilon_{\mathcal{M}} = \max_{0 \leq i \leq l} \{\varepsilon_{\mathcal{M}}(i)\} &\leq \max\{\varepsilon_{\mathcal{M}}(0), e^{\kappa_U \bar{\tau}} \left(\varepsilon_{\mathcal{M}}(0) + O(\Delta)\right)\} \\ &= e^{\kappa_U \bar{\tau}} (\varepsilon_{\mathcal{M}}(0) + O(\Delta)) = e^{\kappa_U \bar{\tau}} (O(\delta^2) + O(\Delta)). \end{aligned}$$

It follows that $\varepsilon_{\mathcal{C}} \to 0$ as $\delta, \Delta \to 0$, which concludes the proof. ■

Remark 2. *The quantity $\bar{\tau}$ depends on the total geodesic distance D, which is fixed during a computation. If one wishes to compute an approximation of $W^u(x_0)$ for bigger D, while maintaining the same global error, then δ and Δ must be chosen smaller, because the global error normally increases with D (except when $U \subset V$).*

Remark 3. *In practice, we do vary Δ_i during a computation. Since the variation of Δ_i is related to the local curvature along geodesics, the overall interpolation error does not increase; see* [13, 21]. *Furthermore, we add and remove mesh points, but any required interpolation is done between points at most $\Delta_{\mathcal{F}}$ apart; see section 3.3.*

5.2. Lower semicontinuity. The main difficulty in proving that $\mathcal{C}(\delta, \Delta)$ converges to $\mathcal{S}_D$ (in the Hausdorff metric) lies in showing that the last $(k-1)$-sphere of $\mathcal{C}(\delta, \Delta)$ actually converges to the geodesic level set S_D. The problem is that the entire mesh changes as δ or Δ vary. Hence, it is not possible to consider the convergence of a specific mesh point and its geodesic distance. However, we believe this to be a technical difficulty, and we have the following.

Conjecture 4 (lower semicontinuity). *For given fixed $D > 0$ let $\mathcal{C}(\delta, \Delta)$ be the simplicial complex approximation of $\mathcal{S}_D$. Then for any $\varepsilon > 0$*

$$\mathcal{S}_D \subset \mathcal{N}_\varepsilon(\mathcal{C}(\delta, \Delta))$$

for all sufficiently small δ and Δ.

If we assume that the vector field, and hence $W^u(x_0)$, is at least C^2, then we can conclude that the algorithm does not bypass a part of $\mathcal{S}_D$. This is the case, because Theorem 1 ensures that $C(\delta, \Delta)$ lies in an ε-neighborhood of the C^2 manifold $\mathcal{S}_{\tilde{D}}$. For example, any bump is C^2, and $C(\delta, \Delta)$ must stay ε-close to the bump for δ and Δ small enough.

From a practical point of view this is the important result: the approximation $\mathcal{C}(\delta, \Delta)$ is ε-close to $W^u(x_0)$ and does not "jump over" parts of the manifold, provided the prespecified accuracy parameters were chosen small enough.

6. Conclusions. We described an algorithm for computing the global k-dimensional stable or unstable manifold of an equilibrium in an n-dimensional vector field for any $1 < k < n$. This method can also be used to compute the stable or unstable manifold of an invariant m-torus, provided this m-torus and an approximation of its linear stable and unstable directions can be found; for a method to compute the latter we refer to [3] and also to [25, 26].

Our algorithm approximates a first piece of the unstable manifold up to a given geodesic distance from the equilibrium (or invariant torus) by continuation of a family of appropriate boundary value problems. The computational error goes to zero as the tolerance parameters of the algorithm go to zero. Hence, the approximants are upper semicontinuous to $W^u(x_0)$. As a result of the smoothness of the manifold, the algorithm does not omit parts of $W^u(x_0)$, provided the tolerance parameters are sufficiently small.

The algorithm has been implemented for the case $k = 2$ and arbitrary n. At present, we solve the boundary value problems by using a shooting approach for which one boundary point lies on a piecewise linear curve. A genuine boundary value problem continuation will have to deal with the absence of smoothness at the mesh points. We are currently working on such an implementation by linking the code to AUTO [6].

While there are no theoretical obstructions to using the algorithm for $k \geq 3$, its implementation would be a serious effort. For the next case $k = 3$ the unstable manifold is approximated by a set of 2-spheres, which are represented by simplicial complexes consisting of triangles. The shells between neighboring 2-spheres consist of tetrahedra. The practical implementation of this data structure, including the operations of adding and removing mesh points, remains a challenge. Moreover, one encounters the general problem of visualizing a three-dimensional object. This makes it difficult to communicate the results, even though some packages, for example, Geomview [28], allow for inspection of three-dimensional manifold data.

It seems a fair assessment that in the foreseeable future growth algorithms for global manifolds will be restricted to the case $k = 2$ of two-dimensional manifolds. In fact, even algorithms for $k = 2$ are largely still in development and have not been used widely. As a consequence, we believe that computing two-dimensional manifolds in problems arising in applications is still in its infancy, with a lot of interesting work to be done.

Acknowledgments. We thank Sebius Doedel, John Guckenheimer, Mike Henderson, Andrew Stuart, and Alexander Vladimirsky for helpful discussions. Both authors gratefully acknowledge the hospitality and support of the Control and Dynamical Systems (CDS) Department at the California Institute of Technology, Pasadena, where part of this work was done.

REFERENCES

[1] R. H. Abraham and C. D. Shaw, *Dynamics—the Geometry of Behavior. Part* 3: *Global Behavior*, Aerial Press, Santa Cruz, CA, 1985.

[2] A. Arneodo, P. Coullet, E. Spiegel, and C. Tresser, *Asymptotic chaos*, Phys. D, 14 (1985), pp. 327–347.

[3] H. W. Broer, H. M. Osinga, and G. Vegter, *Algorithms for computing normally hyperbolic invariant manifolds*, Z. Angew. Math. Phys., 48 (1997), pp. 480–524.

[4] M. Dellnitz and A. Hohmann, *The computation of unstable manifolds using subdivision and continuation*, in Nonlinear Dynamical Systems and Chaos, H. W. Broer, S. A. Van Gils, I. Hoveijn, and F. Takens, eds., Progr. Nonlinear Differential Equations Appl. 19, Birkhäuser, Basel, 1996, pp. 449–459.
[5] M. Dellnitz and A. Hohmann, *A subdivision algorithm for the computation of unstable manifolds and global attractors*, Numer. Math., 75 (1997), pp. 293–317.
[6] E. J. Doedel, A. Champneys, T. Fairgrieve, Y. Kuznetsov, B. Sandstede, and X. Wang, *AUTO* 97: *Continuation and Bifurcation Software for Ordinary Differential Equations*, http://indy.cs.concordia.ca/auto/main.html, 1997.
[7] E. J. Doedel, *private communication*, IMA, Minneapolis, MN, 1997.
[8] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, 2nd ed., Springer-Verlag, New York, 1986.
[9] J. Guckenheimer and A. Vladimirsky, *A Fast Method for Approximating Invariant Manifolds*, preprint, Cornell University, Ithaca, NY, 2003. Available online from http://www.math.cornell.edu/~vlad/papers/InvMfolds/InvMfolds.pdf.
[10] J. Guckenheimer and P. Worfolk, *Dynamical systems: Some computational problems*, in Bifurcations and Periodic Orbits of Vector Fields, D. Schlomiuk, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993, pp. 241–277.
[11] J. Hauser and H. M. Osinga, *On the geometry of optimal control: The inverted pendulum example*, in Proceedings of the American Control Conference Vol. 2, Arlington, VA, 2001, pp. 1721–1726.
[12] M. E. Henderson, *Computing Invariant Manifolds by Integrating Fat Trajectories*, IBM Research Report RC22944, 2003.
[13] D. Hobson, *An efficient method for computing invariant manifolds of planar maps*, J. Comput. Phys., 104 (1993), pp. 14–22.
[14] J. G. Hocking and G. S. Young, *Topology*, 2nd ed., Dover Publications, New York, 1988.
[15] A. Jadbabaie, J. Yu, and J. Hauser, *Unconstrained receding-horizon control of nonlinear systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 776–783.
[16] M. E. Johnson, M. S. Jolly, and I. G. Kevrekidis, *Two-dimensional invariant manifolds and global bifurcations: Some approximation and visualization studies*, Numer. Algorithms, 14 (1997), pp. 125–140.
[17] M. E. Johnson, M. S. Jolly, and I. G. Kevrekidis, *The Oseberg transition: Visualization of global bifurcations for the Kuramoto-Sivashinsky equation*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 11 (2001), pp. 1–18.
[18] O. Junge, *Rigorous discretization of subdivision techniques*, in Proceedings of the International Conference on Differential Equations, B. Fiedler, K. Gröger, and J. Sprekels, eds., World Scientific, Singapore, 2000, pp. 916–918.
[19] E. N. Lorenz, *Deterministic nonperiodic flows*, J. Atmospheric Sci., 20 (1963), pp. 130–141.
[20] B. Krauskopf and H. M. Osinga, *Globalizing two-dimensional unstable manifolds of maps*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 8 (1998), pp. 483–503.
[21] B. Krauskopf and H. M. Osinga, *Growing 1D and quasi 2D unstable manifolds of maps*, J. Comput. Phys., 146 (1998), pp. 404–419.
[22] B. Krauskopf and H. M. Osinga, *Two-dimensional global manifolds of vector fields*, Chaos, 9 (1999), pp. 768–774.
[23] B. Krauskopf and H. M. Osinga, *Visualizing the structure of chaos in the Lorenz system*, Computers and Graphics, 26 (2002), pp. 815–823.
[24] Yu. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, Berlin, 1995.
[25] H. M. Osinga, *Non-orientable manifolds of periodic orbits*, in Proceedings of the International Conference on Differential Equations, B. Fiedler, K. Gröger, and J. Sprekels, eds., World Scientific, Singapore, 2000, pp. 922–924.
[26] H. M. Osinga, *Non-orientable manifolds in three-dimensional vector fields*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 13 (2003), pp. 553–570.
[27] J. Palis and W. De Melo, *Geometric Theory of Dynamical Systems*, Springer-Verlag, Berlin, 1982.
[28] M. Phillips, S. Levy, and T. Munzner, *Geomview: An interactive geometry viewer*, Notices Amer. Math. Soc., 40 (1993), pp. 985–988; software and the accompanying manual are available at http://www.geom.uiuc.edu/.

[29] W. C. Rheinboldt, *Methods for Solving Systems of Nonlinear Equations*, 2nd ed., CBMS-NSF Regional Conf. Ser. in Appl. Math. 70, SIAM, Philadelphia, 1998.
[30] S. H. Strogatz, *Nonlinear Dynamics and Chaos*, Addison–Wesley, Reading, MA, 1994.
[31] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.
[32] M. Spivak, *A Comprehensive Introduction to Differential Geometry, Volume* 1, 2nd ed., Publish or Perish, Houston, TX, 1979.
[33] A. J. Van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, Lecture Notes in Control and Inform. Sci. 218, Springer-Verlag, Berlin, 1994.
[34] P. Worfolk, *Private communication*, Geometry Center, Minneapolis, MN, 1997.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 2, No. 4, pp. 570–608

The Forced van der Pol Equation II: Canards in the Reduced System*

Katherine Bold†, Chantal Edwards‡, John Guckenheimer§, Sabyasachi Guharay¶, Kathleen Hoffman‡, Judith Hubbard‖, Ricardo Oliva§, and Warren Weckesser**

Abstract. This is the second in a series of papers about the dynamics of the forced van der Pol oscillator [J. Guckenheimer, K. Hoffman, and W. Weckesser, *SIAM J. Appl. Dyn. Syst.*, 2 (2003), pp. 1–35]. The first paper described the *reduced system*, a two dimensional flow with jumps that reflect fast trajectory segments in this vector field with two time scales. This paper extends the reduced system to account for *canards*, trajectory segments that follow the unstable portion of the slow manifold in the forced van der Pol oscillator. This extension of the reduced system serves as a template for approximating the full nonwandering set of the forced van der Pol oscillator for large sets of parameter values, including parameters for which the system is chaotic. We analyze some bifurcations in the extension of the reduced system, building upon our previous work in [J. Guckenheimer, K. Hoffman, and W. Weckesser, *SIAM J. Appl. Dyn. Syst.*, 2 (2003), pp. 1–35]. We conclude with computations of return maps and periodic orbits in the full three dimensional flow that are compared with the computations and analysis of the reduced system. These comparisons demonstrate numerically the validity of results we derive from the study of canards in the reduced system.

Key words. van der Pol oscillator, hybrid dynamical system, bifurcations, canards, chaotic invariant set

AMS subject classifications. 37C10, 37C27, 37G15

DOI. 10.1137/S1111111102419130

1. Introduction.

We study the forced van der Pol oscillator [24] in the form

$$\begin{aligned} \varepsilon\dot{x} &= y + x - \tfrac{x^3}{3}, \\ \dot{y} &= -x + a\sin(2\pi\theta), \\ \dot{\theta} &= \omega, \end{aligned} \tag{1.1}$$

where θ is a cyclic variable belonging to $\mathbb{R}/\mathbb{Z}$. The critical manifold of this system is the surface C defined by $\dot{x} = 0$; i.e., $y = x^3/3 - x$. We divide the critical manifold into $C_s = \{(x, y, \theta) \in C \,|\, |x| > 1\}$, $C_u = \{(x, y, \theta) \in C \,|\, |x| < 1\}$, and the two fold curves $S_{\pm 1}$, where

*Received by the editors December 3, 2002; accepted for publication (in revised form) by M. Golubitsky July 21, 2003; published electronically December 22, 2003. This research was partially supported by the National Science Foundation and the Department of Energy. The results were obtained during a Research Experiences for Undergraduates Program at Cornell University sponsored by the National Science Foundation.
http://www.siam.org/journals/siads/2-4/41913.html
†Mathematics Department, University of Texas, Austin, TX 78712.
‡Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250 (khoffman@math.umbc.edu).
§Mathematics Department, Cornell University, Ithaca, NY 14853 (gucken@cam.cornell.edu, RAOliva@lbl.gov).
¶Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544.
‖Geology Department, California Institute of Technology, Pasadena, CA 91125.
**Mathematics Department, Colgate University, Hamilton, NY 13346 (wweckesser@mail.colgate.edu).

$S_\xi = \{(x, y, \theta) \in C \,|\, x = \xi\}$. The slow flow on the critical manifold solves the two dimensional system of differential equations derived from (1.1) by setting $\varepsilon = 0$, differentiating $y = x^3/3 - x$ to obtain $\dot{y} = (x^2 - 1)\dot{x}$ on the critical manifold and rescaling time:

$$\begin{aligned} \theta' &= \omega(x^2 - 1), \\ x' &= -x + a\sin(2\pi\theta). \end{aligned} \tag{1.2}$$

The reduced system is given by augmenting the slow flow with the rules that trajectories jump from the fold curve S_1 to S_{-2} and from the fold curve S_{-1} to S_2, in each case preserving the value of θ. The reduced system with its jumps is a *hybrid* dynamical system [2]. We define a *half return map* H in terms of candidates [4, p. 68] and the symmetry operation $T(x, \theta) = (-x, \theta + 0.5)$ of the slow flow. Specifically, $H : S_2 \to S_2$ is defined by following the slow dynamics from S_2 to S_1, jumping with constant θ from S_1 to S_{-2}, and then applying the symmetry T. Details of these constructions can be found in [14], together with extensive analysis of properties of H. Additionally, in section 4, we formulate a *half Poincaré map* of the full system and a *singular half Poincaré map* of the reduced sytem to further clarify the relationship between solutions of the half return map and solutions of the full forced van der Pol equation.

"Fenichel theory" [10] and a theorem of Levinson [17] imply that the reduced system gives approximations to trajectories of (1.1) that have *nondegenerate slow-fast decompositions* [13]. These are trajectories that approach the fold curves $S_{\pm 1}$ at regular points of the slow flow. When $|a| < 1$, all points on the fold curves $S_{\pm 1}$ are regular and H is a diffeomorphism of the circle S_2. When $|a| > 1$, there are four equilibria of the slow flow on the fold curves, called *folded equilibria*. A symmetric pair of the folded equilibria are always saddles. The properties of the system (1.1) are more complicated in the vicinity of the folded equilibria. Benoît [4] has studied the flow of slow-fast vector fields with two slow variables and one fast variable in the vicinity of folded equilibria. We review his results for folded saddles in the next section; they play an important role in our extension of the reduced system and its return map. The analysis of folded nodes is incomplete, and this leads to a gap in our description of the reduced system for small parameter ranges near $a = 1$ where folded nodes exist.

There are trajectories of the three dimensional system (1.1) that cross the fold curve near the folded saddles and continue along the unstable portion of the slow manifold. The trajectory segments that follow C_u are called *canards*. Canards may end at any place along the unstable portion of the slow manifold, turning abruptly and jumping parallel to the fast direction to one of the stable sheets of the slow manifold. If the trajectories return to the sheet of the slow manifold from which the canard started, we call them *jump back canards*. If they jump to the opposite sheet of the slow manifold from their origin, we call them *jump away canards*. We call the trajectories that follow the unstable portion of the slow manifold to the fold line without jumping *maximal canards*. As $\varepsilon \to 0$, the initial points for a whole family of canards approach the same point. To represent this behavior in the reduced system, we allow the flow to have multiple trajectories emanating from the same point. Specifically, along a canard orbit, there are three possibilities at each time: continuing to flow along the canard or jumping to either side of the unstable sheet of the slow manifold. We describe in section 3 how the dynamics of the canard orbits can be incorporated into extensions of the half return map. Formally, we extend the concept of a discrete dynamical system from maps of a space to

relations for that space. Iterations of one dimensional mappings are sufficiently constrained in their properties that we can still analyze the properties of the dynamical systems defined by these relations. Intuitively, we regard the extensions as segments of the graph of H whose domain has infinitesimal width and large slope.

The extended half return map is essentially a one dimensional discrete dynamical system. The dynamics of one dimensional mappings can be analyzed in terms of *symbolic dynamics* (or kneading theory) [20]. Briefly, the domain of a mapping is partitioned into intervals of monotonicity. These are intervals on which the map is continuous and strictly increasing or decreasing. A symbol is assigned to each interval, and trajectories are mapped to symbol sequences that encode which intervals in the partition contain successive iterates of the initial point. The extreme values of the intervals of monotonicity are called *critical values* and their symbol sequences are called the *kneading data* of the mapping. The full set of symbol sequences associated with trajectories of the mapping can be reconstructed from the kneading data by simple rules. This allows one to almost fully characterize the nonwandering set of mappings from the kneading data, a task that can be completed within certain classes of mappings [12]. Bifurcations in a family of mappings occur when the kneading data change. This happens in a fashion that is constrained by the one dimensionality of the dynamics, and a great deal is known about "universal" patterns that occur in the bifurcations of one dimensional maps [8]. Each extreme value can be regarded as an independent parameter in determining the dynamics of maps with specified critical points and discontinuities. We apply this theory to the extended half return map of the reduced mapping for the forced van der Pol oscillator. Given the number of discontinuities of H, there are fewer types of topological equivalence classes for the extended half return maps than one might expect from the kneading theory. The half return maps may have several points of discontinuity, but there are only two possible values for left and right limits of H as θ approaches a point of discontinuity. Moreover, the endpoints of the canard extensions of the intervals are mapped to the same point by H. Additionally, there is at most one interior local maximum and one interior local minimum of H, so these are the only extreme points that are not points of discontinuity. These points exist precisely when $|a| > 2$ and the points of tangency of the slow flow with $S_{\pm 2}$ do not lie in the stable manifold of the folded saddle.

The existence of the chaotic invariant sets in the forced van der Pol system was discovered by Cartwright and Littlewood [6, 7, 18, 19]—a seminal event in the history of dynamical systems. One of the objectives of this work is to develop a more comprehensive understanding of the nonwandering sets of the forced van der Pol system and how these change with the parameters (a, ω, ε). Vestiges of the chaotic invariant sets remain in the dynamics of the extended half return maps. Based upon analysis of H and its extensions, we give a conjectural description of the complete nonwandering sets of the forced van der Pol oscillator for many parameter values where there are chaotic invariant sets. When the trajectories of the critical values of H approach stable periodic orbits and there are no neutrally stable periodic orbits, we believe that the dynamics of the van der Pol system for nearby parameter values are structurally stable. We identify simple examples where the analysis of the half return map indicates that this behavior occurs and describe the topological equivalence classes of the nonwandering sets of these examples as subshifts of finite type. We present a conjectural picture of the relationship between the nonwandering sets of the extended H and the nonwandering sets of

the system (1.1). Additional analysis is needed to prove these conjectures, but our description gives a detailed geometric perspective on the classical studies of Cartwright and Littlewood that is global in both the phase space and parameter space of the forced van der Pol oscillator.

2. Canards in the reduced system. The reduced system defined in [14] is a hybrid dynamical system given by the slow flow of system (1.2) in the regions of the (θ, x) cylinder with $|x| \geq 1$ together with discrete transition maps from the circles $S_{\pm 1}$ to the circles $S_{\mp 2}$, since $y = x^3/3 - x\,|_{x=\pm 1} = \mp 2/3 = x^3/3 - x\,|_{x=\mp 2}$. When $a > 1$, there are equilibrium points of the slow flow at the points $(\theta, x) = (\pm \sin^{-1}(1/a)/2\pi, \pm 1)$ on the boundary of the region in which the reduced system is defined. Two of these equilibrium points are saddles for the slow flow. The other two equilibrium points are stable nodes if $1 < a < \sqrt{1 + 1/(16\pi\omega)^2}$ and stable foci if $a > \sqrt{1 + 1/(16\pi\omega)^2}$. These equilibria are called folded equilibria because they lie on the fold curves of the critical manifold and do not correspond to equilibria of the van der Pol system (1.1).

We examine the folded saddle $p_{1s} = (\theta_{1s}, 1)$ lying on $x = 1$ and call its stable and unstable manifolds for the slow flow (1.2) W_s and W_u, respectively. We denote the first intersection of $W_u \cap C_s$ by $p_{1u} = (\theta_{1u}, 1)$ and the first intersection of $W_s \cap C_u$ with one of the circles $S_{\pm 1}$ by $p_m = (\theta_{1m}, \pm 1)$. In the van der Pol system (1.1), some trajectories arriving in the vicinity of p_{1s} continue a slow motion near $W_s \cap C_u$ rather than jumping to the stable sheet of the critical manifold near S_{-2}. These trajectories are called *canards*. Benoît [4] analyzed the dynamics of folded saddles in slow-fast systems with two slow variables and one fast variable. His analysis describes the limit properties of canard trajectories as $\varepsilon \to 0$. (See also the discussions of Mischenko et al. [21] and Szmolyan and Wechselberger [23].) We summarize here the results that provide the foundation for our analysis of canards in the reduced system.

Fix a neighborhood of p_{1s} that contains a ball B of radius $O(1)$ centered at the folded saddle. Benoît proves that there is a constant $c > 0$ so that the sets of trajectories that remain within distance $O(\varepsilon)$ from $C \cap B$ form two bundles, each with a cross-section at the folded saddle whose diameter is $O(\exp(-c/\varepsilon))$. As $\varepsilon \to 0$, one of these bundles of slow orbits approaches W_s and the other approaches W_u. Note that the time rescaling used to obtain the slow flow reverses time on C_u, so the trajectory in $W_s \cap C_u$ lies close to trajectories of the van der Pol system flowing away from the folded saddle and not toward it. Benoît further proves that the strong stable manifold of $W_s \cap C_u$ and the strong unstable manifold of $W_s \cap C_u$ meet transversally near the folded saddle. Denote by γ_u the segment of $W_s \cap C_u$ between p_{1s} and p_{1m}. Given a point $p \in \gamma_u$, Benoît's results imply that there are canards in the van der Pol system that emanate from the folded saddle and then jump from γ_u in an $O(\varepsilon)$ neighborhood of C_u near p. Thus we model the limits of canards in the reduced system as the canard *umbra* of γ_u. We define the umbra of γ_u to be a pair of curves on C_s swept out by the intersection of lines parallel to the x axis through points of γ_u. If $p = (x_0, y, \theta)$, the corresponding points on the umbra have the form (x_1, y, θ), where $x_1^3 - 3x_1 = x_0^3 - 3x_0$ or $x_1^2 + x_0 x_1 + x_0^2 - 3 = 0$.

Figure 1 shows structures of the reduced system for parameter values $(a, \omega) = (1.1, 1.57)$. Since trajectories make more than a single circuit in the θ direction, two fundamental domains in θ are plotted, with repetitions of the folded equilibria drawn. The stable and unstable manifolds of the folded saddle are drawn in black. The segment γ_u of W_s is drawn as a heavy black line on one copy of W_u. The two umbra of γ_u are drawn as heavy magenta and

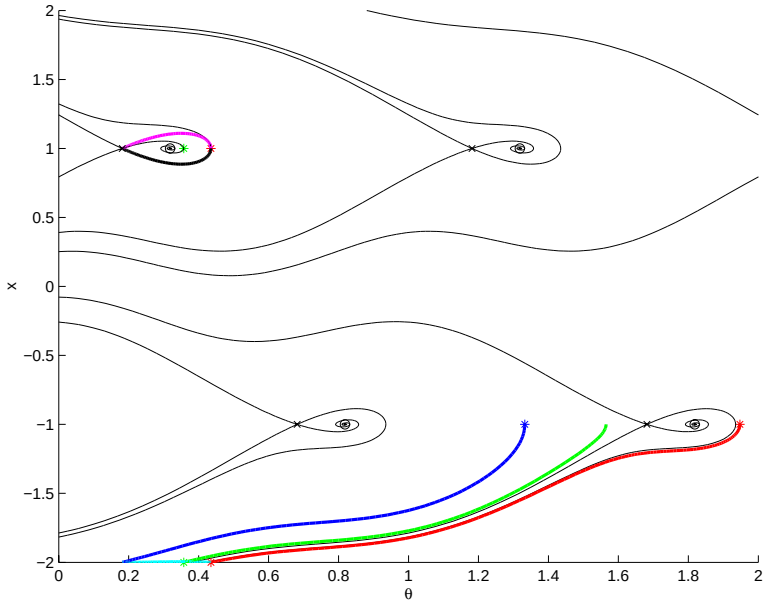


Figure 1. *The phase portrait of the reduced system for the parameter values* $(a, \omega) = (1.1, 1.57)$. *The stable and unstable manifolds of the saddle are drawn in black. The segment* γ_u *is represented by the thick black line and the two umbral curves are drawn in magenta and cyan. The blue and red lines represent the trajectories at the ends of the cyan umbral curve and the green line is the trajectory of* p_{1u}.

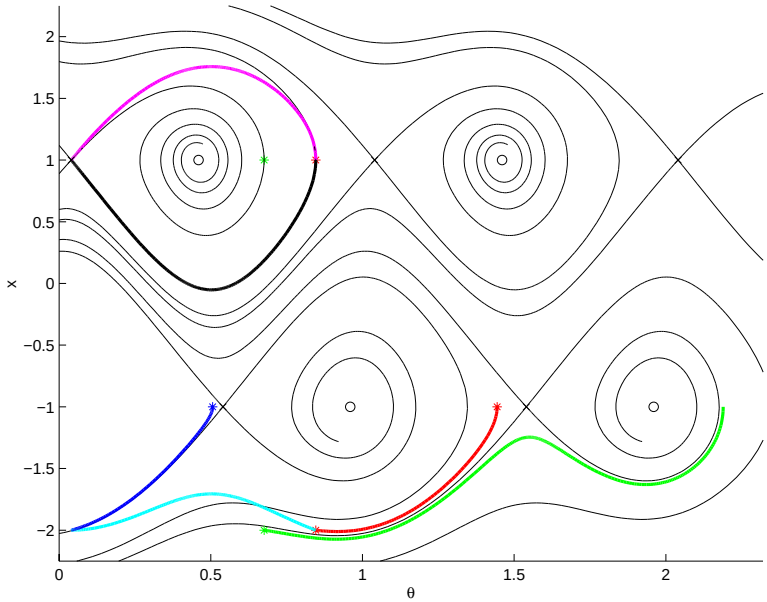


Figure 2. *The phase portrait of the reduced system for the parameter values* $(a, \omega) = (4, 1.51)$. *Note that the color coding is the same as represented in Figure* 1.

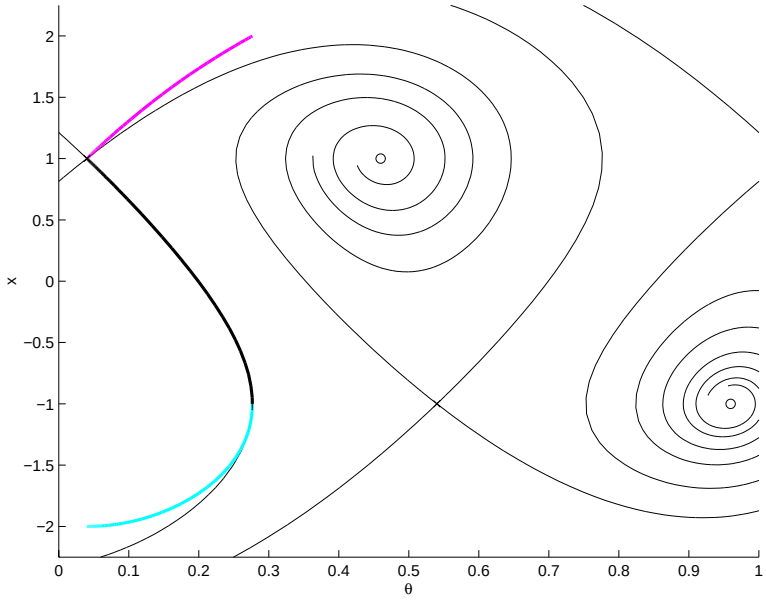


Figure 3. *The phase portrait of the reduced system for the parameter values* $(a, \omega) = (4, 0.5)$. *Note that the color coding is the same as represented in Figure* 1.

cyan curves. The point p_{1u} at the intersection of W_u with S_1 is a green asterisk, and the continuation of p_{1u} in the reduced system is plotted in green to its intersection with S_{-1}. The trajectories starting at the ends of the umbra in the region $x \leq -1$ are drawn in blue and red. Figure 2 shows structures of the reduced system for parameter values $(a, \omega) = (4, 1.51)$. The color coding of the structures is the same as in Figure 1. Finally, Figure 3 shows structures of the reduced system for parameter values $(a, \omega) = (4, 0.5)$. Note that the maximal canard point for these parameter values is on S_{-1} rather than on S_1 as it is in Figures 1 and 2.

3. The extended half return map.

This section defines a multivalued extension to the half return map H whose new values reflect the limit properties of canards in the van der Pol system (1.1). We assume that we are in the parameter region in which the maximal canard point p_m is on S_1 rather than on S_{-1}. The definition of the extension is not completely straightforward because the canards with jumps that decrease the value of x yield trajectories that hit the cross-section S_{-1} before they hit S_1. Therefore, we introduce an artifice of flowing backward along the slow flow from the canard umbra in $x < -1$ to the cross-section S_{-2}, where we can apply the symmetry to obtain a value for the extended half return map. We state and prove a fundamental lemma before describing details of this construction.

Lemma 3.1. *The slow flow vector field is transverse to the umbra of the canard curve* $\gamma_u \subset W_s \cap C_u$.

Proof. Let (θ, x_0) be a point on $\gamma_u \subset W_s \cap C_u$. Then the corresponding points on the umbra have the form (θ, x_1) with $x_1^2 + x_0 x_1 + x_0^2 - 3 = 0$. The slope of the vector field at (θ, x_1) is

$$\sigma_1 = \frac{-x_1 + a \sin(2\pi\theta)}{\omega(x_1^2 - 1)},$$

while the slope of the umbra is

$$\tau_1 = \left(\frac{dx_1}{dx_0}\right)\left(\frac{-x_0 + a\sin(2\pi\theta)}{\omega(x_0^2 - 1)}\right) = \left(\frac{x_1 + 2x_0}{2x_1 + x_0}\right)\left(\frac{-x_0 + a\sin(2\pi\theta)}{\omega(x_0^2 - 1)}\right).$$

Simplifying $\sigma_1 - \tau_1$ (with the help of the computer program Maple), we find that

$$\sigma_1 - \tau_1 = \frac{\pm 3}{\omega\sqrt{12 - 3x_0^2}} \neq 0;$$

the sign depends upon which root x_1 we take. Since the two slopes are different, the vector field at (θ, x_1) is transverse to the umbral curves.

This lemma implies that distinct points on each umbral curve lie on distinct trajectories. To define the extension of the half return map H for the umbral curve that lies in the half cylinder $x \geq 1$, we follow the trajectories of the slow flow from a point (θ_0, x_0) on the umbral curve to its intersection with S_1 at the point $(\theta_1, 1)$ and then define H for this canard trajectory to be $\theta_1 + 0.5$. This gives the same value of θ as jumping from $(\theta_1, 1)$ to $(\theta_1, -2)$ and then applying the symmetry of the slow flow. The extension of H adds a vertical segment to its graph, making H multivalued. The transversality lemma above implies that each value $\theta_1 + 0.5$ on the extended graph of H corresponds to orbits containing a canard with a specific jump point, namely, (θ_0, x_0). Thus we shall regard the extension of H as a monotone segment of infinite slope when analyzing the symbolic dynamics of the extended H in the next section. With this definition, we obtain a limit of trajectories in the van der Pol system consisting of a canard that flows to a jump point along W_s, followed by a jump to $(x_0, x_0^3/3 - x_0, \theta_0)$ on the critical manifold, followed by a trajectory segment of the slow flow that reaches $x = 1$ at $(1, -2/3, \theta_1)$ and then jumps to $x = -2$.

Defining the extension of H for the umbral curve in the half cylinder $x \leq -1$ is less straightforward. When points jump from W_s to a point (θ_0, x_0) on the umbral curve in the half cylinder $x \leq -1$, the trajectory from (θ_0, x_0) has "skipped" an intersection with the circle S_{-2}. The definition of the half return map was based upon trajectories reaching the half cylinder $x \leq -1$ at $x = -2$ so that the symmetry could be used to map them back to S_2. We resolve this difficulty by following the trajectory from (θ_0, x_0) *backward* to its first intersection with the circle S_{-2} at $(\theta_1, -2)$. The extension of H is defined to give a value $\theta_1 + 0.5$ for this point on the umbral curve. The qualitative properties of the backward flow from the umbral curve to S_{-2} depend upon whether there are zero, one, or two tangencies of the slow flow with S_{-2} in the interval (θ_{1s}, θ_m) "covered" by the umbral curve. The tangencies occur at values of $\theta = \arcsin(-2/a)$. If there are no tangencies, then the set of values for $\theta_1 + 0.5$ is the entire interval (θ_{1s}, θ_m). If there is a single point of tangency, then the slow flow has $x'(\theta_m, -2) < 0$, and the set of values for $\theta_1 + 0.5$ is a connected interval whose right endpoint is the first intersection of the backward trajectory of the slow flow through $(\theta_m, -2)$ with S_{-2}. If there are two points of tangency, then the set of $\theta_1 + 0.5$ consists of two intervals with a gap. The gap is an interval whose endpoints both lie on the slow flow trajectory passing through the right-hand tangency point on S_{-2}. In this case, note that the H image of this pair of intervals is connected; the endpoints of the gap both have the same H value. The transversality lemma again implies that each value $\theta_1 + 0.5$ on the extended graph of H

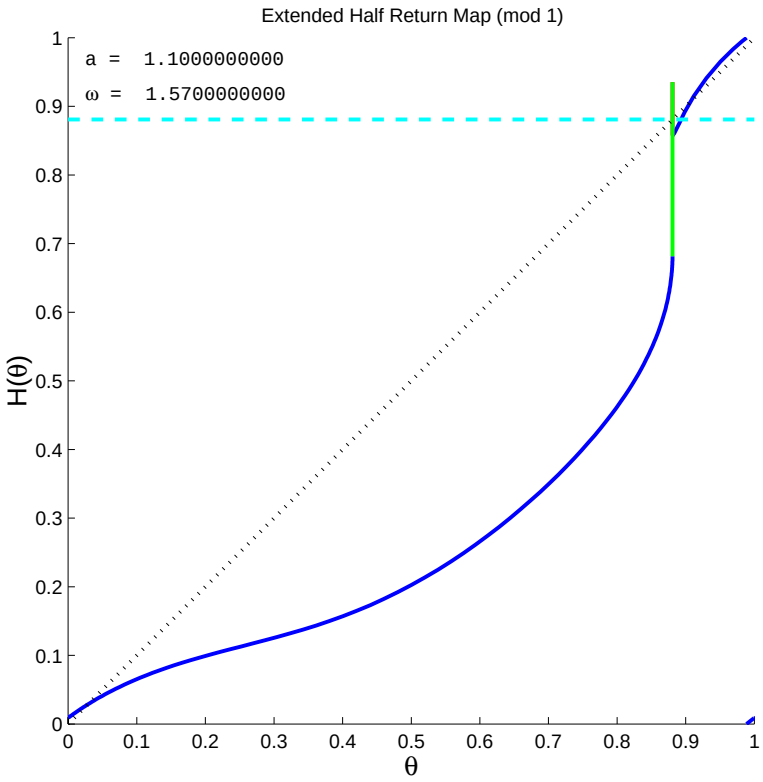


Figure 4. *A plot of the extended half return map for the parameter values* $(a, \omega) = (1.1, 1.57)$. *The canard extensions of* H *are drawn as vertical segments.*

corresponds to orbits containing a canard with a specific jump point. The limit trajectory for the van der Pol system consists of a canard that flows to a jump point along W_s, followed by a jump to $(x_0, x_0^3/3 - x_0, \theta_0)$ on the critical manifold, followed by a trajectory segment that lies on the trajectory of the slow flow through $(-1, 2/3, \theta_1)$, but with the initial point of the segment being $(x_0, x_0^3/3 - x_0, \theta_0)$ and the final point of the segment being on S_{-1}.

Figures 4, 5, and 6 show three examples of the extended half return map for parameter values $(a, \omega) = (1.1, 1.57)$, $(4, 1.51)$, and $(1.02, 1.19)$, respectively. When $1 < a < 2$, there is a single point of discontinuity and no local maximum or minimum of H. Therefore, the values of the maximal canards for the left and right extensions at the discontinuity point have equal values. The canard extensions of H are drawn as (overlapping) red and green vertical segments. When $(a, \omega) = (1.1, 1.57)$, H itself has two fixed points and there are two additional fixed points on the canard extensions.

At parameter values $(a, \omega) = (4, 1.51)$, H has three points of discontinuity. It also has a local maximum and local minimum that are not at points of discontinuity. In the flow of the reduced system, $x' < 0$ at some of the points on the umbra of the canards in $x < -1$. Therefore, the values of the maximal canards on the extensions of H do not agree. However, note that the images of the maximal canard values are the same—a point that lies on the branch of H with a local minimum. Also, the canard extensions at each of the three discontinuity points

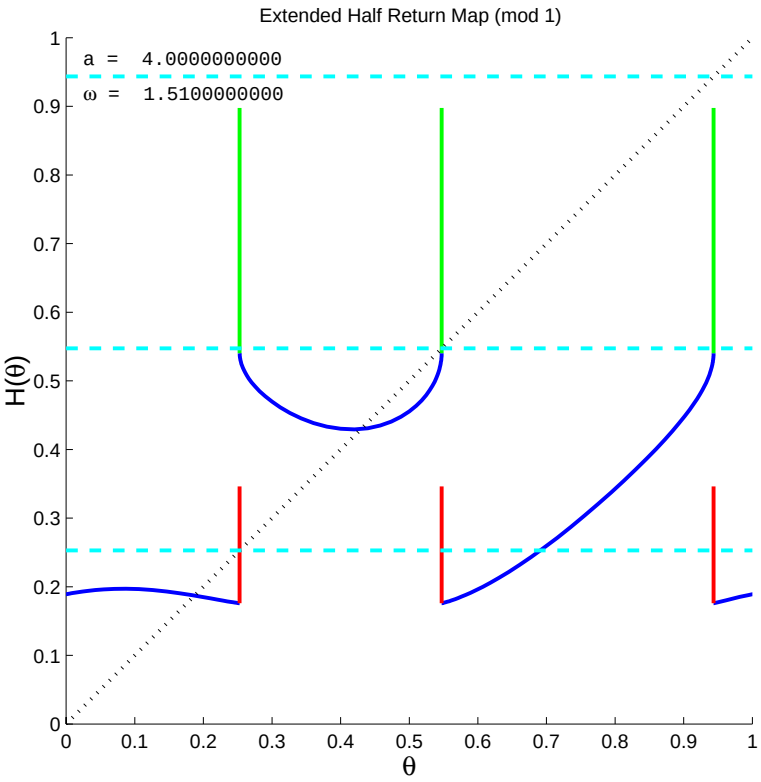


Figure 5. *A plot of the extended half return map for the parameter values* $(a, \omega) = (4, 1.51)$. *The canard extensions of* H *are drawn as red and green vertical segments.*

have the same length. Both the branch with a local minimum and the branch with a local maximum have stable fixed points. In addition there are two fixed points on two of the canard extensions.

At parameter values $(a, \omega) = (1.02, 1.19)$, H has no fixed points at all. However, the second iterate of H, displayed in Figure 7, has features that resemble the half return map for $(a, \omega) = (1.1, 1.57)$. There are four fixed points of H^2 outside its canards, coming from two periodic orbits of period 2. In addition, H^2 has fixed points on each of its canard segments.

4. Poincaré maps. An alternative method for studying canard solutions of the forced van der Pol equation is to use the idea of Poincaré maps. We define a half Poincaré map Φ_ε for (1.1) and a corresponding singular half Poincaré map Φ_0 for the reduced system. Let Σ_1 be the half cylinder $\{(x, y, \theta) \,|\, x = -1, y < 2/3 - \delta\}$ for some small $\delta > 0$ (see Figure 8). The system (1.1) is invariant under the symmetry transformation $T(x, y, \theta) = (-x, -y, \theta + 1/2)$. Let $\Sigma_2 = T(\Sigma_1)$; that is, $\Sigma_2 = \{(x, y, \theta) \,|\, x = 1, y > -2/3 + \delta\}$. As long as we consider ε sufficiently small and parameter values (a, ω) in the interior of the region M_1 (defined in section 6), the flow map from Σ_1 to Σ_2 composed with the symmetry transformation is a well-defined map Φ_ε from Σ_1 to itself. In particular, the trajectories with initial conditions on Σ_1 cross the cylinder $x = 1$ at points of Σ_2; canards that jump away do so with $y > -2/3 + \delta$.

The *singular half Poincaré map* $\Phi_0 : \Sigma_1 \to \Sigma_1$ of the reduced system is composed of a

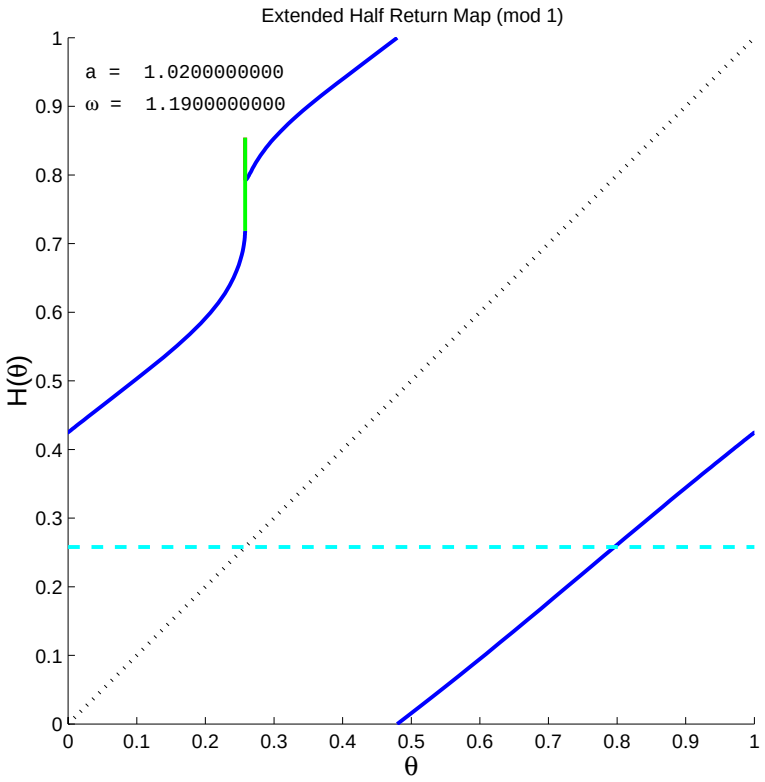


Figure 6. *A plot of the extended half return map for the parameter values* $(a, \omega) = (1.02, 1.19)$. *The canard extensions of* H *are drawn as vertical segments.*

projection along the x direction onto C_s, followed by the slow flow on C_s to the fold line S_{-1}, and a projection along the x direction from S_{-1} to Σ_2. The symmetry transformation T maps Σ_2 back to Σ_1. This map is singular; all points in Σ_1 map to the line $x = -1$, $y = -2/3$ in Σ_1. This definition of Φ_0 breaks down on the stable manifold of the folded saddle. These points do not return, but they have canard extensions that do. So we extend the definition of Φ_0 to account for the canards. As explained above, canards form on trajectories sufficiently close to the stable manifold of the saddle. In the van der Pol flow (1.1), let $\kappa_\varepsilon \subset \Sigma_1$ be the set of points for which canards will form during the flow from Σ_1 to Σ_2. When these trajectories finally cross Σ_2, they have either jumped from a point near C_u to Σ_2 without returning to a neighborhood of C_s (the jump away canards), or they have jumped back to a neighborhood of C_s and then jumped a second time, passing the fold S_{-1} between θ_{1u} and θ_m (the jump back canards). For $0 < \varepsilon \ll 1$, κ_ε is a thin strip. For the reduced system and $\varepsilon = 0$, we define κ_0 to be curve in Σ_1 that projects onto the stable manifold of the saddle. To account for canards in Φ_0, we make Φ_0 a multivalued relation. For any $p \in \kappa_0$, we define $\Phi_0(p)$ to be the curve made up of the projections of γ_u onto Σ_2 (representing jump away canards) and the projection of the segment (θ_{1u}, θ_m) on the fold line S_{-1} onto Σ_2 (representing jump back canards). In Figure 9, we give a qualitative sketch of such a Poincaré map. Figures 10 and 11 show the map for $(a, \omega) = (1.1, 1.57)$ and $(a, \omega) = (4.0, 1.51)$.

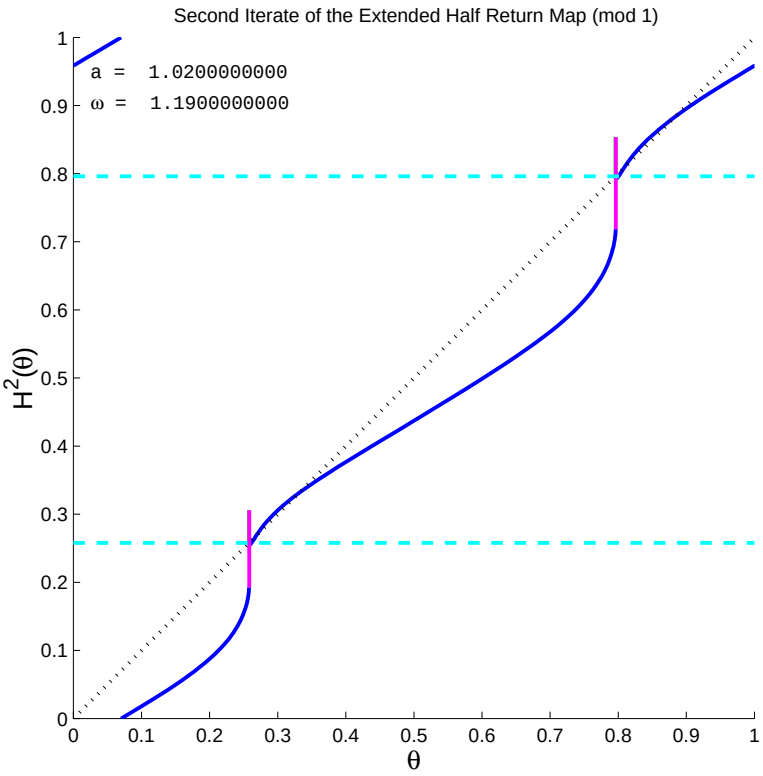


Figure 7. *A plot of the second iterate of the extended half return map for the parameter values* $(a, \omega) = (1.02, 1.19)$.

In Figure 12, we show four examples of singular Poincaré maps where there are two canard fixed points. These come from parameter values (a, ω) where we expect the van der Pol system will have chaotic invariant sets for small $\varepsilon > 0$. The stable manifolds (red) and images of the canards (cyan) give approximations to the expected directions of the stable and unstable bundles in a hyperbolic structure for these invariant sets. In the next section, we use the insight obtained from the singular half Poincaré map in numerical calculations with the three dimensional system with small values of $\varepsilon > 0$.

5. Symbolic dynamics. The extensions to the half return map defined in the previous section consist of two vertical segments that we add to the graph of H at each point of discontinuity. Here we explain how to use symbolic dynamics of one dimensional maps to analyze the extended half return map. The graph of the map is a multivalued function, or relation, so we define trajectories for such objects. We identify sets of trajectories that can be used to reconstruct trajectories with canards in the reduced system that we conjecture are limits as $\varepsilon \to 0$ of trajectories with canards in the forced van der Pol system (1.1). We begin by recalling some of the topological theory of iterations of one dimensional maps, following Milnor and Thurston [20]. We then describe how we extend this theory to account for multivalued maps that have vertical "monotone" segments and discontinuities like those of our extended half return map. We end the section with descriptions of the nonwandering

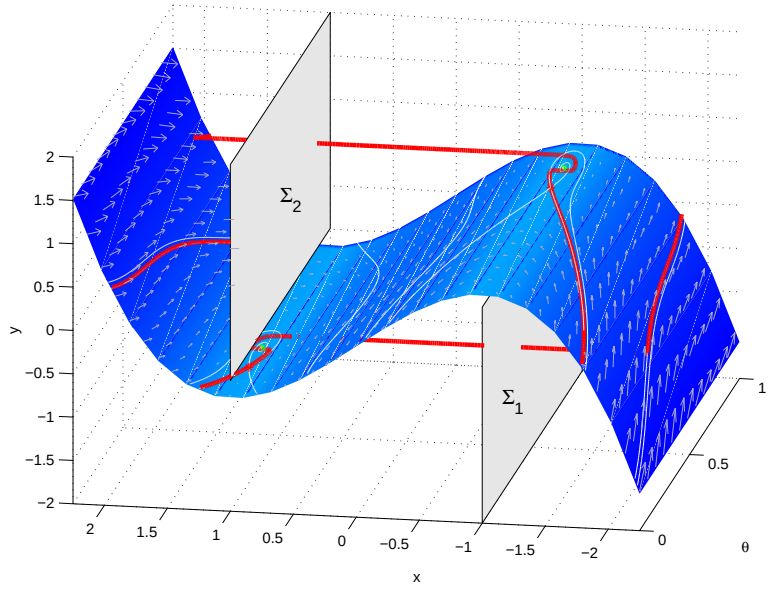


Figure 8. *The blue surface is the critical manifold C. The white lines and grey arrows in C are integral curves and the vector field of the slow flow, respectively. The red curve is an unstable periodic orbit of (1.1), computed by AUTO, for the parameter values $\varepsilon = 0.0001$, $a = 1.0888485842$, and $\omega = 1.55$.*

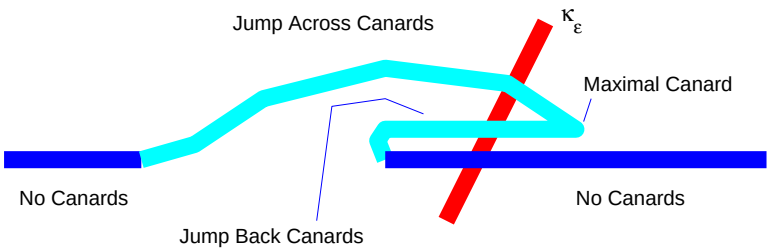


Figure 9. *A cartoon of the image of Σ_1 under the Poincaré map Φ_ε. The union of the dark blue and cyan curves represents $\Phi_\varepsilon(\Sigma_1)$. The red line represents κ_ε, the subset of Σ_1 for which canards form during the flow from Σ_1 to Σ_2, and the cyan segment represents $\Phi_\varepsilon(\kappa_\varepsilon)$. (The vertical separation between the cyan segment labeled "Jump Back Canards" and the dark blue segment on the right is greatly exaggerated.)*

sets of the examples displayed in the previous section.

Let $H : I \to I$ be a *piecewise monotone map* of the interval I. We also allow the possibility that I is the circle S_1. Here piecewise monotone means that there are points $c_0, \ldots, c_n$ such that, on each element of the partition $I = I_1 \cup I_2 \cup \cdots \cup I_n$; $I_j = [c_{j-1}, c_j]$ (with $c_0 = c_n$ in case I is a circle), H has a continuous strictly monotone extension to each I_j. Note that this definition allows H to be discontinuous at the c_j. We assign symbol sequences, called *itineraries*, to each x by $A(x) = \{a_i\}$ with $a_i = j$ when $H^i(x) \in I_j$. The *shift* map σ on sequences is defined by $\sigma(\bar{a}) = \bar{b}$ when $b_i = a_{i+1}$. The best way to treat the mesh points c_j of the partition in this definition is problematic. Since we shall look at multivalued extensions of the map that send endpoints to entire intervals, this issue is discussed at length below. Each

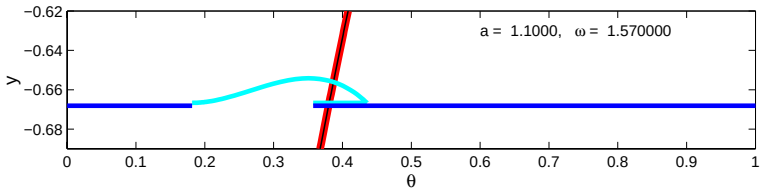


Figure 10. *The singular Poincaré map for* $(a, \omega) = (1.1, 1.57)$. *The red curve is* κ_0.

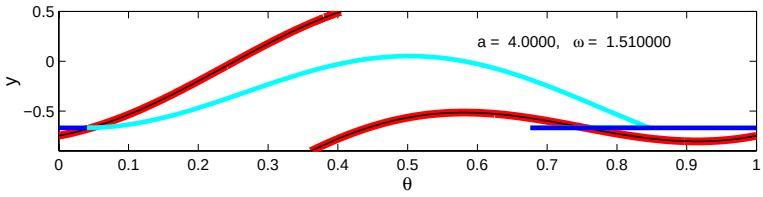


Figure 11. *The singular Poincaré map for* $(a, \omega) = (4.0, 1.51)$. *The red curve is* κ_0.

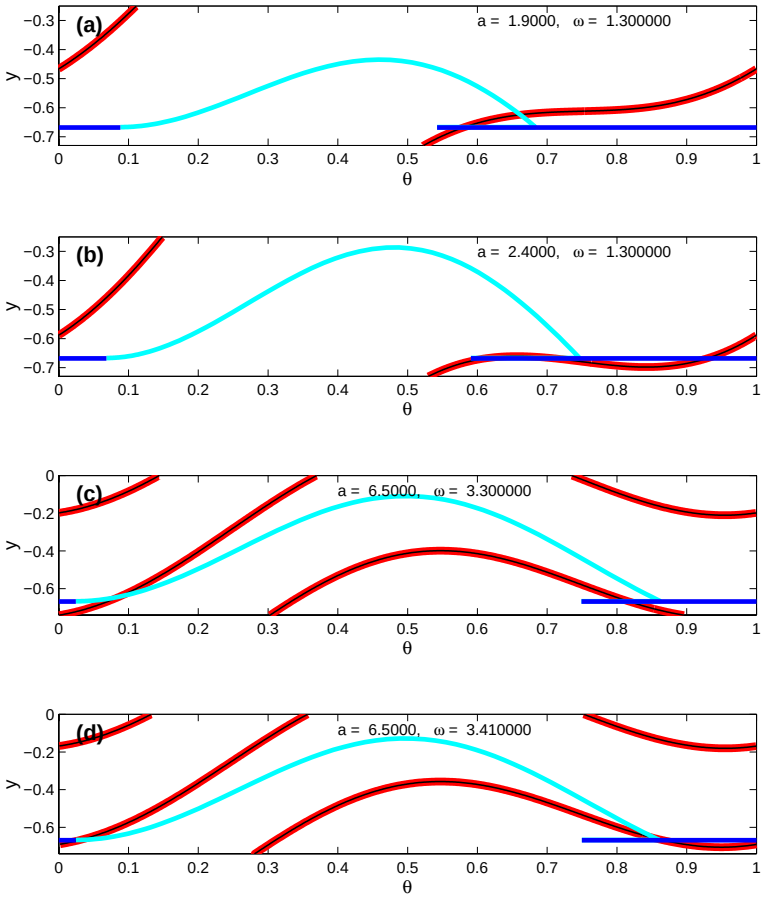


Figure 12. *Singular Poincaré maps that show the four cases where* $\Phi_0(\kappa_0)$ *intersects* κ_0 *twice.*

interval I_j has a sign $s(j) = \pm 1$ defined by whether the map H is increasing or decreasing on I_j. These signs are used to define an order on itineraries: $\{a_i\} < \{b_i\}$ if the sequences differ and for the first index k with $a_k \neq b_k$, $s(a_0)\cdots s(a_{k-1})a_k < s(b_0)\cdots s(b_{k-1})b_k$. Note that the two signs in this inequality are the same since the sequences agree up to their kth terms. The map H^k is increasing or decreasing on the interval

$$\bigcap_{i=0}^{k-1} H^{-i}(I_{A_i})$$

as $s(a_0)\cdots s(a_{k-1}) = \pm 1$. This implies that the order of itineraries is consistent with the order of points of the interval; if $x < y$, then $A(x) \leq A(y)$.

There is a complementary theorem to the result that the ordered itineraries are consistent with the linear order of points on the interval. This theorem almost classifies the map H up to topological equivalence. The points with the same itinerary are called *monotone equivalence classes*. These classes are connected sets: either points or intervals. The itineraries of the critical values $H(c_i)$ are called the *kneading data* for the map. The kneading data constrains the itineraries of other trajectories. If a point $H^i(x) \in I_j$, then $s(j)A(H(c_{j-1})) \leq s(j)A(H^{i+1}(x)) \leq s(j)A(H(c_j))$. The following theorem states that all sequences of addresses that satisfy strong versions of these inequalities are itineraries.

Theorem 5.1 (see [20]). *If a sequence of symbols $\bar{a}$ satisfies the inequalities $s(j)A(H(c_{j-1}))$ $< s(j)\sigma^{i+1}(\bar{a}) < s(j)A(H(c_j))$ when $a_i = j$, then there is a point x whose itinerary is $\bar{a}$.*

Thus the kneading data essentially characterizes the monotone equivalence classes of H. If we further know which monotone equivalence classes are intervals and which are points, this determines the topological equivalence class of the mapping. Thus, in classifying piecewise monotone one dimensional maps up to topological equivalence, the kneading data are the essential invariants.

We want to apply the kneading theory described above to the multivalued extended half return maps defined in the previous section. To accomplish this, we imagine that the extensions of the map are made by inserting intervals of infinitesimal length at the discontinuities of the map and think of the extensions as functions of positive or negative slope on these infinitesimal intervals. Formally, we generalize the definition of a discrete dynamical system to multivalued functions.

Given a set X, the graph of a function $f : X \to X$ is a subset $G \subset X \times X$ with the property that for each $x \in X$, there is exactly one element of G whose first element is x. We relax this requirement and define a discrete dynamical system for *multivalued functions*, defined as relations $G \subset X \times X$ with the property that the projection of G onto its first component is onto; i.e., every $x \in X$ is the first element of at least one pair $(x, y) \in G$. We define the trajectories of G to be sequences $\{x_i\}$ such that $(x_i, x_{i+1}) \in G$ for all $i \geq 0$. If G is the graph of a function $f : X \to X$, then each trajectory is the sequence of f iterates of its first element. In our case, X and G must be chosen in a way that reflects that there are two overlapping vertical segments added to the graph of H at each point of discontinuity. We do this by first setting X to be the circle S_1, but with each critical point p of H "split" into left-hand and right-hand points p_l and p_r. This makes X a non-Hausdorff topological space. Next, the relation G is defined by adding vertical segments (p_l, q_l) and (p_r, q_r) to the graph of H at each discontinuity

point p_l, p_r. If the discontinuity point of H is not one of the two points $\arcsin(2/a)$, then both of the branches of H at the discontinuity point are increasing or decreasing. (If there is a point $(\arcsin(2/a), 2)$ that lies in W_s, then H has a continuous extension at this point.) Where the branches of H are increasing, call p_r the lower discontinuity point and p_l the upper discontinuity point. Where the branches of H are decreasing, call p_l the lower discontinuity point and p_r the upper discontinuity point. The limit value of H at the lower discontinuity points is θ_{1u}, while the limit value of H at the upper discontinuity points is θ_{1s}. At each lower discontinuity point p_*, we add the vertical segment $[\theta_{1u}, \theta_{1m}]$ to G. At each upper discontinuity point we add the vertical segment $[\theta_{1s}, \theta_{2u}]$, where θ_{2u} depends on the sign of θ' at $(-2, \theta_m)$. If $2 + \arcsin(2\pi\theta_m) > 0$, we set $\theta_{2u} = \theta_{1m}$, while if $2 + \arcsin(2\pi\theta_m) < 0$, θ_{2u} is determined so that $(-2, \theta_{2u})$ is the first point on the backward trajectory of the slow flow (1.2) with initial point $(-2, \theta_m)$. This almost completes the definition of the relation G. The one remaining item in its definition is what happens at the inverse images of the critical points. If an inverse image of a critical point p is a regular point q, we include both (q, p_l) and (q, p_r) in G. If an inverse image of a critical point p is a critical point q, then we include only one of (q, p_l) and (q, p_r) in G. If q is the left endpoint of an increasing branch or the right endpoint of a decreasing branch, then we include (q, p_l) in G; otherwise, we include (q, p_r) in G.

We next define a (cyclic) order on the relation G. We begin by extending the order of the circle to X so that $p_l < p_r$ at each critical point. Next, we assign signs to the vertical segments in G: upper discontinuity points of increasing branches and lower discontinuity points of decreasing branches are assigned positive slope with sign $s = +1$, while the extensions of the graph of H at upper discontinuity points of decreasing branches and lower discontinuity points of increasing branches are assigned negative slope with sign $s = -1$. Given points (θ_1, ψ_1) and (θ_2, ψ_2) in G, we say that $(\theta_1, \psi_1) < (\theta_2, \psi_2)$ if either $\theta_1 < \theta_2$ or the points on the same vertical segment with sign slope s and $s\psi_1 < s\psi_2$.

We now define a symbolic dynamics for the relation G. Partition G into segments $I_1, \ldots, I_n$ at critical points. At each critical point, points (p_l, q) and (p_r, q) belong to adjacent elements of the partition. Here a segment is defined relative to the order on G: a segment includes all the pairs between any two pairs in the segment. The partition elements are closed curves in the plane which are "disjoint" if one distinguishes p_l and p_r at each critical point. As for a function, we define the itinerary $\bar{a} = A(\bar{x})$ for a trajectory $\bar{x} = \{x_i\}$ to be the symbol sequence with $a_i = j$ when $x_i \in I_j$. With these definitions, the itineraries of trajectories for the relation G are determined precisely by inequalities on symbol sequences.

Theorem 5.2. *A sequence of symbols $\bar{a} = \{a_i\}$ is the itinerary of a trajectory of G if and only if it satisfies the inequalities $s(j)A(H(c_{j-1})) \le s(j)\sigma^{i+1}(\bar{a}) \le s(j)A(H(c_j))$ with $a_i = j$.*

The proof remains the same as the proof of the original theorem. The key point is the observation that for any sequence satisfying the inequalities $s(j)A(H(c_{j-1})) \le s(j)\sigma^{i+1}(\bar{a}) \le s(j)A(H(c_j))$, the sets

$$\bigcap_{i=1}^{l} H^{-i}(I_{a_i})$$

form a decreasing sequence of nonempty compact connected subsets of G. Their intersection is a nonempty connected subset of the graph.

If the critical points of a piecewise monotone map tend to stable periodic orbits and the nonwandering set has a hyperbolic structure [8], then the map is *structurally stable*. A map is structurally stable if perturbations of the map are all topologically equivalent; i.e., there are homeomorphisms that map trajectories of one map to trajectories of another. We comment here on the stability properties of G. If the trajectories of the critical points of G tend to stable periodic orbits and the two points $(\arcsin(2/a), 2)$ do not lie in W_s, then perturbations of the reduced system will have extensions with the same number of branches and the itineraries of the critical points will be unchanged. If, further, the "map" is expanding on the set of trajectories that do not tend to stable periodic orbits, then G will be structurally stable: its nonwandering set will vary continuously with perturbation. The nonwandering set of a map with hyperbolic nonwandering set is topologically equivalent to a disjoint union of a finite number of stable periodic orbits and expanding subshifts of finite type. The nonwandering set of a structurally stable G is similar, but with one difference that bears discussion. If there are distinct (but possibly coincident) critical points p_1 and p_2 so that the canard pairs (p_1, p_1), (p_1, p_2), (p_2, p_1), and (p_2, p_2) belong to G, then arbitrary sequences of p_1 and p_2 are trajectories for G. This set of trajectories is topologically equivalent to a shift on two symbols. Still, the only points of G visited by these trajectories are the canard pairs above. Thus, there may be large sets of trajectories that visit the same points of G, but in a different temporal order. When we examine the singular limit of the van der Pol system, these sets of trajectories are limits of chaotic invariant sets of the return map for a cross-section of the van der Pol system. Therefore, we still regard these sets of trajectories for G as chaotic invariant sets even though they visit only a finite number of points of G.

We next describe the symbolic dynamics of the examples presented in the last section. When $(a, \omega) = (1.1, 1.57)$, H has a single discontinuity θ_{2s} and is strictly increasing. See Figure 4. There are two fixed points of H, one stable, denoted v_s, and one unstable, denoted v_u. The extensions to H consist of two vertical intervals at the discontinuity, both ending at the maximal canard point, denoted θ_m. The interval (θ_{2s}, θ_m) contains v_u. The interval (v_u, θ_{2s}) is forward invariant, all points iterating to v_s. To obtain the symbolic description of the nonwandering set of the extended H, we select four symbols $\{1, 2, 3, 4\}$ that are assigned to the left-hand "increasing" vertical segment of G, the right-hand "decreasing" vertical segment of G, the interval $[\theta_{2s}, \theta_m]$, and the interval $[v_u, \theta_{2s}]$, respectively. The partition has the properties that $H(I_4) \subset I_4$ and $I_1 \cup I_2 \cup I_3$ is a subset of the "images" of I_j for $j = 1, 2, 3$. It follows that there is an invariant set Λ in $I_1 \cup I_2 \cup I_3$ that is topologically equivalent to the full shift on three symbols. Note that H is expanding on I_3. Points that do not remain in $I_1 \cup I_2 \cup I_3$ converge monotonically to v_s once they map into I_4. We conclude that the nonwandering set of G consists of Λ and v_s.

The second example that we consider has parameter values $(a, \omega) = (4, 1.51)$. Figure 5 shows the graph of H with its extensions for these parameter values. There are three points of discontinuity for H, and H has a local minimum in one subinterval on which it is continuous and a local maximum in another subinterval, and H is increasing on the third subinterval. There are a pair of canard extensions at each point of discontinuity. We partition G into nine subintervals $I_1, \ldots, I_9$ with I_1 the subinterval with a local maximum, I_4 the subinterval with a local minimum, and I_7 the subinterval on which H is increasing. For $j = 1, 4, 7$, we set I_{j+1} to be the canard segment at the right endpoint of I_j. We set I_3 to be the canard segment at

the left endpoint of I_4, I_6 to be the canard segment at the left endpoint of I_7, and I_9 to be the canard segment at the left endpoint of I_1.

There are two stable fixed points of this example, one in I_1 and one in I_4. All of the points in I_1 iterate to the fixed point in I_1, and all of the points in I_4 iterate to the fixed point in I_4. The canard extensions I_2 and I_6 contain fixed points. The images of the "lower" canard segments I_2, I_5, and I_8 contain the canard segments I_2, I_3, and points in I_1 and I_4 that iterate to the stable fixed points. The images of the "upper" canard segments I_3, I_6, and I_9 contain the canard segments I_5, I_6 and points in I_4 and I_7, including a preimage p of the canard segments I_2 and I_3 that lies in I_7. Points other than p in I_7 have images in I_1 or I_4 and iterate to the fixed points. Looking at these images, we find that there is an invariant set Λ topologically equivalent to a subshift with symbols $2, 3, 5, 6, 7$ and that the transition matrix[1] for the subshift is

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix},$$

reflecting that the images of I_2, I_5, and p contain $I_2 \cup I_3$ while the images of I_3 and I_6 contain $I_5 \cup I_6 \cup \{p\}$. The nonwandering set of H consists of the two stable fixed points and Λ.

The third example of a structurally stable nonwandering set we display is for parameter values $(a, \omega) = (1.02, 1.19)$. Figures 6 and 7 show the graphs of the first and second iterates of the half return map H. There are no fixed points of H, but the graph of H^2 resembles the dynamics of H at the parameter values $(a, \omega) = (1.1, 1.57)$. There is a stable periodic orbit of period 2 for H. There are also two regions containing canards in each of which the extension of H^2 has an invariant set that is topologically equivalent to a full shift on three symbols. These two invariant sets of H^2 are interchanged by H. Thus G has an invariant set that is topologically equivalent to a subshift with six symbols whose transition matrix is

$$\begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is the 3×3 zero matrix and $\mathbf{I}$ is the 3×3 matrix with all components 1.

6. Bifurcations.

Bifurcations of the reduced system and its half return map H occur where the system is not structurally stable. A long term goal is to classify the bifurcations that occur as the parameters a and ω vary, but this paper concentrates on several bifurcations that affect changes in the number of fixed points of H and bifurcations that bound a few parameter regions of structurally stable systems with chaotic invariant sets.

Figure 13 reproduces the bifurcation diagram of fixed points of H without its extension [14]. Recall that in the region where there are folded foci, the bifurcations of fixed points

[1]Recall that a subshift of finite type [8] with symbols $s_1, \ldots, s_n$ and transition matrix $T = (T_{ij})$ (an $n \times n$ matrix with components 0 and 1) consists of all sequences $\{a_i\}$ with $a_i \in \{s_1, \ldots, s_n\}$ and $T_{a_i a_{i+1}} = 1$ for all i. One-sided shifts are indexed by $i \geq 0$ while two-sided shifts are indexed by $i \in Z$.

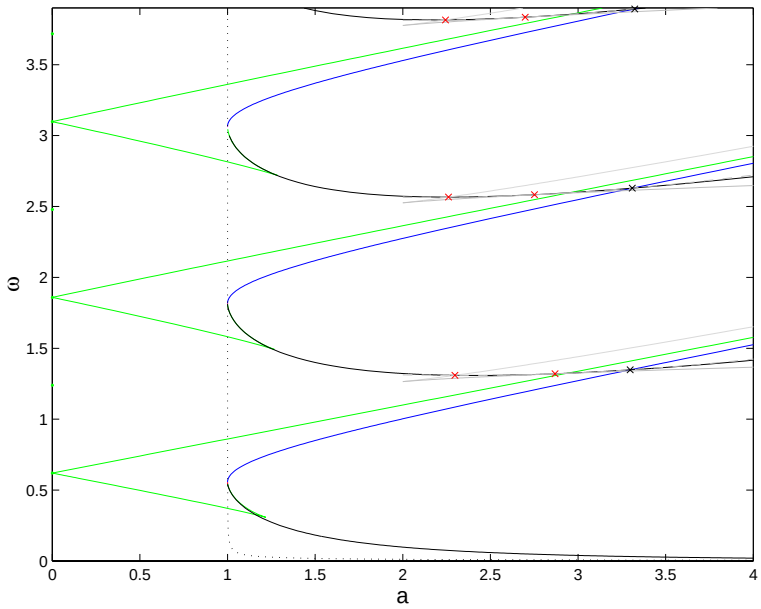


Figure 13. *The bifurcation diagram of fixed points of H. The green curves are saddle-node bifurcations, the black curves are right homoclinic bifurcations, the blue curves are left homoclinic bifurcations, and the gray curves are points where there is a point of W_s tangent to the circle S_2.*

of H that were found are two types of homoclinic orbits (black and blue in the diagram) and saddle-node bifurcations (green). For each circuit number of the corresponding periodic orbits in the reduced system, there is a curve of bifurcations of each of these three types. Discontinuities of H occur at intersections of W_s, the stable manifold of the folded saddle, with the circle $x = 2$ in the phase space, so the number of discontinuities of H changes at points where W_s intersects the circle $x = 2$ tangentially. The locations of these parameter values are drawn as gray curves in the bifurcation diagram. They play an even more important role in the bifurcations of the extended half return map because each discontinuity gives rise to canards and vertical segments in the graph of the extended H. We now turn to the extension of H and its bifurcations.

In the parameter region $a < 2$, there are no tangencies of the vector field with the circle $x = 2$, so H has no local maxima and minima and just a single discontinuity. In the extension of H, the left and right values of the maximal canards are the same. To characterize the dynamics of H, we are especially interested in the kneading data, i.e., the itineraries of the critical values. There are two critical values, namely, θ_m, the value of the maximal canard, and $\theta_r = \theta_{1u} + 1/2$, the right-hand limit of H at the discontinuity. The graphs of the extended half return map are limits of piecewise smooth circle maps with two critical points. Such maps can have quasi-periodic trajectories, but they always have periodic orbits as well [5]. The universal bifurcation diagram for smooth circle maps with two critical values is essentially two dimensional. The properties of G, the relation associated with the extension of H, will be similar to those of these circle maps in the parameter region $1 < a < 2$ with folded foci. We look for bifurcations of G at locations where the itineraries of the two critical values do

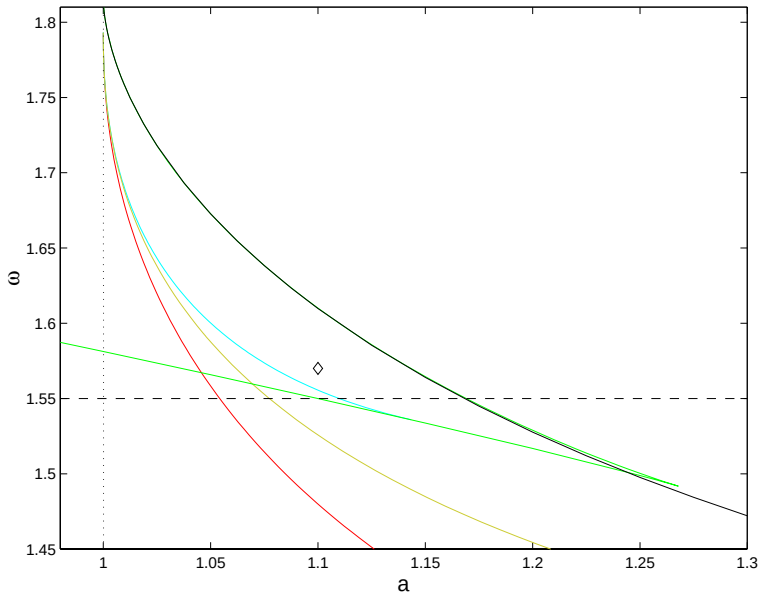


Figure 14. *A detail from the bifurcation diagram. The diamond is the point* $(a, \omega) = (1.1, 1.57)$. *In this figure, we have added an additional bifurcation curve. The yellow curve is the set of parameter values where* $H(\theta_m) = \theta_{2s}$.

not iterate to stable periodic orbits. We begin by looking at the boundary of the region of structurally stable systems in parameter space around $(a, \omega) = (1.1, 1.57)$.

In the previous section, we showed that when $(a, \omega) = (1.1, 1.57)$, the nonwandering set of G consists of a fixed point and an invariant set Λ topologically equivalent to a full shift on three symbols. We discuss here the bifurcations that limit the parameter region in which G has a nonwandering set topologically equivalent to this one. The following properties of H and its extension imply that there is a hyperbolic invariant set Λ equivalent to the shift on three symbols. We first summarize the relevant properties of H:

- H has a single jump discontinuity at $\theta = \theta_{2s}$ with $\theta_r < \theta_{2s}$;
- H is continuous on $S^1 - \{\theta_{2s}\}$;
- H has only two fixed points v_u and v_s, with $\theta_{2s} < v_u < v_s$, where v_u is unstable and v_s is stable;
- and $\theta_m \geq v_u$.

As described in the previous section, the nonwandering set of the extension G of H consists of just the stable fixed point and Λ.

Figure 14 displays the curve μ in cyan along which $\theta_m = v_u$. There is a sequence of bifurcations that accumulate on μ from below (i.e., smaller ω). Below μ, the nonwandering set of G no longer contains an invariant set equivalent to a shift on three symbols. If θ_{2s}^- and θ_{2s}^+ denote the two adjacent points in X where the vertical segments are added, $(\theta_{2s}^-, \theta_{2s}^-)$, $(\theta_{2s}^-, \theta_{2s}^+)$, $(\theta_{2s}^+, \theta_{2s}^-)$, and $(\theta_{2s}^+, \theta_{2s}^+)$ are all in G as long as $\theta_m > \theta_{2s}$. As mentioned earlier, this means that arbitrary sequences of θ_{2s}^- and θ_{2s}^+ are trajectories of G; this set of trajectories is topologically equivalent to a shift on two symbols, so the nonwandering set of G (minus

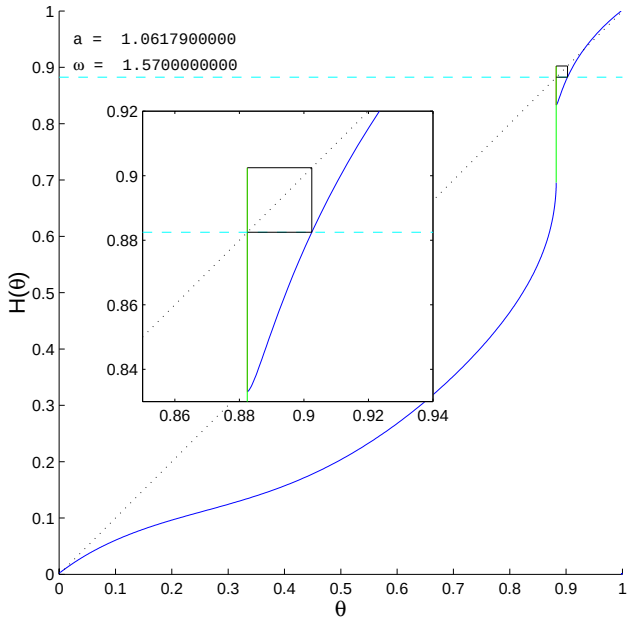


Figure 15. *An example of Proposition 6.1 where* $c_1 = \theta_m$.

the stable fixed point) contains a shift on two symbols and is properly contained in the shift on three symbols. We give a more precise description of these nonwandering sets. Let $c_k = H^{-k}(\theta_{2s})$. We observe that $\lim_{k\to\infty} c_k = v_u$.

Proposition 6.1. *Let H, G, and $\{c_k\}_{k=1}^{\infty}$ be as defined above. Assume that none of the c_k are θ_m. Let $\{c_1, c_2, \ldots, c_n\}$ be the set for which $c_k < \theta_m$. Then the nonwandering set of G consists of the two fixed points of H and a hyperbolic invariant set Λ with $n+2$ symbols $\{\theta_{2s}^-, \theta_{2s}^+, c_1, c_2, \ldots, c_n\}$ and transition matrix*

$$\begin{pmatrix} \mathbf{1}_{2\times 2} & \vdots & \mathbf{1}_{2\times n} \\ \cdots & \cdots & \cdots \\ \mathbf{1}_{1\times 2} & \vdots & \mathbf{0}_{1\times n} \\ \cdots & \cdots & \cdots \\ \mathbf{0}_{(n-1)\times 2} & \vdots & \mathbf{I}_{n-1} \quad \vdots \quad \mathbf{0}_{(n-1)\times 1} \end{pmatrix},$$

where $\mathbf{1}_{m\times n}$ is the $m \times n$ matrix of ones, $\mathbf{0}_{m\times n}$ is the $m \times n$ matrix of zeros, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

The proposition is proved by observing that the points (θ_{2s}^-, c_k) and (θ_{2s}^+, c_k) are in G precisely when $k \le n$. From either of these points, the subsequent trajectory is $c_k, c_{k-1}, \ldots, c_1$. By the construction of G, c_1 can be followed by either θ_{2s}^- or θ_{2s}^+. Note that if $c_1 > \theta_m$, then $\Lambda = \{\theta_{2s}^-, \theta_{2s}^+\}$, and on Λ, G is equivalent to a shift on two symbols; the unextended regions do not play a role in the chaotic invariant sets.

The changes in the structure of the symbolic dynamics occur when $c_k = \theta_m$ for some k; at these points, the kneading data associated with θ_m change. For example, Figure 15 shows

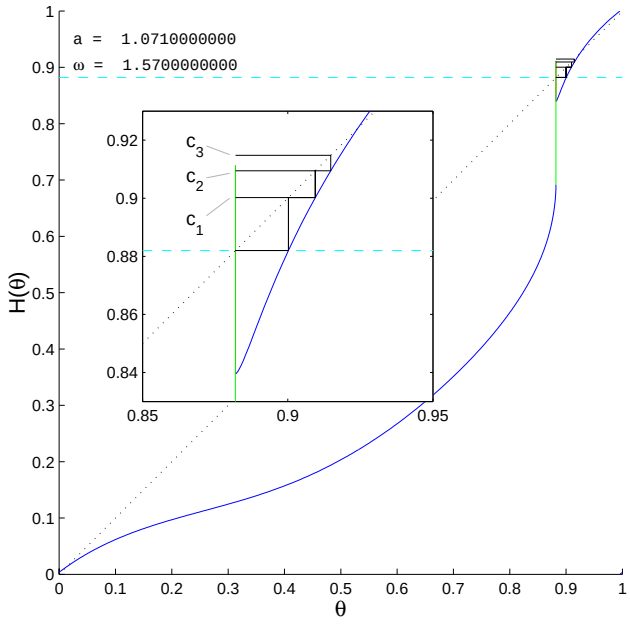


Figure 16. *An example of Proposition 6.1 where $c_2 < \theta_m$ but $c_3 > \theta_m$.*

a case where $c_1 = \theta_m$. This is a point from the yellow curve in Figure 14. On one side of this curve, the chaotic dynamics are equivalent to a shift on $\{\theta_{2s}^-, \theta_{2s}^+\}$. On the other side, the symbols are $\{\theta_{2s}^-, \theta_{2s}^+, c_1\}$, and the transition matrix is

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

If we continue to change the parameters in this direction, more bifurcations will occur. For example, Figure 16 shows a case where $c_2 < \theta_m$ but $c_3 > \theta_m$. In this case $\Lambda = \{\theta_{2s}^-, \theta_{2s}^+, c_1, c_2\}$, and the transition matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Let $\mathcal{U}$ be the connected region in the (a, ω) plane containing $(1.1, 1.57)$ where G has a chaotic invariant set; see Figure 14. There are four curves bounding the region. On the left (near $a = 1$), we have the curve that marks the transition from a folded node to a folded focus (the black dotted line). The upper boundary is the curve of right homoclinic points (the black solid line). At these points, $\theta_r = \theta_{2s}$; i.e., the right side of the discontinuity is a fixed point. Above this curve there is a thin sliver in which the right-hand vertical extension no longer contains a fixed point, and there is a new stable fixed point just to the right of θ_{2s}. The upper boundary of the sliver is a min saddle-node curve (the upper green line), where the

"new" stable fixed point coalesces with v_u, leaving a single stable fixed point. When the new stable fixed point exists, the preimage of θ_{2s} is only a single point that lies in I_2, the left-hand canard segment of G. This fixed point in I_2 is the only point that maps into the canards for G. Consequently, the nonwandering set consists entirely of fixed points—two stable fixed points and two unstable fixed points, one of which lies in I_2. The two other curves that bound $\mathcal{U}$ are the curve where $\theta_m = \theta_{2s}$ (the red curve) and the max saddle-node curve (the lower green curve). After leaving $\mathcal{U}$ by crossing the red curve, $\theta_m < \theta_{2s}$, and there are no fixed points in the vertical extensions. The green curve is the saddle-node curve, and just below this curve, H does not have any fixed points.

Two additional bifurcation curves in $\mathcal{U}$ that divide parameter regions with topologically inequivalent nonwandering sets are plotted. The curve where $c_1 = \theta_m$ is plotted in yellow, and the curve where $v_u = \theta_m$ is plotted in cyan. Thus the region bounded by the yellow, lower green, red, and black dotted curves gives the set of parameters where $c_1 > \theta_m$. The region bounded by the solid black, lower green, cyan, and black dotted lines gives the set of parameters where $v_u < \theta_m$, so the dynamics on Λ can be described as a full shift on three symbols. The infinite sequence of bifurcation curves discussed above lies between the yellow and cyan lines. There is one curve for each equation $c_n = \theta_m$. They are not shown in Figure 14, but they converge to the cyan line as n increases.

This example illustrates some of the new types of bifurcations arising from canards in the extension G of the half return map H. In seeking to classify the different types of bifurcations, we note that extended maps G which are *not* at bifurcation parameters have the property that their critical values tend to stable periodic orbits. Limit sets for the critical values that are not stable periodic orbits can be unstable periodic orbits, subsets of chaotic invariant sets, or quasi-periodic trajectories. There are also parameters at which the number of discontinuities change and where the maximal canards of the reduced system move from the section S_1 to the section S_{-1}. This plethora of possibilities is sufficiently daunting that we do not attempt to determine here the types or locations of more than a few bifurcations of G. We focus upon those whose defining equations are expressed in terms of G itself rather than higher iterates.

One of the main qualitative differences in canards of the reduced systems is in the location of maximal canards. When a is large enough and large enough relative to ω, oscillations of the reduced system induced by the forcing are large enough that the saddle separatrix γ_u intersects the circle S_{-1}. The transition between parameter values for which γ_u intersects the circle S_{-1} and those for which it does not occur when there are heteroclinic orbits that connect the folded saddles on S_1 and S_{-1}. The parameter values with these heteroclinic orbits divide the parameter space into two regions that we call M_1 and M_{-1}. M_1 lies above the heteroclinic curve in the (a, ω) plane and yields maximal canards on the circles S_1. M_{-1} lies below the heteroclinic curve in the (a, ω) plane and yields maximal canards on the circles S_{-1}. In M_1, the maximal canard value associated with the jump back canards (i.e., those that jump to $x > 1$) is θ_m. The maximal canard value associated with the jump away canards (i.e., those that jump to $x < -1$) is either $\theta_m + 0.5$ (when $x'(\theta_m, -2) > 0$) or $\theta_{m2} + 0.5$, where $(\theta_{m2}, -2)$ is the previous point on the slow flow trajectory through $(\theta_m, -2)$ (when $x'(\theta_m, -2) < 0$). Note that $H(\theta_m) = H(\theta_{m2})$ in the second case. Though similar methods can be used to analyze bifurcations in M_{-1}, this paper considers only bifurcations of canards in the region M_1.

Among the new bifurcation types that we expect from the extensions of H are parameters

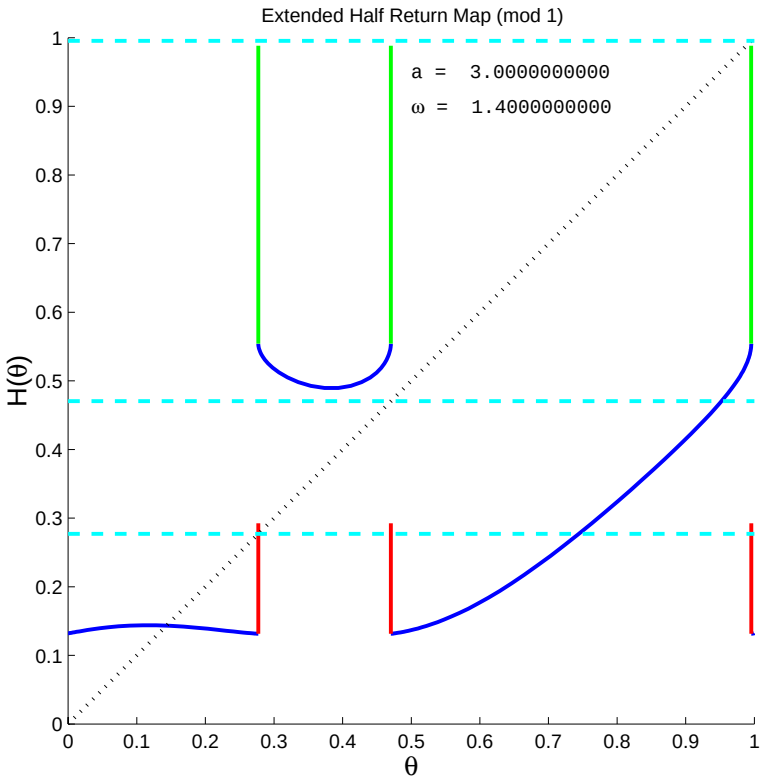


Figure 17. *A plot of the extended half return map for parameter values* $(a, \omega) = (3, 1.4)$.

where $H(\theta_m)$ is a critical point or unstable fixed point; specifically,

1. $H(\theta_m)$ is a discontinuity point, creating a "maximal canard fixed point," or
2. $H(\theta_m) = q$ with q an unstable fixed point of H (not in a canard), or
3. $H(\theta_m) = \arcsin(2/a)$, the local minimum or maximum of H.

If both maximal canard values are $\theta_m + 0.5$, the first of these bifurcations is at the boundary of existence of a chaotic invariant set of trajectories of G that is topologically equivalent to the shift on two symbols. This type of bifurcation cannot occur in families of smooth maps of the circle or interval. In smooth maps, shifts on two symbols are created from maps without chaotic dynamics by an infinite set of bifurcations with intervening parameter intervals in which the critical point tends to stable periodic orbits [11].

Where the maximal canard values differ, it is less apparent what changes in the nonwandering set occur at a bifurcation with a maximal canard fixed point. Figures 17, 18, and 19 show graphs of the extended half return map for $a = 3$ and $\omega = 1.4, 1.409, 1.43$, respectively. The value $\omega = 1.409$ is close to a maximal canard fixed point, but the most visible effect of the bifurcation in the graph of H is to switch which canard branch of H contains the fixed point. The constraint that H have the same image at both maximal canard values prevents the simultaneous existence of two fixed points within canards near these parameter values.

We have seen in the examples with $a = 1.1$ that bifurcations in which a maximal canard value maps into a fixed point can change the structure of chaotic invariant sets without

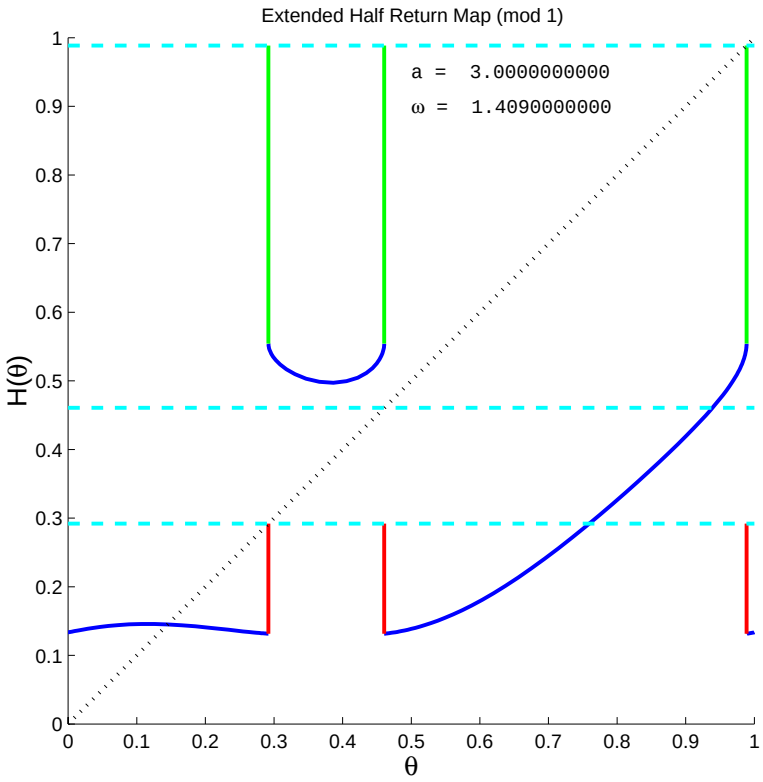


Figure 18. *A plot of the extended half return map for the parameter values* $(a, \omega) = (3, 1.409)$.

changing the number of fixed points of G. The bifurcations in which maximal canard values map into tangency points $\theta = \arcsin(-2/a)/(2\pi)$ do not seem to affect the number of fixed points of H. Instead, these bifurcations are associated with changes in the qualitative structure of the canard extensions of H; there is a transition from parameters at which the two maximal canard values are equal to parameters at which these values differ. The transition occurs when the maximal canard maps into the larger value of the tangency point $\theta = \arcsin(-2/a)/(2\pi)$, as described in section 3.

There are parameter regions where the nonwandering set of G contains an invariant set on which G is equivalent to a subshift of finite type. We have already seen examples, in particular, the region $\mathcal{U}$ in which Proposition 6.1 applies. Here we identify additional regions where this is true, but we do not attempt to completely characterize the nonwandering set as we did in Proposition 6.1. The regions that we consider are those where there are two fixed points in the canard extensions. We have observed several qualitative cases where this occurs.

The example $(a, \omega) = (1.1, 1.57)$ shows a case where the left and right vertical extensions are adjacent and have a fixed point. In this case, the nonwandering set has an invariant subset $\{\theta_{2s}^-, \theta_{2s}^+\}$ on which trajectories are topologically equivalent to a shift on two symbols. The parameter region where this condition holds is bounded above by the black curve of right homoclinic points and bounded below by the red curve of maximal canard fixed points.

There is a region adjacent to the previous region, bounded by the red curve of maximal

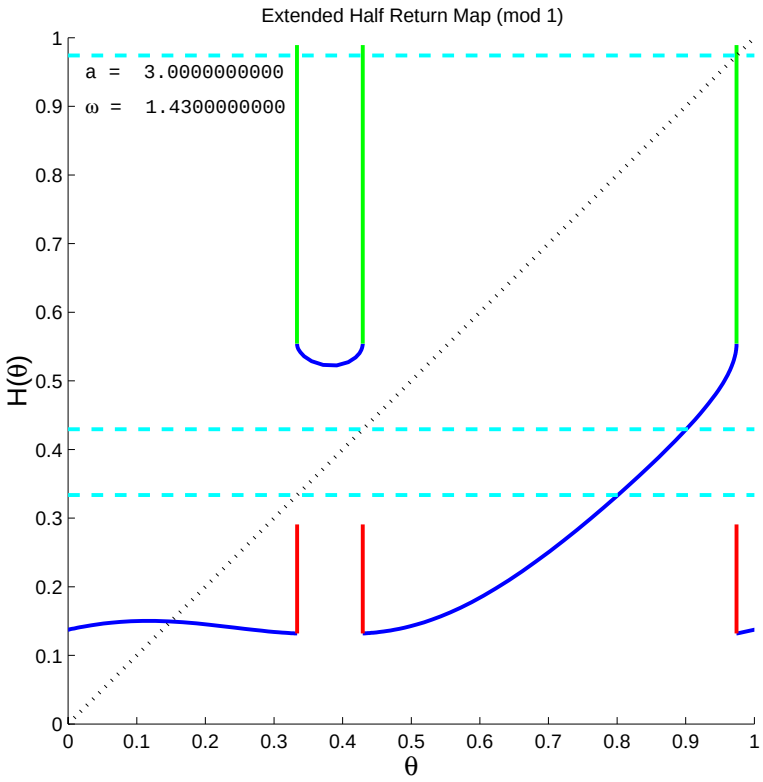


Figure 19. *A plot of the extended half return map for the parameter values* $(a, \omega) = (3, 1.43)$.

canard fixed points, the black curve of right homoclinics, and the gray curve where W_s is tangent to S_2, in which there are two different jump back canard extensions with fixed points. This region is the roughly triangular-shaped region shown in the lower right inset in Figure 23. An example is $(a, \omega) = (2.4, 1.3)$ (see Figure 20). Let d_1, d_2, and d_3 be the θ coordinates of the discontinuities, and let d_i^- and d_i^+ be the left and right adjacent points in X corresponding to these discontinuities. Then there are trajectories made up of only $\{d_1^+, d_2^-\}$ for which G is topologically equivalent to a shift on two symbols.

There is a region where a jump back canard extension and a jump away canard extension have fixed points at different discontinuities. This occurs in the example $(a, \omega) = (4, 1.51)$; see Figure 2. Define $d_i^\pm$ as before. In this case, there are trajectories made up of $\{d_1^-, d_1^+, d_2^-, d_2^+\}$ for which G is topologically equivalent to a subshift with the transition matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

(Note that this transition matrix is a submatrix of the transition matrix derived in the example where $(a, \omega) = (4, 1.51)$.) The region in the (a, ω) plane where this condition holds that also contains $(4, 1.51)$ is bounded below by the black right homoclinic curve and above by the blue

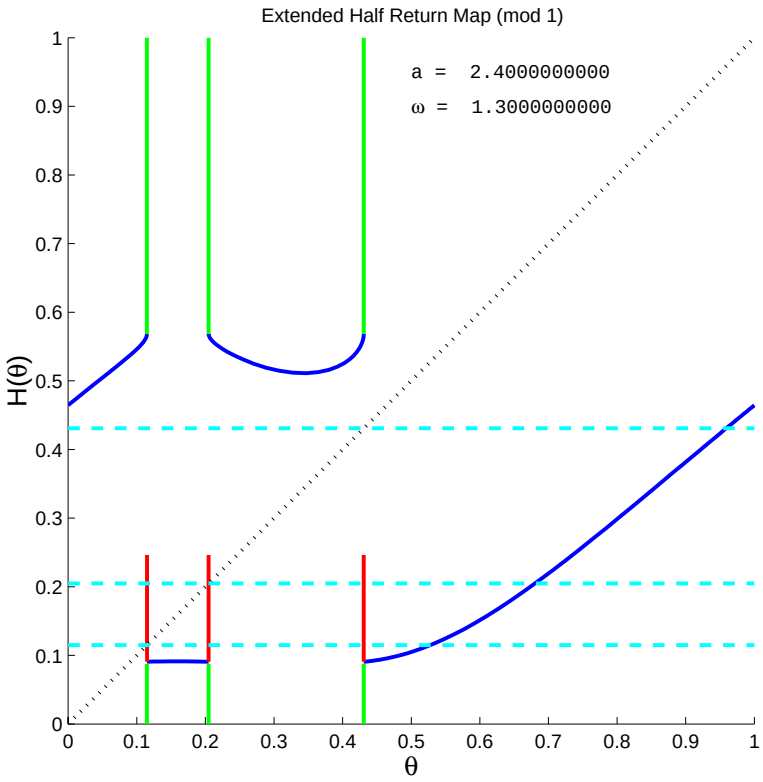


Figure 20. *An example where two different jump back canard extensions have fixed points;* $(a, \omega) = (2.4, 1.3)$.

left homoclinic curve (see Figure 22 or Figure 23). This region appears to be unbounded as a increases.

There are regions where two different jump away canard extensions have fixed points. An example is $(a, \omega) = (6.5, 3.41)$; see Figure 21. As in the case where two different jump back canards have fixed points, on the set $\{d_2^-, d_3^-\}$, G is topologically equivalent to a shift on two symbols. The region where this occurs is bounded below by the red curve of maximal canard fixed points and above by the blue curve of left homoclinic points. We have observed that for circuit number 1, the curve of maximal canard fixed points does not intersect the curve of left homoclinic points, so the case where two different jump away extensions have fixed points occurs only in regions associated with circuit number 2 or higher.

Figure 22 is an augmented bifurcation diagram in which we have added bifurcation curves for the three types of bifurcations involving maximal canard points discussed above. Curves of parameters with maximal canard fixed points are drawn red, curves of parameters with maximal canard points mapping to fixed points are drawn cyan, and curves of parameters with maximal canard points mapping to points of tangency of the slow flow with S_{-2} are drawn yellow. In addition, we draw the parameter curve on which there are heteroclinic connections between the two folded saddles of the reduced system in magenta.

Thus far, we have investigated an extended half return map for the reduced system, seeking to understand the role of canard solutions in the full solution set to the forced van der Pol

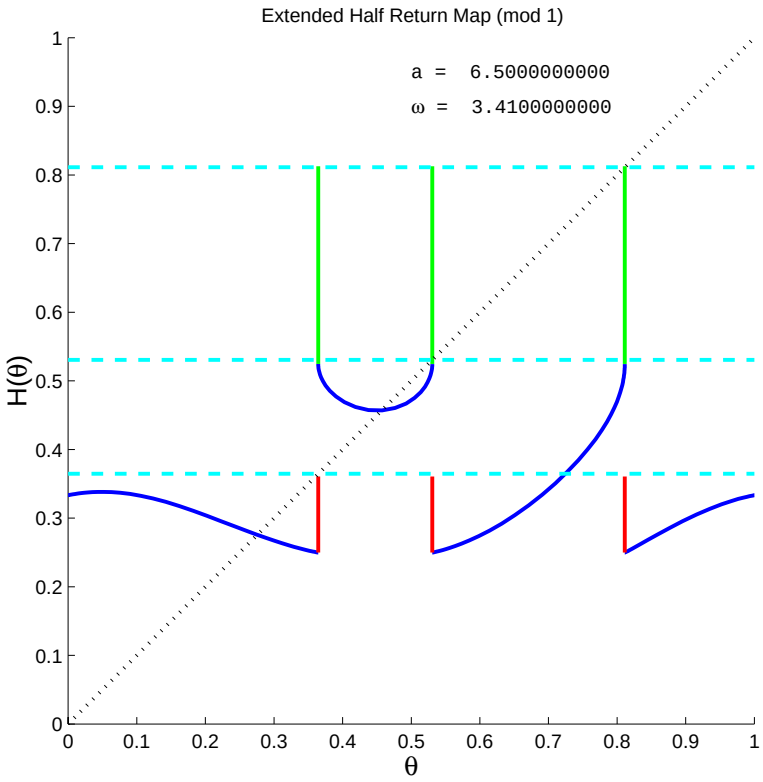


Figure 21. *An example where two different jump away canard extensions have fixed points;* $(a, \omega) = (6.5, 3.41)$.

equation. In particular, we have shown that the canard extensions to the reduced system play an important role in the bifurcations that occur in that system. In the next section, we explore the relationship between the reduced system with its canard extensions and the full system using the half Poincaré map and the singular half Poincaré map. The concept of candidates [4] is useful in formalizing this relationship. *Candidates* are continuous curves consisting of the concatenation of trajectory segments, the slow flow on the stable sheets of the critical manifold, trajectory segments of the slow flow on the unstable sheets of the critical manifold (canards), and segments parallel to the θ axis with the orientation of the fast flow. All of the possible candidates are represented by orbits of our singular half Poincaré map. However, it is not evident that all the candidates are limits of trajectories in the full system as $\varepsilon \to 0$. Though we do not attempt to prove that all candidates are limits of trajectories in the full system, we present numerical computations in the next section that provide evidence that they are.

7. Comparisons of the reduced and full systems: Numerical calculations with $\varepsilon > 0$.

The reduced system is a means of understanding the full van der Pol system (1.1) with $\varepsilon \neq 0$. In this section, we consider the relationship between the full system and the reduced system by comparing computations of Smale horseshoes [22] and bifurcations in the reduced system

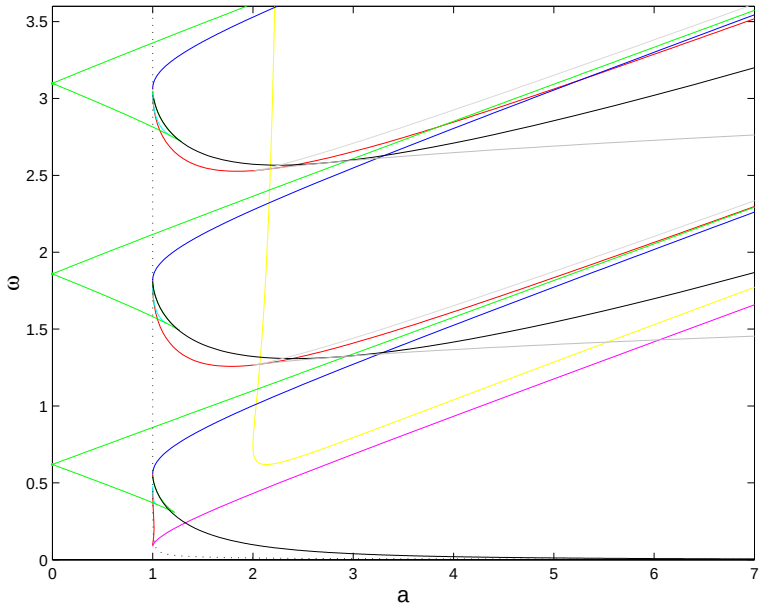


Figure 22. *The bifurcation diagram of fixed points of G, the canard extension of H. The red curves give the location of maximal canard fixed points. The magenta curve shows the location of heteroclinic connections between the two folded saddles of the reduced system. The cyan curves (better seen in Figure 23) show parameters where maximal canard points map to fixed points and the yellow curves show parameters where maximal canard points map to points of tangency of the slow flow with S_{-2}.*

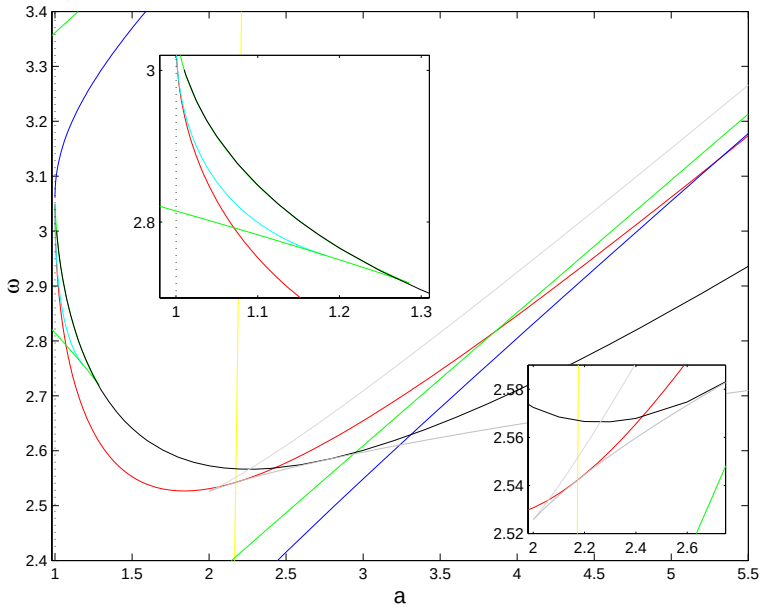


Figure 23. *A closer look at parts of the bifurcation diagram of fixed points of G shown in Figure 22.*

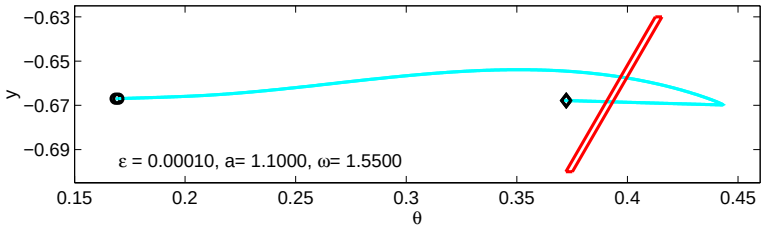


Figure 24. *A Poincaré map that exhibits a horseshoe in (1.1). The parameter values are $\varepsilon = 10^{-4}$, $a = 1.1$, and $\omega = 1.55$. The cyan curves are the images of the horizontal segments at the end of the region plotted in red. Because of the strong contraction on the stable slow manifold, the two cyan curves are indistinguishable in this plot. The ends of the cyan curves are marked with circles on the left and diamonds on the right.*

to computations of the full system with similar parameter values. The results demonstrate that our analysis of canards in the reduced system gives a reliable guide to dynamical behavior associated with canards in the full system.

Numerical computation of trajectories with canards is difficult with initial value solvers [15]. When $\varepsilon > 0$ is small, the exponential growth of deviations from an unstable sheet of the slow manifold is sufficiently rapid that the round-off error in representing an initial point on the slow manifold will give a trajectory that has a jump point close to the initial point. This is an unescapable feature of slow-fast dynamical systems. The use of different initial value solvers, even "stiff" solvers, does not resolve the difficulty. This does not make it hopeless to compute approximations to trajectories with canards, however. Boundary value solvers compute curves that approximately solve a system of differential equations with specified boundary conditions. Their convergence properties are not directly related to the stiffness of the system being solved. We used the program AUTO [9] as a boundary value solver to explore the dynamics of the forced van der Pol system (1.1) for small $\varepsilon > 0$. Our goal was to confirm predictions arising from our analysis of canards in the reduced system and its half return map. This is done with two different types of AUTO computations.

The first AUTO computation seeks a Smale horseshoe in the return map Φ_ε. The cyan curves in Figure 24 are the images under the Poincaré map Φ_ε of the short horizontal line segments at the upper and lower ends of the red parallelogram. The red parallelogram was chosen to include a piece of the strip κ_ε. There are two cyan curves, but the strong contraction on the stable slow manifold makes the resulting curves indistinguishable in this plot. More examples are shown in Figure 25. These examples illustrate the extreme contraction that is associated with the two time scales in the forced van der Pol system. The expansion of horizontal segments within κ_ε by a factor of approximately 50 along the images is also evident in these figures. The combination of the contraction along the long direction of the parallelogram, the expansion along the horizontal direction (resulting from the formation of canards), and the fold that occurs at the maximal canard gives the cyan curves their shape. When $\Phi_\varepsilon(\kappa_\varepsilon)$ intersects κ_ε as in Figures 24 and 25, we see that Φ_ε possesses all the ingredients for the existence of a Horseshoe map with a hyperbolic invariant set.

Singular Poincaré maps for parameter values close to those in Figures 24 and 25 are shown in Figure 26. In particular, Figure 26(b) shows the same values of a and ω as Figure 24, and indeed, the pieces of the plot shown in Figure 24 appear to be well approximated by the

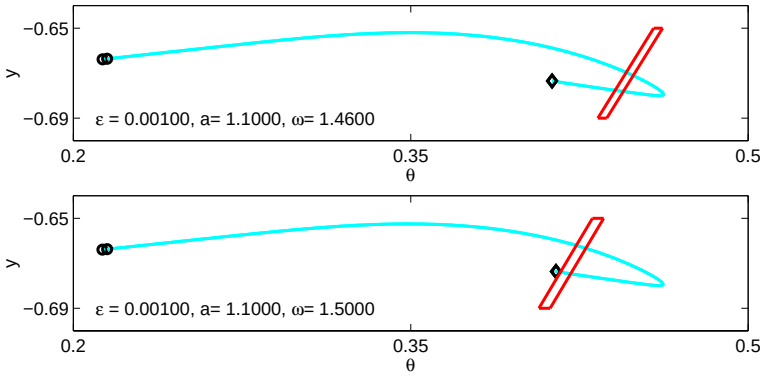


Figure 25. *Poincaré maps that exhibit a horseshoe in* (1.1). *The parameter values are* $\varepsilon = 10^{-3}$, $a = 1.1$, *and* (a) $\omega = 1.46$, (b) $\omega = 1.5$. *The cyan curves are the images of the horizontal red line segments under the Poincaré map* Φ. *Because of the strong contraction on the stable slow manifold, the two cyan curves are indistinguishable in this plot. The ends of the cyan curves are marked with circles on the left and diamonds on the right. (Note that* ε *is ten times larger here than in Figure* 24. *In this case,* ε *is large enough that this map was successfully computed with a standard initial value problem solver in MATLAB.)*

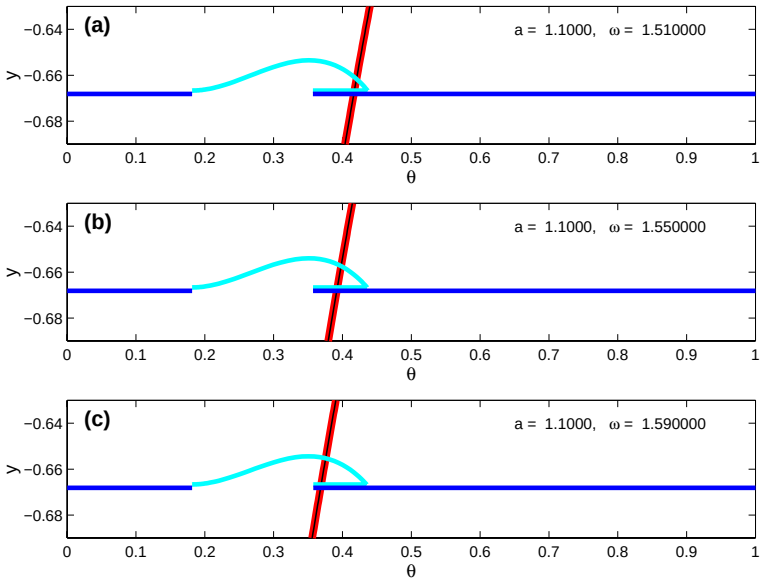


Figure 26. *Singular Poincaré maps with* $a = 1.1$ *and several values of* ω.

singular map shown in Figure 26. This is evidence of the validity of our representation of canards in the reduced system.

We used the continuation capabilities of AUTO to track families of periodic orbits along a line in the parameter space with $\varepsilon = 10^{-4}$ and $\omega = 1.55$ with varying a near 1.1. Our goal is to compare the observed bifurcations of the periodic orbits to the bifurcation diagram for the reduced system displayed in Figure 22. Some of the bifurcations of the van der Pol system (1.1) have different types from the corresponding types in the reduced system due to

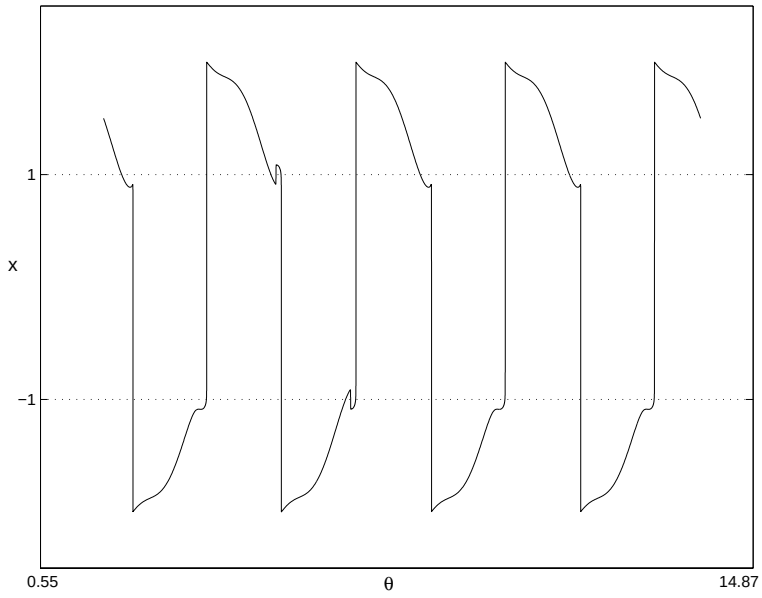


Figure 27. *A graph of x versus θ for a periodic orbit computed by AUTO. This solution is from the curve labeled* "12b:anbbanan" *in Figure 30. The symbol sequence* anbbanan *describes the behavior of the periodic orbit near the folded saddles. The parameter values are $\varepsilon = 10^{-4}$, $a = 1.1$, and $\omega = 1.55$.*

the singular nature of the limit $\varepsilon = 0$. For example, homoclinic bifurcations or bifurcations associated with maximal canards in the reduced system are limits of infinite sets of bifurcations in (1.1) that collapse into one another as $\varepsilon \to 0$. Asymptotic analysis that is not pursued in this paper can be used to predict which bifurcations of periodic orbits of (1.1) will be found close to the bifurcations of the reduced system. Here we present several numerical examples that illustrate relationships between the bifurcations in (1.1) and the reduced system. The data presented here can be compared directly with predictions from asymptotic analysis.

Figure 27 shows a plot of x versus t along a periodic orbit of period 12. (Throughout the remainder of this section, the period of orbits is measured by θ rather than time. Thus a periodic orbit of period k has winding number k in the phase space.) This periodic orbit has five canards that can be identified as segments of the plot with $x \in (-1, 1)$ and moderate slope. There are also three jumps without canards where x crosses the entire interval $(-1, 1)$ with very large slope. The jump back canards occur when the graph enters the region $|x| < 1$ for a short distance but then has a short steep segment that leaves this region without crossing it. This is followed by a jump across $|x| < 1$. In contrast, at the jump away canards, the trajectory enters the region $|x| < 1$ along a segment with moderate slope and then crosses the region along a segment of very steep slope. The order in which the jumps occur is *anbbanan* where the symbol a labels a jump away canard, the symbol b labels a jump back canard, and the symbol n labels a jump without canards.

Recall that the parameters $(a, \omega) = (1.1, 1.55)$ lie in the region where the reduced system has a chaotic invariant set that is somewhere between a shift on two symbols and a shift on three symbols. The symbols are directly related to the a, b, n we used to describe the jumps of

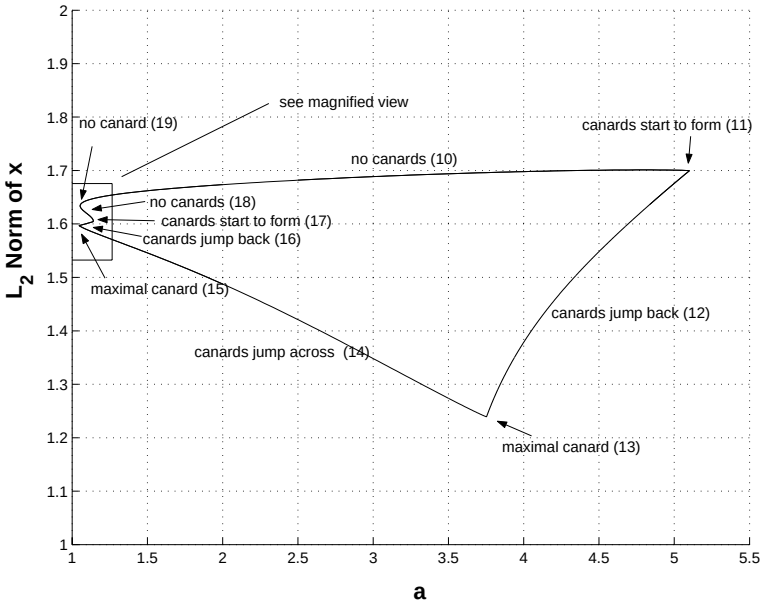


Figure 28. *Symmetric periodic orbits of period* 3 *computed by AUTO. In this example,* $\varepsilon = 10^{-4}$ *and* $\omega = 1.55$.

the periodic orbit in Figure 27. The subshift allows arbitrary sequences of these symbols with the single restriction that there is a limit on the length of consecutive n's. Thus the periodic orbit is compatible with the structure predicted by analysis of the reduced system and can be identified with a specific orbit in the nonwandering set of its extended half return map.

Figures 28, 29, and 30 show bifurcation diagrams for a few periodic orbits of periods 3, 6, and 12 (in θ) computed by AUTO for system (1.1) with $\omega = 1.55$ and $\varepsilon = 10^{-4}$. The plots show the L_2 norm of x along the orbit (normalized by its period) as a function of the continuation parameter a in these calculations. The numbers in parentheses in Figures 28 and 29 are labels attached by AUTO to these solutions. We look at the dynamics of several below. The labels of each solution branch in Figure 30 identify the period and the sequence of jumps displayed by the orbits. In addition, the period 3 orbits are labeled with A for antisymmetric and S for symmetric. The fixed points of the extended half return map G in this region correspond to symmetric periodic orbits of period 3. Asymmetric periodic orbits of period 3 correspond to periodic orbits of G with period 2.

We compare the data produced by these calculations with our analysis of bifurcations of G. Consider the bifurcations of circuit number 3 fixed points encountered along the line $\omega = 1.55$ in Figure 22. The right-hand boundary of the region with circuit number 3 fixed points is a curve of homoclinic orbits. The point labeled (11) in Figure 28 occurs where there is an apparent saddle-node bifurcation. This saddle-node bifurcation is related to the homoclinic bifurcation of the reduced system. The most direct analogue of a homoclinic bifurcation of the reduced system in the three dimensional flow consists of periodic orbits that have a tangency with the surfaces $x = \pm 1$. (Since the orbits are symmetric, there will be tangencies with both $x = 1$ and $x = 1$.) Figure 31 shows a plot of x versus t for the orbit labeled (11). This orbit

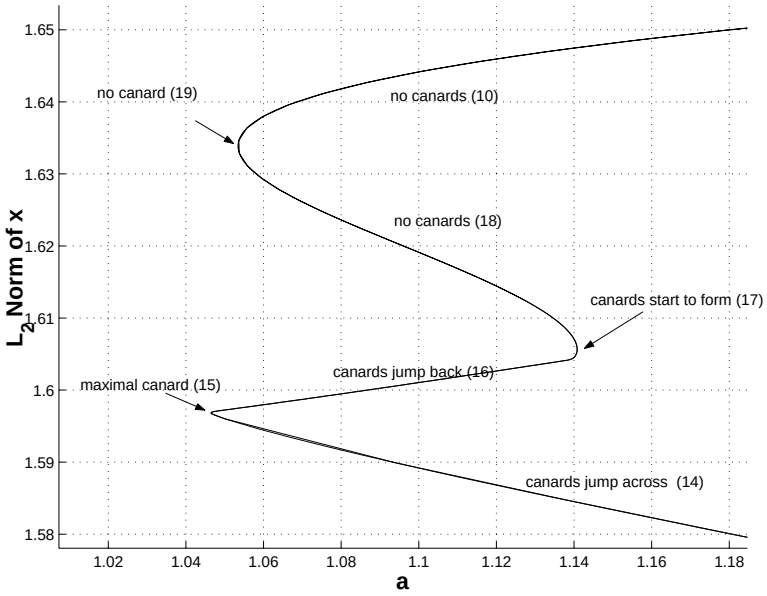


Figure 29. *A magnified view of the boxed region in Figure 28.*

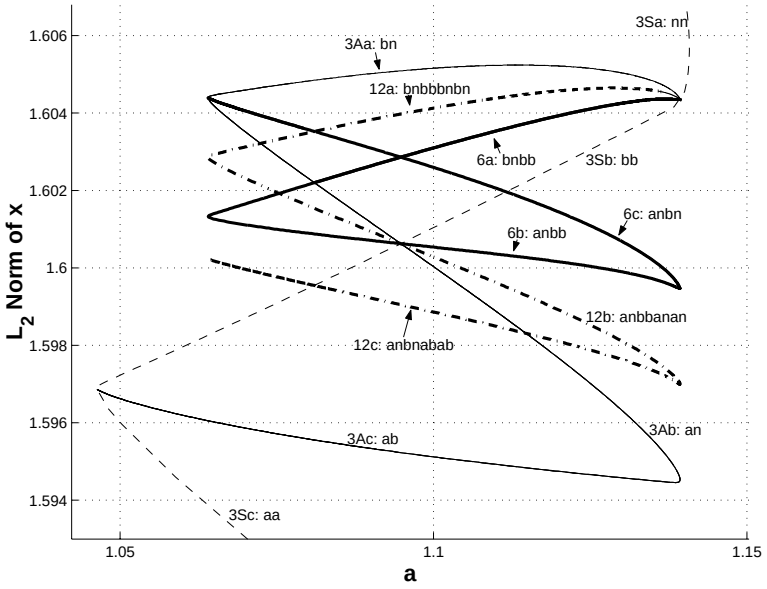


Figure 30. *A partial bifurcation diagram of periodic orbits computed by AUTO. In this example,* $\varepsilon = 10^{-4}$ *and* $\omega = 1.55$.

appears to be very close to one with a tangency with the surfaces $x = \pm 1$. We conjecture that the relationship between a homoclinic bifurcation of the reduced system and the presence of saddle-node bifurcations in the "full" system is a general one [13]. We also note that the apparent vertex of the bifurcation curve at the point labeled (11) seems to be a smooth fold

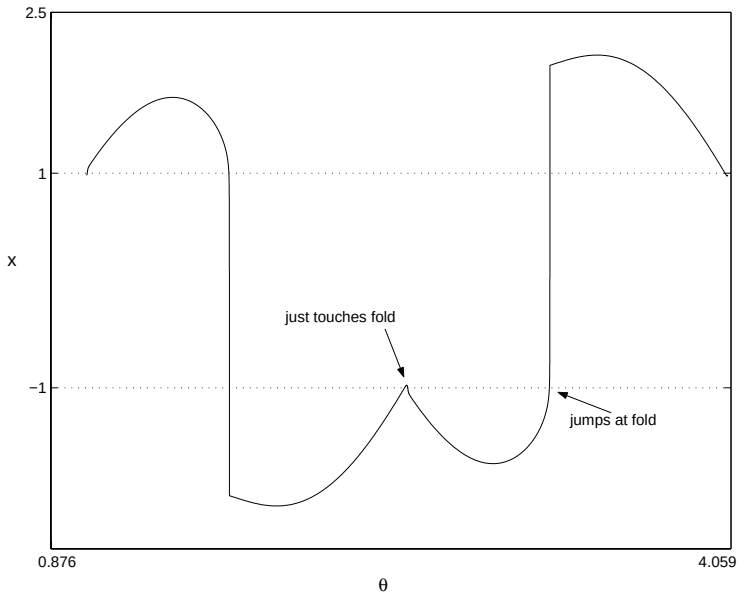


Figure 31. *The periodic orbit labeled* (11) *in Figure* 28*. A jump back canard has just begun to form.*

in a blowup of the region near this turning point in Figure 28.

Decreasing a, the next bifurcation curve of fixed points with circuit number 3 that we encounter is the red curve at which G has a maximal canard fixed point. This corresponds to the point labeled (13) in Figure 28. The periodic orbit is plotted in Figure 32. In this figure, observe that the periodic orbit crosses the surface $x = -1$ without jumping, returning to the surface $x = -1$ before finally jumping. This portion of the periodic orbit is the maximal canard. There is no bifurcation of the family of periodic orbits in this region, but the type of canard contained in the orbit switches from a jump back canard to a jump across canard.

Further bifurcations of the symmetric periodic orbits are best viewed in Figures 14 and 29, which show the regions with finer resolution. The next two bifurcations encountered in the reduced system are very close together. Figure 14 shows that a saddle-node bifurcation curve is crossed (with decreasing a) shortly before a homoclinic bifurcation curve is crossed. In Figure 29, we see a fold at the point labeled (17) near $a = 1.14$ that occurs just before canards begin to form. The homoclinic bifurcation is on the boundary of the region in which G has a chaotic invariant set. For the first time, we find parameters at which there are two periodic orbits with canards. Figure 33 shows an example of a periodic orbit with jump back canards that appear as a is decreased further. There is no apparent bifurcation of the periodic orbits themselves associated with this crossing of the homoclinic curve, but we see below that there are bifurcations in this region giving rise to additional periodic orbits that presumably lie in a chaotic invariant set.

There are two additional bifurcations that we see on the family of periodic orbits. The first of these is another saddle-node bifurcation of a trajectory with no canards. The saddle-node bifurcations of G are located along the green curve in Figure 14. The last bifurcation of the family of periodic orbits occurs where there is a maximal canard. Figure 34 shows this

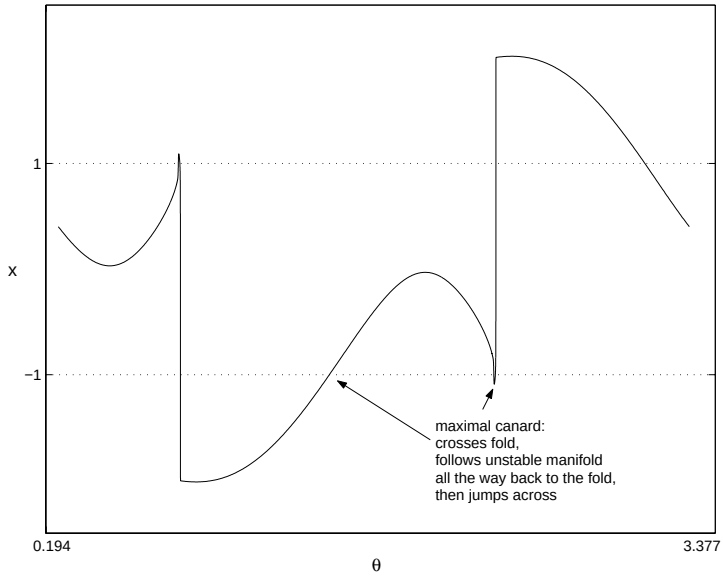


Figure 32. *The periodic orbit labeled* (13) *in Figure* 28*. There is a maximal canard in which the trajectory crosses the surface* $x = -1$ *and then returns to this surface before jumping.*

trajectory. The trajectory crosses the fold curve and continues along the slow manifold until it returns to the fold. There it jumps. This bifurcation corresponds to the maximal canard fixed point curve drawn in red in Figure 14. Here $\theta_m = \theta_{2s}$ is a fixed point for G. This is the boundary of the region where G has a chaotic invariant set. For smaller values of a, G has no canard fixed points.

Let us turn now to Figure 30. The dashed curve is an expanded view of the portion of the curve of symmetric periodic orbits displayed in Figure 29 that contains jump back canards and the point labeled (16). This is the region where we expect to see chaotic invariant sets, including the additional periodic orbits displayed in Figure 30. In the region in which G has a chaotic invariant topologically equivalent to a three shift, there are three fixed points. These correspond to the symmetric periodic orbits in the diagram, with symbol sequences aa, bb, and nn for the full return map. There are also three asymmetric periodic orbits with symbol sequences ab, an, and bn corresponding to the three periodic orbits of period 2 in the shift on three symbols. The diagram also shows a branch of period 6 and a branch of period 12 orbits, each of which has two turning points.

It is notable that many turning points of the branches seem to occur at almost the same value of the continuation parameter a. In generic dynamical systems, there is no reason that several bifurcations should happen simultaneously. However, the very close proximity of these bifurcations to one another is a consequence of the strong contraction and expansion associated with the two time scales in the van der Pol system. The right-hand bifurcation of these branches occurs very close to the homoclinic bifurcation curve that bounds the region where G has a chaotic invariant set. In the van der Pol system, we expect a cascade of bifurcations to occur in a small parameter interval whose length shrinks rapidly to 0 as $\varepsilon \to 0$.

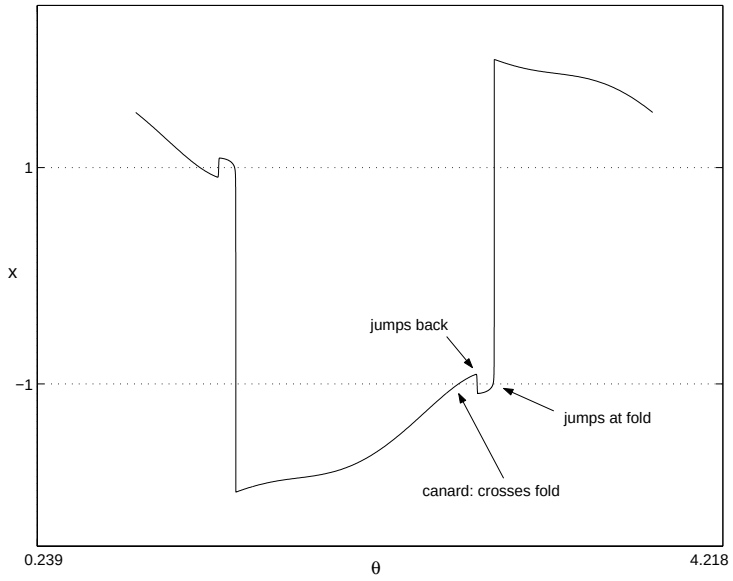


Figure 33. *The periodic orbit labeled* (16) *in Figure* 28*. A jump back canard has just begun to form.*

This explains the location of the right-hand turning points for the asymmetric branches shown in Figure 30. The left-hand turning points seem to occur together at a value of the parameter a where there are no bifurcations of the symmetric period 3 orbits. However, note that all of the branches that are displayed contain a jump without canards. This means that these periodic orbits come from invariant sets that are larger than a two shift. In Figure 14, the yellow curve displays bifurcations in which G^2 maps the maximal canard point back to the discontinuity of G. In section 5 we showed that, to the left of this curve, the chaotic invariant set of G is topologically equivalent to a two shift and consists entirely of G trajectories within the canard extensions of H. Thus we conclude that the left-hand turning points of the asymmetric orbits in Figure 30 correspond to this bifurcation of G.

We make a final observation about Figure 30 that is not related to bifurcations. It appears that all four branches of periodic orbits pass through the same point in the middle of the diagram. In fact, the curves are only close to having a common intersection. The presence of this close intersection is readily explained. The period 6 and period 12 orbits are formed from segments that are almost concatenations of the symmetric and asymmetric period 3 orbits. Therefore, where the two period 3 orbits have the same L_2 norm, the period 6 and period 12 orbits have almost the same norm.

Through these examples, we have shown that there is a *very* close qualitative correspondence between the bifurcations of the extended half return map G and the bifurcations of periodic orbits in the van der Pol system for $\varepsilon = 10^{-4}$. The location of bifurcation curves in the (a, ω) plane has shifted slightly and there is a translation between some types of bifurcations, but the numerical results illustrate that analysis of the reduced system gives a good strategy for the van der Pol system.

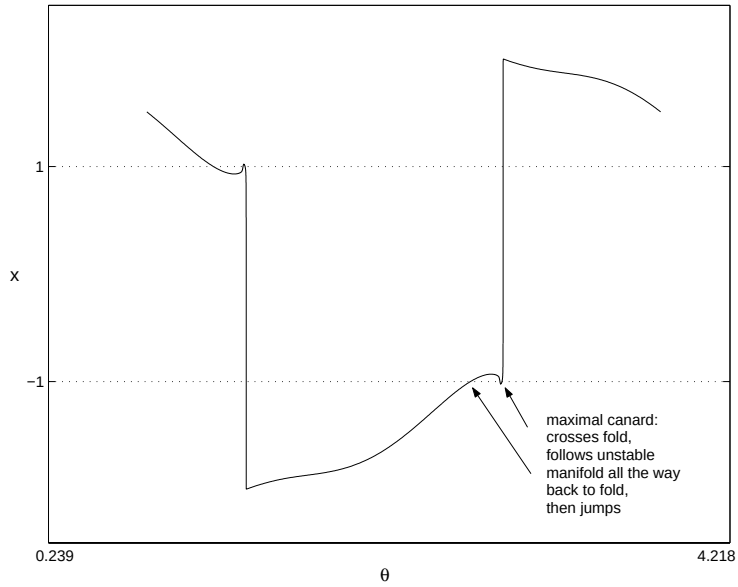


Figure 34. *The periodic orbit labeled* (15) *in Figure* 28*. There is a maximal canard that is much shorter than the maximal canard shown in Figure* 32*.*

8. Concluding remarks. The ultimate goal of this research is to describe the dynamics and bifurcations of the forced van der Pol vector field (1.1). This paper is a step toward that goal, but there is much that remains to be done. We have demonstrated how to extend the half return map of the reduced system to approximate the location of trajectories with canards in the van der Pol system. Furthermore, we have shown how symbolic dynamics can be used to characterize the dynamics on chaotic invariant sets. Each point in the canard extensions of H corresponds to a uniquely determined canard of the reduced system. The trajectories of the reduced system are limits of trajectory families of the van der Pol system (1.1) as $\varepsilon \to 0$. This limit is highly singular, and the analysis here hardly touches upon the nature of this approximation. Instead, we present numerical studies that illustrate the relationship between the two systems.

We have also shown how methods used to analyze bifurcations of one dimensional maps can be applied to the extended half return map G. We have computed curves in parameter space of several types of bifurcations that involve canards. Bifurcations of the van der Pol system (1.1) appear close to the bifurcations of the reduced system that we have identified, but some have a different character. For example, at points where two coincident maximal canard points are fixed, the van der Pol system will have behavior reminiscent of the bifurcations displayed by the Henon mapping [3, 16, 25]. Instead of the immediate creation of a chaotic invariant set, we believe that there is an "exponentially thin" region of trajectories that contains maximal canards. "Normal forms" for maximal canards of a slow-fast system can be derived from the analysis of Arnold et al. [1], but the role that these trajectories play in the global analysis of relaxation oscillations has not been considered. We conjecture that the return map of slow-fast systems near maximal canards have folds that are similar to the Henon mapping with

small positive Jacobian. We hope that systematic investigation with asymptotic methods of the codimension one degenerate slow-fast decompositions in the van der Pol vector field and its reduced system will enable the relationship between these two systems to be made more precise.

REFERENCES

[1] V. I. Arnold, V. S. Afrajmovich, Yu. S. Il'yashenko, and L. P. Shil'nikov, *Dynamical Systems* V, Encyclopaedia Math. Sci. 5, Springer-Verlag, Berlin, 1994.

[2] A. Back, J. Guckenheimer, and M. Myers, *A dynamical simulation facility for hybrid systems*, in Hybrid Systems, Lecture Notes in Comput. Sci. 736, Springer-Verlag, Berlin, 1993, pp. 255–267.

[3] M. Benedicks and L. Carleson, *The dynamics of the Henon map*, Ann. of Math. (2), 133 (1991), pp. 73–169.

[4] É. Benoît, *Canards et enlacements*, Inst. Hautes Études Sci. Publ. Math., 72 (1990), pp. 63–91.

[5] L. Block and J. Franke, *Existence of periodic points for maps of* S^1, Invent. Math., 22 (1973/1974), pp. 69–73.

[6] M. Cartwright and J. Littlewood, *On nonlinear differential equations of the second order* II: *The equation* $\ddot{y} - kf(y,\dot{y})\dot{y} + g(y,k) = p(t) = p_1(t) + kp_2(t)$, $k > 0$, $f(y) \geq 1$, Ann. of Math. (2), 48 (1947), pp. 472–494.

[7] M. Cartwright and J. Littlewood, *Addendum to "On nonlinear differential equations of the second order* II*,"* Ann. of Math. (2), 50 (1949), pp. 504–505.

[8] W. de Melo and S. van Strien, *One-Dimensional Dynamics*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)] 25, Springer-Verlag, Berlin, 1993.

[9] E. Doedel, R. Paffenroth, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. Sandstede, and X. Wang, *AUTO* 2000: *Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, Technical report, California Institute of Technology, Pasadena, CA, 2001.

[10] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21 (1971), pp. 193–225.

[11] J. Guckenheimer, *On the bifurcation of maps of the interval*, Invent. Math., 39 (1977), pp. 165–178.

[12] J. Guckenheimer, *Sensitive dependence to initial conditions for one dimensional maps*, Comm. Math. Phys., 70 (1979), pp. 133–160.

[13] J. Guckenheimer, *Bifurcation and degenerate decomposition in multiple time scale dynamical systems*, in Nonlinear Dynamics and Chaos: Where Do We Go from Here?, J. Hogan, A. Champneys, B. Krauskopf, M. di Bernardo, E. Wilson, H. Osinga, and M. Homer, eds., Institut of Physics Publishing, Bristol, UK, 2002, pp. 1–21.

[14] 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.

[15] J. Guckenheimer and B. Meloon, *Computing periodic orbits and their bifurcations with automatic differentiation*, SIAM J. Sci. Comput., 22 (2000), pp. 951–985.

[16] M. Henon, *A two-dimensional mapping with a strange attractor*, Comm. Math. Phys., 50 (1976), pp. 69–77.

[17] N. Levinson, *Perturbations of discontinuous solutions of non-linear systems of differential equations*, Acta Math., 82 (1950), pp. 71–106.

[18] J. Littlewood, *On nonlinear differential equations of the second order* III: *The equation* $\ddot{y} - k(1-y^2)\dot{y} + y = bk\cos(\lambda t + a)$ *for large* k *and its generalizations*, Acta Math., 97 (1957), pp. 267–308. (*Errata*, Acta Math., 98 (1957), p. 110.)

[19] J. Littlewood, *On nonlinear differential equations of the second order* IV: *The general equation* $\ddot{y} - kf(y)\dot{y} + g(y) = bkp(\phi)$, $\phi = t + a$ *for large* k *and its generalizations*, Acta Math., 98 (1957), pp. 1–110.

[20] J. Milnor and W. Thurston, *On iterated maps of the interval*, in Dynamical Systems (College Park, MD, 1986–1987), Lecture Notes in Math. 1342, Springer-Verlag, Berlin, 1988, pp. 465–563.

[21] E. Mishchenko, Yu. Kolesov, A. Kolesov, and N. Rozov, *Asymptotic Methods in Singularly Perturbed Systems, Translated from the Russian by Irene Aleksanova*, Monogr. Comtemp. Math., Consultants Bureau, New York, 1994.
[22] S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc., 73 (1967), pp. 747–817.
[23] P. Szmolyan and M. Wechselberger, *Canards in R^3*, J. Differential Equations, 177 (2001), pp. 419–453.
[24] B. van der Pol, *The nonlinear theory of electric oscillations*, Proc. IRE, 22 (1934), pp. 1051–1086.
[25] Q. Wang and L.-S. Young, *From invariant curves to strange attractors*, Comm. Math. Phys., 225 (2002), pp. 275–304.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 4, pp. 609–646

Symmetry Groupoids and Patterns of Synchrony in Coupled Cell Networks*

Ian Stewart†, Martin Golubitsky‡, and Marcus Pivato§

Abstract. A coupled cell system is a network of dynamical systems, or "cells," coupled together. Such systems can be represented schematically by a directed graph whose nodes correspond to cells and whose edges represent couplings. A symmetry of a coupled cell system is a permutation of the cells that preserves all internal dynamics and all couplings. Symmetry can lead to patterns of synchronized cells, rotating waves, multirhythms, and synchronized chaos. We ask whether symmetry is the only mechanism that can create such states in a coupled cell system and show that it is not.

The key idea is to replace the symmetry group by the symmetry groupoid, which encodes information about the input sets of cells. (The *input set* of a cell consists of that cell and all cells connected to that cell.) The admissible vector fields for a given graph—the dynamical systems with the corresponding internal dynamics and couplings—are precisely those that are equivariant under the symmetry groupoid. A pattern of synchrony is "robust" if it arises for all admissible vector fields. The first main result shows that robust patterns of synchrony (invariance of "polydiagonal" subspaces under all admissible vector fields) are equivalent to the combinatorial condition that an equivalence relation on cells is "balanced." The second main result shows that admissible vector fields restricted to polydiagonal subspaces are themselves admissible vector fields for a new coupled cell network, the "quotient network." The existence of quotient networks has surprising implications for synchronous dynamics in coupled cell systems.

Key words. coupled systems, synchrony, groupoids, symmetry

AMS subject classifications. 34C14, 34C15, 20L05

DOI. 10.1137/S1111111103419896

1. Introduction.

We use the term *cell* to indicate a system of ODEs. A *coupled cell system* is a set of cells with coupling, that is, a dynamical system whose variables correspond to cells, such that the output of certain cells affects the time-evolution of other cells. The salient feature of a coupled cell system is that the output from each cell is considered to be significant in its own right. A coupled cell system is not merely a system of ODEs but a system of ODEs equipped with canonical observables—the individual cells (see [8]). From a mathematical point of view these output signals can be compared, and this observation leads to a variety of notions of "synchrony." For surveys, see Boccaletti, Pecora, and Pelaez [2] and Wang [14].

In this paper we discuss the *architecture* of a coupled cell system: which cells influence which, which cells are "identical," and which couplings are "identical." We focus on how the

*Received by the editors February 19, 2003; accepted for publication (in revised form) by G. Kriegsmann August 29, 2003; published electronically December 22, 2003.
http://www.siam.org/journals/siads/2-4/41989.html
†Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (ins@maths.warwick.ac.uk).
‡Department of Mathematics, University of Houston, Houston, TX 77204-3476 (mg@uh.edu). The work of this author was supported in part by NSF grants DMS-0071735 and DMS-0244529 and ARP grant 003652-0032-2001.
§Department of Mathematics, Trent University, Peterborough, ON, Canada K9L 1Z6 (pivato@xaravve.trentu.ca).

system architecture leads naturally to synchrony. To do this, we must define carefully when two cells or two couplings are "identical" or "equivalent." Indeed, the main point of this paper is to provide a general mathematical foundation for these ideas. This foundation uses the algebraic structure of groupoids (see Brandt [1] and Higgins [10]) and greatly generalizes the uses of symmetry in coupled cell systems that we have explored previously [7, 8].

Our conventions do not rule out "two-way" coupling, in which cells A and B both influence each other. We represent such a state of affairs by having A coupled to B and B coupled to A. We also do not rule out coupling where cells A and B both influence cell C. Here we consider both A and B as being coupled to C. We do *not* assume the effects of A and B to be additive; in fact, the time-evolution of cell C can in principle be any (smooth) function of the states of C, A, and B.

In this paper we develop an abstract formalism for coupled cell systems, using simple examples that have no particular role in applications, but it is worth noting that coupled cell systems are used to model a variety of physically interesting systems. For examples, see [8] and references therein. We intend to develop applications of the formalism derived here in future work.

In this section we illustrate some central issues by discussing several examples.

Two-cell systems. We begin with the simplest system of two identical cells (with coordinates x_1 and x_2 in $\mathbf{R}^k$). Without making any specific assumption of the form of the "internal dynamics" of each cell or the form of the "coupling between cells," the differential equations for the coupled system have the form

$$\begin{aligned} \dot{x}_1 &= f(x_1, x_2), \\ \dot{x}_2 &= f(x_2, x_1); \end{aligned} \tag{1.1}$$

that is, the same function f governs the dynamics of both cells. There are three issues that we discuss concerning system (1.1): the graph (diagram, network) associated to a coupled cell system, symmetry, and synchrony.

Informally, the "network" of a coupled cell system is a finite directed graph whose nodes represent cells and whose edges represent couplings. Nodes are labeled to indicate "equivalent" cells, which have the same phase space and the same internal dynamic. Edges are labeled to indicate "equivalent" couplings. The graph associated to system (1.1) is given in Figure 1. We think of this graph as representing a pair of systems of differential equations in the following way. The two cells are indicated by identical symbols—so they have the same state variables. That is, the coordinates x_1 of cell 1 and x_2 of cell 2 lie in the same phase space $\mathbf{R}^k$. Since we can interchange cells 1 and 2 without changing the graph, we assume that the same is true for the system of differential equations and that they must have the form (1.1). Note that for this interchange to work, the arrow $1 \to 2$ must be the same as the arrow $2 \to 1$.

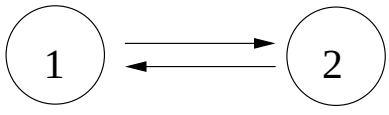


Figure 1. *A two-cell network.*

The discussion in the previous paragraph can be summarized by the following: the permutation $\sigma(x_1, x_2) = (x_2, x_1)$ is a symmetry of the system (1.1). Indeed, more is true: every system of differential equations on $\mathbf{R}^k \times \mathbf{R}^k$ that is equivariant with respect to σ has the form (1.1). That is, abstractly the study of pairs of identical cells that are identically coupled is the same as the study of σ-equivariant systems. Two consequences follow from this remark. First, synchrony in two-cell systems (solutions such that $x_1(t) = x_2(t)$ for all time t) is a robust phenomenon and should not be viewed as surprising. Second, time-periodic solutions can exhibit a kind of generalized synchrony in which the two cells oscillate a half-period out of phase.

The first remark can be restated as follows: the diagonal subspace $V = \{x_1 = x_2\} \subset \mathbf{R}^k \times \mathbf{R}^k$ is flow-invariant for every system (1.1). This remark can be verified in two ways. By inspection restrict (1.1) to V, obtaining

$$\begin{aligned} \dot{x}_1 &= f(x_1, x_1), \\ \dot{x}_1 &= f(x_1, x_1). \end{aligned}$$

It follows that if the initial conditions for a solution satisfy $x_1(0) = x_2(0)$, then $x_1(t) = x_2(t)$ for all time t, and V is flow-invariant. Alternately, we can observe that V is the fixed-point subspace $\mathrm{Fix}(\sigma)$, and fixed-point subspaces are well known to be flow-invariant.

The second remark is related to general theorems about spatio-temporal symmetries of time-periodic solutions to symmetric systems of ODEs. There are two types of theorems here: existence theorems, asserting that certain spatio-temporal symmetries are *possible*, and bifurcation theorems, describing particular scenarios that can generate such solutions. The H/K theorem [4, 7] is an existence theorem; indeed, it states necessary and sufficient conditions for periodic solutions with a given spatio-temporal symmetry group to be possible. In particular, it implies the existence of functions f having time-periodic solutions of period T satisfying

$$x_2(t) = x_1(t + T/2) \tag{1.2}$$

as long as the phase space of each cell has dimension $k \geq 2$. So states with this type of spatio-temporal pattern can exist. Indeed, they can exist robustly (that is, they can persist when f is perturbed) and are therefore typical in the appropriate coupled cell systems. In this case, we can say more: such solutions can arise through Hopf bifurcation. This is a consequence of the general theory of symmetric Hopf bifurcation, [7, 8, 9]. (Note that when $k = 1$, nonconstant periodic solutions satisfying (1.2) must intersect the diagonal V and hence be in V for all time: this is a contradiction.)

A three-cell network. Consider the three-cell network illustrated in Figure 2. The systems of differential equations corresponding to this network have the form

$$\begin{aligned} \dot{x}_1 &= f(x_1, x_2), \\ \dot{x}_2 &= g(x_2, x_1, x_3), \\ \dot{x}_3 &= f(x_3, x_2), \end{aligned} \tag{1.3}$$

where $g(x_2, x_1, x_3) = g(x_2, x_3, x_1)$, $x_1, x_3 \in \mathbf{R}^k$, and $x_2 \in \mathbf{R}^\ell$. Note that all such systems are equivariant with respect to the permutation $\tau(x_1, x_2, x_3) = (x_3, x_2, x_1)$ and that synchronous

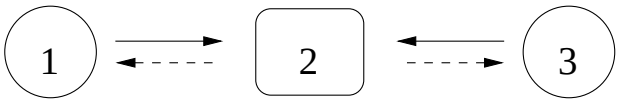


Figure 2. *A three-cell network with transposition symmetry.*

solutions (where $x_1(t) = x_3(t)$ for all time t) occur robustly because the "polydiagonal" subspace $W = \{x : x_1 = x_3\}$ is flow-invariant for (1.3).

There are two differences between the three-cell network in Figure 2 and the two-cell network in Figure 1. First, not all τ-equivariant systems on $\mathbf{R}^k \times \mathbf{R}^\ell \times \mathbf{R}^k$ have the form (1.3), since in the general τ-equivariant system f can depend nontrivially on both x_1 and x_3. So there can be additional structure in coupled cell systems that does not correspond directly to symmetry. Second, the half-period, out of phase, time-periodic solutions satisfy

$$x_3(t) = x_1(t + T/2) \quad \text{and} \quad x_2(t) = x_2(t + T/2). \tag{1.4}$$

In particular, the oscillations in cell 2 are forced by symmetry to occur at twice the frequency of those in cells 1 and 3. So multirhythms [7] can be forced by the architecture of coupled cell networks.

Another three-cell network. We now show that robust synchrony is possible in networks that have no symmetry. Consider the three-cell network in Figure 3. Here we have used two distinct symbols (square and circle) for cells and three types of arrows for couplings. The role of these symbols can be seen in the form of the ODE: identical symbols correspond to identical functions in the appropriate variables.

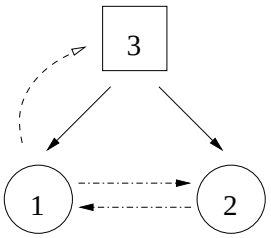


Figure 3. *A three-cell network without symmetry.*

This network has no symmetry, but the network structure forces the "polydiagonal" subspace $Y = \{x : x_1 = x_2\}$ to be flow-invariant. To verify this point observe that the coupled cell systems associated with this network have the form

$$\begin{aligned} \dot{x}_1 &= f(x_1, x_2, x_3), \\ \dot{x}_2 &= f(x_2, x_1, x_3), \\ \dot{x}_3 &= g(x_3, x_1), \end{aligned} \tag{1.5}$$

where $x_1, x_2 \in \mathbf{R}^k$ and $x_3 \in \mathbf{R}^\ell$. Restricting the first two equations to Y yields

$$\begin{aligned} \dot{x}_1 &= f(x_1, x_1, x_3), \\ \dot{x}_2 &= f(x_1, x_1, x_3), \end{aligned}$$

implying that Y is a flow-invariant subspace.

There is a precise sense in which cells 1 and 2 are equivalent within this network, and it is this observation that will enable us to prove the flow-invariance of subspaces like Y in a more abstract (and general) setting. Define the "input set" of a cell j to be the cell j and all cells i that connect to cell j. Also include the arrows from cells i to j. See Figure 4.

We can now explain why Y is flow-invariant, in terms of a permutation that acts on the network. This permutation is not a symmetry of the whole network, but it preserves enough structure to create a flow-invariant subspace. The key property is that the input sets of cells 1 and 2 are isomorphic via the permutation σ that maps $(1\ 2\ 3) \to (2\ 1\ 3)$.

If the system (1.5) were equivariant with respect to σ, then the fixed-point space of σ would be flow-invariant by [9, 7]. Moreover, the fixed-point space of σ is Y. However, (1.5) is *not* equivariant with respect to σ. Indeed, if we apply σ, then the equation transforms into

$$\begin{aligned} \dot{x}_2 &= f(x_2, x_1, x_3), \\ \dot{x}_1 &= f(x_1, x_2, x_3), \\ \dot{x}_3 &= g(x_3, x_2). \end{aligned} \tag{1.6}$$

The first two equations are the same as in (1.5), but the third equation is not. However, the third *is* the same on the space Y, where $x_2 = x_1$. So the restriction of the equations to Y is σ-equivariant, and this is enough to make Y flow-invariant.

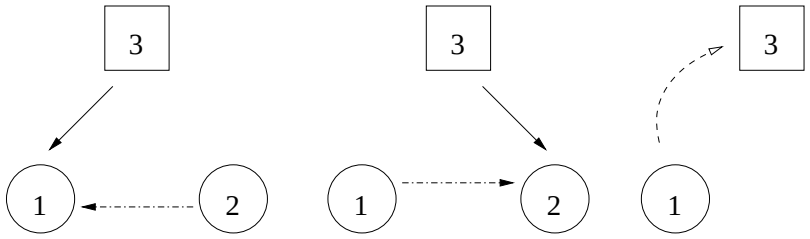


Figure 4. *Input sets for three-cell network without symmetry.*

Traveling waves in a seven-cell network. Consider the seven-cell linear network in Figure 5. The corresponding differential equations have the general form

$$\begin{aligned} \dot{x}_1 &= B(x_1), & \dot{x}_2 &= A(x_2, x_1), & \dot{x}_3 &= A(x_3, x_2), \\ \dot{x}_4 &= A(x_4, x_3), & \dot{x}_5 &= A(x_5, x_4), & \dot{x}_6 &= A(x_6, x_5), \\ & & \dot{x}_7 &= A(x_7, x_6). \end{aligned} \tag{1.7}$$

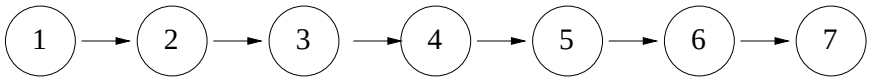


Figure 5. *Seven-cell linear network.*

It does not seem to be a simple matter to determine whether traveling waves are present in this network. If the cell phase spaces are all one-dimensional, there are no nontrivial

time-periodic states so no traveling waves. With higher-dimensional phase spaces, special assumptions are needed to produce traveling waves. However, if we introduce back coupling from cell 3 to cell 1, as shown in Figure 6, traveling waves can typically be expected, even in the one-dimensional case, as explained below. This is curious, because informally Figure 6 would normally be considered as being less regular in form than Figure 5. So the issue of "regular form" for a coupled cell network is fairly subtle. The key feature here is that all input sets for cells in the network in Figure 6 are isomorphic, whereas this is not true for the cells in Figure 5. It is this additional "symmetry" on the groupoid level that makes traveling waves typical. Indeed, Figure 6 has many groupoid symmetries (42 in all).

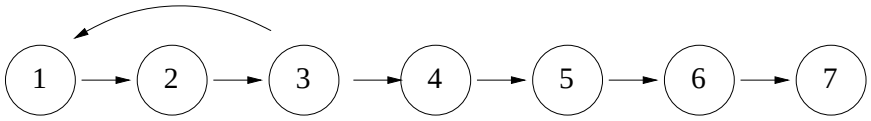


Figure 6. *Seven-cell linear network with back connection.*

We discuss why traveling wave solutions arise in two ways. First, the assumption that all of the cells and arrows in Figure 6 are identical implies that the first equation in (1.7) is now

$$\dot{x}_1 = A(x_1, x_3).$$

If we set

$$\begin{aligned} x_7 &= x_4 = x_1 = y_1, \\ x_6 &= x_3 = y_3, \\ x_5 &= x_2 = y_2, \end{aligned} \tag{1.8}$$

then the system of seven equations reduces to a three-equation system

$$\begin{aligned} \dot{y}_1 &= A(y_1, y_3), \\ \dot{y}_2 &= A(y_2, y_1), \\ \dot{y}_3 &= A(y_3, y_2), \end{aligned} \tag{1.9}$$

which is the general form associated with the directed ring of coupled cells in Figure 7. It is not hard to show using Hopf bifurcation (see [7, 9]) that the system (1.9) can support a discrete rotating wave $y(t)$, where

$$y_2(t) = y_1\left(t - \frac{T}{3}\right), \quad y_3(t) = y_2\left(t - \frac{T}{3}\right),$$

and y_1 is periodic of period T. This solution yields a *traveling wave* solution for the network in Figure 6. See a sample simulation in Figure 8. (The number 7 is not significant here: the same ideas work for any chain containing three or more cells and with feedback from any cell other than the first.)

More importantly, the three-cell ring in Figure 7 is a quotient network of the one in Figure 6, where the quotient map β takes x_1, x_4, x_7 to y_1, x_2, x_5 to y_2, and x_3, x_6 to y_3. We define "quotient" in section 8, but the key point is that solutions for the three-cell ring

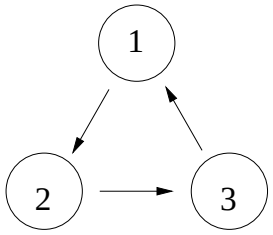


Figure 7. *Three-cell directed ring: Quotient of the network in Figure 6.*

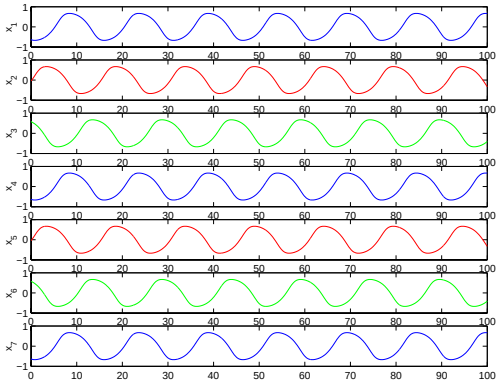


Figure 8. *Traveling wave solution in seven-cell chain.*

naturally "lift" to solutions for the seven-cell network via (1.8). The crucial features here are that β induces an isomorphism from each input set in the first network to an input set in the second network, and every coupled cell system of differential equations in the quotient lifts to a coupled cell system in the first network.

We seek to isolate the abstract structural features that are responsible for the behavior described in the above examples and to place the discussion in a rigorous formal context. We structure the paper as follows. Coupled cell networks are rigorously defined in terms of nodes and arrows in section 2. The key concept, the groupoid structure of a coupled cell network, is defined in section 3 in terms of input sets. The phase space and admissible vector fields associated to a coupled cell network are discussed in section 4. Basically, the intuitive ideas presented in this introduction are formalized as equivariance with respect to the symmetry groupoid of the network. Section 5 describes an extended example, which motivates the rest of the paper. In section 6 we introduce two different notions of robust synchrony: flow-invariant subspaces and balanced equivalence relations. We prove that these notions are equivalent. Quotient maps and quotient networks, which constitute a fourth equivalent notion for synchrony, are discussed in section 8. Quotient networks are an especially useful concept because they illuminate the generic dynamics of vector fields restricted to synchronous invariant subspaces, which can include phase-locked states and synchronized chaos. Examples illustrating these points are discussed in section 7. The relationship between the dynamics on a synchronous subspace and the induced dynamics on the quotient network is discussed in section 9.

2. Coupled cell networks. We begin by formally defining a coupled cell network. At this stage we consider only the abstract network architecture (a labeled directed graph). The associated family of ODEs is discussed in section 4.

Definition 2.1. *A* coupled cell network G *consists of the following:*

(a) *A finite set* $\mathcal{C} = \{1, \dots, N\}$ *of* nodes *or* cells.

(b) *A finite set of ordered pairs* $\mathcal{E} \subseteq \mathcal{C} \times \mathcal{C}$ *of* directed edges *or* arrows. *Each edge* (c, d) *has a* tail c *and a* head d.

(c) *An equivalence relation* $\sim_C$ *on cells in* $\mathcal{C}$. *The* type *or* cell label *of cell* c *is the* $\sim_C$*-equivalence class* $[c]_C$ *of* c.

(d) *An equivalence relation* $\sim_E$ *on edges in* $\mathcal{E}$. *The* type *or* coupling label *of edge* e *is the* $\sim_E$*-equivalence class* $[e]_E$ *of* e. *An edge* (c, c) *is an* internal edge*; a cell is* active *if it has an internal edge. We assume that every cell is active, that is,*

$$\Delta_{\mathcal{C}} = \{(c, c) : c \in \mathcal{C}\} \subset \mathcal{E}. \tag{2.1}$$

In addition, we require the following compatibility conditions*:*

(e) *Equivalent edges have equivalent tails and heads. That is, if* $(i, c) \sim_E (j, d)$*, then* $i \sim_C j$ *and* $c \sim_C d$.

(f) *Internal edges are equivalent if their tails are equivalent. Internal edges and noninternal edges are never equivalent. That is, for all* $c, d, d' \in \mathcal{C}$*,*

$$(c, c) \sim_E (d, d') \iff d = d' \text{ and } d \sim_C c.$$

Formally, the coupled cell network G *is the quadruple* $G = (\mathcal{C}, \mathcal{E}, \sim_C, \sim_E)$.

We represent a coupled cell network G by a diagram constructed as follows.

(1) For each $\sim_C$-equivalence class of cells choose a distinct *node symbol* $\bigcirc, \square, \triangle$, and so on.

(2) For each $\sim_E$-equivalence class of noninternal edges, choose a distinct *arrow* $\rightarrow, \Rightarrow, \rightsquigarrow$, and so on.

The compatibility conditions in Definition 2.1 state that arrows between distinct cells can be identical only when the nodes at the heads are identical and the nodes at the tails are identical, and that node symbols can be interpreted as arrows from a cell to itself.

The above definition is essentially the standard concept of a directed graph (or digraph) in graph theory (see, for example, Tutte [13], Wilson [16]) modified to incorporate labeling of nodes and edges. We assume that the graph is finite because this makes the associated dynamical systems (discussed in section 4) finite-dimensional. However, most of the theory generalizes to infinite graphs. The assumption that all cells are active can be removed, at the expense of notational complications, but the details are routine and we do not treat this case here.

Example 2.2. Suppose that the network G is defined by

$$\begin{aligned}
&\mathcal{C} = \{1, 2, 3, 4\},\\
&\mathcal{E} = \{(1, 2), (1, 3), (2, 4), (3, 1), (3, 4)(4, 1)\},\\
\sim_C\ &\text{has equivalence classes } \{1\}, \{2, 3\}, \{4\},\\
\sim_E\ &\text{has equivalence classes } \{(1, 2), (1, 3)\}, \{(2, 4), (3, 4)\}, \{(3, 1)\}, \{(4, 1)\},
\end{aligned}$$

and all nodes are active. Then the diagram of G has the form shown in Figure 9 for the given choices of symbols.

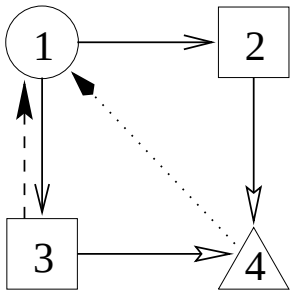


Figure 9. *Example of the diagram of a coupled cell network.*

3. Input sets and groupoids. In this section we define the basic algebraic structure of a coupled cell network—its symmetry groupoid. Some preliminary concepts are required.

Input sets. As discussed in more detail in section 4 the variables that appear in a given component f_c of the vector fields f associated to a coupled cell network depend only on those cells that are linked to cell c by an arrow. This observation is abstracted as the following definition.

Definition 3.1. *The* input set $I(c)$ *of a cell* c *is*

$$I(c) = \{i \in \mathcal{C} : (i, c) \in \mathcal{E}\}.$$

Since all cells are active, $c \in I(c)$*. We call* c *the* base cell *of* $I(c)$*.*

Two cells with isomorphic input sets will be called "input equivalent."

Definition 3.2. *The relation* $\sim_I$ *of* input equivalence *on* $\mathcal{C}$ *is defined by* $c \sim_I d$ *if and only if there exists a base cell preserving bijection*

$$\beta : I(c) \to I(d) \tag{3.1}$$

(by which we mean that $\beta(c) = d$*) such that for all* $i \in I(c)$

$$(i, c) \sim_E (\beta(i), d). \tag{3.2}$$

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*.*

On setting $i = c$ in (3.2), we see that c and d have the same type ($c \sim_C d$) if they are input equivalent ($c \sim_I d$). The converse is easily seen to be false. Moreover, $B(c, d)$ is empty unless $c \sim_I d$.

An important class of networks is one in which all cells are input equivalent. So we define the following.

Definition 3.3. *A* homogeneous *network is a coupled cell network such that* $B(c, d) \neq \emptyset$ *for every pair of cells* c, d*.*

Examples of homogeneous networks are given in Figures 14 and 15.

Example 3.4. We return to Example 2.2 whose diagram is shown in Figure 9. The input sets are shown in Figure 10. There are three $\sim_I$-equivalence classes: $\{1\}, \{2,3\}$, and $\{4\}$. The isomorphism between $I(2)$ and $I(3)$ is the bijection $\tau : \{2,1\} \to \{3,1\}$ for which $\tau(2) = 3$ and $\tau(1) = 1$.

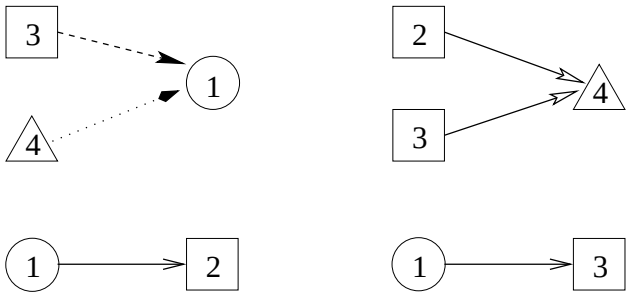


Figure 10. *Input sets for Figure 9.*

The symmetry groupoid. We now introduce the central concept of this paper, the "symmetry groupoid" of a coupled cell network. The symmetry groupoid is a generalization of the symmetry group of a symmetric network. It includes not just symmetries of the whole network, but symmetries between particular subgraphs—namely, the input sets.

Definition 3.5. *The* symmetry groupoid *of a coupled cell network G is the disjoint union*

$$\mathcal{B}_G = \dot{\bigcup}_{c,d \in \mathcal{C}} B(c,d).$$

The term "groupoid" was introduced by Brandt [1] and is developed at length in Higgins [10]. The term refers to an algebraic structure that is similar to a group, with the exception that products of elements may not always be defined. Different authors formalize groupoids in slightly different (but mostly equivalent) ways. Essentially, a groupoid must satisfy three conditions:

(1) The product operation is associative in the sense that whenever one of $\alpha(\beta\gamma)$ and $(\alpha\beta)\gamma$ is defined, then so is the other, and they are equal.
(2) There are distinguished elements ε_j that act as identity elements, in the sense that $\varepsilon_j\alpha = \alpha$ and $\alpha\varepsilon_j = \alpha$ whenever these are defined. (Here the indices j correspond to the "objects" of the groupoid, which in our case are the cells.)
(3) Every element α has an inverse α^{-1}, in the sense that both $\alpha\alpha^{-1}$ and $\alpha^{-1}\alpha$ are identities.

In the case of $\mathcal{B}_G$, the groupoid structure is captured by the following:

(1) We define the product of $\beta_1 \in B(c,d)$ and $\beta_2 \in B(c',d')$ if and only if $c' = d$, and then we set $\beta_2\beta_1 = \beta_2 \circ \beta_1 \in B(c,d')$, where $\circ$ denotes composition of maps. Composition is of course associative when it is defined.
(2) The identity elements $\mathrm{id}_{I(c)}$ for $c \in \mathcal{C}$ are the groupoid identity elements.
(3) For inverses, observe that $\beta \in B(c,d)$ if and only if $\beta^{-1} \in B(d,c)$.

It follows in particular that $B(c,c)$ is a group, the *vertex group* corresponding to c. Vertex groups are important in groupoid theory and play a key role in this paper.

Remark 3.6. The term "disjoint union" in Definition 3.5 is used in a technical sense. The sets $B(c,d)$ for different pairs (c,d) are not necessarily disjoint. For example, if $\mathcal{C} = \{1,2\}$, where cells 1 and 2 are inequivalent, each coupled to the other by inequivalent arrows, then $B(1,1)$ and $B(2,2)$ both consist of the identity map on $\{1,2\}$. It is convenient to *make* them disjoint. One way to do this is to replace each bijection $\beta \in B(c,d)$ by the triple (β,c,d). Then β defines the permutation, c is an index specifying its "domain," and d is an index specifying its "range." As far as the groupoid structure is concerned, the product $\beta_2\beta_1$ is defined only when $c' = d$. This occurs when the "range" of β_1 is equal to the "domain" of β_2, in the sense just specified. However, the set-theoretic ranges and domains of the corresponding bijections may permit the composition of β_1 and β_2 as *functions* in cases where we do not wish to permit them to be multiplied in the groupoid.

The point here is that we are not dealing merely with bijections on sets but with base point preserving bijections on based sets. Composition must respect the base points as well as the sets.

For simplicity, we use β to denote an input isomorphism, rather than the cumbersome (β,c,d), because the appropriate c,d are usually obvious.

Alternatively, it would be possible to represent the input structure of a given cell as consisting of that cell, all cells connected to it, *and* the associated arrows. Input isomorphisms then necessarily preserve the base point (the cell at the head of all arrows). Technically, however, this choice causes other complications—for example, input isomorphisms now act on subgraphs and not on subsets of cells—so we shall not use it here. However, it is a useful informal way to visualize input isomorphisms.

Example 3.7. Again, we return to Example 2.2. The nonempty sets $B(c,d)$ are as follows:

$B(1,1)$: The identity map on $\{1,3,4\}$.

$B(2,2)$: The identity map on $\{2,1\}$.

$B(3,3)$: The identity map on $\{3,1\}$.

$B(4,4)$: The identity map on $\{4,2,3\}$ and the permutation σ on $\{4,2,3\}$ for which $\sigma(2) = 3, \sigma(3) = 2, \sigma(4) = 4$.

$B(2,3)$: The map $\tau : \{2,1\} \to \{3,1\}$ for which $\tau(1) = 1, \tau(2) = 3$.

$B(3,2)$: The inverse τ^{-1} of τ.

Subgroupoids and connected components. For the basics of groupoids see Brandt [1], Brown [3], Higgins [10], and MacLane [11]. For applications see Weinstein [15]. Groupoids combine several features of groups with features of graphs, and we discuss one of each now. The group-theoretic notion is that of a subgroupoid; the graph-theoretic one is that of a connected component.

A subset $\mathcal{S} \subset \mathcal{B}_G$ is a *subgroupoid* if $\mathcal{S}$ is closed under products (when defined) and taking inverses.

The *connected components* of the groupoid $\mathcal{B}_G$ are in one-to-one correspondence with $\sim_I$-equivalence classes on $\mathcal{C}$. Specifically, let A be a $\sim_I$-equivalence class. Then the subgroupoid

$$\mathcal{S}(A) = \dot{\bigcup}_{c,d \in A} B(c,d) \tag{3.3}$$

is a *connected component* of $\mathcal{B}_G$. Moreover, we have the following lemma.

Lemma 3.8. *The groupoid $\mathcal{B}_G$ is the disjoint union of its connected components. That is,*

$$\mathcal{B}_G = \dot{\bigcup}_A \mathcal{S}(A),$$

where A runs through the $\sim_I$-equivalence classes in $\mathcal{C}$. Moreover, if A, A' are two such distinct classes, and $\beta \in \mathcal{S}(A), \beta' \in \mathcal{S}(A')$, then the product $\beta\beta'$ is not defined.

We say that two cells $c, d \in \mathcal{C}$ are in the *same connected component* of $\mathcal{B}_G$ if and only if $c \sim_I d$. The mental image here is that associated with any groupoid there is a graph, whose elements are the vertices of the groupoid and whose (directed) edges are the groupoid elements. It is the connected components of this graph that are being described. See Higgins [10, Chapter 3].

If c, d belong to the same connected component, then the vertex groups $B(c, c)$ and $B(d, d)$ are *conjugate*, in the sense that there exists $\gamma \in B(c, d)$ such that

$$B(c, c) = \gamma^{-1} B(d, d)\gamma.$$

In particular, $B(c, c)$ and $B(d, d)$ are isomorphic groups.

Structure of $B(c, d)$. For later use, we determine the general structure of the sets $B(c, d)$.

(1) If $c \not\sim_I d$, then $B(c, d) = \emptyset$.

(2) If $c = d$, then we define an equivalence relation $\equiv_c$ on $I(c)$ by

$$j_1 \equiv_c j_2 \quad \Longleftrightarrow \quad (j_1, c) \sim_E (j_2, c)$$

for $j_1, j_2 \in I(c)$. Let the $\equiv_c$-equivalence classes of $I(c)$ be $K_0, \dots, K_r$ for $r = r(c)$ so that

$$I(c) = K_0 \,\dot{\cup} \cdots \dot{\cup} K_r. \tag{3.4}$$

We may choose $K_0 = \{c\}$ by Definition 2.1(f). Let

$$k_s = |K_s| \qquad (0 \le s \le r).$$

Then $B(c, c)$ is a group given by

$$B(c, c) = \mathbf{S}_{k_1} \times \cdots \times \mathbf{S}_{k_r},$$

where each $\mathbf{S}_{k_s}$ comprises all permutations of K_s, extended by the identity on $I(c) \setminus K_s$.

(3) If $c \sim_I d$ and $c \neq d$, define $\equiv_d$ on $I(d)$ in the same way. Let $\beta \in B(c, d)$, and let

$$L_s = \beta(K_s) \qquad (0 \le s \le r(c)).$$

Then β and β^{-1} preserve $\sim_E$, so the $\equiv_d$-equivalence classes of $I(d)$ are the $L_s (0 \le s \le r(c))$, and $r(d) = r(c)$.

Choose a fixed but arbitrary $\beta_0 \in B(c, d)$, having the above property. Then

$$B(c, d) = B(d, d)\beta_0 = \beta_0 B(c, c).$$

Conversely, any $\beta_0 : I(c) \to I(d)$ such that $L_s = \beta_0(K_s)$ for $0 \le s \le r(c)$ lies in $B(c, d)$.

4. Vector fields on a coupled cell network. We now define the class $\mathcal{F}_G^P$ of vector fields corresponding to a given coupled cell network G. This class consists of all vector fields that are "compatible" with the labeled graph structure or, equivalently, are "symmetric" under the groupoid $\mathcal{B}_G$. It also depends on a choice of "total phase space" P, which we assume is fixed throughout the subsequent discussion. For example, in the two-cell system (1.1) we have $P = \mathbf{R}^k \times \mathbf{R}^k$, which depends on the choice of k.

For each cell in $\mathcal{C}$ define a *cell phase space* P_c. This must be a smooth manifold of dimension ≥ 1, which for simplicity we assume is a nonzero finite-dimensional real vector space. We require

$$c \sim_C d \quad \Longrightarrow \quad P_c = P_d,$$

and we employ the same coordinate systems on P_c and P_d. Only these identifications are canonical. If $P_c = P_d$ or P_c is isomorphic to P_d when $c \not\sim_C d$, then the identification of P_c and P_d will be deemed accidental and will have no significance for our present purposes. Although the relation $c \sim_C d$ means that cells c and d have the same phase space, it does not imply that they have isomorphic (that is, conjugate) dynamics.

Define the corresponding *total phase space* to be

$$P = \prod_{c \in \mathcal{C}} P_c$$

and employ the coordinate system

$$x = (x_c)_{c \in \mathcal{C}}$$

on P.

The *cell projection* corresponding to cell c is the natural projection

$$\pi_c : P \to P_c.$$

More generally, suppose that $\mathcal{D}$ is any subset of $\mathcal{C}$. Define

$$P_{\mathcal{D}} = \prod_{c \in \mathcal{D}} P_c$$

and let

$$\pi_{\mathcal{D}} : P \to P_{\mathcal{D}}$$

be the natural projection. Further, write

$$x_{\mathcal{D}} = \pi_{\mathcal{D}}(x)$$

and suppress braces when $\mathcal{D}$ is a singleton. That is, $\pi_c(x) = x_c = x_{\{c\}}$.

Finally, suppose that $\mathcal{D}_1, \mathcal{D}_2$ are subsets of $\mathcal{C}$ and that there is a bijection $\beta : \mathcal{D}_1 \to \mathcal{D}_2$ such that $\beta(d) \sim_C d$ for all $d \in \mathcal{D}_1$. Define the *pullback map*

$$\beta^* : P_{\mathcal{D}_2} \to P_{\mathcal{D}_1}$$

by

$$(\beta^*(z))_j = z_{\beta(j)} \qquad \forall j \in \mathcal{D}_1, z \in P_{\mathcal{D}_2}. \tag{4.1}$$

By direct calculation it is easy to verify three simple properties of the pullback:

$$\begin{array}{rcl} (\beta\gamma)^* & = & \gamma^*\beta^*, \\ \mathrm{id}^* & = & \mathrm{id}, \\ (\gamma^{-1})^* & = & (\gamma^*)^{-1}. \end{array} \tag{4.2}$$

Note the reversed order in the first of these equations.

We use pullback maps to relate different components of the vector field associated with a given coupled cell network. Specifically, the class of vector fields that is encoded by a coupled cell network will be defined using the following concept.

Definition 4.1. *A vector field $f : P \to P$ is* $\mathcal{B}_G$-equivariant *or* G-admissible *if the following hold:*

(a) *For all $c \in \mathcal{C}$ the component $f_c(x)$ depends only on $x_{I(c)}$; that is, there exists $\hat{f}_c : P_{I(c)} \to P_c$ such that*

$$f_c(x) = \hat{f}_c(x_{I(c)}). \tag{4.3}$$

(b) *For all $c, d \in \mathcal{C}$ and $\beta \in B(c, d)$ (so that, in particular, $d = \beta(c)$)*

$$\hat{f}_d(x_{I(d)}) = \hat{f}_c(\beta^*(x_{I(d)})) \qquad \forall x \in P. \tag{4.4}$$

For brevity, we write this condition as

$$f_{\beta(c)}(x) = f_c(\beta^*(x)) \qquad \forall x \in P. \tag{4.5}$$

When using (4.5) it is necessary to bear in mind the constraint that $f_d(x)$ depends only on $x_{I(d)}$. Otherwise, $\beta^*(x)$ is not defined.

We call (a) the *domain condition* and (b) the *equivariance condition* on f.

Remark 4.2. If β belongs to the vertex group $B(c, c)$, then (4.5) implies that

$$f_c(\beta^*(x)) = f_c(x) \qquad \forall x \in P. \tag{4.6}$$

That is, f_c is $B(c, c)$-*invariant*. It is easy to check that this property is the same as the usual property of invariance under a group, provided we consider $B(c, c)$ as acting on $P_{I(c)}$.

Definition 4.3. *For a given choice of the P_c we define the class $\mathcal{F}_G^P$ to consist of all G-admissible vector fields on P.*

These are the most general vector fields on P that are consistent with the coupled cell network.

Example 4.4. We describe $\mathcal{F}_G^P$ for the diagram of Figure 9. There are three cell types $\bigcirc, \square, \triangle$, and we choose three corresponding phase spaces U, V, W. Then the state variable is

$x = (x_1, x_2, x_3, x_4)$, where $x_1 \in U$, $x_2, x_3 \in V$, $x_4 \in W$. There are four arrow types. We claim that the G-admissible vector fields f are those of the form

$$\begin{array}{rcll} f_1(x) & = & A(x_1, x_3, x_4), & \text{where } A : U \times V \times W \to U, \\ f_2(x) & = & B(x_2, x_1), & \text{where } B : V \times U \to V, \\ f_3(x) & = & B(x_3, x_1), & \\ f_4(x) & = & C(x_4, x_2, x_3), & \text{where } C : W \times V \times V \to W, \end{array}$$

and C is symmetric in x_2, x_3.

To prove this, we consider the equivariance condition (4.5) for all the bijections β listed in Example 3.7. There are two nontrivial cases: $B(2,3)$ and $B(4,4)$. First, suppose that $c = 2 and d = 3$, and consider the bijection $\tau : I(2) \to I(3)$ for which $\tau(2) = 3, \tau(1) = 1$. Suppose that we define the function $B : P_{I(2)} \to P_2$ by

$$B(x_2, x_1) = f_2(x)$$

so that $B = \hat{f}_2$. Then $f_3(x) = \hat{f}_3(x_3, x_1)$, and we wish to express this in terms of B.

It is easy to work out the pullback of τ. If we write the elements of $P_{I(3)}$ in the form $x = (x_3, x_1)$, then $y = \tau^*(x)$ takes the form $y = (y_2, y_1) \in P_{I(2)}$, where

$$\begin{aligned} y_2 &= (x)_{\tau(2)} = x_3, \\ y_1 &= (x)_{\tau(1)} = x_1. \end{aligned}$$

Then

$$\tau^*(x_3, x_1) = (x_3, x_1),$$

and condition (4.5) tells us that

$$\hat{f}_3(x) = B(x_3, x_1)$$

as claimed. (The pullback τ^* is not the identity, because its range and domain are different. It is an *identification*.)

Similarly, if we consider $\sigma \in B(4,4)$, then we have a function C defined by $f_4(x) = C(x_4, x_2, x_3)$. Now the pullback $\sigma^* : P_{I(4)} \to P_{I(4)}$ acts as

$$\sigma^*(x_4, x_2, x_3) = (x_4, x_3, x_2),$$

and condition (4.5) tells us that

$$C(x_4, x_2, x_3) = \hat{f}_4(x) = C(x_4, x_3, x_2)$$

so that C is symmetric in x_2, x_3.

Here and from now on we adopt the convention that x_c is the first variable listed in the argument of $\hat{f}_c$. We can show that $\hat{f}_c$ is symmetric in some subset of variables by putting a bar over that set so that here

$$f_4(x) = C(x_4, \overline{x_2, x_3}).$$

(To do this, we have to order the variables suitably, and in some cases this cannot be done consistently. The use of a bar is convenient for the purposes of this paper.) Note that the network G is *not* symmetric under the 2-cycle (2 3), because the arrow from cell 3 to cell 1 does not correspond to an arrow from cell 2 to cell 1.

Admissible vector fields. The proofs of the main theorems of this paper rely on the construction of certain special G-admissible vector fields. In this subsection we describe these constructions.

As motivation, consider Example 4.4. Here, the most general G-admissible vector field is specified by three functions A, B, C. These functions can be assigned independently of each other. There is one such function for each $\sim_I$-equivalence class of cells, that is, each connected component of $\mathcal{B}_G$. If $c \in \mathcal{C}$, then f_c is $B(c,c)$-invariant; if $d \sim_I c$, then f_d is uniquely defined by f_c through the condition of $\mathcal{B}_G$-equivariance. We now give a formal statement of these properties and show that they are valid in general.

The main point is that Lemma 3.8 implies that $\mathcal{B}_G$-equivariance imposes conditions relating components f_c, f_d of f when c, d lie in the same connected component of $\mathcal{B}_G$, but not otherwise. We can therefore construct G-admissible vector fields g on P whose components g_c are zero for all c outside a fixed $\sim_I$-equivalence class. We will prove that such vector fields span $\mathcal{F}_G^P$.

Let $Q \subseteq \mathcal{C}$ be a $\sim_I$-equivalence class. Define

$$\mathcal{F}_G^P(Q) = \{f \in \mathcal{F}_G^P : f_s(x) = 0 \quad \forall\, s \notin Q\}. \tag{4.7}$$

Vector fields in $\mathcal{F}_G^P(Q)$ are *supported on* Q. The subset $\mathcal{F}_G^P(Q)$ is a linear subspace of $\mathcal{F}_G^P$.

The key constraint on a vector field in $\mathcal{F}_G^P(Q)$ is $B(q,q)$-equivariance for some fixed but arbitrary $q \in Q$. In fact we have the following lemma.

Lemma 4.5. *Given a $\sim_I$-equivalence class $Q \subseteq \mathcal{C}$, let $q \in Q$ and let $g_q : P_{I(q)} \to P_q$ be any $B(q,q)$-invariant mapping. Then g_q extends uniquely to a vector field in $\mathcal{F}_G^P(Q)$.*

Proof. For any $r \in Q$, choose $\beta_0 \in B(q,r)$ (which exists since $r \sim_I q$) so that $\beta_0(q) = r$. Equivariance forces us to define

$$g_r(y) = g_q(\beta_0^*(y)) \qquad \forall y \in P_{I(r)}, \tag{4.8}$$

so the extension to $r \in Q$ is unique if it exists. It is easy to show that g_r does not depend on the choice of β_0. Finally, if $r \notin Q$, we define $g_r(x) = 0$.

We have now extended g_q to a vector field g on the whole of P. We claim that $g \in \mathcal{F}_G^P(Q)$. Clearly, the components g_r of g with $r \notin Q$ vanish. It is therefore sufficient to show that if $r \in Q$, $\gamma \in B(r,s)$, and $z \in P_{I(s)}$, then

$$g_s(z) = g_r(\gamma^*(z)). \tag{4.9}$$

The component g_s is defined by choosing $\beta_1 \in B(q,s)$ and setting

$$g_s(y) = g_q(\beta_1^*(y)) \qquad \forall y \in P_{I(s)}. \tag{4.10}$$

To establish (4.9), let

$$\delta = \beta_1^{-1}\gamma\beta_0 \in B(q,q)$$

so that

$$\gamma = \beta_1 \delta \beta_0^{-1}.$$

Then, using (4.2), we compute

$$\begin{aligned} g_r(\gamma^*(z)) &= g_r((\beta_1\delta\beta_0^{-1})^*(z)) \\ &= g_r((\beta_0^*)^{-1}\delta^*\beta_1^*(z)) \\ &= g_q(\beta_0^*(\beta_0^*)^{-1}\delta^*\beta_1^*(z)) \\ &= g_q(\delta^*(\beta_1^*(z))) \\ &= g_q(\beta_1^*(z)) \\ &= g_s(z) \end{aligned}$$

(where $g_q(\delta^*(\beta_1^*(z))) = g_q(\beta_1^*(z))$ because $\beta_1^*(z) \in I(q)$ and g_q is $B(q,q)$-invariant). This calculation proves (4.9). ■

The importance of such vector fields g stems from the following proposition.

Proposition 4.6.

$$\mathcal{F}_G^P = \bigoplus_Q \mathcal{F}_G^P(Q),$$

where Q runs over the $\sim_I$-equivalence classes of G.

Proof. Suppose that $f \in \mathcal{F}_G^P$, so that f is $\mathcal{B}_G$-equivariant. Let Q be a $\sim_I$-equivalence class, and pick $q \in Q$. Define $g \in \mathcal{F}_G^P(Q)$ by setting

$$g_q(x) = f_q(x) \qquad \forall x \in P,$$

which is $B(q,q)$-invariant since f is $\mathcal{B}_G$-equivariant. For the same reason,

$$g_r(x) = f_r(x) \qquad \forall x \in P, r \in Q,$$

where g_r is defined as in Lemma 4.5. Recall that $g_s(x) = 0$ for all $s \notin Q$. Repeating this construction for all $\sim_I$-equivalence classes Q we see that

$$\mathcal{F}_G^P = \sum \mathcal{F}_G^P(Q).$$

However, the definition of $\mathcal{F}_G^P(Q)$ shows that

$$\mathcal{F}_G^P(Q) \cap \sum_{R \neq Q} \mathcal{F}_G^P(R) = \{0\}$$

(where R ranges over $\sim_I$-equivalence classes other than Q), so the sum is direct. ■

5. Patterns of synchrony: Example. There are many kinds of synchrony in coupled cell systems: for surveys see Boccaletti, Pecora, and Pelaez [2] and Wang [14]. Most notions of synchrony depend on specific dynamics of cells and couplings. Some notions are model-independent; that is, they are valid for any vector field consistent with the given cell architecture. We believe that it is useful to distinguish model-independent properties from model-dependent ones, because this separates the effect of the general architecture of the system from that of the specific model equations employed, which clarifies the role of the model and its parameters.

We now approach the central issue of this paper: conditions under which certain cells in a coupled cell network can synchronize *as a consequence of the network architecture*. Because the theoretical issues are somewhat abstract, we first discuss a motivating example.

Example 5.1. Consider the ten-cell network G_1 of Figure 11. There are two cell types. Cells 0 and 1 have type ◯, and cells 2, 3, 4, 5, 6, 7, 8, 9 have type □. There are three arrow types $\longrightarrow, \Longrightarrow, - - \to$. The shading on the nodes divides $\mathcal{C}$ into three classes:

$$\{0,1\},\{2,3,6,8\},\{4,5,7,9\}. \tag{5.1}$$

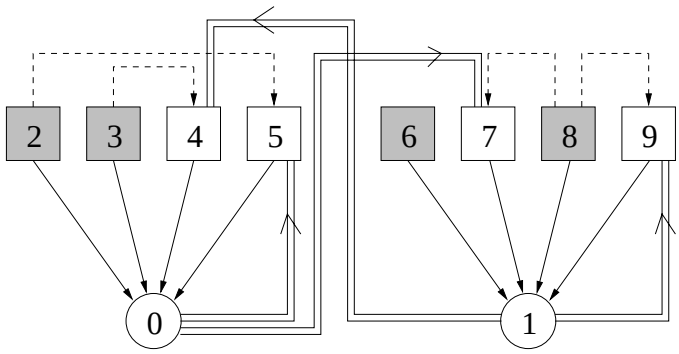


Figure 11. *A* 10*-cell system. The shading indicates a possible pattern of synchronous cells.*

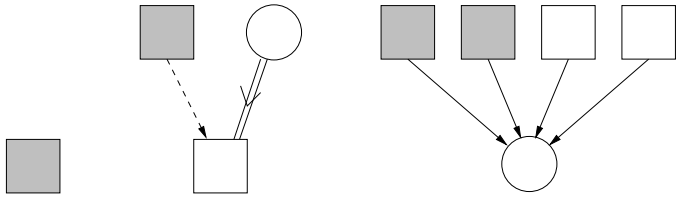


Figure 12. *Input types for the* 10*-cell system, including shading.*

There are three distinct input types, illustrated in Figure 12. Cells 2, 3, 6, 8 have no inputs except themselves. Cells 4, 5, 7, 9 have two inputs: one is of type ◯, and the other is of type □ and in the class $\{2,3,6,8\}$. Cells 0, 1 have four inputs, all of type □; of these, two are in the class $\{2,3,6,8\}$, and the other two are in the class $\{4,5,7,9\}$.

(We have deliberately included some cells without inputs—in this case, cells 2, 3, 6, and 8—to make it clear that such cells can be considered synchronous in our formalism. Of course such a form of synchrony is dynamically unstable, but in this paper we are studying existence, not stability, and we do not wish to rule out unstable synchrony because it still corresponds to an invariant subspace for the dynamics. It is easy to modify this example to add further connections that provide inputs to these cells: for example, each of cells 2, 3, 6, and 8 can receive one input from any of the cells 4, 5, 7, and 9.)

With appropriate choice of phase spaces, a vector field $f \in \mathcal{F}^P_{G_1}$ takes the form

$$\begin{array}{llll} f_0 = A(x_0, \overline{x_2, x_3, x_4, x_5}), & & f_5 = C(x_5, x_0, x_2), \\ f_1 = A(x_1, \overline{x_6, x_7, x_8, x_9}), & & f_6 = B(x_6), \\ f_2 = B(x_2), & & f_7 = C(x_7, x_0, x_8), \\ f_3 = B(x_3), & & f_8 = B(x_8), \\ f_4 = C(x_4, x_1, x_3), & & f_9 = C(x_9, x_1, x_8). \end{array} \tag{5.2}$$

Consider the space

$$Y = \{(u, u, v, v, w, w, v, w, v, w)\}$$

determined by making entries constant on the classes (5.1). On Y the vector field f restricts to

$$\begin{array}{llll} g_0 = A(u, \overline{v, v, w, w}), & & g_5 = C(w, u, v), \\ g_1 = A(u, \overline{v, w, v, w}), & & g_6 = B(v), \\ g_2 = B(v), & & g_7 = C(w, u, v), \\ g_3 = B(v), & & g_8 = B(v), \\ g_4 = C(w, u, v), & & g_9 = C(w, u, v). \end{array} \tag{5.3}$$

By symmetry g_0, g_1 are identical. Bearing this in mind, we see that Y is flow-invariant for f.

Identifying elements of Y with triples (u, v, w), we obtain an induced vector field $\overline{f}$ of the form

$$\begin{aligned} \overline{f}_0 &= A'(u, \overline{v, w}), \\ \overline{f}_1 &= B(v), \\ \overline{f}_3 &= C(w, u, v), \end{aligned}$$

where

$$A'(u, \overline{v, w}) = A(u, \overline{v, v, w, w}).$$

This is the class of admissible vector fields for the simpler coupled cell network G_2 shown in Figure 13. Here cells v, w have the same *type*, but we have shaded cell v to show which equivalence class it corresponds to.

The coupled cell network G_2 is an example of a quotient network. What structure in G_1 makes Y flow-invariant for all $f \in \mathcal{F}^P_{G_1}$ and permits this reduction to G_2 on Y? The key feature is how the three classes (5.1) relate to input isomorphisms. In section 6 we develop the theory of flow-invariant subspaces, and in section 8 we develop the general theory of such reductions.

6. Patterns of synchrony: Theory. We now isolate the abstract features of Example 5.1 that make Y flow-invariant. The classes (5.1) can be represented as the equivalence classes corresponding to an equivalence relation. The properties of this equivalence relation, relative to the symmetry groupoid of the network, turn out to control the existence of the flow-invariant subspace Y and the quotient network G_2. We begin by considering the equivalence relation.

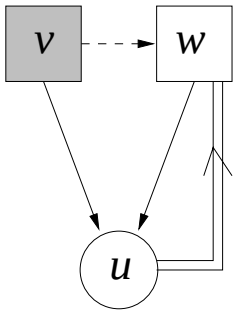


Figure 13. *Quotient network G_2 of the 10-cell system G_1. Shading relates cells to those in G_1.*

Let $G = (\mathcal{C}, \mathcal{E}, \sim_C, \sim_E)$ be a coupled cell network. Choose a total phase space P, and let $\bowtie$ be an equivalence relation on $\mathcal{C}$, partitioning the cells into equivalence classes. We assume that $\bowtie$ is a refinement of $\sim_{\mathcal{C}}$; that is, if $c \bowtie d$, then c and d have the same cell labels. It follows that the *polydiagonal subspace*

$$\Delta_{\bowtie} = \{x \in P : x_c = x_d \text{ whenever } c \bowtie d \quad \forall c, d \in \mathcal{C}\}$$

is well defined since x_c and x_d lie in the same space $P_c = P_d$. The polydiagonal $\Delta_{\bowtie}$ is a linear subspace of P.

For instance, in Example 5.1 we can define $\bowtie$ to have equivalence classes (5.1), in which case $\Delta_{\bowtie} = Y$.

Definition 6.1. *A trajectory $x(t)$ of $f \in \mathcal{F}_G^P$ is* $\bowtie$-polysynchronous *if its components are constant on $\bowtie$-equivalence classes. That is,*

$$c \bowtie d \Longrightarrow x_c(t) = x_d(t) \qquad \forall t \in \mathbf{R}$$

or $x(t) \in \Delta_{\bowtie}$ for all $t \in \mathbf{R}$.

Polysynchronous states are *patterns of synchrony*. Trivially, any trajectory is polysynchronous with respect to the relation of equality (which partitions $\mathcal{C}$ into its individual cells). Only nontrivial polysynchrony is interesting.

Robust polysynchrony.

Definition 6.2. *Let $\bowtie$ be an equivalence relation on $\mathcal{C}$. Then $\bowtie$ is* robustly polysynchronous *if $\Delta_{\bowtie}$ is invariant under every vector field $f \in \mathcal{F}_G^P$. That is,*

$$f(\Delta_{\bowtie}) \subseteq \Delta_{\bowtie} \quad \forall f \in \mathcal{F}_G^P.$$

Equivalently, if $x(t)$ is a trajectory of any $f \in \mathcal{F}_G^P$, with initial condition $x(0) \in \Delta_{\bowtie}$, then $x(t) \in \Delta_{\bowtie}$ for all $t \in \mathbf{R}$.

We now find necessary and sufficient conditions on $\bowtie$ to ensure that $\bowtie$ is robustly polysynchronous. We begin by showing that robust polysynchrony can occur only between cells that have isomorphic input sets. This is intuitively clear because these are the only cells that involve the "same" function in the corresponding components of admissible vector fields, and the proof bears out this intuition.

Lemma 6.3. *If $\bowtie$ is robustly polysynchronous, then $\bowtie$ refines $\sim_I$. That is, for all $c, d \in \mathcal{C}$*

$$c \bowtie d \Longrightarrow c \sim_I d.$$

Proof. By the discussion immediately preceding Lemma 3.8 we need to show that if $c \bowtie d$, then c and d are in the same connected component Q of $\mathcal{B}_G$. Suppose they are not; then we will show that $\Delta_\bowtie$ is not flow-invariant. Choose $x(0) \in \Delta_\bowtie$ so that $x_c(0) = x_d(0) \neq 0$, and choose $f \in \mathcal{F}_G^P(Q)$, where $d \notin Q$. Let $x(t)$ be the solution to the differential equation f. Since $f \in \mathcal{F}_G^P(Q)$ implies that f_d vanishes, $x_d(t) = x_d(0)$ for all t. If we can choose f so that $f_c(x(0)) \neq 0$, then $x_c(t) \neq x_c(0)$ for small t, so $\Delta_\bowtie$ is not flow-invariant, and we are finished.

It remains to choose such an f. By Lemma 4.5 we need only find a $B(c,c)$-invariant mapping $g_c : P_{I(c)} \to P_c$ such that $g_c(x(0)) \neq 0$, since such an invariant mapping extends to a vector field in $\mathcal{F}_G^P(Q)$. For example, we may take $g_c(x) = x_c(0) \neq 0$, which is $B(c,c)$-invariant. ∎

Balanced equivalence relations. In order to motivate our characterization of robustly synchronous equivalence relations $\bowtie$, we repeat the analysis of Example 5.1 with a slightly different equivalence relation. Suppose that we partition the ten cells into the classes

$$\{0,1\}, \{2,3,8\}, \{4,5,6,7,9\}$$

so that the color of cell 6 is now gray instead of white, all other colors remaining as in Figure 11. Now the associated polydiagonal is

$$Y' = \{(u,u,v,v,w,w,w,w,v,w)\}.$$

The general vector field (5.2) remains unchanged, but its restriction (5.3) to Y changes in just one component: now

$$g_1 = A(u, \overline{w,w,v,w}).$$

This is no longer the same as g_0, so no reduction to the three-cell network is possible.

What is the source of this difference? The symmetry property of A (that is, its $B(1,1)$-invariance) implies that the order of the v's and w's does not matter, but there are three occurrences of w in g_1 and only two occurrences in g_0. Similarly there is one occurrence of v in g_1, but there are two occurrences in g_0. This difference in "multiplicity" makes g_1 differ from g_0 and so destroys the possibility of Y' being an invariant subspace.

This and similar examples lead to the following concept.

Definition 6.4. *An equivalence relation $\bowtie$ on $\mathcal{C}$ is* balanced *if for all $c, d \in \mathcal{C}$ with $c \bowtie d$ and $c \neq d$, there exists $\gamma \in B(c,d)$ such that $i \bowtie \gamma(i)$ for all $i \in I(c)$.*

In particular, $B(c,d) \neq \emptyset$ implies $c \sim_I d$. Therefore, balanced equivalence relations refine $\sim_I$.

The equivalence relation for Example 5.1 is balanced; the modified equivalence relation is not balanced. It turns out that this is the crucial distinction when it comes to constructing a quotient network; see Theorem 6.5 below.

There is a relatively simple graphical way to test whether a given equivalence relation $\bowtie$ is balanced. Color the cells in a network so that two cells have the same color precisely when they are in the same $\bowtie$-equivalence class. Then $\bowtie$ is balanced if and only if every pair of $\bowtie$-equivalent cells is connected by a color preserving groupoid element.

For example, consider the seven-cell network in Figure 14. Let $\bowtie$ be the equivalence relation with equivalence classes

$$\{1,4,7\}, \quad \{2,5\}, \quad \{3,6\},$$

as indicated by the colors in Figure 14. Observe that the pink (light gray) cells have input sets "white to pink," the white cells have input sets "blue (dark gray) to white," and the blue cells have input sets "pink to blue." So $\bowtie$ is a balanced equivalence relation, since all cells in the same equivalence class have identically colored input sets.

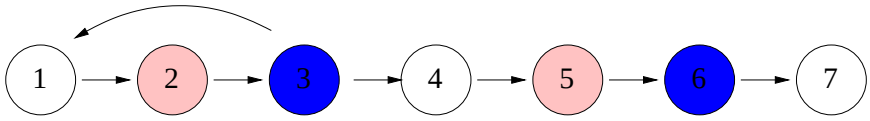


Figure 14. *Seven-cell linear network with $\bowtie$-equivalence classes indicated by color.*

The main theorem on polysynchrony. An examination of these examples leads to the following general result.

Theorem 6.5. *Let $\bowtie$ be an equivalence relation on a coupled cell network. Then $\bowtie$ is robustly polysynchronous if and only if $\bowtie$ is balanced.*

Proof. If $\bowtie$ is balanced, then Definition 6.4 implies that $\Delta_{\bowtie}$ is invariant under any admissible vector field; that is, $\bowtie$ is robustly polysynchronous. This is obvious in the "color" interpretation: as we have seen, $\bowtie$ is balanced if and only if every pair of cells of the same color are related by a color preserving input isomorphism. This implies that if $f \in \mathcal{F}_G^P$ and $c \bowtie d$, then $f_c(x) = f_d(x)$ for all $x \in \Delta_{\bowtie}$. That is, $\Delta_{\bowtie}$ is flow-invariant for f.

To prove the converse, suppose that $\bowtie$ is robustly polysynchronous. Then every $f \in \mathcal{F}_G^P$ maps $\Delta_{\bowtie}$ to $\Delta_{\bowtie}$. We wish to prove that $\bowtie$ is balanced; that is, if $c \bowtie d$ and $c \neq d$, then there exists $\gamma \in B(c,d)$ such that $i \bowtie \gamma(i)$ for all $i \in I(c)$. Since $\bowtie$ refines $\sim_I$ (Lemma 6.3), the set $B(c,d)$ is nonempty.

Define $K_0, \ldots, K_r$ as in (3.4) so that there is a partition

$$I(c) = K_0 \,\dot\cup \cdots \dot\cup\, K_r,$$

where i, i' belong to the same K_s if and only if $(i,c) \sim_E (i',c)$. Similarly, there is a partition

$$I(d) = L_0 \,\dot\cup \cdots \dot\cup\, L_r$$

with the corresponding property. We may choose the numbering so that

$$i \in K_s, j \in L_s \Longrightarrow (i,c) \sim_E (j,d), \qquad 1 \leq s \leq r. \tag{6.1}$$

As before, we may take $K_0 = \{c\}, L_0 = \{d\}$. (Because $c \sim_I d$, the sets K_s and L_s have the same cardinality for $0 \leq s \leq r$, and the same r occurs for $I(c)$ and $I(d)$.)

Suppose that we can prove that for any s with $0 \le s \le r$ and any $\bowtie$-equivalence class $U \subseteq \mathcal{C}$,

$$|U \cap K_s| = |U \cap L_s|. \tag{6.2}$$

Then we can define a bijection $\gamma : I(c) \to I(d)$ such that

$$\gamma(U \cap K_s) = U \cap L_s \qquad (0 \le s \le r) \tag{6.3}$$

for all U. By (6.1), $\gamma \in B(c,d)$. Moreover, (6.3) implies that $\gamma(i) \bowtie i$ for all $i \in I(c)$. For we may take U such that $i \in U$, and then $\gamma(i) \in U$ as well.

Thus it remains to prove the cardinality condition (6.2). To do so, we introduce a $\mathcal{B}_G$-equivariant map h, which depends on s, and apply it to an element $y \in \Delta_\bowtie$ that depends on U, as follows.

Let $M : P_i \to P_c$ be a nonzero linear map, where $i \in K_s$. Let $h_c : P_{I(c)} \to P_c$ be defined by

$$h_c(x) = M\left(\sum_{i \in K_s} x_i\right), \tag{6.4}$$

which is $B(c,c)$-invariant, since K_s is a $B(c,c)$-orbit. By Lemma 4.5 we may then define, for all other $c' \in \mathcal{C}$,

$$h_{c'}(x) = h_c(\beta^*(x)) = \sum_{j \in \beta(K_s)} M(x_j),$$

where β is some (hence any) element of $B(c,c')$ and $c' \sim_I c$, and

$$h_{c'}(x) = 0$$

otherwise. Moreover, the resulting h is $\mathcal{B}_G$-equivariant. Since $\bowtie$ is polysynchronous, h maps $\Delta_\bowtie$ to itself.

Next, define $y \in P$ by

$$y_j = \begin{cases} v, & j \in U, \\ 0, & j \notin U, \end{cases}$$

for some fixed $v \in P_a$ for which $M(v) \ne 0$, where $a \in U \cap K_s$.

We are assuming that $c \sim_I d$ so that $B(c,d) \ne \emptyset$. Let $\beta \in B(c,d)$, which implies that $\beta(K_s) = L_s$. Clearly

$$h_d(x) = \sum_{j \in L_s} M(x_j) \qquad \forall x \in \Delta_\bowtie. \tag{6.5}$$

Since h preserves $\Delta_\bowtie$,

$$h_c(x) = h_d(x) \qquad \forall x \in \Delta_\bowtie. \tag{6.6}$$

Apply (6.5),(6.6) to y:

$$h_c(y) = |U \cap K_s| M(v),$$
$$h_d(y) = |U \cap L_s| M(v).$$

By (6.6), since $M(v) \neq 0$, we deduce that

$$|U \cap K_s| = |U \cap L_s|$$

for all U and all $0 \leq s \leq r$. However, this is (6.2), so $\bowtie$ is balanced. ■

7. Dynamics on polysynchronous subspaces. As illustrated in Example 5.1 the restriction of a coupled cell vector field to a polysynchronous subspace has itself a special structure. The restriction is an admissible vector field for an associated "quotient" coupled cell network. In this section we construct the quotient network for a given polysynchronous subspace and illustrate some of the implications for the dynamics of the restriction. We begin with an example.

Example 7.1. Consider the five-cell network illustrated in Figure 15 (left). All cells are cell-equivalent, so a phase space for this network has the form $P = (\mathbf{R}^k)^5$ for some k. Since all cells are also input-equivalent (that is, have isomorphic input sets) the diagonal (x, x, x, x, x) is polysynchronous. Recall that a network is *homogeneous* when all of its cells are input isomorphic.

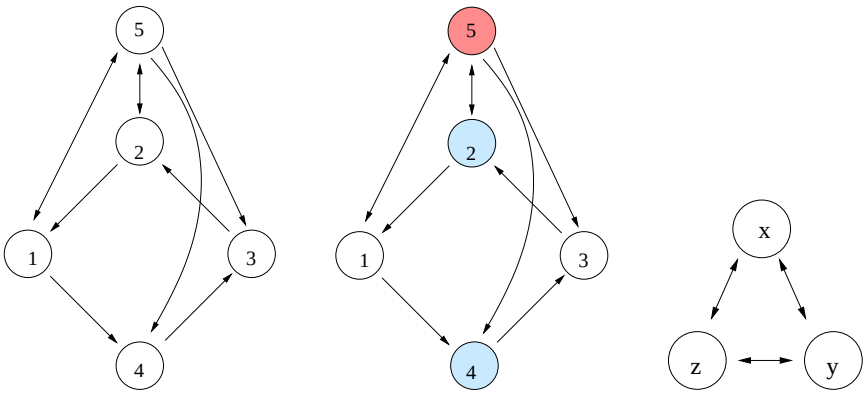


Figure 15. *A homogeneous five-cell network with a balanced relation leading to a quotient three-cell bidirectional ring.*

There is, however, a more interesting $3k$-dimensional polysynchronous subspace $\Delta_{\bowtie}$ associated to the balanced relation illustrated in Figure 15 (center). That subspace is

$$\Delta_{\bowtie} = \{(x, y, x, y, z) : x, y, z \in \mathbf{R}^k\}.$$

Next we discuss the structure of the restriction of an admissible coupled cell vector field

to $\Delta_{\bowtie}$. The general admissible vector field has the form

$$\begin{array}{rcl} \dot{x}_1 &=& f(x_1, \overline{x_2, x_5}), \\ \dot{x}_2 &=& f(x_2, \overline{x_3, x_5}), \\ \dot{x}_3 &=& f(x_3, \overline{x_4, x_5}), \\ \dot{x}_4 &=& f(x_4, \overline{x_1, x_5}), \\ \dot{x}_5 &=& f(x_5, \overline{x_1, x_2}), \end{array} \tag{7.1}$$

where $f : (\mathbf{R}^k)^3 \to \mathbf{R}^k$ is symmetric in the last two arguments. The restriction of (7.1) to $\Delta_{\bowtie}$ has the form

$$\begin{array}{rcl} \dot{x} &=& f(x, \overline{y, z}), \\ \dot{y} &=& f(y, \overline{z, x}), \\ \dot{z} &=& f(z, \overline{x, y}). \end{array} \tag{7.2}$$

Observe that (7.2) is the general vector field associated to the three-cell bidirectional ring illustrated in Figure 15 (right). We will show that there is a general construction that leads to this three-cell quotient, but first we discuss some implications for the dynamics of the five-cell system.

Observe also that the restriction (7.2) has $\mathbf{D}_3$ symmetry and is, in fact, the general $\mathbf{D}_3$-equivariant vector field on $(\mathbf{R}^k)^3$. So it is possible for a quotient network to have symmetry even when the original network has none. It is known that when $k \geq 2$, such vector fields can support discrete rotating waves and solutions where two cells are out of phase, while the third cell has twice the frequency of the other two [9, 7]. These solutions are also solutions to the original five-cell system. Typical simulations are shown in Figure 16. The middle and right simulations are obtained just by changing initial conditions.

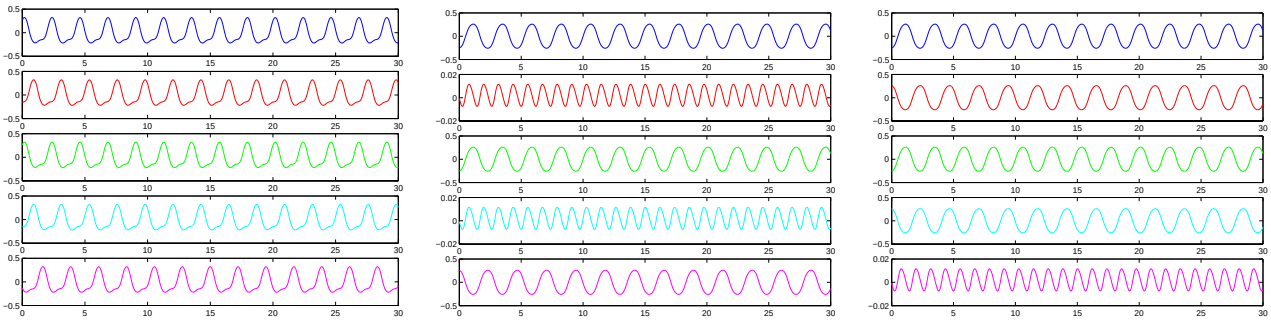


Figure 16. *Simulations in homogeneous five-cell network in Figure 15. (Left) rotating wave; (middle) double frequency in cells 2 and 4; (right) double frequency in cell 5.*

It is also possible for the restricted system to exhibit symmetric chaos, as illustrated in the five-cell simulations in Figures 17 and 18.

Perhaps the simplest example of a network that has no symmetry but does have a quotient network with symmetry is the three-cell network in Figure 19. This is the same as the "master-slave" network of Pecora and Carroll [12]. This network is part of the same family of networks as the seven-cell network described in section 1 (Figure 6) and again in section 6 (Figure 14).

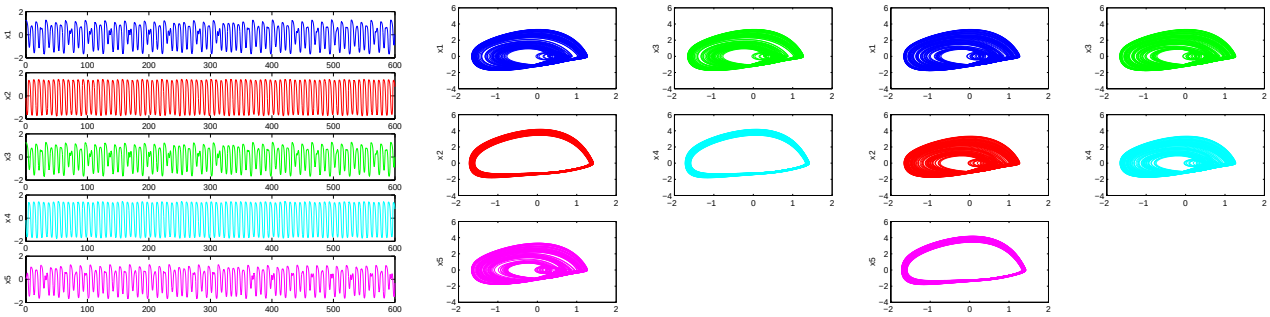


Figure 17. *Simulations in homogeneous five-cell network in Figure 15. (Left) time series for chaotic attractor with* $\mathbf{Z}_2$ *symmetry; (middle) phase plane with cells* $1, 3, 5$ *and cells* $2, 4$ *exhibiting symmetry on average; (right) double frequency in cell* 5*.*

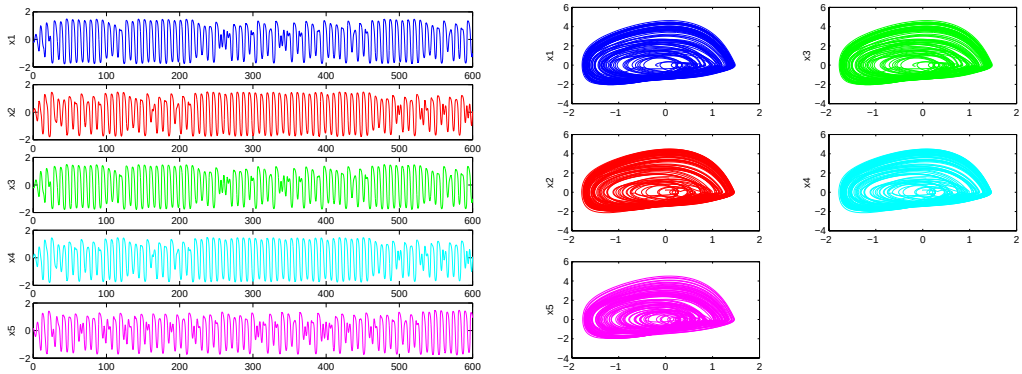


Figure 18. *Simulations in homogeneous five-cell network in Figure 15. (Left) time series for chaotic attractor with* $\mathbf{Z}_3$ *symmetry; (right) phase planes with all cells exhibiting symmetry on average.*

Construction of the natural quotient network. Let $\bowtie$ be a balanced equivalence relation on a coupled cell network $G = (\mathcal{C}, \mathcal{E}, \sim_C, \sim_E)$. In a series of steps we construct the quotient network $G_{\bowtie}$ corresponding to the polysynchronous subspace $\Delta_{\bowtie}$. To do this we need to define the cells and edges of the quotient network and the equivalence relations on them; that is, we must define $\mathcal{C}_{\bowtie}, \sim_{C_{\bowtie}}, \mathcal{E}_{\bowtie}, \sim_{E_{\bowtie}}$. Most steps are straightforward, but those related to edge-equivalence are more complicated.

(A) Let $\overline{c}$ denote the $\bowtie$-equivalence class of $c \in \mathcal{C}$. The cells in $\mathcal{C}_{\bowtie}$ are the $\bowtie$-equivalence classes in $\mathcal{C}$; that is,

$$\mathcal{C}_{\bowtie} = \{\overline{c} : c \in \mathcal{C}\}.$$

Thus we obtain $\mathcal{C}_{\bowtie}$ by forming the *quotient* of $\mathcal{C}$ by $\bowtie$; that is, $\mathcal{C}_{\bowtie} = \mathcal{C}/\bowtie$.

(B) Define

$$\overline{c} \sim_{C_{\bowtie}} \overline{d} \iff c \sim_C d.$$

This is well defined since $\bowtie$ refines $\sim_C$.

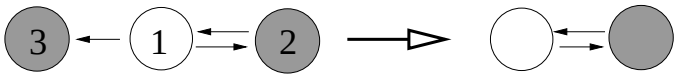


Figure 19. *A three-cell network with no symmetry having a quotient two-cell network with* $\mathbf{Z}_2$ *symmetry.*

(C) The edges in the quotient network are the projection of edges in the original network that do not link distinct but $\bowtie$-equivalent cells. That is,

$$\mathcal{E}_{\bowtie} = \{(\overline{i}, \overline{c}) : (i, c) \in \mathcal{E}, i \not\bowtie c\} \cup \{(\overline{c}, \overline{c}) : c \in \mathcal{C}\}.$$

(D) We now define $\sim_{E_{\bowtie}}$. Suppose that $(j, d) \in \mathcal{E}_{\bowtie}$, and let $c \in \mathcal{C}$ satisfy $\overline{c} = d$. Define

$$\Omega_c(j) = \{i \in I(c) : \overline{i} = j\}. \tag{7.3}$$

Now let $(j_1, d_1), (j_2, d_2) \in \mathcal{E}_{\bowtie}$. We say that

$$(j_1, d_1) \sim_{E_{\bowtie}} (j_2, d_2)$$

if and only if for some $c_1, c_2 \in \mathcal{C}$ with $\overline{c_1} = d_1, \overline{c_2} = d_2$ there exists $\gamma \in B(c_1, c_2)$ such that

$$\gamma(\Omega_{c_1}(j_1)) = \Omega_{c_2}(j_2). \tag{7.4}$$

Remark 7.2. We interrupt our discussion to provide a word picture of the construction of edge-equivalence $\sim_{E_{\bowtie}}$. Suppose that the cells of $\mathcal{C}$ are colored by $\bowtie$-equivalence classes, as discussed previously. So every cell in $\mathcal{C}_{\bowtie}$ can be identified with a unique color. The set $\Omega_c(j)$ consists of those cells in the input set $I(c)$ having color j. The edges (j_1, d_1) and (j_2, d_2) are $\sim_{E_{\bowtie}}$-equivalent if there is an input equivalence of $I(c_1)$ to $I(c_2)$ that maps cells of color j_1 to cells of color j_2. In particular, the number of cells in $I(c_1)$ of color j_1 must equal the number of cells in $I(c_2)$ of color j_2.

This completes the construction of $G_{\bowtie} = (\mathcal{C}_{\bowtie}, \mathcal{E}_{\bowtie}, \sim_{C_{\bowtie}}, \sim_{E_{\bowtie}})$, except for one final technical remark. As stated, the definition of $\sim_{E_{\bowtie}}$ appears to depend on the choice of c_1, c_2 in (D). In fact, it does not.

Lemma 7.3. *Suppose that* $\bowtie$ *is balanced. Let* $c_1, c_2, c_1', c_2' \in \mathcal{C}$*, where* $c_1 \bowtie c_1'$ *and* $c_2 \bowtie c_2'$*. Let* $j_1, j_2 \in \mathcal{C}_{\bowtie}$*. Suppose that there exists* $\beta \in B(c_1, c_2)$ *such that*

$$\beta(\Omega_{c_1}(j_1)) = \Omega_{c_2}(j_2).$$

Then there exists $\beta' \in B(c_1', c_2')$ *such that*

$$\beta'(\Omega_{c_1'}(j_1)) = \Omega_{c_2'}(j_2).$$

Proof. Since $\bowtie$ is balanced, there exists (for $k = 1, 2$) an element $\gamma_k \in B(c_k, c_k')$ such that $\gamma_k(i) \bowtie i$ for all $i \in I(c_k)$. Therefore,

$$\gamma_k(\Omega_{c_k}(j_k)) = \Omega_{c_k'}(j_k).$$

Clearly $\beta' = \gamma_2 \beta \gamma_1^{-1}$ is an input isomorphism and by construction $\beta'(\Omega_{c_1'}(j_1)) = \Omega_{c_2'}(j_2)$. ■

Lemma 7.3 implies that if (D) holds for some choice of c_1, c_2 satisfying the required conditions, then it holds for any choice of c_1, c_2.

Finally, we show that $G_{\bowtie}$ is a coupled cell network. To do so, we must verify the compatibility conditions in Definition 2.1(e,f).

(E) If $(j_1, d_1) \sim_{E_{\bowtie}} (j_2, d_2)$, then $j_1 \sim_{C_{\bowtie}} j_2$ and $d_1 \sim_{C_{\bowtie}} d_2$.

Choose $c_1, c_2 \in \mathcal{C}$ such that $\overline{c_1} = d_1$ and $\overline{c_2} = d_2$. The definition of $\sim_{E_{\bowtie}}$ implies there exists $\gamma \in B(c_1, c_2)$ such that $\gamma(\Omega_{c_1}(j_1)) = \Omega_{c_2}(j_2)$. Since γ is an input isomorphism, it preserves cell type, so $c_1 \sim_C c_2$. However, now the definition of $\sim_{C_{\bowtie}}$ shows that $d_1 \sim_{C_{\bowtie}} d_2$. Next choose any $i \in \Omega_{c_1}(j_1)$. Then $\gamma(i) \in \Omega_{c_2}(j_2)$, and $i \sim_C \gamma(i)$. Therefore, $j_1 \sim_{C_{\bowtie}} j_2$.

(F) Internal edges are never equivalent to noninternal ones; that is,

$$(j_1, j_1) \sim_{E_{\bowtie}} (j_2, d_2) \iff j_2 = d_2 \text{ and } j_2 \sim_{C_{\bowtie}} j_1$$

for all $j_1, j_2, d_2 \in \mathcal{C}_{\bowtie}$.

We prove $\Longrightarrow$. Assume that $(j_1, j_1) \sim_{E_{\bowtie}} (j_2, d_2)$ and choose $c_1, c_2 \in \mathcal{C}$ such that $\overline{c_1} = j_1$ and $\overline{c_2} = d_2$. The definition of $\sim_{E_{\bowtie}}$ implies that there exists $\gamma \in B(c_1, c_2)$ satisfying (7.4): $\gamma\Omega_{c_1}(j_1) = \gamma\Omega_{c_2}(j_2)$. Suppose that $\Omega_{c_1}(j_1) = \{c_1, i_1, \dots, i_p\}$. Then $\Omega_{c_2}(j_2) = \{\gamma(c_1), \gamma(i_1), \dots, \gamma(i_p)\} = \{c_2, \gamma(i_1), \dots, \gamma(i_p)\}$ and $d_2 = \overline{c_2} = \overline{\gamma(i_1)} = \cdots = \overline{\gamma(i_p)} = j_2$ by definition of $\Omega_{c_2}(j_2)$. Thus $j_2 = d_2$. As $j_2 = \overline{c}_2, j_1 = \overline{c}_1$, and $c_1 \sim_c c_2$ (since $B(c_1, c_2) \neq \emptyset$), then $j_1 \sim_{C_{\bowtie}} j_2$.

The converse is obtained by direct calculation.

It remains to prove that the restriction of each G-admissible vector field to $\Delta_{\bowtie}$ is a $G_{\bowtie}$-admissible vector field. This result follows from Theorem 9.2, whose proof uses "quotient maps," which are introduced in section 8.

Remark 7.4 (on the symmetry groupoid of the natural quotient). It is reasonable to ask for a characterization of the symmetry groupoid of the natural quotient $G/\bowtie$ in terms of the symmetry groupoid of G and its relation to $\bowtie$.

Define

$$\Sigma^{\bowtie}(c,d) = \{\sigma \in B(c,d) : \sigma(i) \bowtie i \qquad \forall i \in I(c)\},$$
$$\mathrm{T}^{\bowtie}(c,d) = \{\tau \in B(c,d) : i \bowtie j \iff \tau(i) \bowtie \tau(j) \qquad \forall i, j \in I(c)\}.$$

Then define two subgroupoids of $\mathcal{B}_G$ by

$$\Sigma^{\bowtie} = \dot{\bigcup}_{c,d \in \mathcal{C}} \Sigma^{\bowtie}(c,d),$$
$$\mathrm{T}^{\bowtie} = \dot{\bigcup}_{c,d \in \mathcal{C}} \mathrm{T}^{\bowtie}(c,d).$$

It turns out that $\mathrm{T}^{\bowtie}$ consists precisely of the $\bowtie$-compatible elements of $\mathcal{B}_G$, a concept that we introduce later in Definition 8.6. It follows that $\mathcal{B}_{G/\bowtie}$ consists precisely of the bijections induced on $\mathcal{C}/\bowtie$ by the subgroupoid $\mathrm{T}^{\bowtie}$ of $\mathcal{B}_G$.

Moreover, the elements of $\Sigma^{\bowtie}$ act as the identity on $\mathcal{C}/\bowtie$. In fact, they form the isotropy subgroupoid of any generic element of the polydiagonal $\Delta_{\bowtie}$ (that is, an element $x \in \Delta_{\bowtie}$ such that $x_i = x_j \Leftrightarrow i \bowtie j$). By analogy with the group-symmetric case, we expect $\mathcal{B}_{G/\bowtie}$ to be

equal to the quotient groupoid $\mathrm{T}^{\bowtie}/\Sigma^{\bowtie}$. Moreover, $\mathrm{T}^{\bowtie}$ ought to be the "normalizer groupoid" of $\Sigma^{\bowtie}$ in $\mathcal{B}_G$.

Dias and Stewart [5] prove the above statements. We omit the proofs here because they involve technicalities about quotient groupoids that would take us too far afield.

Remark 7.5 (on the lifting of $G_{\bowtie}$ admissible vector fields). In symmetric dynamics the issue of "hidden symmetry" arises. Here, the restriction of an equivariant vector field onto the fixed-point space of a subgroup Σ is always equivariant under the normalizer of Σ, but sometimes it obeys extra constraints. See [9, 7]. The next example shows that the same issue arises in the groupoid context. In particular, vector fields that are admissible with respect to the quotient network $G_{\bowtie}$ do not always lift to vector fields that are admissible with respect to the original coupled cell network G.

Consider the four-cell network in Figure 20 (left). The equivalence relation $\bowtie$ indicated by color is balanced, and consider the natural quotient three-cell network in Figure 20 (right).

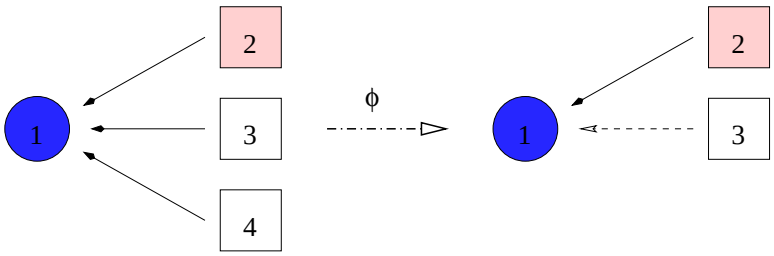


Figure 20. *A four-cell example.*

Admissible vector fields of the four-cell network have the form

$$\begin{aligned} \dot{x}_1 &= A(x_1, \overline{x_2, x_3, x_4}), \\ \dot{x}_2 &= B(x_2), \\ \dot{x}_3 &= B(x_3), \\ \dot{x}_4 &= B(x_4). \end{aligned} \tag{7.5}$$

Admissible vector fields of the three-cell network have the form

$$\begin{aligned} \dot{u}_1 &= f(u_1, u_2, u_3), \\ \dot{u}_2 &= g(u_2), \\ \dot{u}_3 &= g(u_3). \end{aligned} \tag{7.6}$$

If we identify (x_1, x_2, x_3, x_4) with (u_1, u_2, u_3, u_3), we induce a vector field from (7.5), and we thereby obtain one of the form

$$\begin{aligned} \dot{u}_1 &= A(u_1, \overline{u_2, u_3, u_3}), \\ \dot{u}_2 &= B(u_2), \\ \dot{u}_3 &= B(u_3), \end{aligned} \tag{7.7}$$

which is admissible by (7.6).

Note that not every $G_{\bowtie}$-admissible vector field (7.6) can be extended to a G-admissible vector field. Compare the linear terms in A (namely, $\alpha u_1 + \beta(u_2 + 2u_3)$) with the linear terms

in f in (7.6) (namely, $\alpha u_1 + \beta u_2 + \gamma u_3$). This is a groupoid analogue of hidden symmetry and raises similar issues. Dias and Stewart [5] give a complete groupoid-theoretic characterization of the cases when every $G_{\bowtie}$-admissible vector field extends to a G-admissible vector field.

8. Quotient maps. In this section we give a formal definition of a quotient map $\phi : G_1 \to G_2$, where G_1 and G_2 are coupled cell networks. The definition is purely graph-theoretic.

Definition 8.1. *Let $G_i = (\mathcal{C}_i, \mathcal{E}_i, \sim_{C_i}, \sim_{E_i})$ be coupled cell networks. The map*

$$\phi : \mathcal{C}_1 \to \mathcal{C}_2$$

is a quotient map *if the following hold:*

(a) Cells lift: *ϕ is surjective.*

(b) Input arrows lift: *If $(i,c) \in \mathcal{E}_1$, then $(\phi(i), \phi(c)) \in \mathcal{E}_2$. Conversely, if $(j,d) \in \mathcal{E}_2$ and $c \in \mathcal{C}_1$ such that $\phi(c) = d$, then there exists $i \in \mathcal{C}_1$ such that $\phi(i) = j$ and $(i,c) \in \mathcal{E}_1$.*

(c) Input isomorphisms lift: *Let $d, d' \in \mathcal{C}_2$ and $\beta_2 \in B(d,d')$. Choose $c, c' \in \mathcal{C}_1$ such that $\phi(c) = d$ and $\phi(c') = d'$. Then there exists $\beta_1 \in B(c,c')$ such that*

$$\beta_2(\phi(i)) = \phi(\beta_1(i)) \tag{8.1}$$

for all $i \in I(c)$.

There are several observations that follow directly from the definition of a quotient map $\phi : \mathcal{C}_1 \to \mathcal{C}_2$. Define the equivalence relation $\bowtie_\phi$ on $\mathcal{C}_1$ by

$$c \bowtie_\phi c' \quad \Longleftrightarrow \quad \phi(c) = \phi(c'). \tag{8.2}$$

That is, any two cells in $\mathcal{C}_1$ that project by ϕ onto the same cell in $\mathcal{C}_2$ have the same color (that is, are $\bowtie_\phi$-equivalent).

Lemma 8.2. *Let $\phi : \mathcal{C}_1 \to \mathcal{C}_2$ be a quotient map. Then the following hold:*

(a) *If $\phi(c) = d$, then*

$$\phi(I(c)) = I(d). \tag{8.3}$$

(b) *For every $c, c' \in \mathcal{C}_1$ such that $\phi(c) = \phi(c')$ there is an input isomorphism $\beta \in B(c,c')$ such that*

$$\phi(i) = \phi(\beta(i)) \tag{8.4}$$

for all $i \in I(c)$.

(c) *The equivalence relation $\bowtie_\phi$ is balanced.*

Proof. Part (a) follows directly from Definition 8.1 (b). Part (b) follows from Definition 8.1 (c) by setting $d = d'$ and $\beta_2 = \mathrm{id}$ on $I(d)$. The existence of $\beta \in B(c,c')$ in Part (b) implies that $c \sim_{I_1} c'$ and hence $c \sim_{C_1} c'$. Using (8.2), identity (8.4) is equivalent to $i \bowtie_\phi \beta(i)$ for all $i \in I(c)$, which is the definition of "balanced" in Definition 6.4. Thus Part (c) holds. ∎

Quotient networks are examples of quotient maps.

Theorem 8.3. *Assume that $\bowtie$ is a balanced equivalence relation on $\mathcal{C}$, and let $\mathcal{C}_{\bowtie}$ be the natural coupled cell network whose cells are the equivalence classes of $\bowtie$. Let $\overline{c}$ denote the $\bowtie$-equivalence class of the cell $c \in \mathcal{C}$. Then the map $\phi : \mathcal{C} \to \mathcal{C}_{\bowtie}$ defined by $c \mapsto \overline{c}$ is a quotient map.*

Proof. We verify that $\phi : \mathcal{C} \to \mathcal{C}_{\bowtie}$ is a quotient map; that is, we verify Definition 8.1 (a)–(c).

(a) Cells lift since ϕ is onto by construction.

(b) Input arrows lift by definition. See part (C) in the construction of the network $\mathcal{C}_{\bowtie}$.

(c) We show that input isomorphisms lift. Recall that Definition 8.1 (c) states the following: For every $d, d' \in \mathcal{C}_{\bowtie}$, $c, c' \in \mathcal{C}$ such that $\overline{c} = d, \overline{c'} = d'$, and $\beta_2 \in B(d, d')$, there exists $\beta_1 \in B(c, c')$ such that $\beta_2(\overline{i}) = \overline{\beta_1(i)}$ for all $i \in I(c)$. So we must construct the input isomorphism $\beta_1 : I(c) \to I(c')$.

We first show that input sets lift; that is, $\overline{I(c)} = I(\overline{c})$ for each $c \in \mathcal{C}$. Suppose that $j \in I(\overline{c})$. Since input arrows lift, there exist $i', c' \in \mathcal{C}$ such that $\overline{i'} = j$, $\overline{c'} = \overline{c}$, and $(i', c') \in \mathcal{E}$. By construction of $\mathcal{C}_{\bowtie}$, c' and c are $\bowtie$-equivalent. Since $\bowtie$ is balanced, there exists $\gamma \in B(c', c)$ for which $i = \gamma(i') \bowtie i'$. It follows that $\gamma(i', c') = (i, c) \in \mathcal{E}$ and that $\overline{i} = j$. Thus each input arrow in $I(\overline{c})$ lifts to an input arrow in $I(c)$.

Next we show that β_1 exists. The set $\Omega_c(j)$, defined in (7.3), consists of cells in $I(c)$ that are $\bowtie$-equivalent and project onto the node $j \in \mathcal{C}_{\bowtie}$. Therefore, we can choose a finite set J of j such that

$$I(c) = \dot{\bigcup}_{j \in J} \Omega_c(j).$$

Since input sets lift, the existence of β_2 implies that

$$I(c') = \dot{\bigcup}_{j \in J} \Omega_{c'}(\beta_2(j)).$$

We construct the permutation β_1 by finding bijections

$$\beta_1|_{\Omega_c(j)} : \Omega_c(j) \to \Omega_{c'}(\beta_2(j))$$

for all $j \in J$ and letting β_1 be their union. The existence of $\beta_2 \in B(d, d')$ implies that $(j, d) \sim_{E_{\bowtie}} (\beta_2(j), d')$. Recall from (7.4) that the definition of $\sim_{E_{\bowtie}}$ implies that there exists $\gamma \in B(c, c')$ such that $\gamma(\Omega_c(j)) = \Omega_{c'}(\beta_2(j))$. Thus $\beta_2(\overline{i}) = \overline{\gamma(i)}$ for all $i \in \Omega_c(j)$. Now set $\beta_1|_{\Omega_c(j)} = \gamma$. ∎

Nonuniqueness and universality.

Theorem 8.4. *Assume that $\bowtie$ is a balanced equivalence relation on $\mathcal{C}$, and let $G_{\bowtie}$ be the associated natural quotient network with quotient map ϕ. Then the pair $(G_{\bowtie}, \phi)$ is universal. That is, if G' is a coupled cell network with a quotient map ϕ' with $\bowtie_{\phi'} = \bowtie$, then there is a quotient map $\xi : G_{\bowtie} \to G'$ such that $\phi'(c) = \xi(\phi(c))$ for all $c \in \mathcal{C}$.*

In this situation we say that (G', ϕ') *factors through* $(G_{\bowtie}, \phi)$. Note that with the definition of ϕ given in Theorem 8.3, $c \bowtie d$ if and only if $c \bowtie_{\phi} d$.

First, we give an example to show that quotient networks need not be unique. Then we prove Theorem 8.4, which shows that the natural quotient is universal. That is, all other quotient networks are quotients of $G_{\bowtie}$ of a rather trivial kind: distinct cells remain distinct.

Example 8.5. Figure 21 shows three coupled cell networks. The network G is the seven-cell chain of Figure 6. The network $G_{\bowtie}$ is the three-cell ring of Figure 7 in which all three arrows are equivalent. The network G' is another three-cell ring, in which the arrows are *not* equivalent. It is easy to see that there exist three quotient maps $\phi : G \to G_{\bowtie}$, $\phi' : G \to G'$, and $\xi : G_{\bowtie} \to G'$, shown by the coloring of the figure. Moreover,

$$\phi'(c) = \xi(\phi(c)) \qquad \forall c \in \mathcal{C}.$$

Clearly ϕ and ϕ' induce the same equivalence relation on G; that is, $\bowtie_{\phi} = \bowtie_{\phi'}$. However, $G_{\bowtie}$ and G' are not isomorphic.

In fact, there are three other quotient networks with the same equivalence relation. Namely, form a three-cell ring and define two arrows to be equivalent but the third to be different. These three networks can be inserted between $G_{\bowtie}$ and G'.

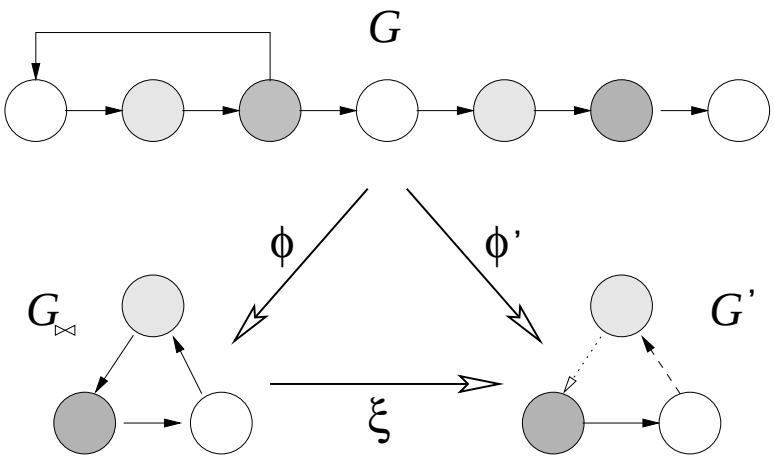


Figure 21. *Two distinct quotients with the same equivalence relation.*

The essential point now is that Example 8.5 exhibits the only way in which uniqueness fails. The natural quotient $G_{\bowtie}$ defined above is the one in which as many arrows as possible are edge-equivalent. All other quotients are obtained from the natural one by employing the same cells and refining $\sim_E$.

It will be helpful to introduce the following concept.

Definition 8.6. *Let $\gamma : J \to K$ be a bijection between subsets $J, K \subset \mathcal{C}$, and let $\bowtie$ be an equivalence relation on $\mathcal{C}$. Say that γ is* $\bowtie$-compatible *if for all $j_1, j_2 \in J$*

$$j_1 \bowtie j_2 \Leftrightarrow \gamma(j_1) \bowtie \gamma(j_2). \tag{8.5}$$

Essentially, the point here is that γ *permutes* $\bowtie$-equivalence classes. Note that in the definition of "balanced" we have the stronger condition $i \bowtie \gamma(i)$, in which γ *fixes* $\bowtie$-equivalence classes.

Such maps arise for the following reason. Suppose that $\phi : G \to G'$ is any quotient map of coupled cell networks, and let $\beta \in \mathcal{B}_{G'}$ be an input isomorphism. The definition of "quotient" requires there to exist a lift $\tilde{\beta} \in \mathcal{B}_G$. The definition of "lift" clearly implies that

$$\tilde{\beta} \text{ is } \bowtie_\phi \text{-compatible} \tag{8.6}$$

with $J = I(c), K = I(d)$, whence $\tilde{\beta} \in B(c,d)$ (if $\beta \in B(\phi(c), \phi(d)) \subseteq \mathcal{B}_{G'}$).

Proof of Theorem 8.4. Suppose that G is a coupled cell network and $\bowtie$ is a balanced equivalence relation on $\mathcal{C}$. Let $(G_\bowtie, \phi)$ be the natural quotient by $\bowtie$ so that $\bowtie_\phi = \bowtie$. Let (G', ϕ') be any quotient network with $\bowtie_{\phi'} = \bowtie$. We claim that ϕ' factors through ϕ.

Define $\xi : G_\bowtie \to G'$ as follows. Let $\overline{c} \in \mathcal{C}_\bowtie$ be a cell of $G_\bowtie$. Define

$$\xi(\overline{c}) = \phi'(c).$$

The map ξ is well defined because $\bowtie_{\phi'} = \bowtie = \bowtie_\phi$. It is a bijection $\xi : \mathcal{C}_\bowtie \to \mathcal{C}'$.

We claim that ξ is a quotient map. The defining properties are obvious, except for the condition that input isomorphisms lift from G' to $G_\bowtie$. Suppose that $\beta \in \mathcal{B}_{G'}$. Then β lifts from G' to G, yielding an input isomorphism $\tilde{\beta} \in \mathcal{B}_G$. By (8.6), $\tilde{\beta}$ is $\bowtie$-compatible. Therefore, it induces a bijection γ on $\mathcal{C}_\bowtie$ defined by

$$\gamma(\overline{c}) = \overline{\tilde{\beta}(c)}.$$

The definition of edge-equivalence in the construction of $G_\bowtie$ implies that γ is an input isomorphism in $G_\bowtie$. Therefore, every $\beta \in \mathcal{B}_{G'}$ lifts to some $\gamma \in \mathcal{B}_{G_\bowtie}$. ■

Several other properties follow directly from this proof. To state them, we need the following definition:

Definition 8.7. *Let $\xi : G \to G'$ be bijective on cells. Then G' is an* edge-refinement *of G if*

$$\xi(i,c) \sim_{E'} \xi(j,d) \quad \Rightarrow \quad (i,c) \sim_E (j,d).$$

Corollary 8.8.

(a) *Every quotient network corresponding to a given balanced equivalence relation $\bowtie$ is an edge-refinement of the natural quotient $G_\bowtie$.*

(b) *Conversely, every edge-refinement of $G_\bowtie$ is a quotient network corresponding to $\bowtie$.*

(c) *Let G', G'' be edge-refinements of $G_\bowtie$. Then G'' is an edge-refinement of G' if and only if $\mathcal{B}_{G'} \supseteq \mathcal{B}_{G''}$.*

(d) *The condition $\mathcal{B}_{G'} \supseteq \mathcal{B}_{G''}$ is equivalent to $\mathcal{F}^P_{G'} \subseteq \mathcal{F}^P_{G''}$ for any choice of phase space P on cells, where cells in G', G'' are identified if they correspond to the same $\bowtie$-class of cells in G.*

The above corollary establishes that the phenomena described in Example 8.5 are typical of the general case.

9. Induced vector fields are admissible. Now we come to the second main theorem of this paper. We show that any quotient map $\phi : G_1 \to G_2$ converts G_1-admissible vector fields into G_2-admissible vector fields in a natural way.

The basic idea is the following. Let Δ_ϕ denote the polydiagonal subspace corresponding to the equivalence relation $\bowtie_\phi$ (previously denoted $\Delta_{\bowtie_\phi}$). We claim that the space of

G_1-admissible vector fields restricted to Δ_ϕ can be naturally identified with a subspace of the space of G_2-admissible vector fields. The main consequence of this observation is that interesting dynamics (rotating waves, symmetric chaos) in this subspace for the cell system G_2 corresponds to the same dynamics in the cell system G_1, in which $\bowtie$-equivalent cells are synchronous.

We first choose cell phase spaces P_c for $c \in \mathcal{C}_1$. Then $\phi(c) \in \mathcal{C}_2$, and we let the corresponding cell phase space be $\overline{P}_{\phi(c)} = P_c$. The space $\overline{P}_{\phi(c)}$ is well defined since quotient maps preserve the relation $\sim_c$.

Choose a set of representatives $\mathcal{R}$ for the map ϕ. That is, $\mathcal{R} \subseteq \mathcal{C}_1$ and for each $d \in \mathcal{C}_2$ there exists a unique $c \in \mathcal{R}$ such that $\phi(c) = d$. Thus the set of all $\phi(c)$ runs through the elements of $\mathcal{C}_2$ without duplication when c runs through $\mathcal{R}$. Then define

$$\overline{P} = \prod_{c\in\mathcal{R}} \overline{P}_{\phi(c)} = \prod_{c\in\mathcal{R}} P_c.$$

If $x = (x_c)_{c\in\mathcal{C}_1}$ defines coordinates on P, we can consider $y = (y_{\phi(c)})_{\phi(c)\in\mathcal{C}_2}$ as defining coordinates on $\overline{P}$. Moreover, for each $c \in \mathcal{C}_1$ there exists a unique $r \in \mathcal{R}$ such that $\phi(c) = \phi(r)$, and then $y_{\phi(c)}$ is identified with $y_{\phi(r)}$.

In section 8 we introduced the notion of a quotient map between coupled cell networks. The key property that we wish to ensure is that a quotient map $\phi : G_1 \to G_2$ induces a natural mapping $\hat{\phi} : \mathcal{F}^P_{G_1} \to \mathcal{F}^{\overline{P}}_{G_2}$, where $\overline{P}$ is obtained by identifying the relevant factors of P.

Quotients preserve admissibility. We now establish an important property of quotient maps: they induce admissible vector fields.

Suppose that $\phi : G_1 \to G_2$ is a quotient map. There is an injective map $\alpha : \overline{P} \to P$ defined by

$$\alpha(y)_c = y_{\phi(c)} \qquad \forall c \in \mathcal{C}_1, y \in \overline{P}. \tag{9.1}$$

Note that $\Delta_\phi = \alpha(\overline{P})$, so $\alpha : \overline{P} \to \Delta_\phi$ is a bijection. Replacing y by $\alpha^{-1}x$, for $x \in \Delta_\phi$, (9.1) becomes

$$(\alpha^{-1}x)_{\phi(c)} = x_c \qquad \forall c \in \mathcal{C}_1. \tag{9.2}$$

Definition 9.1. *Since $f \in \mathcal{F}^P_{G_1}$ leaves Δ_ϕ invariant, we can define a vector field $\overline{f}$ on $\overline{P}$, the* induced vector field*, by restricting f to Δ_ϕ and projecting the result onto $\overline{P}$ by α^{-1}. More precisely,*

$$\overline{f}(y) = \alpha^{-1}(f(\alpha(y))) \qquad \forall y \in \overline{P}. \tag{9.3}$$

We will also denote $\overline{f}$ by $\hat{\phi}(f)$.

The main result of this section is Theorem 9.2 below. Several applications of this theorem can be found in [6].

Theorem 9.2. *For any $f \in \mathcal{F}^P_{G_1}$, the induced vector field $\overline{f}$ lies in $\mathcal{F}^{\overline{P}}_{G_2}$.*

Proof. Since $f \in \mathcal{F}^P_{G_1}$ is G_1-admissible, it satisfies the two conditions of Definition 4.1: the domain condition (4.3) and the equivariance condition (4.5).

The domain condition states the following: For all $c \in \mathcal{C}_1$ there exists a function $\hat{f}_c : P_{I(c)} \to P_c$ such that

$$f_c(x) = \hat{f}_c(x_{I(c)}). \tag{9.4}$$

Equivalently, f_c depends only on variables from cells in $I(c)$.

$\mathcal{B}_{G_1}$-equivariance states the following: For all $c, c' \in \mathcal{C}_1$ and for all $\gamma \in B(c, c')$ we have

$$\hat{f}_{c'}(x) = \hat{f}_c(\gamma^*(x)) \qquad \forall x \in P_{I(c')}, \tag{9.5}$$

where

$$(\gamma^*(x))_i = x_{\gamma(i)} \qquad \forall i \in I(c) \tag{9.6}$$

and is undefined elsewhere.

We must verify Definition 4.1. That is, we must show the following:

(a) The vector field $\overline{f}$ satisfies the domain condition for G_2.

(b) The vector field $\overline{f}$ satisfies the equivariance condition for G_2.

To prove (a), let $d \in \mathcal{C}_2$, and suppose that $\phi(c) = d$. We must show that $\overline{f}_d(y)$ depends only on $y_{I(d)}$ for $y \in \overline{P}$. We have

$$\begin{array}{rcll} \overline{f}_d(y) & = & \overline{f}_{\phi(c)}(y) & \\ & = & (\overline{f}(y))_{\phi(c)} & \\ & = & (\alpha^{-1}(f(\alpha(y))))_{\phi(c)} & \text{by (9.3)} \\ & = & (f(\alpha(y)))_c & \text{by (9.2)} \\ & = & f_c(\alpha(y)) & \text{by definition.} \end{array}$$

By (9.4), the value of $f_c(\alpha(y))$ depends only on $(\alpha(y))_{I(c)}$. However, by (9.1),

$$(\alpha(y))_{I(c)} = y_{\phi(I(c))} = y_{I(d)}$$

since $\phi(I(c)) = I(d)$ by (8.3). This proves (a).

We interrupt this proof to verify a lemma.

Lemma 9.3. *Let $d, d' \in \mathcal{C}_2$ and let $\beta \in B(d, d')$. Choose $c, c' \in \mathcal{C}_1$ such that $\phi(c) = d, \phi(c') = d'$. Suppose that β lifts to $\tilde{\beta} \in B(c, c')$. Then for all $y \in \overline{P}$ we have*

$$\tilde{\beta}^*(\alpha(y)) = \alpha(\beta^*(y)). \tag{9.7}$$

Proof. For all $i \in I(c)$ we have

$$\begin{array}{rcll} (\tilde{\beta}^*(\alpha(y)))_i & = & (\alpha(y))_{\tilde{\beta}(i)} & \text{by (9.6)} \\ & = & y_{\phi(\tilde{\beta}(i))} & \text{by (9.1)} \\ & = & y_{\beta(\phi(i))} & \text{by (8.1)} \\ & = & (\beta^*(y))_{\phi(i)} & \text{by (9.6)} \\ & = & (\alpha(\beta^*(y)))_i & \text{by (9.1),} \end{array}$$

which proves (9.7). ∎

Now we return to the proof of Theorem 9.2. To prove (b) we must show that for all $d, d' \in \mathcal{C}_2$ and $\beta \in B(d, d')$,

$$\overline{f}_{d'}(y) = \overline{f}_d(\beta^*(y)) \qquad \forall y \in P_{I(d')}, \tag{9.8}$$

where by definition

$$(\beta^*(y))_i = y_{\beta(i)}.$$

Choose $c, c' \in \mathcal{C}_1$ such that $\phi(c) = d, \phi(c') = d'$. Use Definition 8.1 (c) to lift $\beta : I(d) \to I(d')$ to $\tilde{\beta} : I(c) \to I(c')$. That is, by (8.1),

$$\beta(\phi(i)) = \phi(\tilde{\beta}(i)) \qquad \forall i \in I(c).$$

We know that

$$f_{c'}(x) = f_c(\tilde{\beta}^*(x)) \qquad \forall x \in P.$$

Therefore, setting $x = \alpha(y)$, we have

$$f_{c'}(\alpha(y)) = f_c(\tilde{\beta}^*(\alpha(y))) \qquad \forall y \in \overline{P}. \tag{9.9}$$

By Lemma 9.3,

$$f_{c'}(\alpha(y)) = f_c(\alpha(\beta^*(y))). \tag{9.10}$$

Now, by definition (9.3), $\overline{f}(y) = \alpha^{-1}(f(\alpha(y)))$. Therefore,

$$\begin{array}{rcll} \overline{f}_{d'}(y) & = & (\alpha^{-1}(f(\alpha(y))))_{d'} & \\ & = & (\alpha^{-1}(f(\alpha(y))))_{\phi(c')} & \\ & = & (f(\alpha(y)))_{c'} & \text{by (9.2)} \\ & = & f_{c'}(\alpha(y)). & \end{array}$$

Similarly,

$$\begin{array}{rcll} \overline{f}_d(\beta^*(y)) & = & (\overline{f}(\beta^*(y)))_d & \\ & = & (\alpha^{-1}(f(\alpha(\beta^*(y)))))_d & \\ & = & (f(\alpha(\beta^*(y))))_c & \text{by (9.2)} \\ & = & f_c(\alpha(\beta^*(y))) & \\ & = & f_c(\tilde{\beta}^*(\alpha(y))) & \text{by (9.7)}, \end{array}$$

and the result follows from (9.9). ■

Theorem 9.2 is valid for all of the quotients in Example 8.5. However, it is clear that $\mathcal{B}_{G'}$ is a proper subset of $\mathcal{B}_{G_{\bowtie}}$. Therefore, $\mathrm{im}(\hat{\phi}) = \mathrm{im}(\hat{\phi}') \subseteq \mathcal{F}_{G_{\bowtie}} \subset \mathcal{F}_{G'}$, so we gain more information about induced vector fields $\overline{f}$ and their lifts f if we work with $(G_{\bowtie}, \phi)$ rather than (G', ϕ').

Note that Example 7.5 is the natural quotient, so $\hat{\phi}$ need not be surjective when ϕ is natural. It is never surjective when ϕ is not the natural quotient map.

10. Final comments. The formalism of symmetry groupoids proposed in this paper can be set up for many analogous systems that possess a network structure. Here, we have associated to each cell (node of the network) a continuous-time dynamical system defined on a manifold, and to each directed edge a coupling between such systems. We briefly consider variations on this theme.

Extra constraints can be imposed, an important one being to make the system Hamiltonian; see [8]. An analogous formalism can be introduced for discrete-time dynamics (coupled map lattices and generalizations to networks), or discrete-time discrete-space dynamics (cellular automata), and groupoid-equivariance implies constraints on the dynamics (in particular, on patterns of synchrony). If cells represent states of a stochastic process and edges represent transitions, then the network corresponds to a Markov chain, and now the symmetry groupoid implies constraints on the stationary probability density function. We can also extend the groupoid formalism to stochastic differential equations and delay-differential equations.

The theory developed here is a preliminary step toward a formal understanding of pattern-formation in general, not necessarily symmetric, coupled cell networks. Its main focus is robust synchrony. Many other questions about the dynamics of coupled cell networks can be tackled within the groupoid framework; indeed, work is in progress on several of these. In all cases, the central role of the symmetry groupoid as a formal algebraic structure that captures the constraints imposed by the network topology is paramount.

Acknowledgments. We thank Ronnie Brown for advice about groupoids and suggestions for appropriate literature, Ana Dias for commenting on an early version of the paper, and Andrew Török for identifying the $\mathbf{D}_3$ symmetry in the quotient in Example 7.1. The work of Ian Stewart was carried out in part during visits to the Banach Center, Warsaw and the Mathematics Institute of the Jagiellonian University, Kraków, organized, respectively, by Staszek Janeczko and Krzysztof Cesielski; he thanks them and their institutions for hospitality and financial support. Finally, we thank all three referees of this paper for their careful scrutiny, which led to substantial improvements.

REFERENCES

[1] H. Brandt, *Über eine Verallgemeinerung des Gruppenbegriffes*, Math. Ann., 96 (1927), pp. 360–366.
[2] S. Boccaletti, L. M. Pecora, and A. Pelaez, *Unifying framework for synchronization of coupled dynamical systems*, Phys. Rev. E (3), 63 (2001), 066219.
[3] R. Brown, *From groups to groupoids: A brief survey*, Bull. London Math. Soc., 19 (1987), pp. 113–134.
[4] P. L. Buono and M. Golubitsky, *Models of central pattern generators for quadruped locomotion* I. *Primary gaits*, J. Math. Biol., 42 (2001), pp. 291–326.
[5] A. Dias and I. Stewart, *Symmetry groupoids and admissible vector fields for coupled cell networks*, submitted.
[6] M. Golubitsky, M. Nicol, and I. Stewart, *Some curious phenomena in coupled cell networks*, submitted.
[7] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, Progr. Math. 200, Birkhäuser Verlag, Basel, 2002.
[8] M. Golubitsky and I. Stewart, *Patterns of oscillation in coupled cell systems*, in Geometry, Dynamics, and Mechanics: 60th Birthday Volume for J. E. Marsden, P. Holmes, P. Newton, and A. Weinstein, eds., Springer-Verlag, New York, 2002, pp. 243–286.

[9] 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.
[10] P. J. Higgins, *Notes on Categories and Groupoids*, Van Nostrand Reinhold Mathematical Studies 32, Van Nostrand Reinhold, London, 1971.
[11] S. MacLane, *Categories for the Working Mathematician*, Springer-Verlag, New York, 1971.
[12] L. M. Pecora and T. L. Carroll, *Synchronization in chaotic systems*, Phys. Rev. Lett., 64 (1990), pp. 821–824.
[13] W. T. Tutte, *Graph Theory*, Encyclopedia Math. Appl. 21, G.-C. Rota, ed., Addison–Wesley, Reading, MA, 1984.
[14] X. F. Wang, *Complex networks: Topology, dynamics and synchronization*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002) pp. 885–916.
[15] A. Weinstein, *Groupoids: Unifying internal and external symmetry*, Notices Amer. Math. Soc., 43 (1996), pp. 744–752.
[16] R. J. Wilson, *Introduction to Graph Theory*, 3rd ed., Longman, Harlow, UK, 1985.

SIAM J. Applied Dynamical Systems
Vol. 2, No. 4, pp. 647–671

Pulse Dynamics in an Actively Mode-Locked Laser*

John B. Geddes†, Willie J. Firth‡, and Kelly Black§

Abstract. We consider pulse formation dynamics in an actively mode-locked laser. We show that an amplitude-modulated laser is subject to large transient growth and we demonstrate that at threshold the transient growth is precisely the Petermann excess noise factor for a laser governed by a nonnormal operator. We also demonstrate an exact reduction from the governing PDEs to a low-dimensional system of ODEs for the parameters of an evolving pulse. A linearized version of these equations allows us to find analytical expressions for the transient growth below threshold. We also show that the nonlinear system collapses onto an appropriate fixed point, and thus in the absence of noise the ground-mode laser pulse is stable. We demonstrate numerically that, in the presence of a continuous noise source, however, the laser destabilizes and pulses are repeatedly created and annihilated.

Key words. mode-locked laser, transient growth, excess noise, Hermite polynomials

AMS subject classifications. 37-xx, 35-xx

DOI. 10.1137/S1111111102416599

1. Introduction.

Lasers play a crucial role in a number of optical devices and technologies. They are at the core of various consumer devices, and they are responsible for the rate of increase in the speed of our optical communication networks [20]. The need for stable pulsed lasers continues, and this continues to fuel research into either improving current devices or inventing new ones [37].

While the variety of lasers on the marketplace today is quite astounding, a typical laser is composed of an optical resonator, a laser gain medium, and a pump source [33]. The resonator can in principle support a large number of longitudinal (along the resonator axis) and transverse (transverse to the resonator axis) modes. The resonator also introduces losses, which are frequency-dependent. The frequency band over which laser oscillation can occur is determined by the frequency region over which the gain of the laser medium exceeds the resonator losses.

In terms of the longitudinal modes, most lasers have many modes within the gain band, and the laser output consists of radiation at a number of closely spaced frequencies. In the absence of any further control elements, random fluctuations and nonlinear effects in the laser result in an output that varies in an unpredictable way. If, on the other hand, the oscillating modes are forced to maintain equal frequency spacing and a fixed phase relationship, it is possible

*Received by the editors October 30, 2002; accepted for publication (in revised form) by C. Jones July 2, 2003; published electronically December 22, 2003. A preliminary report of the results presented here was given at the IQEC/CLEO Europe in Nice, France in September 2000 (paper CThJ 0007) and at the SIAM Conference on Applications of Dynamical Systems in Snowbird, Utah in May 2001 (session CP25).
http://www.siam.org/journals/siads/2-4/41659.html
†Franklin W. Olin College of Engineering, Needham MA 02492 (john.geddes@olin.edu).
‡Department of Physics, University of Strathclyde, Glasgow G4 0NG, Scotland (willie@phys.strath.ac.uk).
§Department of Mathematics and Statistics, University of New Hampshire, Durham NH 03824 (kelly.black@unh.edu).

to obtain a variety of mode-locked behaviors, including a frequency-modulated output and a continuous train of laser pulses. The technique of mode-locking was proposed theoretically by Lamb in 1964 [25].

In the same year a number of experimental studies were published on the use of internal modulation to obtain mode-locking in gas lasers. Hargrove, Fork, and Pollack [16] experimentally obtained a train of pulses from an He-Ne laser by mode-locking with an internal acoustic loss modulator. Harris and Targ [17] used an internal frequency modulator to mode-lock the same laser, resulting in a frequency-modulated output. In the following years, the techniques of mode-locking were demonstrated in other laser systems, including the argon ion laser [7], the ruby laser [4], and the neodymium-doped yttrium aluminum garnet (Nd:YAG) laser [6]. In each case, the pulse-widths achieved were on the order of tens of pico-seconds, but the pulse trains were subject to fluctuations and instabilities. Other means of mode-locking were also employed, including self-locking [3] and the use of saturable absorbers [8].

Theoretical work on active mode-locking developed in concert with experimental work, with emphasis on either internal amplitude modulation (AM) or frequency modulation (FM). DiDomenico [5] showed theoretically that mode-locking could be obtained by internal loss modulation, while Harris and McDuff [18] developed an internal FM theory. Haken and Pauthier [15] showed later that mode-locked laser pulses are described by a harmonic-oscillator eigenfunction. Nelson [29] extended this work and showed that the steady-state laser pulse is described by a Gaussian envelope modulated by polynomials. These efforts were all carried out in the frequency domain.

The first theoretical results in the time domain were obtained by Kuizenga and Siegman [23], who were able to predict the width of mode-locked pulses in AM and FM lasers. These predictions were confirmed experimentally by the same authors for FM laser operation [24] and by Smith et al. [36] for an AM laser. The theory of Kuizenga and Siegman was based on the assumption that a circulating Gaussian pulse exists in the resonator. By imposing a self-consistency condition after one round-trip, they showed in a simple way that a steady-state Gaussian pulse could exist. Kim, Marathe, and Rabson [22] later showed that the Gaussian pulse was in fact a solution of an integral equation and that higher-order solutions also existed—the Hermite–Gaussian modes. Haus [19] demonstrated two years later that the higher-order modes are linearly unstable and hence unrealizable in the steady-state.

The technique of internal modulation introduces two parameters—the depth of the modulation and the degree to which the modulation period is detuned from the round-trip time of the laser cavity. While the steady-state effects of detuning were considered in the early years of mode-locked lasers, it is only more recently that the dynamic effects of detuning have come under scrutiny. In 1998, Morgner and Mitschke [27] investigated the influence of detuning on the pulse formation dynamics in mode-locked lasers. They found experimentally that in an actively mode-locked laser the equilibrium position of the laser pulse shifts linearly with respect to the detuning parameter—in close agreement with the steady-state predictions of Kuizenga and Siegman [23]. More importantly, they found that perturbations to the steady-state laser pulse grow and "drift" through the pulse with a constant velocity. As the detuning is increased, their findings suggest that no stable pulse can be maintained as perturbations can grow large enough to compete with the existing laser pulse. This results in a destabilizing of the laser which they identify as a drift instability.

More recently, Kärtner, Zumbühl, and Matuschek [21] suggested that this instability exhibits a transition to turbulence. They showed that the detuned laser is an example of a nonnormal system and, as a result, perturbations to the ground-mode pulse are subject to strong transient growth analogous to that experienced by the laminar state in fluid pipe flow [38]. Furthermore, they quantified the effect of detuning and predicted that the transient growth scales with the exponential of the detuning parameter. The precise nature of the instability was not considered, however; nor was the connection to the drift velocity of Morgner and Mitschke.

Similar results have been predicted for the frequency-modulated laser by Longhi and Laporta [26]. In addition to showing that perturbations are subject to large transient amplification due to the nonorthogonality of the laser modes, they also showed that the laser exhibits a strongly enhanced sensitivity to external noise. The existence of so-called excess noise was first predicted by Petermann [30] in 1979 in the context of a gain-guided laser amplifier but has since been generalized to a wide array of optical systems, including general nonorthogonal optical systems [34, 35] and even nonorthogonal polarization modes [39]. As a result, the response of any laser system to noise is often characterized by the Petermann excess noise factor.

The purpose of this paper is to conduct a careful investigation of the dynamics of pulse formation in an amplitude-modulated laser. We show that the transient growth expected at the lasing threshold is equivalent to the Petermann excess noise factor, and we find exact expressions for the transient growth factor away from threshold. Moreover, we find an exact reduction from the governing PDEs to a low-dimensional system of ODEs for the parameters of a circulating pulse. We show that the evolving ground-mode pulse is a globally attracting solution of this system. In the presence of noise, however, we demonstrate numerically that the laser destabilizes and undergoes a repeating cycle of pulse creation and annihilation.

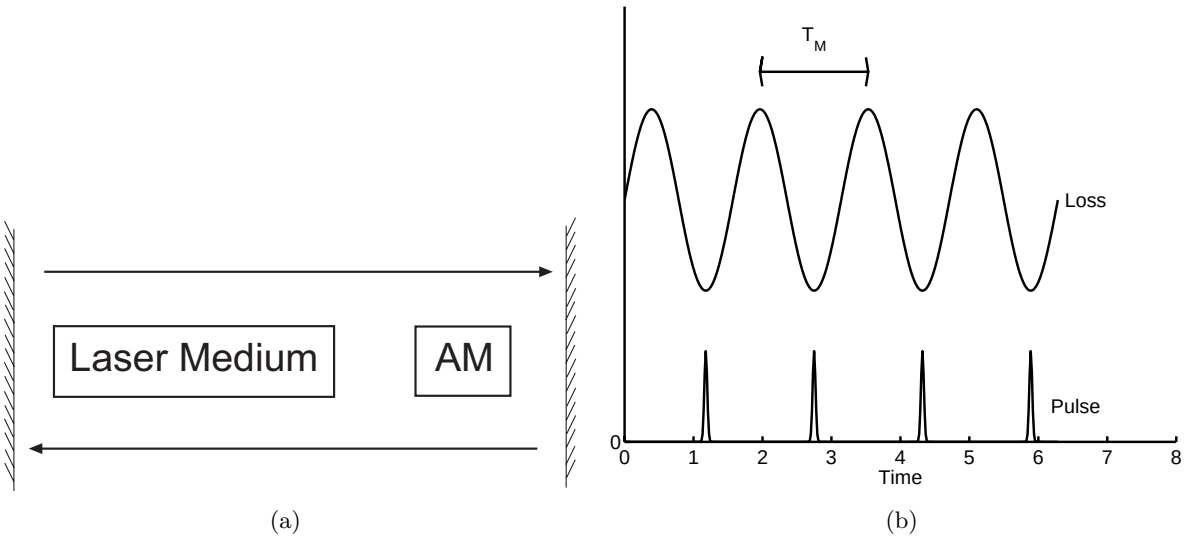


Figure 2.1. (a) *An actively mode-locked laser consisting of an optical resonator, a laser medium, and an amplitude modulator.* (b) *Laser pulses can form only when the losses are minimum.*

2. Governing equations. A schematic of an actively mode-locked laser is shown in Figure 2.1. An amplitude modulator is placed inside a laser cavity, with a modulation period T_M that closely matches the round-trip time T_R of the laser cavity. As a result of the losses introduced by the modulator, the laser pulses are expected to have a pulse-width much shorter than the round-trip time of the cavity, which allows us to describe the laser on two independent time scales. The cavity time T is sampled on the cavity round-trip time T_R, which is typically on the order of nano-seconds. In contrast, the local time t resolves the resulting pulse shape, which may be on the order of pico-seconds. The evolution equations for the complex electric field envelope $A(T,t)$ and the laser gain $g(T)$ for $t \in (-\infty,\infty)$ and $T \in [0,\infty)$ are [21, 10]

$$T_R \frac{\partial A(T,t)}{\partial T} = \left(g(T) - l - \mu t^2\right) A(T,t) + \Gamma \frac{\partial A(T,t)}{\partial t} + D\frac{\partial^2 A(T,t)}{\partial t^2}\ , \tag{2.1}$$

$$\frac{dg(T)}{dT} = \alpha - \gamma g(T) - \beta g(T)||A(T,*)||^2\ , \tag{2.2}$$

where the pulse energy, $||A(T,*)||^2$, is defined in terms of the standard inner-product on $L^2(-\infty,\infty)$, i.e.,

$$||A(T,*)||^2 = \langle A(T,*), A(T,*)\rangle = \int_{-\infty}^{\infty} |A(T,t)|^2 dt.$$

The evolution equation (2.1) for the pulse envelope includes a number of relevant physical effects. The pulse experiences both gain, $g(T)$, and loss, $D\partial_{tt} - l - \mu t^2$, as it propagates in the cavity. The parameter l represents the fixed cavity losses which are frequency-independent, while D represents the curvature of the intracavity losses in the frequency domain which limits the bandwidth of the laser. The parameter μ is proportional to the depth of the loss modulation. While the losses are actually modulated periodically, the modulation depth is relatively large so that radiation can only build up during the time of low intracavity loss, which is much shorter than the modulator period. In that case, the cosine modulation is approximated by a parabola so that μ actually represents the curvature of the loss modulation at the point of minimum loss. In addition, the loss modulation period is not perfectly matched to the round-trip time and this detuning is captured by $\Gamma = T_M - T_R$.

The dynamics of the gain medium is captured in (2.2). The gain in the cavity, $g(T)$, depends on the amount of energy driving the laser, the rate at which the gain medium can radiate the laser photons, and the number of photons present in the cavity. The parameter α accounts for the rate of energy gain due to pumping, γ represents the gain relaxation rate, and β is related to the saturation power of the gain medium. The number of photons in the cavity at any given time T is proportional to the pulse energy, $||A(T,*)||^2$.

The evolution equations, (2.1) and (2.2), are of course supplemented with both initial, $A(T=0,t)$ and $g(T=0)$, and boundary conditions. Our search for pulse-like solutions requires us to satisfy vanishing or Dirichlet boundary conditions for the pulse amplitude, i.e.,

$$\lim_{t\to\pm\infty} A(T,t) = 0. \tag{2.3}$$

Finally, there are several restrictions on the parameters, all of which are assumed real in this paper. Γ may be positive or negative, while l, μ, D, α, γ, and β are all positive. For

the purpose of the simulations presented in this paper, the following fixed values were used: $l = 0.01$, $\mu = 1$, $D = 0.0001$, $\gamma = 0.01$, $\beta = 0.1$, and $T_R = 1$. The parameters Γ and α were varied as needed.

3. Steady-state solutions. The steady-state solutions of (2.1)–(2.2) are defined when both $A(T,t) = A_s(t)$ and $g(T) = g_s$ are independent of T, i.e.,

$$\left(g_s - l - \mu t^2 + \Gamma\partial_t + D\partial_{tt}\right) A_s = 0, \tag{3.1}$$

$$\alpha - \gamma g_s - \beta g_s ||A_s||^2 = 0. \tag{3.2}$$

A cursory inspection of (3.1)–(3.2) reveals that there is a trivial steady-state solution, $A_s = 0$ and $g_s = \alpha/\gamma$. This corresponds to the laser being "off." There are, of course, steady-state solutions which correspond to the laser being on and *mode-locked* and which can be found by looking for nontrivial solutions to (3.1) subject to the boundary conditions (2.3). These nontrivial solutions can be determined by looking for a solution of the form

$$A_s(t) = \psi(t) \exp\left(-\frac{(t-a)^2}{2\sigma^2}\right), \tag{3.3}$$

where a and σ are free parameters to be determined. Substituting ansatz (3.3) into (3.1) results in

$$\begin{aligned} &D\psi_{tt} + \left(\frac{\Gamma\sigma^2 - 2D(t-a)}{\sigma^2}\right)\psi_t \\ &\quad + \left(g_s - l - \mu t^2 + \frac{D(t-a)^2 - D\sigma^2 - \Gamma\sigma^2(t-a)}{\sigma^4}\right)\psi = 0\,. \end{aligned} \tag{3.4}$$

This equation can be transformed into the Hermite equation by a judicious choice of a and σ. If we define

$$\sigma^2 = \sqrt{D/\mu}, \tag{3.5}$$

$$a = -\sigma^2\Gamma/2D \tag{3.6}$$

and transform the variable t by $t \to \sigma u$, the resulting equation is

$$\psi_{uu} - 2u\psi_u + \frac{1}{\sqrt{\mu D}}\left(g_s - l - \sqrt{\mu D} - \frac{\Gamma^2}{4D}\right)\psi = 0\,. \tag{3.7}$$

This is precisely the Hermite equation if

$$\frac{1}{\sqrt{\mu D}}\left(g_s - l - \sqrt{\mu D} - \frac{\Gamma^2}{4D}\right) = 2n; \quad n = 0, 1, 2, \ldots\,, \tag{3.8}$$

which must be true in order to satisfy the boundary conditions (2.3). In other words, a nontrivial solution exists for the pulse amplitude $\psi(t)$ if the steady-state gain, $g_s = g_n$, is given by

$$g_n = l + \left(2\sqrt{\mu D}\left(\Delta^2 + n + \frac{1}{2}\right)\right), \quad n = 0, 1, 2, \ldots\,, \tag{3.9}$$

where Δ is a normalized detuning, $\Delta = \sigma\Gamma/\sqrt{8}D$. If this is the case, then (3.7) is the Hermite equation with the usual Hermite polynomials $H_n(u)$ as solutions. Each value of n in (3.9) gives us a steady-state pulse $A_s(t) = A_n(t)$, which is a Hermite–Gaussian mode of the form

$$A_n(t) = K_n\, H_n\left(\frac{t}{\sigma}\right) \exp\left(-\frac{(t-a)^2}{2\sigma^2}\right). \tag{3.10}$$

The pulse amplitudes, K_n, are determined via (3.2), which gives

$$||A_n||^2 = \frac{\alpha\ -\gamma\ g_n}{\beta g_n}. \tag{3.11}$$

Since the sequence g_n increases linearly with n, the threshold for lasing operation is given by $\alpha = \alpha_{th} = \gamma\ g_0$, as this is the first value of α which permits the existence of a pulse. We expect the ground-mode ($n = 0$) to dominate, at least close to the lasing threshold. The basic lasing solution is a Hermite–Gaussian pulse of width σ centered at $t = a = -\sqrt{2}\sigma\Delta$. If the modulator period matches the round-trip period perfectly, i.e., $\Delta = 0$, then the pulse is centered at $t = 0$, which corresponds to the minimum loss point. If, on the other hand, there is any slight detuning, i.e., $\Delta \neq 0$, then the pulse will be shifted away from the minimum loss point by an amount proportional to the detuning.

4. Linear stability analysis. Haus [19] demonstrated that the trivial solution loses stability to the ground-mode at $\alpha = \alpha_{th} = \gamma g_0$, and to the nth-order mode when $\alpha = \gamma g_n$. He also showed that the ground-mode is linearly stable to the higher-order modes and that the higher-order modes are linearly unstable to the lower-order modes. As mentioned earlier, we would therefore expect that, close to threshold, the ground-mode will be favored.

These linear stability results are limited, however, for reasons that will be detailed in the following section on transient growth. Both for completeness and as a means of motivating this work, we include here a detailed linear stability analysis. Linearization of the governing equations, (2.1) and (2.2), about one of the steady-state solutions ($A = A_s$, $g = g_s$) leads to

$$T_R\frac{\partial}{\partial T}A(T,t) = \left(g_s - l - \mu t^2 + \Gamma\partial_t + D\partial_{tt}\right)A(T,t) + g(T)A_s(t), \tag{4.1}$$

$$\frac{d}{dT}g(T) = -\frac{\alpha}{g_s}g(T) - \beta g_s\left(\langle A_s, A\rangle + \langle A, A_s\rangle\right), \tag{4.2}$$

where $A(T,t)$ and $g(T)$ now represent perturbations to the steady-state solution.

4.1. Trivial steady-state. We begin by considering the stability of the trivial solution, $A_s = 0$ and $g_s = \alpha/\gamma$. In this case the linearized equations reduce to

$$T_R\frac{\partial}{\partial T}A(T,t) = \left(\frac{\alpha}{\gamma} - l - \mu t^2 + \Gamma\partial_t + D\partial_{tt}\right)A(T,t), \tag{4.3}$$

$$\frac{d}{dT}g(T) = -\gamma g(T). \tag{4.4}$$

Note that the pulse envelope and the gain are now decoupled. Perturbations to the gain clearly decay to zero on a time scale of $1/\gamma$. Turning to the pulse envelope equation, the

eigenvalues, λ_n, and eigenfunctions, $\phi_n(t)$, of (4.3) subject to the boundary conditions (2.3) follow naturally from the steady-state solutions (3.10), and we have

$$\left(\frac{\alpha}{\gamma} - l - \mu t^2 + \Gamma\partial_t + D\partial_{tt}\right)\phi_n(t) = \lambda_n\phi_n(t), \tag{4.5}$$

$$\lambda_n = \frac{\alpha}{\gamma} - g_n, \quad n = 0, 1, 2, \dots, \tag{4.6}$$

$$\phi_n(t) = H_n\left(\frac{t}{\sigma}\right)\exp\left(-\frac{(t-a)^2}{2\sigma^2}\right). \tag{4.7}$$

Recalling the nature of g_n, (3.9), reveals that as the pumping parameter α is increased the trivial solution first loses stability to the $n = 0$ mode at $\alpha_{th} = \gamma g_0$. Increasing α further will result in the trivial solution also losing stability to the $n = 1$ mode at $\alpha = \gamma g_1$, the $n = 2$ mode at $\alpha = \gamma g_2$, and so forth.

4.2. Nontrivial steady-state. Linearizing the governing equations about one of the nontrivial solutions, $A = A_n$ and $g = g_n$, leads to the set of equations

$$T_R\frac{\partial}{\partial T}A(T,t) = \left(g_n - l - \mu t^2 + \Gamma\partial_t + D\partial_{tt}\right)A(T,t) + g(T)A_n(t), \tag{4.8}$$

$$\frac{d}{dT}g(T) = -\frac{\alpha}{g_n}g(T) - \beta g_n\left(\langle A_n, A\rangle + \langle A, A_n\rangle\right), \tag{4.9}$$

where $A(T,t)$ and $g(T)$ represent perturbations to the nontrivial solution.

We now seek a solution for $A(T,t)$ in the form of a time-dependent linear combination of the first $N+1$ Hermite–Gaussian modes,

$$A(T,t) = \sum_{j=0}^{N} C_j(T)H_j\left(\frac{t}{\sigma}\right)\exp\left(-\frac{(t-a)^2}{2\sigma^2}\right), \tag{4.10}$$

where a and σ take the values shown in (3.5) and (3.6). If we substitute (4.10) into (4.8) and (4.9) and use the recursive properties of the Hermite polynomials and the linear independence of the Hermite–Gaussians, we arrive at the set of linear ODEs

$$T_R\frac{d}{dT}C_n(T) = K_n g(T), \tag{4.11}$$

$$T_R\frac{d}{dT}C_j(T) = (g_n - g_j)C_j(T), \quad j = 0, 1, 2, \dots, N, \quad j \neq n, \tag{4.12}$$

$$\frac{d}{dT}g(T) = -\frac{\alpha}{g_n}g(T) - 2\beta g_n K_n\sum_{j=0}^{N} M_{nj}C_j(T), \tag{4.13}$$

where the interaction matrix elements, M_{nj}, are given by

$$M_{nj} = \int_{-\infty}^{\infty} H_n\left(\frac{t}{\sigma}\right)H_j\left(\frac{t}{\sigma}\right)\exp\left(-\frac{(t-a)^2}{\sigma^2}\right)dt.$$

Let us first consider the stability of the ground-mode, $n = 0$. The eigenvalues, $\{\lambda_j\}_{j=1}^{N+2}$, of this system of $N + 2$ linear differential equations can be determined explicitly and are

$$\lambda_1 = -\frac{\alpha T_R}{2g_0}\left(1 + \sqrt{1 - \frac{8(\alpha - \gamma g_0)g_0^2}{\alpha^2 T_R}}\right), \tag{4.14}$$

$$\lambda_2 = -\frac{\alpha T_R}{2g_0}\left(1 - \sqrt{1 - \frac{8(\alpha - \gamma g_0)g_0^2}{\alpha^2 T_R}}\right), \tag{4.15}$$

$$\lambda_j = g_0 - g_{j-2}, \quad j = 3, 4, \ldots, N + 2. \tag{4.16}$$

The first two eigenvalues involve the coupling between perturbations to the ground-mode and to the gain. Recalling that $\alpha - \gamma g_0 > 0$ for the ground-mode to exist implies that these perturbations are always damped since both eigenvalues have negative real parts. The third eigenvalue governs the dynamics of perturbations along the $n = 1$ mode and the coupling between it, the ground-mode, and the gain. These perturbations are always damped since $g_1 > g_0$. The other eigenvalues follow suit, and perturbations along the higher-order modes are more heavily damped. We therefore conclude that the ground-mode is linearly stable to arbitrary perturbations. If we consider the stability of the higher-order modes, $n > 0$, then (4.12) shows that these modes are always unstable to each of the lower-order modes.

In summary, then, we have confirmed that if $\alpha < \alpha_{th}$, the trivial solution is linearly stable. If, on the other hand, $\alpha > \alpha_{th}$, the trivial solution is linearly unstable, and the $n = 0$ mode begins to grow. The other modes also turn on as α is further increased. The $n = 0$ mode is linearly stable to arbitrary perturbations if $\alpha > \alpha_{th}$, and the higher-order modes are linearly unstable to the $n = 0$ mode. Linear analysis suggests, therefore, that below threshold the laser will remain "off," while above threshold the laser will "turn on" and form a mode-locked pulse.

5. Transient growth. Linear stability theory guarantees only that infinitesimal perturbations will die off asymptotically; short-term growth is a possibility. If the linear operator is *normal* [1, 28], then this can be strengthened further and we can conclude that perturbations cannot grow. On the other hand, nonnormal operators allow for short-term transient growth of perturbations.

The importance of nonnormal operators emerged recently in the context of viscous shear flow [11, 31, 32, 38] as a possible mechanism for the instability of the laminar flow below threshold. The nonorthogonality of the eigenfunctions of the linear operator results in short-term transient growth of perturbations to the laminar state. The subsequent interplay between transient growth and nonlinear mixing may be enough to then destabilize the system. The calculation of the magnitude of transient growth expected [31] and of the types of perturbations which experience maximum transient growth [11] is complicated by the nature of the linear operator.

In the context of actively mode-locked lasers, Kärtner, Zumbühl, and Matuschek [21] pointed out that the linearized amplitude-modulated laser is governed by a nonnormal operator. They suggested that the degree of nonnormality depends strongly on the detuning Δ and

that transient growth on the order of $\exp(4\Delta^2)$ could be expected. Longhi and Laporta [26] have reported similar findings for the frequency-modulated laser.

The relevance of nonnormal operators to laser systems is much older, however, than the dates of these studies might suggest. It has been known since the seminal work of Petermann in 1979 that, in the presence of stochastic forcing, laser systems governed by nonnormal operators are subject to excess noise [30, 34, 35]. (Laser operation depends on the presence of such stochastic forcing in the form of spontaneous emission noise.) The Petermann excess noise factor (K) is a measure of the enhancement of the noise source and is given by

$$K = \frac{\langle u, u\rangle \, \langle v, v\rangle}{\langle u, v\rangle^2}, \tag{5.1}$$

where u is the eigenfunction of the linear operator, v is the adjoint eigenfunction, and $\langle *, *\rangle$ is an appropriate inner-product.

We will show in section 5.1 that the growth sustained by perturbations at the lasing threshold is precisely the Petermann excess noise factor. Furthermore, it will become clear that the perturbation which leads to maximum growth is the adjoint eigenmode, in agreement with known results in viscous shear flow [11]. Below threshold we compute the optimal growth curve [32, 38] by evaluating the norm of the solution operator to the linearized equation; the solution operator is based on the truncated Hermite expansion discussed in [2]. We find that the transient growth below threshold is bounded in magnitude by $\exp(4\Delta^2)$ and that the perturbation which experiences most growth is closely related to the adjoint mode. In section 6 we confirm these numerical results by finding explicit analytical expressions for the transient growth below threshold as well as the form of the relevant perturbation. In section 5.2 we compute the optimal growth curve for the linear operator above threshold and find again that the growth is bounded by $\exp(4\Delta^2)$. The perturbation which gives rise to the maximum transient growth in this case is the adjoint mode.

5.1. Below threshold. In light of this, let us reconsider the linear stability of the nonlasing solution. Recall that the nonlasing solution is defined by $A_s = 0$ and $g_s = \alpha/\gamma$. The linearized equations, already introduced in section 4, are

$$T_R \frac{\partial}{\partial T} A(T,t) = \mathcal{L} A(T,t), \tag{5.2}$$

$$\frac{d}{dT} g(T) = -\gamma g(T), \tag{5.3}$$

$$\mathcal{L} = \frac{\alpha}{\gamma} - l - \mu t^2 + \Gamma \partial_t + D \partial_{tt}, \tag{5.4}$$

where again $g(T)$ and $A(T,t)$ represent perturbations to the nonlasing solution. We have already determined the eigenvalues, λ_n, and eigenfunctions, $\phi_n(t)$, of $\mathcal{L}$ subject to boundary conditions (2.3),

$$\mathcal{L}\phi_n(t) = \lambda_n \phi_n(t), \tag{5.5}$$

$$\lambda_n = \frac{\alpha}{\gamma} - g_n, \quad n = 0, 1, \dots, \tag{5.6}$$

$$\phi_n(t) = H_n\left(\frac{t}{\sigma}\right) \exp\left(-\frac{(t-a)^2}{2\sigma^2}\right). \tag{5.7}$$

We will also require the adjoint modes. The adjoint operator, defined as usual by

$$\langle u, \mathcal{L}v\rangle = \langle \mathcal{L}^* u, v\rangle, \tag{5.8}$$

is given by

$$\mathcal{L}^* = \frac{\alpha}{\gamma} - l - \mu t^2 - \Gamma\partial_t + D\partial_{tt}. \tag{5.9}$$

The adjoint operator $\mathcal{L}^*$ can be obtained from $\mathcal{L}$ by making the transformation $\Gamma \to -\Gamma$. The eigenvalues μ_n and eigenfunctions ψ_n of $\mathcal{L}^*$ subject to the boundary conditions (2.3) are given by

$$\mathcal{L}^*\psi_n(t) = \mu_n\psi_n(t), \tag{5.10}$$

$$\mu_n = \frac{\alpha}{\gamma} - g_n, \quad n = 0, 1, \dots, \tag{5.11}$$

$$\psi_n(t) = H_n\left(\frac{t}{\sigma}\right)\exp\left(-\frac{(t+a)^2}{2\sigma^2}\right). \tag{5.12}$$

The dependence of g_n on Γ^2 leaves the eigenvalues unchanged under the transformation, and the linear dependence of a on Γ means that the adjoint modes are the Hermite–Gaussians with the Gaussian centered at $t = -a$.

The general solution to (5.2) is given by

$$A(T,t) = \sum_{n=0}^{\infty} a_n\phi_n(t)e^{\lambda_n T/T_R},$$

where the a_n's are chosen to satisfy the initial condition

$$A(0,t) = \sum_{n=0}^{\infty} a_n\phi_n(t).$$

Recalling that the eigenfunctions and their adjoints form a biorthogonal set [28, 33], we can determine the coefficients a_n by projecting onto the adjoint to give

$$a_n = \frac{\langle A(0,t), \psi_n(t)\rangle}{\langle \phi_n(t), \psi_n(t)\rangle},$$

where $\langle *, *\rangle$ is again the standard inner-product on $L^2(-\infty, \infty)$. The general solution to the linearized problem is then

$$A(T,t) = \sum_{n=0}^{\infty} \frac{\langle A(0,t), \psi_n(t)\rangle}{\langle \phi_n(t), \psi_n(t)\rangle}\ \phi_n(t)\ e^{\lambda_n T/T_R}.$$

Below the lasing threshold, all of the eigenvalues are negative. As we approach threshold, however, the leading eigenvalue approaches zero. Let us first consider the perturbation dynamics at threshold. Since $\lambda_0 = 0$ and $\dots \lambda_3 < \lambda_2 < \lambda_1 < 0$, in the limit as $T \to \infty$ we find that

$$A_\infty(t) = \lim_{T\to\infty} A(T,t) = \frac{\langle A(0,t), \psi_0(t)\rangle}{\langle \phi_0(t), \psi_0(t)\rangle}\ \phi_0(t)$$

so that the dynamics eventually collapse onto the ground-mode as we would expect since the ground-mode is neutrally stable at threshold and all other modes are unstable. The energy of this solution is then

$$||A_\infty(t)||^2 = \langle A_\infty(t), A_\infty(t)\rangle = \frac{\langle A(0,t),\psi_0(t)\rangle^2\langle\phi_0(t),\phi_0(t)\rangle}{\langle\phi_0(t),\psi_0(t)\rangle^2}.$$

For a given ground-state mode this depends only on the initial condition, and the energy is maximized if the initial condition is parallel to the adjoint ground-mode, i.e.,

$$A(0,t) \propto \psi_0(t). \tag{5.13}$$

In this case the energy growth factor is

$$\text{Growth} = \frac{||A_\infty(t)||^2}{||A(0,t)||^2} = \frac{\langle\phi_0(t),\phi_0(t)\rangle\langle\psi_0(t),\psi_0(t)\rangle}{\langle\phi_0(t),\psi_0(t)\rangle^2}, \tag{5.14}$$

which is precisely the Petermann excess noise factor of (5.1). The inner-products are given by

$$\begin{aligned}
\langle\phi_0(t),\phi_0(t)\rangle &= \int_{-\infty}^{\infty} \exp\left(-\frac{(t-a)^2}{\sigma^2}\right) dt = \sigma\sqrt{\pi},\\
\langle\psi_0(t),\psi_0(t)\rangle &= \int_{-\infty}^{\infty} \exp\left(-\frac{(t+a)^2}{\sigma^2}\right) dt = \sigma\sqrt{\pi},\\
\langle\phi_0(t),\psi_0(t)\rangle &= \int_{-\infty}^{\infty} \exp\left(-\frac{(t-a)^2}{2\sigma^2}\right)\exp\left(-\frac{(t+a)^2}{2\sigma^2}\right) dt,\\
&= \exp\left(-\frac{a^2}{\sigma^2}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{t^2}{\sigma^2}\right) dt,\\
&= \sigma\sqrt{\pi}\exp\left(-\frac{a^2}{\sigma^2}\right),
\end{aligned}$$

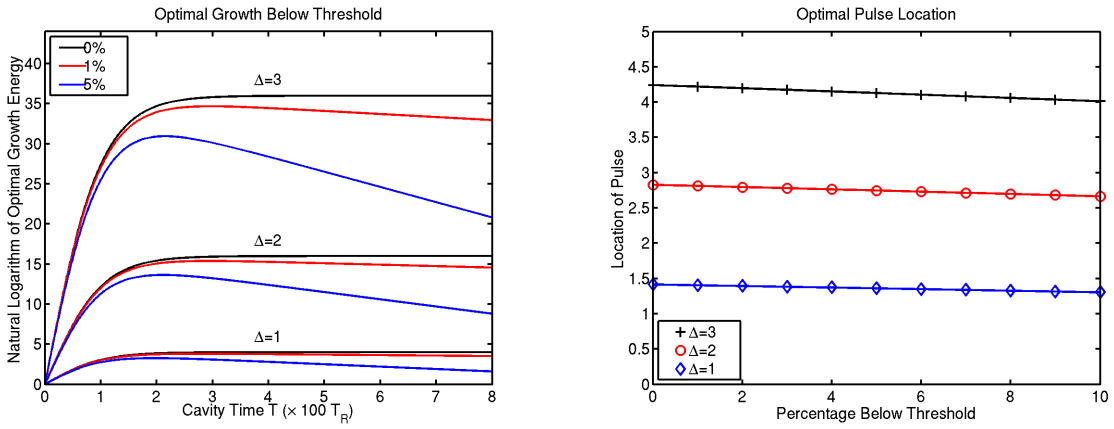
and the energy growth is then

$$\text{Growth} = \exp\left(\frac{2a^2}{\sigma^2}\right) = \exp\left(4\Delta^2\right),$$

which agrees with the results of Kärtner, Zumbühl, and Matuschek [21].

We have shown that if we linearize the governing equations about the trivial solution, then the resulting linear system can be solved exactly. Since the leading eigenvalue is zero at threshold, any perturbation eventually collapses onto the ground-mode, which is a Gaussian pulse located at $t = a$. If we ask which perturbation leads to the most growth in energy, it is the adjoint ground-mode, which corresponds to a Gaussian pulse located at $t = -a$. We show in section 6 how such a perturbation evolves into the ground-mode, but in any case the linear growth experienced by such a perturbation has an exponential dependence on Δ^2.

Below threshold, we compute the optimal growth curve in order to identify the optimal perturbation which leads to maximum transient growth [32, 38]. The optimal growth energy is defined as the norm of the solution operator to (5.2), i.e., $||\exp(\mathcal{L}t/T_R)||^2$, where $\mathcal{L}$ is defined



(a) (b)

Figure 5.1. (a) *Optimal growth curves below threshold for various values of* α *and* Δ. (b) *Location of optimum pulse.*

in (5.4). Our algorithm is based on a spectral Hermite approximation, as discussed in [2]. In each computation discussed below we use 128 basis functions. We vary α and Δ and set all other parameters to the values given in section 2.

In Figure 5.1(a) we show the natural logarithm of the optimal growth energy against cavity time T for different values of Δ and α. For each value of Δ we vary α from its threshold value to 5% below threshold. (Recall that α_{th} depends on Δ.) At threshold, the optimal growth energy increases exponentially before leveling off at the expected value of $\exp(4\Delta^2)$. Below threshold, the optimal growth energy increases exponentially, reaches a maximum value, and then decays exponentially. While the maximum energy obtained in each case is less than that obtained at threshold, it is still of the same order. In addition, the time taken to reach maximum energy depends on Δ and decreases as we move below threshold. The precise relationship between α, Δ, the maximum energy, and the time taken to reach maximum energy can be obtained exactly and is discussed in section 6.

The optimal perturbation which leads to maximum transient growth is determined by first finding the time of maximum growth and then computing the SVD of the matrix exponential; the leading singular vector gives the optimal perturbation [32, 38]. We find numerically that the optimal perturbation is a single Gaussian pulse of equilibrium width. In Figure 5.1(b) we graph the optimal pulse location determined numerically (symbols) and the exact optimal location (solid lines) obtained in section 6. These results demonstrate that at threshold the pulse occupies the adjoint location as expected. Below threshold, however, the optimal pulse location moves toward the origin.

The transient pulse formation dynamics are captured in Figure 5.2, which shows a contour plot of the pulse amplitude as a function of local time t and cavity time T for $\Delta = 1$ at 5% below threshold. An initial optimum pulse grows exponentially as it sweeps across the domain into position. The time taken to attain maximum energy roughly coincides with the time taken to sweep into position. After this point, the pulse-amplitude decays exponentially to zero. In section 6, we will find an explicit solution for the pulse-location and pulse-amplitude as a

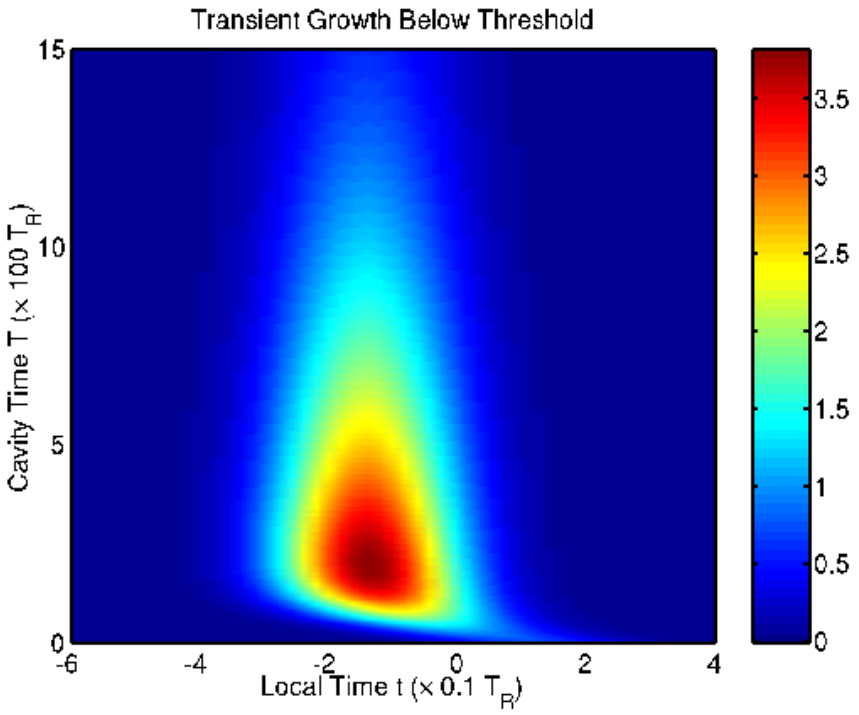


Figure 5.2. *Contour plot of pulse-amplitude versus local time t and cavity time T for $\Delta = 1$ at 5% below threshold. The initial condition corresponds to an optimum pulse centered at $t = \sqrt{1.85} \times 0.1 T_R$.*

function of T which will confirm these numerical experiments.

5.2. Above threshold. Above threshold, the growth of an optimal perturbation to the trivial solution is no longer transient since the trivial solution itself becomes unstable to the ground-mode pulse. In this regime, we need to consider the equations linearized about the ground-mode, A_0 and g_0. In this case they are

$$T_R \frac{\partial}{\partial T} A(T,t) = \left(g_0 - l - \mu t^2 + \Gamma \partial_t + D\partial_{tt}\right) A(T,t) + g(T) A_0(t), \tag{5.15}$$

$$\frac{d}{dT} g(T) = -\frac{\alpha}{g_0} g(T) - \beta g_0 \left(\langle A_0, A\rangle + \langle A, A_0\rangle\right), \tag{5.16}$$

where $A(T,t)$ and $g(T)$ represent perturbations to the ground-mode solution.

In Figure 5.3(a) we show the optimal growth results for several different values of Δ with α held at 5% above threshold. In each case, we compute the norm of the solution operator to the linearized equations, (5.15) and (5.16). Above threshold, the optimal energy initially grows exponentially before dying off. The nature of the die-off changes as we increase Δ since the leading eigenvalue becomes complex. The oscillation period matches that predicted by linear stability analysis. In Figure 5.3(b) we set $\Delta = 3$ and show the results as we change α from its threshold value to 5% above threshold. At threshold, the linear operator governing perturbations to an equilibrium ground-mode pulse matches the linear operator which governs perturbations to the trivial solution. The growth at threshold should therefore be $\exp(4\Delta^2)$,

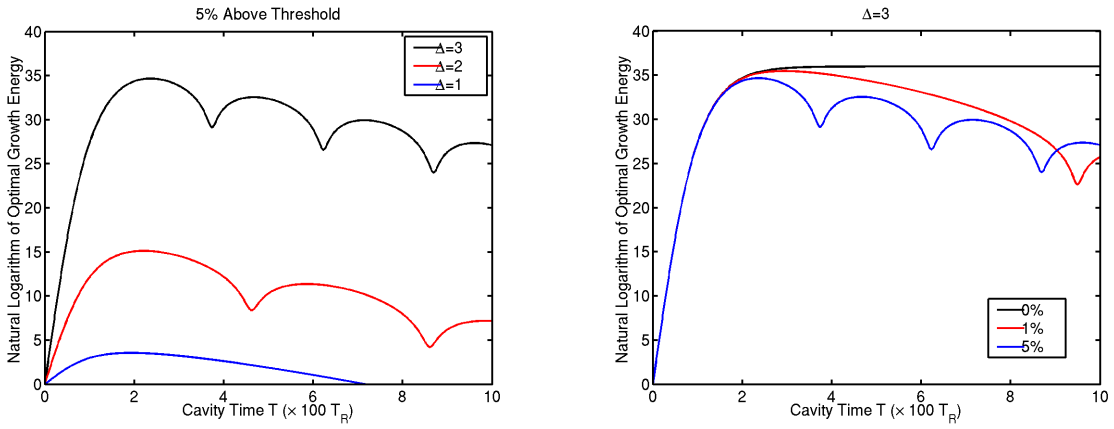


Figure 5.3. *Optimal growth curves above threshold:* (a) *Natural logarithm of optimal growth energy versus cavity time as a function of* Δ *at* 5% *above threshold;* (b) *natural logarithm of optimal growth energy versus cavity time as a function of* α *for* $\Delta = 3$.

and this is indeed the case. The growth above threshold is bounded by this value, and computing the SVD of the matrix exponential at the time of maximum energy reveals that the adjoint mode is in fact the optimal perturbation. We have checked this result for various values of Δ and up to 50% above threshold.

6. Nonlinear dynamics.

The linear stability analysis of section 4 and the details on transient growth of section 5 provide us with a general picture of the linear dynamics of perturbations to both the trivial solution and the ground-mode pulse. We now consider the full nonlinear dynamics of pulse formation. We begin by considering the evolution of a single pulse, for which an exact reduction to a low-dimensional system of nonlinear ODEs is available—the so-called pulse-parameter equations. We use a linearized version of these equations to extend our results on transient growth, and we show analytically that above threshold a ground-mode pulse is an attracting fixed point of the nonlinear system. These results are confirmed by direct simulation of the ODEs using an implicit method [2].

We conclude this section with a series of full simulations of the governing PDEs in the presence of noise. In the first set of simulations, we consider the interaction between a ground-mode pulse and a single noise pulse. In concert with our transient growth predictions, we find that a ground-mode pulse undergoes a strong interaction with a noise pulse but that a ground-mode pulse eventually re-emerges from this interaction. In the second and final set of simulations, we include a stochastic noise source and find that the ground-mode pulse repeatedly undergoes a process of creation and annihilation which could be interpreted as turbulent dynamics [21].

6.1. Single pulse evolution.

In this section we address the question of how a single ground-mode pulse evolves from the trivial solution. A reduction to a set of pulse-parameter equations is achieved by seeking a ground-mode pulse whose pulse parameters depend on T. This is motivated by the steady-state circulating Gaussian pulse analysis work of Kuizenga

and Seigman [23], but we derive dynamic pulse-parameter equations in a systematic manner.

We first assume a pulse solution of the form

$$A(T,t) = \psi(T)\exp\left(-\frac{(t-a(T))^2}{2\sigma^2(T)}\right), \tag{6.1}$$

where the pulse parameters $\psi(T)$, $a(T)$, and $\sigma(T)$ represent the pulse-amplitude, pulse-location, and pulse-width, respectively. In light of our choice of real parameters, we may choose the pulse parameters to be real without loss of generality. The following derivatives are required:

$$\begin{aligned}
\Gamma\frac{\partial A}{\partial t} &= -\frac{\Gamma\psi(t-a)}{\sigma^2}\exp\left(-\frac{(t-a)^2}{2\sigma^2}\right),\\
D\frac{\partial^2 A}{\partial t^2} &= \left(-\frac{D\psi}{\sigma^2}+\frac{D\psi(t-a)^2}{\sigma^4}\right)\exp\left(-\frac{(t-a)^2}{2\sigma^2}\right),\\
T_R\frac{\partial A}{\partial T} &= \left(T_R\psi_T+\frac{T_R\psi a_T(t-a)}{\sigma^2}+\frac{T_R\psi(\sigma^2)_T(t-a)^2}{2\sigma^4}\right)\exp\left(-\frac{(t-a)^2}{2\sigma^2}\right),
\end{aligned}$$

and we write $\mu t^2 = \mu(t-a)^2 + 2a\mu(t-a) + \mu a^2$. Substituting these derivatives into the governing equation (2.1) and grouping terms into the form of standard polynomials give

$$\begin{aligned}
&T_R\left(\psi_T P_0(u)+\frac{a_T}{\sigma}\psi P_1(u)+\frac{\sigma_T}{\sigma}\psi P_2(u)\right)\exp\left(-\frac{u^2}{2}\right)\\
&=\left(g-l-\mu a^2-\frac{D}{\sigma^2}\right)\psi P_0(u)\exp\left(-\frac{u^2}{2}\right)\\
&-\left(2\mu\sigma a+\frac{\Gamma}{\sigma}\right)\psi P_1(u)\exp\left(-\frac{u^2}{2}\right)\\
&+\left(\frac{D}{\sigma^2}-\mu\sigma^2\right)\psi P_2(u)\exp\left(-\frac{u^2}{2}\right),
\end{aligned}$$

where $u=(t-a)/\sigma$ and $P_j(u)=u^j$. Linear independence of the polynomials $P_j(u)$ leads to the following set of differential equations for the pulse-parameters:

$$\begin{aligned}
T_R\frac{d}{dT}\sigma^2 &= 2(D-\mu\sigma^4),\\
T_R\frac{d}{dT}a &= -\Gamma-2\mu\sigma^2 a,\\
T_R\frac{d}{dT}\psi &= \left(g-l-\mu a^2-\frac{D}{\sigma^2}\right)\psi.
\end{aligned}$$

Turning our attention to the gain equation, (2.2), the pulse ansatz (6.1) leads to

$$\begin{aligned}
\frac{d}{dT}g &= \alpha-\gamma g-\beta g\psi^2\int_{-\infty}^{\infty}\exp\left(-\frac{(t-a)^2}{\sigma^2}\right)dt\\
&= \alpha-\gamma g-\beta g\psi^2\sigma\sqrt{\pi}.
\end{aligned}$$

The pulse-parameter equations are then

$$T_R \frac{d}{dT}\sigma^2 = 2(D - \mu\sigma^4), \tag{6.2}$$

$$T_R \frac{d}{dT}a = -\Gamma - 2\mu\sigma^2 a, \tag{6.3}$$

$$T_R \frac{d}{dT}\psi = \left(g - l - \mu a^2 - \frac{D}{\sigma^2}\right)\psi, \tag{6.4}$$

$$\frac{d}{dT}g = \alpha - \gamma g - \sqrt{\pi}\beta\sigma g\psi^2. \tag{6.5}$$

The system of equations (6.2)–(6.5) forms a hierarchy in the following sense. The pulse-width, $\sigma(T)$, evolves independently of the other pulse-parameters. The pulse-location, $a(T)$, is driven by the dynamics of the pulse-width. The pulse-amplitude, $\psi(T)$, and gain, $g(T)$, are mutually coupled and depend on the pulse-width and pulse-position. There are two fixed points of this dynamical system. In both cases, the equilibrium pulse-width and pulse-location take the values

$$\sigma_s^2 = \sqrt{D/\mu}, \quad a_s = -\frac{\Gamma}{2\mu\sigma_s^2}.$$

One of the fixed points corresponds to the trivial solution,

$$g_s = \alpha/\gamma, \quad \psi_s = 0,$$

while the other corresponds to the ground-mode pulse,

$$g_s = l + \mu a_s^2 + D/\sigma_s^2, \quad \psi_s^2 = \frac{\alpha - \gamma g_s}{\beta g_s \sqrt{\pi}\sigma_s}.$$

These agree precisely with the results derived in section 3.

6.1.1. Linear dynamics. Before discussing the general solution of the pulse-evolution equations, let us first reconsider the dynamics of transient linear pulse formation below threshold. If we linearize the pulse-amplitude and gain equations about the trivial solution ($\psi_s = 0$ and $g_s = \alpha/\gamma$), the result is

$$T_R \frac{d}{dT}\sigma^2 = 2\left(D - \mu\sigma^4\right), \tag{6.6}$$

$$T_R \frac{d}{dT}a = -\Gamma - 2\mu\sigma^2 a, \tag{6.7}$$

$$T_R \frac{d}{dT}\psi = \left(\frac{\alpha}{\gamma} - l - \mu a^2 - \frac{D}{\sigma^2}\right)\psi, \tag{6.8}$$

$$\frac{d}{dT}g = -\gamma g, \tag{6.9}$$

where $\psi(T)$ and $g(T)$ represent perturbations to the trivial solution. Note that the pulse-amplitude and gain are now decoupled. The perturbation to the steady-state gain dies off

on the time scale of $1/\gamma$, while the pulse-amplitude perturbation is driven only by the pulse-location and pulse-width.

In order to simplify the resulting analysis, let us assume that the pulse-width is in steady-state, i.e., $\sigma = \sigma_s$. Equation (6.7) is a nonhomogeneous first-order differential equation and has the solution

$$a(T) = a_s + (a_0 - a_s)\exp\left(-\frac{2\sqrt{\mu D}\,T}{T_R}\right), \tag{6.10}$$

where a_0 is the initial location of the pulse. The pulse therefore sweeps into position monotonically on a time scale of $T_R/2\sqrt{\mu D}$. Note that while the *velocity* of the pulse, da/dT, depends on its current location, it depends linearly on the detuning Γ, in agreement with the drift velocity findings of Morgner and Mitschke [27].

Recalling the definition of g_0, (3.9), the pulse-amplitude equation, (6.8), can be written as

$$T_R\frac{d}{dT}\psi = \left(\left(\frac{\alpha}{\gamma} - g_0\right) + \mu(a_s^2 - a^2)\right)\psi \tag{6.11}$$

and is therefore separable. An explicit solution in terms of T is possible, but it is more revealing to write the solution as a function of pulse-location $a(T)$,

$$\psi(T) = \psi_0\exp\left(\frac{(\alpha - \gamma g_0)T}{\gamma T_R}\right)\exp\left(\frac{(a(T)+a_s)^2 - (a_0+a_s)^2}{4\sigma_s^2}\right), \tag{6.12}$$

where ψ_0 is the initial amplitude of the pulse.

Let us first consider the linearized dynamics at threshold, in which case $\alpha = \gamma g_0$. The first exponential in (6.12) therefore drops out. Moreover, since $a(T)$ changes monotonically, we see that $\psi(T)$ must also change monotonically, and we also know that $\lim_{T\to\infty} a(T) = a_s$. Therefore,

$$\psi_s = \lim_{T\to\infty}\psi(T) = \psi_0\exp\left(\frac{4a_s^2 - (a_0+a_s)^2}{4\sigma_s^2}\right). \tag{6.13}$$

What then should the initial location of the pulse be in order to maximize the growth in pulse-amplitude? The quadratic nature in a_0 of the exponent above reveals that the optimum value is $a_0 = -a_s$, i.e., the adjoint location. In this case, the pulse experiences a growth of

$$\text{Growth} = \frac{\psi_s^2}{\psi_0^2} = \exp\left(\frac{2a_s^2}{\sigma_s^2}\right) = \exp(4\Delta^2), \tag{6.14}$$

which agrees with our results from section 5.

We can extend our analysis further by considering the transient growth below threshold, for which $\alpha < \gamma g_0$. In this case, the first exponential in (6.12) no longer vanishes but represents a decaying exponential function of T. Multiplication by the other exponential, which changes monotonically in T, shows that the growth experienced by a pulse below threshold is less than that at threshold. The maximum amplitude obtained by the pulse can be determined by finding the time T^* at which $\psi_T = 0$. Let us denote the pulse-amplitude and pulse-location

at this time as $\bar{\psi}$ and $\bar{a}$, respectively, i.e., $\bar{\psi} = \psi(T^*)$ and $\bar{a} = a(T^*)$. Then (6.11) shows that the maximum amplitude is obtained when

$$\left(\frac{\alpha}{\gamma} - g_0\right) + \mu\left(a_s^2 - \bar{a}^2\right) = 0. \tag{6.15}$$

If we denote $\alpha = \alpha_{th}(1-\varepsilon)$ so that ε represents the % below threshold, then the pulse-location corresponding to maximum amplitude is given by one of the roots of

$$\bar{a}^2 = a_s^2\left(1 - \frac{\varepsilon g_0}{\mu a_s^2}\right), \tag{6.16}$$

which shows that close to threshold the pulse-location decreases linearly with ε.

There are, of course, two solutions to (6.16). A simple calculation shows that one of these corresponds to a maximum amplitude and the other to a minimum amplitude. The maximum energy growth is therefore achieved if we place the initial pulse in the minimum location, i.e., $a_0 = -\bar{a}$. At threshold, this optimum pulse is just the adjoint pulse, but below threshold this optimum pulse is located closer to the origin as discussed in section 5.

The time T^* at which the maximum amplitude is obtained can now be determined from (6.10). If we assume that the pulse begins in the optimum location, $a_0 = -\bar{a}$, then

$$\frac{2\sqrt{\mu D}}{T_R}T^* = \ln\left(\frac{\varepsilon g_0}{\mu a_s^2}\right) - 2\ln\left(1 - \sqrt{1 - \frac{\varepsilon g_0}{\mu a_s^2}}\right). \tag{6.17}$$

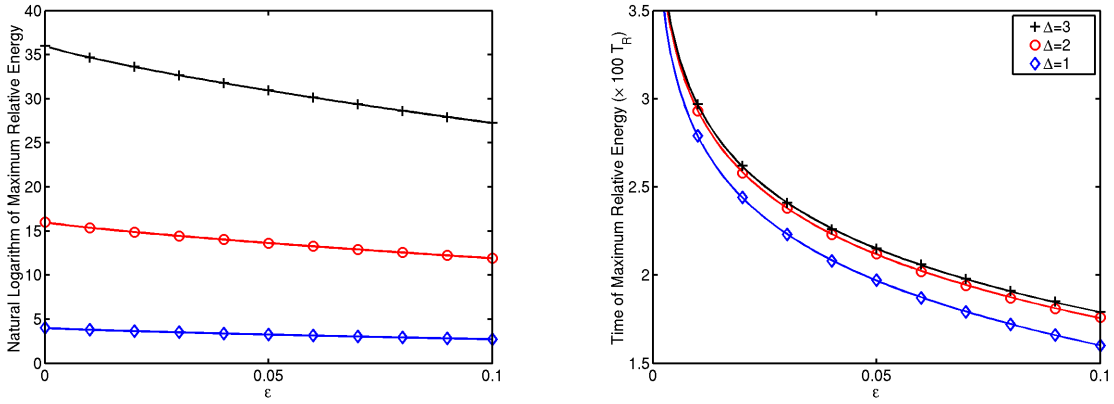
The time taken to reach maximum energy therefore decreases according to $\ln(\varepsilon)$. The maximum energy can now be determined directly from (6.12) if we evaluate this expression with $T = T^*$, $a(T^*) = \bar{a}$, and $a_0 = -\bar{a}$. The result is

$$\ln\left(\frac{\bar{\psi}^2}{\psi_0^2}\right) = 4\Delta^2\sqrt{1 - \frac{\varepsilon g_0}{\mu a_s^2}} - \frac{2\varepsilon g_0 T^*}{T_R}, \tag{6.18}$$

which shows that the maximum energy obtained by the pulse below threshold is always less than that at threshold.

These results are summarized in Figure 6.1, in which we graph both the exact results obtained above (solid lines) and those from direct numerical simulation of the governing linearized PDE (symbols). In Figure 6.1(a) we show the maximum relative energy attained by the pulse as a function of ε, while in Figure 6.1(b) we show the time taken to achieve maximum relative energy versus ε. These results demonstrate that close to threshold the initial optimum pulse experiences growth on the order of $\exp(4\Delta^2)$, and that this growth shows a weak drop-off versus ε.

6.1.2. Nonlinear dynamics. Now let us consider the general solution of the pulse-parameter equations, (6.2)–(6.5). The equations for the pulse-width and pulse-location can be solved exactly. In the limit as $T \to \infty$ both σ and a approach their steady-state values, σ_s and a_s. In terms of the long-term behavior we can therefore consider the pulse-amplitude and gain equations with both the pulse-width and pulse-location in steady-state. Under these conditions



(a) (b)

Figure 6.1. *Transient growth of an optimal perturbation to the trivial solution below threshold:* (a) *Natural logarithm of the maximum relative pulse energy versus percentage below threshold;* (b) *time taken to reach maximum relative energy versus percentage below threshold.*

the equations are

$$T_R \frac{d}{dT}\psi = (g - g_0)\psi, \tag{6.19}$$

$$\frac{d}{dT}g = \alpha - \gamma g - \sqrt{\pi}\beta\sigma_s g\psi^2. \tag{6.20}$$

It is convenient at this stage to use the pulse-energy $E = \sqrt{\pi}\sigma_s\psi^2$ and rewrite the equations as

$$E_T = 2\frac{(g - g_0)}{T_R}E, \tag{6.21}$$

$$g_T = \alpha - \gamma g - \beta g E. \tag{6.22}$$

This system represents a set of nonlinear differential equations in the plane. As a result, there is a limited number of solutions available, including evolution to a fixed point and limit cycles [14]. As we already know, the fixed point corresponding to the trivial solution, $g = \alpha/\gamma$, $E = 0$, exists for all α, is linearly stable for $\alpha < \alpha_{th}$, and is linearly unstable for $\alpha > \alpha_{th}$. In addition, the fixed point corresponding to the ground-mode lasing solution, $g = g_0$, $E = (\alpha - \gamma g_0)/\beta g_0$, exists only if $\alpha > \alpha_{th}$ and is linearly stable there. In what follows we will show that closed orbits are not possible. The system therefore evolves so as to collapse onto the appropriate fixed point. Below threshold this is the trivial solution, while above threshold this is the ground-mode lasing solution.

In order to rule out the existence of closed orbits, we will make use of *Dulac's criterion* [14] in the first quadrant, $E > 0, g > 0$. In this context, Dulac's criterion amounts to finding a continuously differentiable real-valued function $u(E, g)$ such that $\partial_E(u(E, g)E_T) + \partial_g(u(E, g)g_T)$ is of one sign throughout the first quadrant. It suffices to choose $u(E, g) = 1/Eg$, and we find

that

$$\begin{aligned}\partial_E(u(E,g)E_T)+\partial_g(u(E,g)g_T) &= \partial_E\left(2\frac{(g-g_0)}{T_R E g}E\right)+\partial_g\left(\frac{(\alpha-\gamma g-\beta g E)}{Eg}\right)\\ &= \partial_E\left(2\frac{(g-g_0)}{T_R g}\right)+\partial_g\left(\frac{\alpha}{Eg}-\frac{\gamma}{E}-\beta\right)\\ &= 0+\partial_g\left(\frac{\alpha}{Eg}\right)\\ &= -\frac{\alpha}{Eg^2}\\ &< 0 \quad \text{in the first quadrant.}\end{aligned}$$

There are therefore no closed orbits in the first quadrant, and the system must collapse onto the appropriate fixed point.

In order to confirm these predictions, we have run simulations of the governing ODEs (6.2)–(6.5) both above and below threshold. In Figures 6.2 and 6.3 we show the results for the case of $\Delta = 3$. The initial conditions correspond to the pulse-width in steady-state ($\sigma(0) = \sigma_s$), the pulse-location in the optimum location ($a(0) = -\bar{a}$), an initial pulse amplitude of 10^{-8}, and an initial gain of g_0.

In Figure 6.2(a)–(d) we show the results when we operate at 5% below threshold. Figure 6.2(a) shows that the pulse sweeps into position exponentially with T, as expected from (6.10). For these parameter values, our linear analysis suggests that we should see amplitude growth on the order of $\exp(2\Delta^2) \approx 6.5 \times 10^7$ and that the maximum amplitude should be reached when $T \approx 2.2 \times 100 T_R$. Figure 6.2(c) shows that this is still true in the presence of the nonlinear coupling between the gain and the pulse-amplitude. The gain dynamics are shown in Figure 6.2(b). After initially decaying, the gain rebounds and assumes its equilibrium value of $\alpha/\gamma = 0.1805$. In Figure 6.2(d) we show the corresponding contour plot of the nonlinear pulse formation process, reconstructed from the ODE simulation.

In Figure 6.3(a)–(b) we show the pulse-amplitude dynamics when we operate the laser at 5% above threshold. After it overshoots its mark, the pulse-amplitude oscillates briefly before reaching its equilibrium value of $\psi_0 \approx 0.168$. The contour plot of the pulse evolution is shown in Figure 6.3(b).

6.2. Single pulse and noise. In the previous section we showed that, above threshold, an initial optimum pulse evolves nonlinearly into a ground-mode pulse. In the early stages of its evolution it experiences large energy growth on the order of $\exp(4\Delta^2)$. After it has swept into position, the nonlinear interaction between the pulse amplitude and the gain arrests this growth and allows a ground-mode pulse of the correct energy to form.

The linear stability of this ground-mode pulse was addressed earlier. We showed that perturbations to the ground-mode pulse can experience similar levels of linear transient growth. In the presence of the nonlinearity the question of the outcome of these perturbations arises. In order to answer this question we have run a series of full PDE simulations. Our algorithm is based on a spectral Hermite approximation, as detailed in [2]. In the results that follow, we used 128 spectral modes and a time step of 0.01. Unless stated otherwise, all parameters are set to those given in section 2.

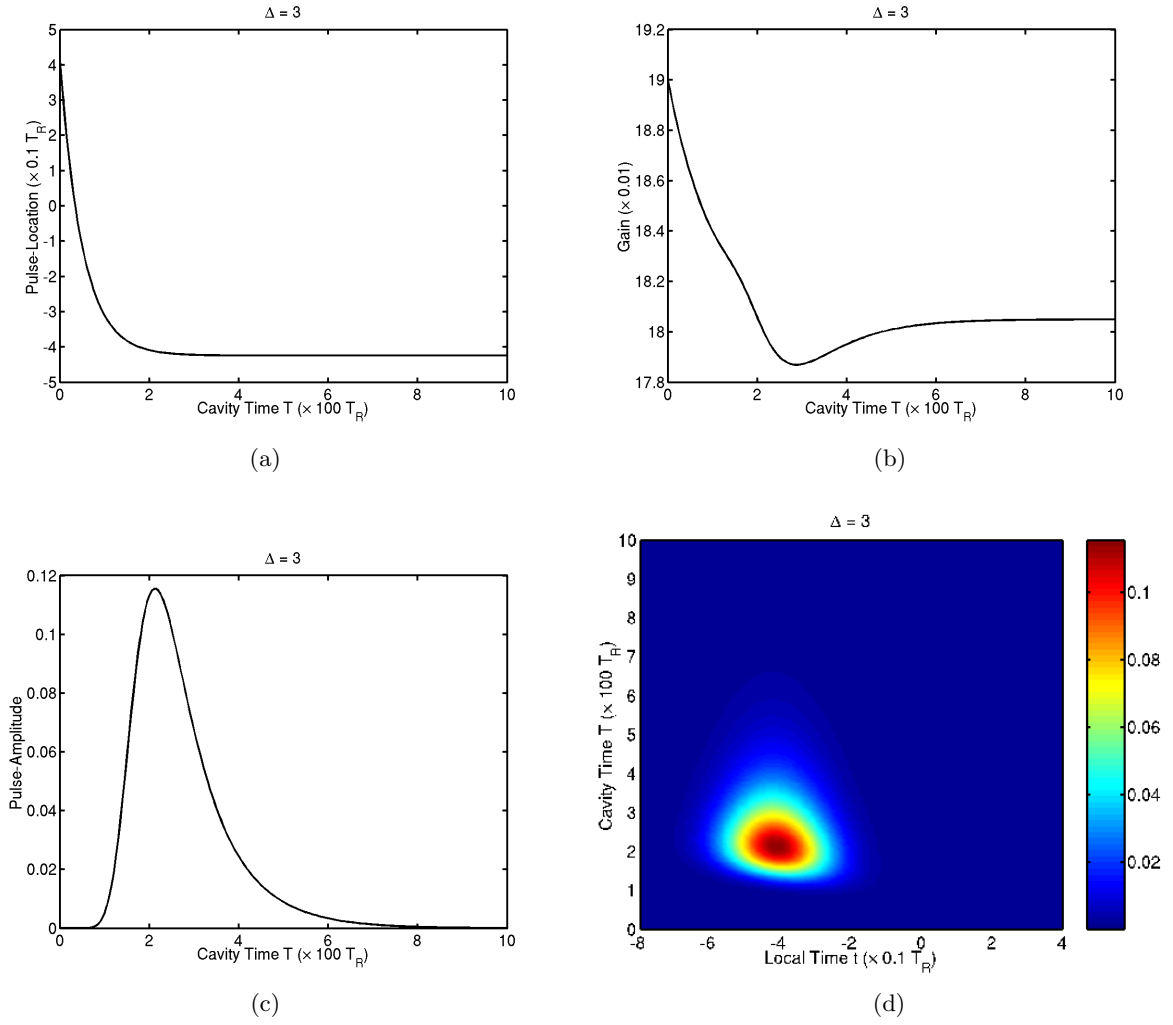
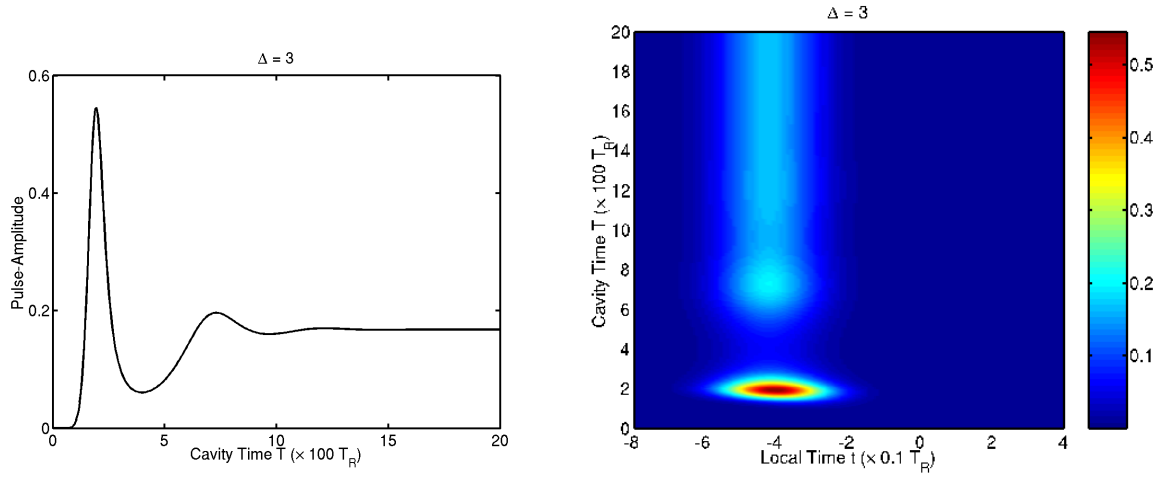


Figure 6.2. *Nonlinear evolution of a pulse at* 5% *below threshold for* $\Delta = 3$ *as governed by* (6.2)–(6.5). (a) *Pulse-location,* (b) *gain,* (c) *pulse-amplitude, and* (d) *contour plot of the pulse amplitude. The initial conditions correspond to an optimum pulse of equilibrium width and amplitude* 10^{-8} *and an initial gain of* g_0.

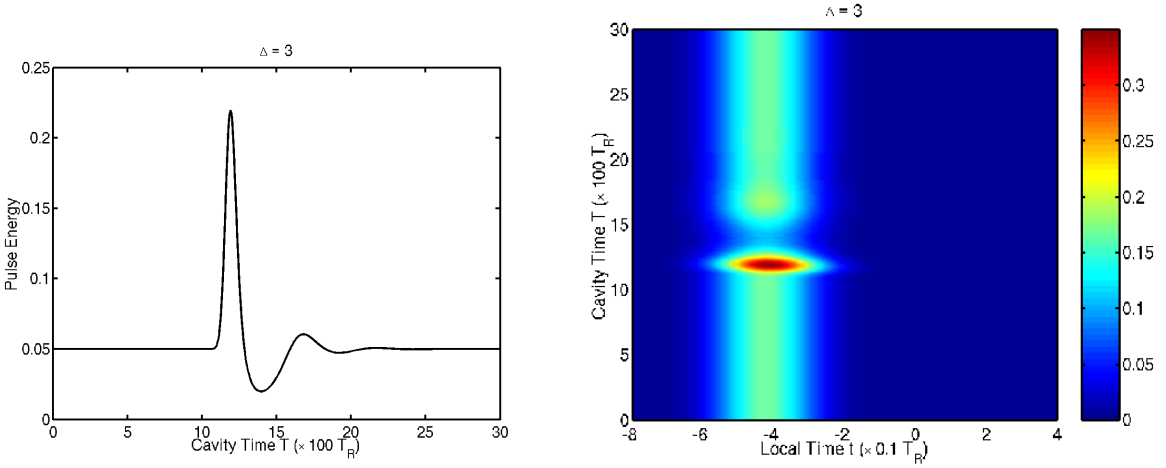
In the first set of simulations, we consider the nonlinear interaction between a ground-mode pulse and a single perturbation pulse which initially occupies the adjoint location. In Figure 6.4 we show the results of a simulation for the case of $\Delta = 3$ at 5% above threshold. The adjoint pulse has an initial amplitude of 10^{-8} and is injected at $T = 10 \times 100T_R$. In Figure 6.4(a) we show the energy in the cavity as a function of T. The amplitude of the noise pulse grows very quickly as it sweeps across the domain and, by the time it arrives in position, its amplitude is comparable to that of the existing ground-mode pulse! It then begins to interact with the ground-mode pulse, and out of this interaction emerges a new ground-mode pulse, as shown in Figure 6.4(b). After some relaxation oscillations, the new ground-mode pulse settles into equilibrium.

In the second set of simulations, we consider the full nonlinear dynamics (2.1)–(2.2) in



(a) (b)

Figure 6.3. *Nonlinear evolution of a pulse at* 5% *above threshold for* $\Delta = 3$ *as governed by* (6.2)–(6.5). (a) *Pulse-amplitude and* (b) *contour plot of the pulse amplitude. The initial conditions correspond to an optimum pulse of equilibrium width and amplitude* 10^{-8} *and an initial gain of* g_0.

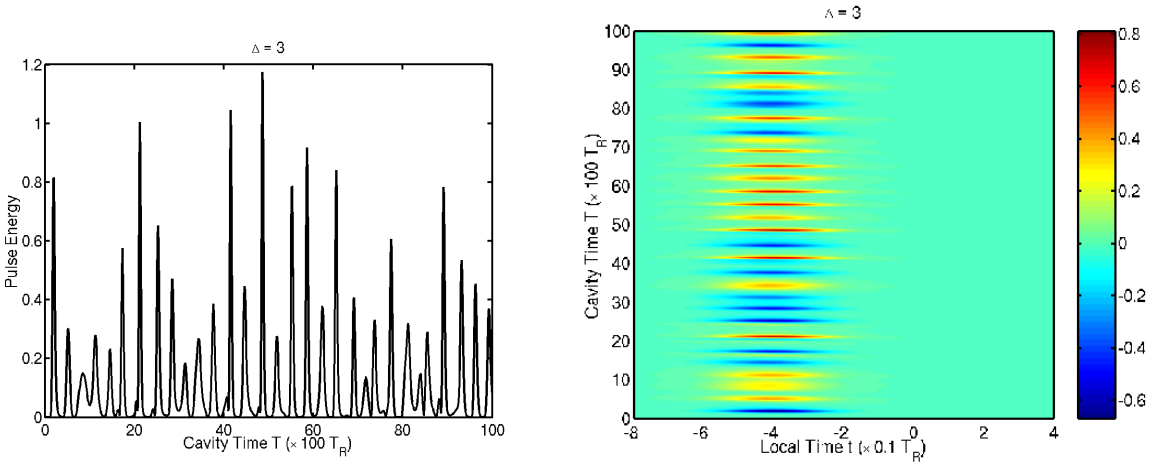


(a) (b)

Figure 6.4. *Pulse-dynamics at* 5% *above threshold for* $\Delta = 3$ *as governed by* (2.1)–(2.2). (a) *Pulse-energy and* (b) *contour plot of the pulse-amplitude. With the pulse-amplitude and gain in equilibrium, a single adjoint noise pulse of amplitude* 10^{-8} *is injected at* $T = 10 \times 100T_R$.

the presence of stochastic noise. Our noise source is white with an amplitude of 10^{-8}. In Figure 6.5 we show the results of a simulation for $\Delta = 3$ at 5% above threshold. Figure 6.5(a) shows the energy in the cavity as a function of T, while Figure 6.5(b) shows the contour plot of the pulse-amplitude.

Figure 6.5 suggests that in the presence of a continuous noise source, the process of adjoint noise pulse growth occurs over and over again in the following sense. Out of the noise



(a) (b)

Figure 6.5. *Pulse dynamics at* 5% *above threshold for* $\Delta = 3$ *as governed by* (2.1)–(2.2) *in the presence of white noise of amplitude* 10^{-8}. (a) *Pulse-energy and* (b) *contour plot of* $A(T, t)$.

spectrum, a pulse that begins in the adjoint location dominates as it experiences the most transient growth. As it grows and sweeps into position, it begins to compete with an existing ground-mode pulse. Out of this interaction emerges a new ground-mode pulse, which is now susceptible to the growth of another adjoint pulse. This process then repeats endlessly, with the time between successive pulses dictated by the transient growth time.

7. Conclusion. In this paper we have considered the linear and nonlinear dynamics of pulse formation in an actively mode-locked laser. The trivial solution of the laser equations loses stability at the lasing threshold to an off-center Gaussian pulse. The delay in the laser pulse with respect to the minimum loss point is proportional to the normalized modulator detuning Δ. The lowest-order Hermite–Gaussian mode is linearly stable to perturbations, while the higher-order modes are linearly unstable to the lower-order modes. One would therefore expect that at threshold the laser would turn on and emit stable ground-mode laser pulses.

As a result of the nonnormality of the linear operator, however, the laser is subject to large transient growth on the order of $\exp(4\Delta^2)$ at the lasing threshold, in agreement with previous findings by Kärtner, Zumbühl, and Matuschek [21]. We also find that the transient growth at threshold is precisely the Petermann excess noise factor [30, 34, 35] for a laser governed by a nonnormal operator. Below threshold we found exact expressions for the transient growth, and we showed that the perturbation which experiences maximum growth is closely related to the adjoint laser mode.

Our reduction of the governing PDEs to a low-dimensional system of ODEs for an evolving pulse is novel. It provides us with an insight into the pulse formation dynamics. The linearized version allowed us to determine exact results on transient growth, such as the maximum energy attained and the time taken to reach maximum energy. We also showed that the nonlinear system collapses onto the appropriate fixed point and that no limit cycles are possible. In the

absence of a continuous noise source, there are no interesting dynamics; an initial perturbation grows and sweeps into position and is stable there. In the presence of noise, however, the laser effectively destabilizes as new laser pulses are repeatedly formed out of the noise spectrum.

The importance of nonnormal operators is well known to both the fluids [38] and optics [35] communities. In viscous shear flow, much of the focus has been on the transient growth experienced by a single perturbation; the resulting dynamics are often a matter of conjecture, although simple models have been proposed which couple transient growth and nonlinear mixing [13]. In laser physics, the presence of a stochastic noise source, in the form of spontaneous emission noise, is vital for laser operation. The Petermann excess noise factor is a measure of the influence that the nonorthogonal laser modes have on the strength of the noise source. At the lasing threshold, these two factors—transient growth and excess noise—are identical. This results from the fact that the transient growth experienced at threshold is determined by the projection of the initial perturbation onto the ground-mode laser pulse. If the initial perturbation is the adjoint laser mode, then this is precisely the excess noise factor.

Comparable results have been reported by Longhi and Laporta [26] in the context of a frequency-modulated laser. The results obtained here could be extended to cover the frequency-modulated case by allowing the modulation parameter μ to become complex. In addition, an immediate extension of the work presented here would be to spatio-temporal as opposed to temporal laser pulses. In this case, Dunlop, Firth, and Wright [9] have derived the appropriate spatio-temporal master equation that describes pulse formation and have applied it to a Kerr lens mode-locked laser. Allowing for both AM and FM and including both space and time in the laser operator would represent a complete generalization of the model studied here.

Acknowledgments. We thank the referees for their excellent feedback, particularly the comments relating to the optimal growth curves.

REFERENCES

[1] G. Bachman and L. Narici, *Functional Analysis*, Academic Press, New York, 1966.
[2] K. Black and J. B. Geddes, *Spectral Hermite approximations for the actively mode-locked laser*, J. Sci. Comput., 16 (2001), pp. 81–120.
[3] M. H. Crowell, *Characteristics of mode-coupled lasers*, IEEE J. Quantum. Electronics, 1 (1965), pp. 12–20.
[4] T. Deutsch, *Mode-locking effects in an internally modulated ruby laser*, Appl. Phys. Lett., 7 (1965), pp. 80–82.
[5] M. DiDomenico, *Small-signal analysis of internal (coupling type) modulation of lasers*, J. Appl. Phys., 35 (1964), pp. 2870–2876.
[6] M. DiDomenico, H. M. Marcos, J. E. Geusic, and R. E. Smith, *Generation of ultrashort optical pulses by mode locking the YAG:Nd laser*, Appl. Phys. Lett., 8 (1966), pp. 180–183.
[7] A. J. DeMaria and D. A. Stetser, *Laser pulse-shaping and mode-locking with acoustic waves*, Appl. Phys. Lett., 7 (1965), pp. 71–73.
[8] A. J. DeMaria, D. A. Stetser, and H. Heynau, *Self mode-locking of lasers with saturable absorbers*, Appl. Phys. Lett., 8 (1966), pp. 174–176.
[9] A. M. Dunlop, W. J. Firth, and E. M. Wright, *Master equation for spatio-temporal beam propagation and Kerr lens mode-locking*, Opt. Comm., 138 (1997), pp. 211–226.
[10] A. M. Dunlop, W. J. Firth, and E. M. Wright, *Time-domain master equation for pulse evolution and laser mode-locking*, Opt. Quantum Electronics, 32 (2000), pp. 1131–1146.

[11] B. F. Farrell, *Optimal excitation of perturbations in viscous shear flow*, Phys. Fluids, 31 (1988), pp. 2093–2102.
[12] B. F. Farrell and P. J. Ioannou, *Stochastic forcing of the linearized Navier-Stokes equations*, Phys. Fluids A, 5 (1993), pp. 2600–2609.
[13] T. Gebhardt and S. Grossmann, *Chaos transition despite linear stability*, Phys. Rev. E (3), 50 (1994), pp. 3705–3711.
[14] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 1983.
[15] H. Haken and M. Pauthier, *Nonlinear theory of multi-mode action in loss modulated lasers*, IEEE J. Quantum Electronics, 4 (1968), pp. 454–459.
[16] L. E. Hargrove, R. L. Fork, and M. A. Pollack, *Locking of He-Ne laser modes induced by synchronous intracavity modulation*, Appl. Phys. Lett., 5 (1964), pp. 4–5.
[17] S. E. Harris and R. Targ, *Fm oscillation of the He-Ne laser*, Appl. Phys. Lett., 5 (1964), pp. 202–204.
[18] S. E. Harris and O. P. McDuff, *Theory of FM laser oscillation*, IEEE J. Quantum Electronics, 1 (1965), pp. 245–262.
[19] H. A. Haus, *A Theory of Forced Mode Locking*, IEEE J. Quantum Electronics, 11 (1975), pp. 323–330.
[20] J. M. Hopkins and W. Sibbett, *Ultra-short pulse lasers: Big payoffs in a flash*, Scientific American, 283 (2000), pp. 72–79.
[21] F. X. Kärtner, D. M. Zumbühl, and N. Matuschek, *Turbulence in mode-locked lasers*, Phys. Rev. Lett., 82 (1999), pp. 4428–4431.
[22] D. M. Kim, S. Marathe, and T. A. Rabson, *Eigenfunction analysis of mode-locking process*, J. Appl. Phys., 44 (1973), pp. 1673–1675.
[23] D. J. Kuizenga and A. E. Siegman, *FM and AM mode-locking of the homogeneous laser* I: *Theory*, IEEE J. Quantum Electronics, 6 (1970), pp. 694–708.
[24] D. J. Kuizenga and A. E. Siegman, *FM and AM mode-locking of the homogeneous laser* II: *Experimental results*, IEEE J. Quantum Electronics, 6 (1970), pp. 709–715.
[25] W. E. Lamb Jr., *Theory of an optical laser*, Phys. Rev., 134 (1964), pp. A1429–A1450.
[26] S. Longhi and P. Laporta, *Excess noise in intracavity laser frequency modulation*, Phys. Rev. E. (3), 61 (2000), pp. R989–R992.
[27] U. Morgner and F. Mitschke, *Drift instabilities in the pulses from cw mode-locked lasers*, Phys. Rev. E. (3), 58 (1998), pp. 187–192.
[28] P. M. Morse and H. Feshbach, *Methods of Theoretical Physics, Vol.* 1, McGraw–Hill, New York, 1953.
[29] T. J. Nelson, *A coupled-mode analysis of mode locking in homogeneously broadened lasers*, IEEE J. Quantum Electronics, 8 (1972), pp. 29–33.
[30] K. Petermann, *Calculated spontaneous emission factor for double-heterostructure injection lasers with gain-induced waveguiding*, IEEE J. Quantum Electronics, 15 (1979), pp. 566–570.
[31] S. C. Reddy and D. S. Henningson, *Energy growth in viscous channel flows*, J. Fluid Mech., 252 (1993), pp. 209–238.
[32] P. J. Schmid and D. S. Henningson, *Stability and Transition in Shear Flows*, Appl. Math. Sci. 142, Springer-Verlag, New York, 2001.
[33] A. E. Siegman, *Lasers*, University Science Books, Sausalito, CA, 1986.
[34] A. E. Siegman, *Excess spontaneous emission in non-Hermitian optical systems.* I. *Laser amplifiers*, Phys. Rev. A, 39 (1989), pp. 1253–1263.
[35] A. E. Siegman, *Excess spontaneous emission in non-Hermitian optical systems.* II. *Laser oscillators*, Phys. Rev. A, 39 (1989), pp. 1264–1268.
[36] P. W. Smith, T. J. Bridges, E. G. Burkhardt, and O. R. Wood, *Mode-locked high pressure waveguide CO_2 laser*, Appl. Phys. Lett., 21 (1972), pp. 470–472.
[37] G. Stix, *The triumph of the light*, Scientific American, 284 (2001), pp. 80–87.
[38] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll, *Hydrodynamic stability without eigenvalues*, Science, 261 (1993), pp. 578–584.
[39] A. M. van der Lee, A. L. Mieremet, M. P. van Exter, N. J. van Druten, and J. P. Woerdman, *Quantum noise in a laser with nonorthogonal polarization modes*, Phys. Rev. A, 61 (2000), 033812.